Machine Translated by Google

Энциклопедия

Автор и редактор

Эльзейн Салах Эльзейн

Член IEEE

- Статьи по электронике

Издательство IEEE, Великобритания

Этот контент является конфиденциальным

Machine Translated by Google

электроника

MDPI

Статья

Выборочный захват деталей сложной формы с использованием топологического Извлечение скелета

Андреа Пенниси [1], Моника Силео [2], Доменико Даниэле Блуази [1] и Франческо Пьерри 2,*

[1] Департамент международных гуманитарных и социальных наук, Международный университет UNINT в Риме, 00147 Рим, Италия; andrea.pennisi@gmail.com (AP); domenico.bloisi@unint.eu (DDB)
[2] Инженерный факультет Университета Базиликаты, 85100 Потенца, Италия; monica.sileo@unibas.it
* Переписка: francesco.pierri@unibas.it; Тел.: +39-0971-205020

Аннотация: Для повышения автономности и гибкости робототехнических систем решающую роль играет способность воспринимать и захватывать объекты. Более подробно, роботы-манипуляторы должны обнаружить наличие объектов в своем рабочем пространстве, определить точку захвата и рассчитать траекторию приближения к объектам с положением рабочего органа, подходящим для выполнения задачи. Это могут быть сложные задачи при наличии сложной геометрии, где можно обнаружить несколько кандидатов на точку захвата. В этой статье мы представляем новый подход к работе с автомобильными деталями сложной формы, состоящий из основанного на глубоком обучении метода извлечения топологического скелета и механизма выбора позы активного захвата. В частности, мы используем модифицированную версию известного алгоритма Lightweight OpenPose для оценки топологического скелета реальных автомобильных деталей. Предполагаемый скелет используется для выбора наилучшей позы для захвата подручного объекта . Наш подход разработан так, чтобы быть более эффективным в вычислительном отношении по сравнению с другими существующими методами обнаружения позы захвата. Количественные эксперименты, проведенные с помощью манипулятора с 7 степенями свободы на различных реальных автомобильных компонентах, демонстрируют эффективность предложенного подхода с вероятностью успеха 87,04%.

Ключевые слова: компьютерное зрение для производства; глубокое обучение хватанию и манипулированию; визуальное обучение

check for updates

Цитирование: Пенниси А.; Силео, М.; Блуази, Д.Д.; Пьерри, Ф. Выборочный захват деталей сложной формы с использованием извлечения топологического скелета. Электроника 2024, 13, 3021. https://doi.org/10.3390/electronics13153021 .

Академические редакторы: Кшиштоф Окарма и Петр Лех

Поступила: 4 июня 2024 г.
Пересмотрено: 23 июля 2024 г.
Принято: 29 июля 2024 г.
Опубликовано: 31 июля 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Детали сложной геометрической формы широко используются в обрабатывающей промышленности, например, в автомобилестроении. Использование роботов для обработки деталей сложной формы по-прежнему остается сложной задачей из-за проблем с восприятием, планированием и рассуждением. В частности, неопределенность в положении объекта для захвата и восприятие шума из-за отражающих материалов являются распространенными проблемами в промышленных сценариях.

Захват объектов сложной геометрии можно грубо разделить на методы, основанные на моделях, основанные на уже существующих 3D-моделях, и методы, основанные на обучении, которые используют машинное обучение для прогнозирования точек захвата. Оба подхода должны воспринимать внешнюю среду с использованием алгоритмов, основанных на зрении, на основе камер и облаков точек для обнаружения объектов, сегментации и оценки позы, как в [1,2], или стратегий, основанных на тактильных ощущениях, как в [3]. , требующие датчиков для измерения силы, тактильной обратной связи и обнаружения скольжения. Гибридные подходы сочетают в себе несколько методов обеспечения надежности, включая объединение нескольких датчиков и активное восприятие.

В этой статье мы представляем полный конвейер для обработки автомобильных деталей сложной формы с использованием робота-манипулятора с 7 степенями свободы. В частности, мы применяем подход, основанный на глубоком обучении, для разработки многообъектного детектора, предназначенного для извлечения топологического скелета, принадлежащего детали, которую нужно ухватить, чтобы точно оценить ее положение. После оценки позы осуществляется выбор лучшей позы для захвата, чтобы увеличить вероятность успешного захвата объекта.

Вклад этой работы тройной.

Machine Translated by Google

1. Процесс извлечения скелета обеспечивает представление позы объекта в трехмерном пространстве. Следовательно, это также позволяет точно оценить ориентацию объекта.
2. Мы используем MobileNetV3 для замены исходной магистрали MobileNetV1 в сети извлечения скелетов и настраиваем ее для обнаружения скелетов промышленных объектов со сложной геометрией как спереди, так и сзади, даже если объекты имеют разные формы с обеих сторон.
3. Выбор хватательной позы осуществляется с использованием динамического подхода. Это означает, что ошибка при извлечении скелета обнаруживается автономно, и робот активно меняет свое положение, чтобы лучше выполнить захват.

Мы провели несколько экспериментов с реальными автомобильными деталями, чтобы проверить наш подход, который позволяет без каких-либо ограничений обнаруживать ключевые точки объекта в перевернутом виде .

Оставшаяся часть статьи организована следующим образом. Раздел 2 содержит краткое описание существующих соответствующих работ. Наш метод представлен в разделе 3. Эксперименты , демонстрирующие эффективность предложенного подхода, показаны в разделе 4. Наконец, выводы и будущие направления сделаны в разделе 5.

2. Сопутствующая работа

За последние несколько лет, благодаря наличию мощных графических процессоров, методы глубокого обучения стали пригодными для обнаружения точек захвата. Они доказали свою способность заменить традиционные аналитические подходы, основанные на геометрических свойствах, физических моделях и силовом анализе. Архитектура сверточной нейронной сети (CNN) под названием GraspNet, способная сегментировать доступные области на поверхности объектов, была представлена в [4]. В [5] CNN в сочетании с информацией, предоставленной камерой глубины , использовалась для обнаружения присутствия объекта и наилучшей позы захвата.
Было предложено несколько подходов для повышения точности глубоких CNN, см., например, [6,7], но они обычно требуют длительного времени вычислений (т. е. порядка секунд).

Более эффективные подходы, требующие только изображений глубины, были предложены в [8,9]. Более подробно в [8] глубокая сверточная нейронная сеть была обучена в моделируемой среде для изучения функций, связанных с захватом, и возврата единого решения для каждого объекта. В [9] была предложена так называемая генеративная захватывающая сверточная нейронная сеть (GG-CNN) . Он позволяет напрямую оценивать качество захвата и положение захвата для каждого пикселя входного изображения глубины и достаточно быстр для выполнения захвата в динамических средах. Производительность GG-CNN была улучшена за счет внедрения GG-CNN2 [10], который представляет собой CNN, основанную на архитектуре семантической сегментации [11].
Общей характеристикой методов обнаружения точек захвата, основанных на глубоком обучении, является необходимость расчета значения качества захвата для каждого пикселя изображения, что занимает чрезвычайно много времени.

Когда знания об объекте и кандидаты в позы захвата недоступны, можно аппроксимировать объект, используя примитивы формы, например, используя многоракурсные измерения [12] или идентифицируя особенности в сенсорных данных [13]. Метод, предложенный в [14], заключается в выборе кандидатов на позу захвата после определения областей, где уже был осуществлен успешный захват . В [1] предлагается захват частично известных объектов в неструктурированной среде на основе расширения для промышленного контекста хорошо известного метода вычитания фона [15]. Благодаря распространению недорогих датчиков глубины, многие алгоритмы 3D-регистрации были использованы для решения проблемы захвата объекта. Например , в [2] модель захватываемого объекта создается с использованием набора облаков точек, полученных из разных положений, и фиксируется номинальная поза захвата. Впоследствии эта модель сравнивается с представлением объекта во время выполнения, чтобы вычисли

В этой работе мы предлагаем основанный на скелетах подход для обнаружения поз захвата, который по своей сути менее требует вычислительных затрат из-за компактного представления объекта через скелет.

Machine Translated by Google

Качественное и количественное сравнение нашего подхода с наиболее значимыми работами, описанными в этом разделе, показано в Таблице 1. Это сравнение учитывает учитывать не только результаты, но и ограничения каждой работы.

Таблица 1. Таблица сравнения различных подходов к захвату объектов.

Методы	Приложения	Количественные результаты	Ограничения
Архитектура CNN с DDF модуль [4]	Роботизированный захват в реальном времени	90% точности по Корнеллу получить набор данных	Ошибка в прогнозировании ориентации для некоторые объекты
Структура, основанная на более быстром R-CNN и DACAB [5]	Захват предмета с помощью мобильного телефона манипулятор	86,3% успеха	Неэффективный метод поиска
GQ_CNN для классификации надежных хватательный [6]	Хватание предметов домашнего обихода	99% точности	Длительное время вычислений
DCNN на основе изображений глубины предсказать позу захвата [8]	Схватывание неизвестных предметов	92% (70%) точности с цилиндрический (коробчатые) предметы	Генерация единого захвата решение для одного объекта
HC для изучения прототипа детали [14]	Захват похожих предметов	NA	Захват предметов сложной формы объекты с ранее невиданными функции.
Топологический скелет извлечение (эта работа)	Захват предметов сложной формы автомобильные детали	87,04% успеха	Нужна хорошая камера калибровка

3. Предлагаемый метод

На рисунке 1 показана общая функциональная архитектура нашего подхода. Он состоит из четырех основные модули, а именно сбор визуальных данных, извлечение топологического скелета, захват Выбор позы и захват робота. Подробное описание каждого модуля приведено ниже.

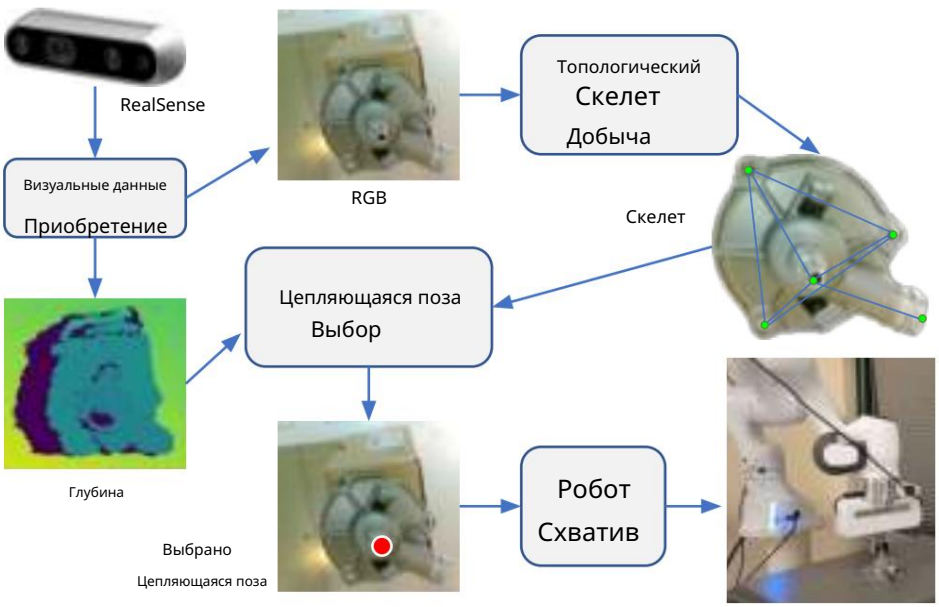


Рисунок 1. Функциональная архитектура предлагаемого подхода.

Machine Translated by Google

3.1. Сбор визуальных данных

Визуальные данные (как RGB, так и глубина) собираются с помощью камеры Intel Realsense D435 RGBD, установленной на концевом эффекторе с помощью опоры, напечатанной на 3D-принтере, в так называемой конфигурации «глаз в руке» . На рисунке 2 показаны опорные кадры, прикрепленные к рабочему органу робота Fe и камере Fc.

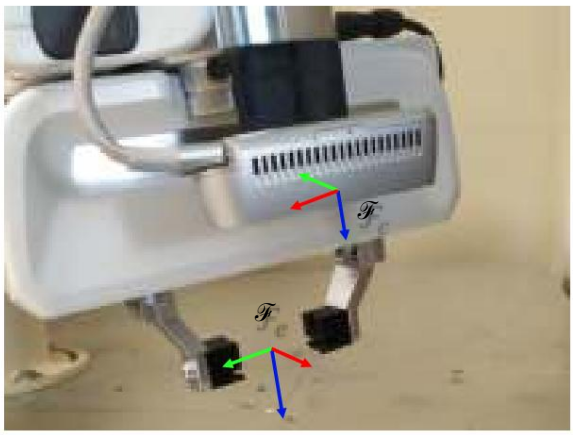


Рисунок 2. Концевой эффектор и опорные кадры камеры.

Камера Intel Realsense D435 имеет минимальное расстояние по глубине, за пределами которого она не может обеспечить измерение глубины, примерно равное 28 см, а для сбора данных камеры требуется библиотека realsense-ros, поскольку связь между модулями происходит через операционную систему робота (POC). Камера была предварительно откалибрована с использованием 30 изображений плоского 2D-шахматного рисунка. Процесс калибровки включает как внутреннюю, так и внешнюю калибровку. Первый предназначен для определения параметров камеры, которые описывают, как камера преобразует трехмерные координаты сцены в двумерные координаты изображения, такие как фокусное расстояние, главная точка и оптические искажения, а второй предоставляет параметры которые описывают жесткое преобразование, которое сопоставляет трехмерные координаты реального мира с трехмерными координатами системы отсчета камеры. Реализация процедуры калибровки, предложенная библиотекой VISP [16] на основе [17,18,], была принята с использованием шахматной доски, составленной из квадратов 9 × 6 размерами 0,02645 м.

Стоит помнить, что хорошая процедура калибровки имеет решающее значение для успеха процедуры захвата, поскольку она обеспечивает точное восприятие окружающей среды, позволяя точно идентифицировать и позиционировать точки в трехмерном пространстве для успешных манипуляций.

3.2. Извлечение топологического

скелета. Предлагаемый метод разработан для объектов, которые:

• являются жесткими, так как это неприменимо к деформируемым
объектам; • не являются совершенно симметричными, поскольку, хотя можно определить несимметричный
топологический скелет, детектор может запутаться в процессе извлечения из-за симметричных особенностей.

В этой работе мы концентрируемся на реальных автомобильных деталях, в том числе на двух крышках маслоотделителя картера из чугуна и пластика и воздушной трубке. Выбранные объекты имеют возрастающий уровень сложности. Первый объект, чугунная крышка маслоотделителя картера, демонстрирует высокую степень симметрии с несколькими точками захвата и может быть захвачен цилиндрической частью, что снижает влияние ошибок ориентации робота вокруг оси пальца . Второй объект, пластиковая крышка маслоотделителя картера, также демонстрирует высокую степень симметрии с различными точками захвата, но его необходимо захватывать в определенной Наконец, воздушная труба имеет сложную форму, лишена симметрии и имеет только две точки захвата, что представляет собой наиболее сложную задачу для робота.

Machine Translated by Google

Мы решили смоделировать их скелеты, учитывая несколько ключевых точек, некоторые из которых соответствуют потенциальным точкам захвата для подъема объекта с помощью робота-манипулятора (см. правую верхнюю часть рисунка 1). Чтобы обнаружить топологический скелет (TS) объектов, которые необходимо захватить, мы рассмотрим Lightweight OpenPose [19], архитектура которого состоит из трех основных компонентов: экстрактора признаков, средства оценки TS и сети полей сходства частей (PAF).

По сравнению с исходным OpenPose [20] мы выбрали Lightweight OpenPose, потому что высокие вычислительные требования первого метода делают его менее применимым в приложениях реального времени на устройствах с небольшой вычислительной мощностью. OpenPose использует двухветвевую многоступенчатую архитектуру CNN. Первая ветвь прогнозирует карты достоверности частей (PCM) для частей тела, а вторая ветвь прогнозирует поля аффинности частей (PAF) для моделирования связей между частями тела. Архитектура включает в себя несколько этапов сверток для итеративного уточнения этих прогнозов, что приводит к высокой точности за счет увеличения вычислительной нагрузки. Light OpenPose, с другой стороны, изменяет исходную архитектуру, чтобы уменьшить сложность и повысить эффективность. Этот подход уменьшает количество сверточных слоев и этапов и использует отделимые по глубине свертки вместо стандартных сверток для уменьшения количества параметров и операций. Более того, магистральная сеть использует MobileNet или ShuffleNet вместо более тяжелых VGG19 или ResNet, используемых в исходном OpenPose, и оптимизирует вычисление полей сходства частей, чтобы обеспечить баланс между точностью и эффективностью.

Извлечение функций. Оригинальный Lightweight OpenPose использует сеть MobileNetV1 , оптимизированную для извлечения функций в реальном времени. MobileNet — это семейство архитектур нейронных сетей, предназначенных для эффективного развертывания на мобильных и встроенных устройствах с ограниченными вычислительными ресурсами. Ключевой особенностью MobileNet является использование разделенных по глубине сверток, что позволяет значительно сократить количество необходимых параметров и вычислений при сохранении высокой точности.

Хотя MobileNetV1 является высокоэффективной нейронной сетью, у нее есть некоторые ограничения и недостатки, которые следует учитывать. Например, он имеет ограниченную точность, поскольку разработан для обеспечения баланса между размером и точностью модели. Он может не достичь того же уровня точности, что и более крупные и сложные нейронные сети, особенно на сложных объектах, где ключевые точки (соединения) не очевидны. Операция глубинно-разделимой свертки, используемая в MobileNetV1, может быть менее выразительной, чем традиционные сверточные операции, и может быть не в состоянии уловить все важные особенности изображения.

По вышеуказанным причинам в этой работе мы предлагаем заменить MobileNetV1 на MobileNetV3 [21] на этапе извлечения признаков. MobileNetV3 был разработан для устранения ограничений MobileNetV1 при сохранении эффективности. Архитектура сети MobileNetV3, использованная в данной работе, показана на рисунке 3.

MobileNet V3 имеет два основных варианта: (1) MobileNet V3-Large, предназначенный для приложений с более высокой точностью, с большим количеством уровней и каналов, и (2) MobileNet V3-Small, оптимизированный для сред с ограниченными ресурсами, обеспечивающий некоторую точность в обмен на сокращение объема вычислений. - циональный спрос. Наш выбор пал на последнее. MobileNet V3 представляет несколько новых компонентов, таких как инвертированные остаточные блоки для поддержания высокой степени эффективности, модули сжатия и возбуждения (SE) для улучшения репрезентативной мощности модели за счет повторной калибровки откликов функций по каналам и активации H-Swish. Функция.

Функция жесткого взмаха — это нелинейная функция активации, которая разработана так, чтобы быть более эффективной, чем традиционные функции активации, такие как ReLU. Функция жесткого взмаха определяется как

$$\text{swish}(x) = x\,6 \quad \frac{\text{ReLU6}(x+3)\ \text{h-}}{\qquad}, \tag{1}$$

где ReLU6(x) = min(max(x, 0), 6) — это урезанная функция ReLU, которая выводит значения от 0 до 6, по-прежнему обеспечивая нелинейное поведение и одновременно увеличивая скорость вычислений по сравнению со стандартной функцией ReLU .

Еще одной важной особенностью MobileNetV3 является использование модуля сжатия и возбуждения (SE). Модуль SE — это простой и эффективный способ улучшить репрезентативную мощность.

Machine Translated by Google

сети (т. е. способность изучать и представлять сложные закономерности и функции в входные данные). Он работает путем изучения коэффициентов масштабирования по каналам, которые используются для выборочного повысить информативные возможности в сети. Модуль SE добавляется к каждому узкому месту блок в архитектуре MobileNetV3, способствующий повышению точности по отношению к MobileNetV1.

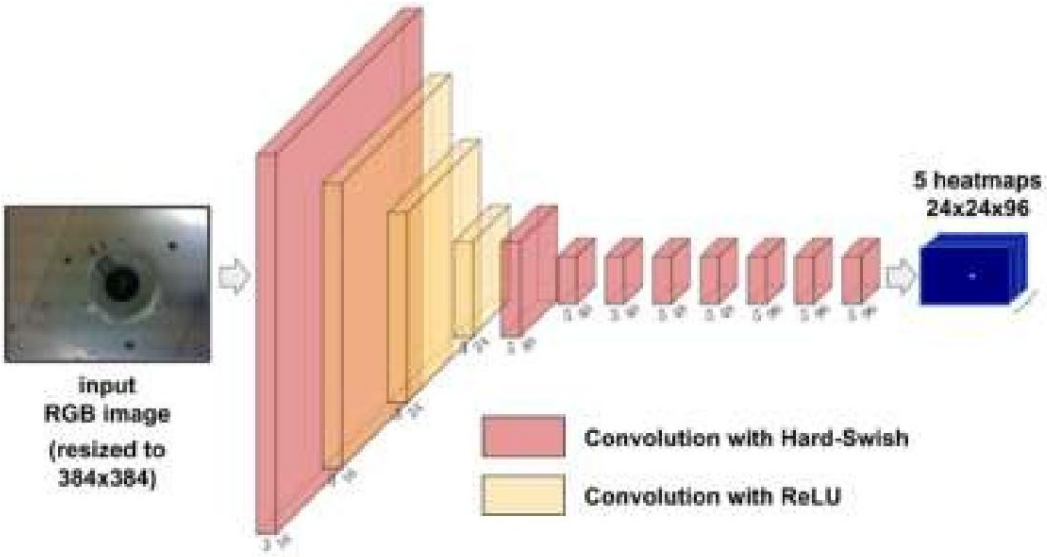


Рисунок 3. Архитектура MobileNetV3.

MobileNetV3 также представляет новую технику, называемую мобильным перевернутым узким местом. свертка (MBConv), которая представляет собой модифицированную форму отделимой по глубине свертки. используется в MobileNetV1. Блок MBConv состоит из трёх типов свёрток: 1 × 1 свертка для расширения количества каналов, глубинная свертка для выполнения пространственных фильтрация и свертка 1 × 1 для уменьшения количества каналов до исходного размера. Блок MBConv также включает в себя соединение быстрого доступа, которое позволяет градиенту течь. напрямую от входа к выходу. Блок MBConv помогает повысить выразительность модели относительно MobileNetV1.

Наконец, MobileNetV3 включает в себя блок среднего потока, который используется для поддержания высокого уровень точности при минимизации количества необходимых вычислений. Он также использует операция динамической свертки, которая адаптируется к входным данным. Подробности о параметрах используемые для каждого блока, описаны в Таблице 2. Размер входного изображения изменяется до 384 × 384, Выходные данные представляют собой набор карт объектов размером 24 × 24 × 96, по одной для каждой ключевой точки и одной для фон.

Таблица 2. Архитектура сети MobileNetV3, используемая в этой статье. HS = жесткий взмах, RE = ReLU, s = шаг.

Ввод	Оператор	Размер опыта	#вне	ЮВ	Надпримеры	с
3842 × 3	conv2d, 3 × 3	1922 × 16	-	16	-	HS	2
узких, 3 × 3	962 × 16	16	16	Икс	РЭ	2
узких, 3 × 3	482 × 24	72	24	-	РЭ	2
узких, 3 × 3	482 × 24	88	24	-	РЭ	1
узких, 5 × 5	242 × 40	96	40	Икс	HS	2
гриф, 5 × 5	242 × 40	240	40	Икс	HS	1
гузка, 5 × 5	242 × 48	240	40	Икс	HS	1
гузка, 5 × 5	242 × 48	120	48	Икс	HS	1
глыба, 5 × 5	242 × 96	144	48	Икс	HS	1
глыба, 5 × 5	242 × 96	288	96	Икс	HS	1
гриф, 5 × 5		576	96	Икс	HS	1
		576	96	Икс	HS	1

Machine Translated by Google

Оценка ТС. Карты объектов из MobileNetV3 являются входными данными для создания набора возможных ключевых точек для каждой части объекта на изображении. Фактически, карты объектов фиксируют пространственную информацию во входном изображении и обеспечивают расширенное представление изображения, которое можно использовать для обнаружения ключевых точек. MobileNetV3 добавляет специальную головку для прогнозирования местоположения ключевых точек, которая состоит из серии сверточных слоев, генерирующих тепловые карты, уточняющих функции, извлеченные магистралью, и генерирующих тепловые карты для каждой ключевой точки. На рисунке 4 показан пример выходных данных средства оценки TS для чугунной крышки маслоотделителя картера, который состоит из пяти тепловых карт, по одной для к Каждая тепловая карта имеет то же пространственное разрешение, что и карты объектов, и нормализована до значений от 0 до 1. Каждый пиксель тепловой карты указывает вероятность того, что соответствующая часть тела присутствует в этом месте изображения.

Сеть ПАФ. Он принимает карты признаков, сгенерированные экстрактором признаков, в качестве входных данных и выводит набор карт признаков PAF, по одной для каждой пары обнаруженных ключевых точек. Карты объектов PAF кодируют направление и силу связей между ключевыми точками, используя двухканальное представление, где каждый канал кодирует разные аспекты связи . В частности, один канал кодирует единичный вектор, который представляет направление соединения, а другой канал кодирует показатель достоверности, который представляет силу соединения.

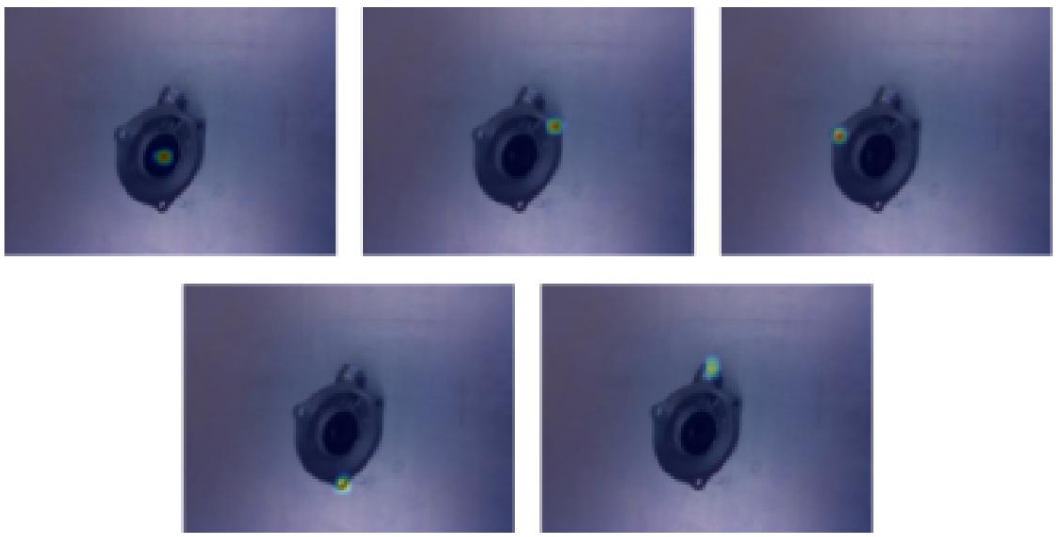


Рисунок 4. Примеры тепловой карты чугунной крышки маслоотделителя картера. Рассмотренным ключевым точкам соответствует пять тепловых карт .

Окончательный расчет ТС. После создания PAF и тепловых карт они используются вместе для группировки отдельных ключевых точек в окончательный TS. Окончательный TS получается путем сначала идентификации соединений-кандидатов с использованием PAF, а затем оценки соединений на основе вероятности того, что они образуют действительное соединение. Затем соединения используются для построения окончательного TS путем соединения отдельных ключевых точек в полный TS объ

После расчета ключевых точек мы используем информацию о глубине для построения окончательного 3D TS с учетом набора ключевых точек из Lightweight OpenPose. На рисунке 5 показаны некоторые примеры окончательных TS для трех рассмотренных объектов, где можно отметить устойчивость предложенного подхода к извлечению TS по отношению к различным ракурсам объекта , к фотохромным изменениям и частичным перекрытиям. Этот подход также работает с несколькими экземплярами объекта.

Machine Translated by Google

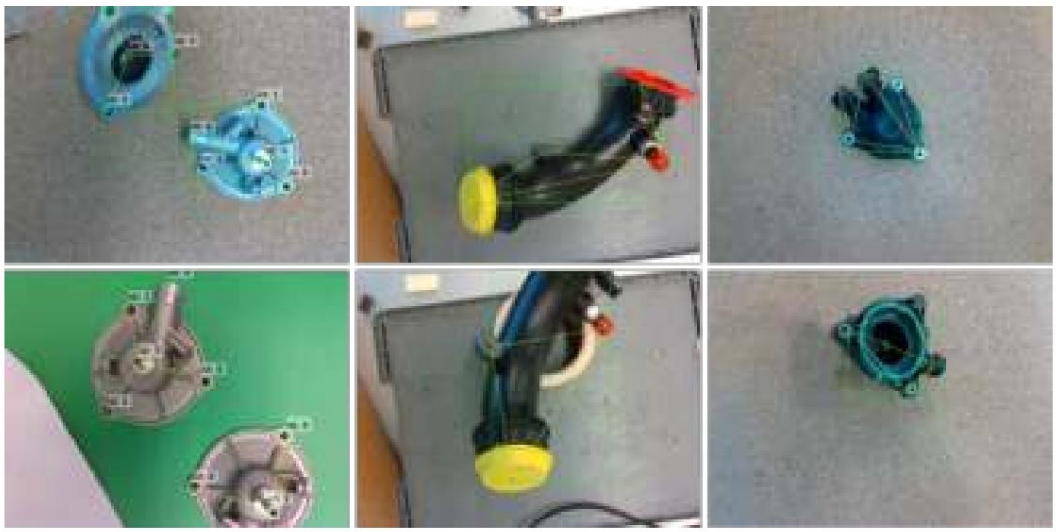


Рисунок 5. Примеры извлечения ТС на разных объектах: слева чугунная крышка маслоотделителя картера, посередине воздухопровод, а справа пластиковая крышка маслоотделителя картера. Наш подход устойчив к различным видам одного и того же объекта, к фотохромным изменениям и частичной окклюзии.

3.3. Захват выбора позы После

выбора ключевых точек Nk для извлечения TS эти ключевые точки также идентифицируются в CAD-модели объекта с помощью программного обеспечения для 3D-моделирования. В результате создается номинальное трехмерное представление TS, TSN, в системе координат CAD Ff . Более того, положения в Ff всех возможных Ng , захватывающих системы отсчета (см. рисунок 6), выраженные через (4 × 4) однородную матрицу преобразования [22], f T

gj , j = 1, . . . , Ng, могут быть локализованы на модели. Для ясности предположим, что каждая точка схватывания совпадает с ключевой точкой.

Рис. 6. Захватные рамки для рассматриваемых объектов: чугунная крышка картера маслоотделителя (верхний ряд), пластиковая крышка картерного маслоотделителя (средний ряд) и воздуховод (нижний ряд).

Machine Translated by Google

Тогда, учитывая все возможные комбинации трех ключевых точек

$$St = ti , i = 1, \ldots , Нт = \overset{Нк}{3} : ti = (Pj , Pl , Pm), j, l, m \quad \{1, \ldots , Nk\}, j/ = l/ = m ,$$

для каждой тройки ti плоскость идентифицируется посредством прикрепленной к ней системы координат, положение которой f обозначается однородной матрицей преобразования T_{ti} . Для каждой системы отсчета захвата относительное положение относительно i-й плоскости можно определить как

$$T^{ти} = T\,gj\,ти^{ф} \overset{-1}{ж} T \underset{гдж}{} . \quad (2)$$

Таким образом, для каждой точки захвата может быть вычислен список Nt матриц преобразования $T\,gj^{ти}$, представляющих положения кадра захвата в плоском кадре. Этот набор операций, обобщенный в алгоритме 1, выполняется только один раз.

Алгоритм 1: Алгоритм предварительной обработки

Ввод : TCH, Нг, Нт , $T \underset{гдж}{ж}$ (j = 1, . . . , Ng)

Выход: $T \underset{гджай}{ти}$ (i = 1,..., Nt ; j = 1,..., Ng) 1 для каждой тройки ti

ключевых точек в TSN do f 2 Вычислить T ti 3 end для 4 для i = 1, . . . , Nt

do для j = 1, . . . , Ng do 1 ff ti

T = TT gj ti gj

5

6

7 конец для

8 конец для

ти 9 возврат T gj

Во время выполнения выполняются следующие шаги:

1. Для различения объектов используется детектор YOLO [23] . YOLO был выбран, поскольку он быстрее, чем системы на основе классификаторов, но с аналогичной точностью и делает прогнозы с помощью одной сетевой оценки, рассматривая обнаружение объектов как единую задачу регрессии, что приводит к высокой точности. Более того, YOLO может обнаруживать и классифицировать несколько объектов одновременно на изображении.
2. Извлекается текущий 3D TS, TSC .
3. Точка захвата, ближайшая к камере, стр. 4a. Если видны как $\underset{копия,}{с}$ выбран лучшим.

минимум 3 ключевые точки, набор из трех ключевых точек, tk , соответствующая плоскость используется для в кадре камеры, T tk матрица преобразования, T, процедура $\underset{ТК,}{с}$ вычисления и выбора однородного продолжается с шага 5.4b . Если $_{ГЦК,}$ который идентифицирует позу захвата в плоском кадре. Затем, видны только 2 или меньше ключевых точек, робот

начинает двигаться по кругу вокруг центра ограничивающей рамки объекта, чтобы получить новое изображение с другой точки зрения . Затем процедура возвращается к шагу 1.

5. Поза захвата в кадре камеры рассчитывается как

$$T \underset{gcc}{с} \quad tk = TT \underset{gcc}{с} tk .$$

Эта процедура обобщена в алгоритме 2.

Определим однородную матрицу преобразования $T \underset{}{е}$ с , т. е. постоянная однородная матрица, выполняющая преобразование между кадром камеры и кадром конечного эффектора , полученная методом калибровки, описанным в разделе 3.1.

Machine Translated by Google

Алгоритм 2: Алгоритм времени выполнения

Ввод: для каждого объекта gj obj T ti

Выход: T c GCC

1 Получить изображение с камеры

2 Обнаружить текущий объект по изображению

3 TSC извлечь TS ti 4 T =

T получить список преобразований относительно текущего gj i=1,...,Nt ; j=1,...,Ng объект извлечь точку захвата, ближайшую к

камере

5 p c копия

6 , если видимые_ключевые точки 3,

7 | то tk извлечь тройку видимых ключевых точек в TSC

8. Вычислить T tk из T c

9 | Извлечь T gcc TC

10 еще

11 переведите робота на другую точку обзора

12 ПЕРЕХОД К 1

13 конец, если

14 T c gcc tk T c TK T ГСС

15 возврат T gcc c

Чтобы захватить позу захвата в инерциальной системе отсчета, T c gcc преобразуется следующим образом:

$$T_{gcc} = T_e \, {}^{e}_{c}T \, {}^{c}_{gcc}T, \quad (3)$$

где Te — однородная матрица, представляющая положение рабочего органа в инерциальной системе отсчета.

Замечание 1. Стоит отметить, что если точка захвата не совпадает с ключевой точкой, описанная выше процедура по-прежнему применима, но необходимо применить дальнейшее постоянное преобразование, чтобы связать точку захвата с одной из ключевых точек, принадлежащих плоскости.

3.4. Захват робота Чтобы

выполнить захват, исполнительному органу должна быть дана команда выровнять свою систему отсчета по системе отсчета захвата. Траектория планируется путем задания последовательности из трех точек: первая — это поза обзора робота, промежуточная — точка подхода , т. е. точка, расположенная по оси z захватывающей системы отсчета на расстоянии 10 см до начала координат, а последнее — это расчетное положение захвата, pˆ. Желаемое положение конечного рабочего органа, pe,d (t), определяется как г . Подробнее,

$$p_{e,d}(t) = \begin{cases} (p_a - p_0)\,\dfrac{s_1(t)}{t_a}\,p_0 + p_a - p_0 & \text{для } 0 \le t \le t_a \\ \hat{p} - \dfrac{s_2(t)}{p_{a_г}}\,(\hat{p} - p_a) + {}_{г}p_a & \text{для } t_a < t \le t_f \end{cases}. \quad (4)$$

где p0 — позиция обзора, а pa — позиция точки подхода, s1(t) (s2(t)) — длина дуги от p0 до pa (от точек пути pa до pˆ , как для s1(t), так и для s2(t)), временной г). Чтобы обеспечить постоянное ускорение и скорость на

закон можно представить как полином пятой степени.

Что касается моментов времени, tf — это продолжительность движения, а ta — промежуточный момент времени в точке сближения, который выбран для быстрого движения до точки сближения и медленного движения в непосредственной близости от объекта.

Machine Translated by Google

Что касается ориентации рабочего органа, планируется достичь той же ориентации расчетного положения захвата Rg в точке подхода и поддерживать такую ориентацию постоянной на протяжении последней части пути.

Запланированная траектория с точки зрения положения и ориентации является входными данными алгоритма обратной кинематики с обратной связью [22], направленного на вычисление опорных значений положений и скоростей суставов. Обозначим через pe (t) и Re(t) положение и ориентацию рабочего органа соответственно, а через Re,d (t) — желаемую ориентацию рабочего органа. Привязка скорости суставов робота , ˙qr (t), вычисляется как

$$q_p^{\cdot}(\tau) = Дж^{\dagger}(q(t))(v^{\cdot}_{e,d}(t) + Ke(t)), \tag{5}$$

где J † (q(t)) обозначает правую псевдообратную матрицу Якобиана робота, K IR6×6 положительно определенный матричный коэффициент $^{T}_{e, д}$ $\omega T_{e,d}^{T}$ — желаемый линейный концевой эффектор и усиления, ve,d = угловая скорость p˙, а e — ошибка отслеживания, определенная как

$$e = \begin{matrix} pe,d & pe \\ \eta\epsilon\ e,d & \eta e,d\epsilon\ e \end{matrix} \quad S(\epsilon\ e,d)\epsilon\ e, \tag{6}$$

где Qe = {ηe , ϵ e} и Qe,d = {ηe,d , ϵ e,d} — единичный кватернион, извлеченный из Re и Re,d соответственно, а S(·) — кососимметричный матричный оператор, выполняющий перекрестное произведение [2

Блок-схема всего процесса представлена на рисунке 7.

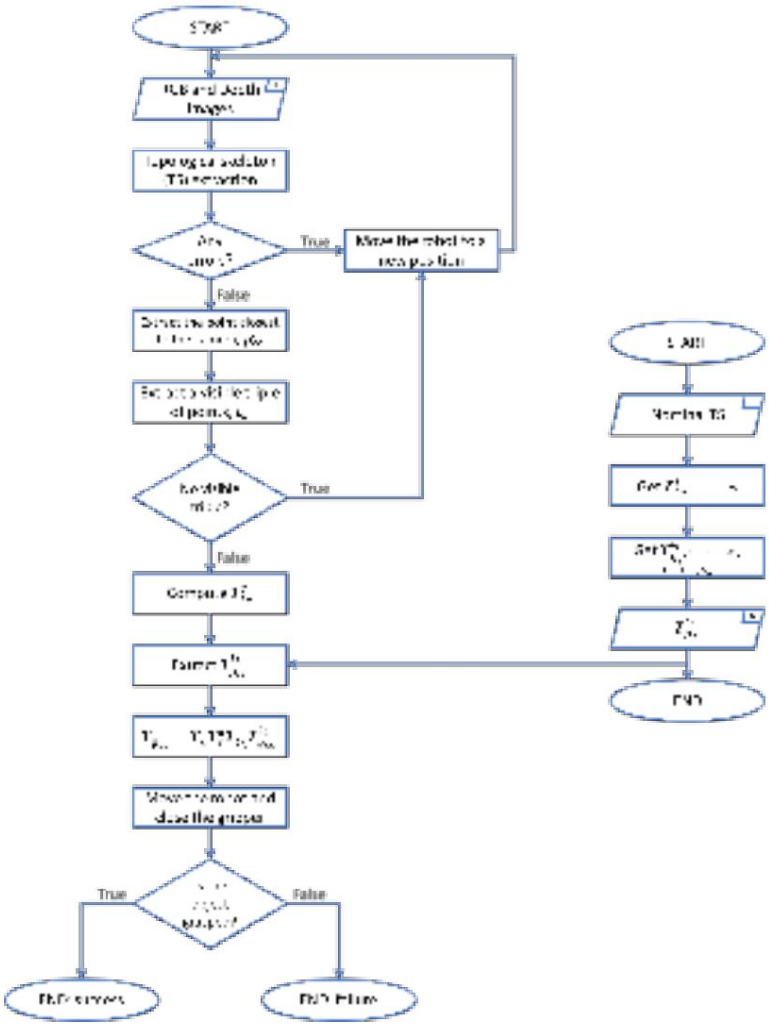


Рисунок 7. Блок-схема всего процесса.

Machine Translated by Google

4. Результаты экспериментов

Экспериментальная установка состоит из камеры Intel RealSense D435, установленной на Робот-манипулятор Franka Emika Panda с 7 вращающимися шарнирами. Робот можно управлять с помощью интерфейса управления Franka (FCI) и библиотеки libfranka C++. библиотека с открытым исходным кодом, которая напрямую управляет роботом с внешней рабочей станции через подключение к сети Ethernet. В этой работе метапакет Franka_ros , который интегрирует libfranka в ROS. Рабочая станция работает под управлением Ubuntu 18.04 LTS и поддерживает режим реального времени. ядро на процессоре Intel Xeon 3,7 ГГц с 32 ГБ оперативной памяти. Мы провели эксперименты с тремя рассматриваемыми объектами, показанными на рисунке 6, а количественные результаты сообщается ниже.

4.1. Результаты извлечения TS

С помощью Coco Annotator [24] было аннотировано 5992 изображения. Маркированные данные имеют были разделены на наборы для обучения, проверки и тестирования, состоящие из 4618, 229 и 1145 изображений. соответственно. В таблице 3 показано количество изображений в обучающем, проверочном и тестовом наборах. для каждого рассматриваемого объекта.

Таблица 3. Количество образцов изображений, используемых в обучающем, проверочном и тестовом наборах для рассматриваемых объектов.

Объект	Обучение	Проверка	Тест
Чугунная крышка маслоотделителя картера	1440	60	300
Воздушная труба	1406	46	228
Пластиковая крышка маслоотделителя картера	1772	123	617

Метрикой, которую мы использовали для оценки обнаружения TS, является сходство ключевых точек объекта. (OKC) [25], определяемый следующим образом:

$$OKC = \frac{i\varepsilon[0,N \quad 1] \exp \quad \frac{-d_{я}^{2}}{2c\ 2к_{я}^{2}} \quad \delta(vi > 0)}{i\varepsilon[0,N \quad 1]\ \delta(vi > 0)} \tag{7}$$

где:

s – масштаб объекта;
di — расстояние предсказанной ключевой точки i от основной истины;
ki — константа для каждой ключевой точки, которая контролирует спад;
vi — флаг видимости.

OKS рассчитывается для каждой выборки, представляющей объект. Флаг видимости принимает учитывать, видна точка или нет: если ключевая точка помечена, δ(vi > 0) равно 1, в противном случае — 0 без учета закрытых ключевых точек.

В нашем сценарии мы использовали OKS для расчета истинно положительного (TP), ложноотрицательного (FN), Ложноположительное (FP) и истинно отрицательное (TN) обнаружение. Если обнаружение имеет OKS > порога, это считается ТП; в противном случае, как FP. В частности, мы рассмотрели два порога: а именно 0,5 и 0,75, и рассчитал следующие показатели: точность, отзыв, показатель F1, и средняя точность (AP). В таблице 4 показаны результаты нашего алгоритма для тестового набора 1145 изображений.

Таблица 4. Результаты детектора TS при различных порогах для тестового набора из 1145 изображений.

Порог	Точность	Отзывать	F1-Оценка	АП
0,5	0,92	0,90	0,91	0,82
0,75	0,86	0,89	0,87	0,72

Чтобы вычислить производительность нашего экстрактора TS во время выполнения, мы протестировали его на подмножестве 60 изображений с использованием NVIDIA RTX A5500, среднее время выполнения 0,012 с и

Machine Translated by Google

стандартное отклонение 0,0018 с. На подмножестве из 40 изображений с использованием NVIDIA QUADRO T2000 среднее время выполнения составляет 0,019 с, а стандартное отклонение — 0,0025 с.

4.2. Результаты детектора

объектов . Для обучения детектора объектов мы аннотировали 750 изображений размером 640 × 480 с помощью инструмента аннотаций LabelImg [26]. Мы разделили набор данных на наборы для обучения, проверки и тестирования, состоящие из 450, 150 и 150 изображений соответственно. После этапа обучения средняя точность на тестовом наборе составляет 97,32%, а вероятность успеха — 96,70%. Вывод по изображениям был выполнен на NVIDIA QUADRO T2000. На подмножестве из 40 изображений среднее время выполнения составляет 0,323 с, а стандартное отклонение — 0,0615 с.

4.3. Робот захватывает результаты

Определим ошибки определения положения и ориентации оценки как

$$e_{e\,п} = п\,г^{e} - \hat{p}_{г}^{e}, \quad (8)$$

$$e_{e\,\varphi} = \varphi\,г^{e} - \hat{\varphi}_{г}^{e}, \quad (9)$$

где $p_{г}^{e}$ — фактическое положение захвата, а $\hat{p}_{г}^{e}$ это оценка, предоставленная визуальным алгоритм. Что касается ориентации, $\varphi_{г}^{e}$ ($\hat{\varphi}_{г}^{e}$) — углы Эйлера, извлеченные из фактического (оценочная) поза захвата. Принятие углов Эйлера вместо кватернионов, как в (6), обеспечивает более четкую физическую интерпретацию ошибок ориентации. Верхний индекс e означает , что переменные выражены в кадре конечного эффектора (см. рисунок 2).

Для получения статистически значимых результатов было проведено 54 теста на захват (20 за чугунную крышку маслоотделителя картера, 19 за воздуховод и 15 за пластиковую крышку маслоотделителя картера) путем размещения объектов в различных конфигурациях и в разных условиях освещенности. и с разным опытом, чтобы робот мог изучить все возможные позы для захвата. Испытание на захват считается успешным, если захватывающий удерживает предмет стабильно в течение 10 с. Набор снимков процедуры захвата показан на рисунке 8, где верхний ряд относится к успешному тестированию, а нижний ряд — к неудачному.

(а)

(б)

Рисунок 8. Снимки двух случаев захвата. а) Успешный захват. (б) Неудача.

Только 7 экспериментов (2 с чугунной крышкой маслоотделителя картера, 3 с воздушной трубкой и 2 с пластиковой крышкой маслоотделителя картера) оказались неудачными. Таким образом, был получен показатель успеха 87,04%. В таблицах 5–7 показаны средние ошибки положения и ориентации , а также соответствующее стандартное отклонение для успешных испытаний.

Machine Translated by Google

Таблица 5. Средние погрешности чугунной крышки маслоотделителя картера.

Успешные испытания	$e_{\text{линейный}}^{e}$ [мм] e	$e_{\text{пи}}^{e}$ [мм] e	$e_{\text{фх}}^{e}$ [градус]	$e_{\text{фу}}^{e}$ [градус]
1	0,734	9.471	3.128	3.515
2	6,867	0,717	3,284	3.091
3	1,081	5,857	6.092	12.433
4	13.032	10,752	3.132	6.256
5	1,796	0,775	6.801	0,77
6	3,832	0,759	3.913	3.546
7	3,629	1,531	2,981	3.416
8	1,324	6,747	0,674	0,92
9	3.073	7.437	0,62	35.308
10	1,293	1,021	0,952	1,605
11	3,916	1,631	0,391	6.217
12	4,62	2,697	3.044	4.508
13	2.368	4.09	9.12	3.953
14	0,463	6.314	2.456	10.075
15	1,816	2,214	2,217	5.806
16	2,958	4,751	1,842	7,763
17	2,86	8,736	1,354	4.138
18	3,018	0,747	13,855	2.16
Иметь в виду	3,26	4,236	3,659	6.415
Среднеквадратичное отклонение	2,820	3,278	3,329	7.607

Таблица 6. Средние ошибки для воздуховода.

Успешные испытания	$e_{\text{линейный}}^{e}$ [мм]	$e_{\text{пи}}^{e}$ [мм]	$e_{\text{фх}}^{e}$ [градус]	$e_{\text{фу}}^{e}$ [градус]
1	0,043	0,989	2,439	8.109
2	9,586	17,223	4,597	0,555
3	1,098	2,612	16,723	13.204
4	3,777	6,188	19,545	10.702
5	7,213	0,227	3,663	1,042
6	3,882	1,516	1,684	0,315
7	2,897	2,176	11,183	1,733
8	0,144	1,932	10.255	7,891
9	2.41	1.412	12.057	7.392
10	5.042	0,797	12,674	20,755
11	2,989	2,809	11.516	26.216
12	14,5	4.903	0,988	2.294
13	11.149	1,207	5.192	0,53
14	3,736	5.841	9.211	25,563
15	0,818	8.494	12.022	20.307
16	0,636	14.538	15.122	7.086
Иметь в виду	4.37	4.554	9.304	9.606
Среднеквадратичное отклонение	4.085	4.843	5.447	8.801

Для всех объектов ошибки положения по оси z рамы рабочего органа имеют не сообщалось, поскольку они незначительны из-за геометрии объекта. По той же причине, ошибки ориентации вокруг оси z рамы концевого эффектора могут быть незначительными для чугунная крышка маслоотделителя картера и воздуховод.

В некоторых испытаниях наблюдались большие ошибки, в основном по оси Y рамы концевого эффектора, но объект был успешно захвачен, поскольку захват характеризуется параллельными пальцами, и ошибки в направлении закрытия более допустимы.

Системные сбои можно разделить на две основные категории:

1. Ошибки, связанные с отсутствием (см. рисунок 9a,d) или неточной (см. рисунок 9b,g,e) ключевой точкой. обнаружение или прогнозирование, а также неправильная оценка глубины.
2. Ошибки в оценке позы, которые могут стать причиной скольжения объекта.

Machine Translated by Google

Таблица 7. Средние погрешности пластиковой крышки маслоотделителя картера.

Успешные испытания	$e_{____}$ [мм] e	$e_{\text{пи}}$ [мм] e	$e^{e}_{\varphi x}$ [градус]	$e^{e}_{\text{e }\varphi y}$ [градус]	$e^{e}_{\varphi z}$ [градус]
1	13.453	1,036	7,726	4.594	10.505
2	1,795	8.396	11.419	16,797	1,977
3	1,387	5.306	2.214	3.451	11,761
4	7.493	5.909	3.41	5.945	7.285
5	0,799	6,753	3.439	12,744	2,62
6	2,997	1,521	2.03	4,885	9.134
7	5,897	9.452	10.56	1,247	4,62
8	0,973	1,319	7.221	0,007	4.146
9	2.246	5,577	2.104	3.025	2.231
10	1,268	13,631	3.325	4.165	26.108
11	4.547	13,786	2,265	1,418	7,61
12	0,575	1,472	8,638	2.531	10,629
13	3,606	2.318	1,721	6.289	23.799
Иметь в виду	3,618	5,883	5.082	5.161	9.417
Среднеквадратичное отклонение	3,487	4.281	3.381	4.527	7.367

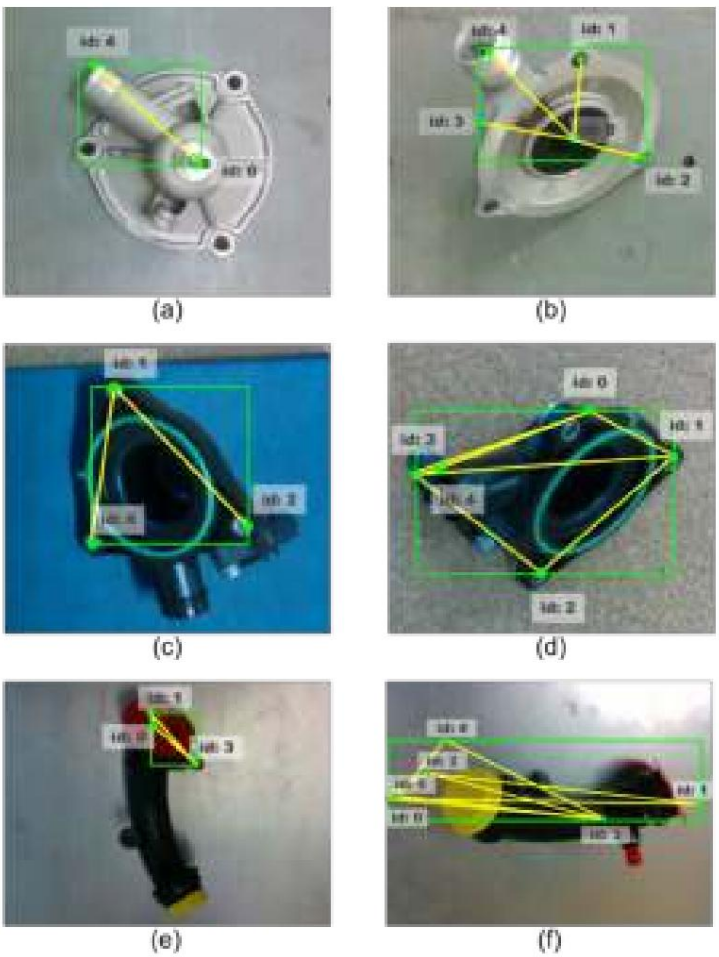


Рисунок 9. Примеры отсутствующего (a,e) и неточного (b,d,f) обнаружения ключевых точек в TS. Пример отсутствие обнаружения ключевой точки, которое может привести к успешному захвату объекта (c).

В случае отсутствия обнаружения ключевой точки, например, из-за относительного объекта-камеры положении, сбоем можно управлять, перемещая точку зрения камеры и получая новый прогноз (см. раздел 3.3). В остальных случаях процедура захвата завершается. с провалом. Поскольку робот может обнаружить нарушение захвата, весь процесс повторяется.

Machine Translated by Google

Стоит отметить, что в соответствии с процедурой, описанной в Алгоритме 2, захват объекта возможен даже при правильном обнаружении только трех видимых ключевых точек (см. Рисунок 9в) , при условии, что одна из них является точкой захвата, расположенной в месте, которое можно захватить имеющимся концевым эффектором.

5. Выводы

В этой работе был предложен надежный метод захвата деталей сложной геометрии в промышленном сценарии. В такой среде проблемы с захватом возникают из-за неопределенности положения объекта, который нужно схватить, а также из-за восприятия шума из-за его материала. В частности, мы сосредоточились на реальных автомобильных деталях со сложной геометрией и отражающими поверхностями, которые вызывают шум на карте глубины. Предлагаемое решение основано на сети извлечения TS, которая может создавать графическое представление объекта в реальном времени. Этап рассуждения используется, чтобы решить, достаточно ли текущий вид объекта для фактического захвата или манипулятору необходимо двигаться, чтобы лучше захватывать объект. Количественные эксперименты были проведены с роботом с 7 степенями свободы и тремя различными автомобильными деталями сложной формы, продемонстрировав, что предлагаемый подход является быстрым и надежным. Высокая точность и возможность работы в режиме реального времени этого предлагаемого подхода делают его подходящим решением для промышленных приложений, где требуется быстрая и точная работа.

Ввиду сложности рассматриваемых объектов полное количественное сравнение эффективности с другими подходами, представленными в литературе, вряд ли может быть проведено. Однако набор тестовых данных общедоступен, что позволяет проводить сравнения в будущем.

В будущих направлениях мы намерены изучить интеграцию данных о глубине в процесс извлечения TS, чтобы напрямую получать трехмерные положения ключевых точек. Более того, этап обнаружения объекта также может быть интегрирован в процедуру извлечения TS.

Вклад автора: концептуализация, AP, MS, DDB и FP; Методология, AP, MS, DDB и FP; Программное обеспечение, AP, MS и DDB; Расследование, AP и MS; Валидация, AP и MS; Формальный анализ, AP, MS и DDB; Письмо — первоначальный вариант, AP и MS; Написание — просмотр и редактирование, DDB и FP; Надзор, ДДБ и ФП; Администрирование проекта, DDB и FP; Финансирование приобретения, FP Все авторы внесли равный вклад в эту работу. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Данное исследование было поддержано Министерством университетов и исследований Италии в рамках гранта PRIN 2022 PNRR MELODY (Совместная работа нескольких роботов для поддержки задач по разборке) n. P2022XALNS.

Заявление о доступности данных. Исходный код метода обнаружения TS общедоступен по адресу https://github.com/apennisi/CoGP-TS. Исходный код нашего подхода на основе ROS общедоступен по адресу https://github.com/sileom/graspingWithSkeleton.git. Несколько видеороликов экспериментов доступны по адресу https://tinyurl.com/bdhyf493. Изображения тестового набора доступны по адресу https://tinyurl. ком/bdxs8n7z. Все модели, используемые в описанной стратегии, общедоступны и их можно скачать по адресу https://tinyurl.com/3a4nnc88. (Все ссылки доступны 23 июля 2024 г.).

Благодарности: авторы хотели бы поблагодарить Алессандро Лоренцо, Симону Д'Амато и Антонио Джардиелло за помощь в процессе аннотирования изображений.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Сокращения

В данной рукописи использованы следующие сокращения:

Си-Эн-Эн	Сверточная нейронная сеть
ГРИП	Степень свободы
ФЦИ	Интерфейс управления Франка
ФКН	Полностью сверточная сеть
ФН	Ложноотрицательный результат

Machine Translated by Google

ФП	Ложно положительный
Генеративная захватывающая сверточная нейронная сеть GG-CNN	
MBConv Mobile: инвертированная свертка «узких мест»	
ОКС	Сходство ключевых точек объекта
ПАФ	Поля сходства деталей
РОС	Операционная система робота
ТН	Истинно отрицательный
ТП	Истинный Позитив
ТС	Топологический скелет

Рекомендации

1. Силео, М.; Блуази, Д.Д.; Пьерри, Ф. Обнаружение и захват объектов в реальном времени с использованием вычитания фона в промышленном сценарии. В материалах 6-го Международного форума IEEE по исследованиям и технологиям для общества и промышленности (RTSI) 2021 г., виртуально, 6–9 сентября 2021 г.; стр. 283–288.
2. Силео, М.; Блуази, Д.Д.; Пьерри, Ф. Захват твердых промышленных объектов с использованием 3D-регистрации. Машины 2023, 11, 396. [CrossRef]
3. Костанцо, М.; Де Мария, Г.; Леттера, Г.; Натале, К. Могут ли роботы заполнить полку супермаркета?: Планирование движения и контроль захвата. IEEE-робот. Автомат. Маг. 2021, 28, 61–73. [Перекрестная ссылка]
4. Асиф, У.; Тан, Дж.; Харрер, С. GraspNet: эффективная сверточная нейронная сеть для обнаружения захвата в реальном времени для маломощных устройств. В материалах IJCAI, Стокгольм, Швеция, 13–19 июля 2018 г.; стр. 4875–4882.
5. Чжан Х.; Тан, Дж.; Чжао, К.; Лян, З.; Лю, Л.; Чжун, Х.; Фан С. Метод быстрого обнаружения и захвата мобильного манипулятора. на основе улучшенного более быстрого R-CNN. Индийский робот. Межд. Джей Робот. Рез. Прил. 2020, 47, 167-175. [Перекрестная ссылка]
6. Малер Дж.; Лян, Дж.; Нияз, С.; Ласки, М.; Доан, Р.; Лю, Х.; Охеа, JA; Голдберг, К. Dex-net 2.0: Глубокое обучение для планирования надежного захвата с помощью синтетических облаков точек и аналитических показателей охвата. В Proceedings of the Robotics: Science and Systems (RSS), Кембридж, Массачусетс, США, 12–16 июля 2017 г.
7. Пинто, Л.; Гупта, А. Самоконтроль над увеличением: учимся схватывать за 50 тысяч попыток и 700 робото-часов. В материалах международной конференции IEEE по робототехнике и автоматизации (ICRA) 2016 г., Стокгольм, Швеция, 16–21 мая 2016 г.; стр. 3406–3413.
8. Шмидт П.; Варенкамп, Н.; Вехтер, М.; Асфур, Т. Схватывание неизвестных объектов с использованием глубоких сверточных нейронных сетей на основе изображений глубины. В материалах международной конференции IEEE по робототехнике и автоматизации (ICRA) 2018 г., Брисбен, Австралия, 21–25 мая 2018 г.; стр. 6831–6838.
9. Моррисон, Д.; Корк, П.; Лейтнер, Дж. Замыкание цикла роботизированного захвата: подход к генеративному синтезу захвата в реальном времени. В Proceedings of the Robotics: Science and Systems (RSS), Питтсбург, Пенсильвания, США, 26–30 июня 2018 г.
10. Моррисон, Д.; Корк, П.; Лейтнер, Дж. Обучение надежному, реактивному роботизированному хватанию в реальном времени. Межд. Джей Робот. Рез. 2020, 39, 027836491985906. [Перекрестная ссылка]
11. Ю, Ф.; Колтун В. Многомасштабная агрегация контекста с помощью расширенных сверток. arXiv 2015, arXiv:1511.07122.
12. Дюна, К.; Маршан, Э.; Коллоуэт, К.; Леру, К. Активная приблизительная оценка формы неизвестных объектов. В материалах Международной конференции IEEE/RSJ 2008 г. по интеллектуальным роботам и системам, Ницца, Франция, 22–26 сентября 2008 г.; стр. 3622–3627.
13. Крафт Д.; Пюжо, Н.; Баёсески, Э.; ПОПОВИЧ, М.; Крагич, Д.; Калкан, С.; Вергёттер, Ф.; Крюгер, Н. Рождение объекта: обнаружение предметности и извлечение формы объекта через комплексы объект-действие. Межд. Дж. Робот-гуманоид. 2008, 5, 247–265. [Перекрестная ссылка]
14. Детри, Р.; Эк, СН; Мадри, М.; Пиатер, Дж.; Крагич, Д. Обобщение схватывания частично схожих объектов. В материалах Международной конференции IEEE по робототехнике и автоматизации 2012 г., Сент-Пол, Мичиган, США, 14–18 мая 2012 г.; стр. 3791–3797.
15. Блуази, Д.Д.; Пенниси, А.; Иокки, Л. Фоновое моделирование в морской сфере. Мах. Вис. Прил. 2014, 25, 1257–1269. [Перекрестная ссылка]
16. Маршан, Э.; Шпиндлер, Ф.; Chaumette, F. ViSP для визуального сервопривода: универсальная программная платформа с широким классом навыков управления роботами. IEEE-робот. Автомат. Маг. 2005, 12, 40–52. [Перекрестная ссылка]
17. Каннала, Дж.; Брандт, С.С. Общая модель камеры и метод калибровки для обычных, широкоугольных объективов и объективов типа «рыбий глаз». IEEE Пер. Паттерн Анал. Мах. Интел. 2006, 28, 1335–1340. [Перекрестная ссылка] [ПабМед]
18. Цай, Р.Ю.; Ленц, Р.К. Новая технология полностью автономной и эффективной трехмерной калибровки руки и глаза робототехники. IEEE Транс. Робот. Автомат. 1989, 5, 345–358. [Перекрестная ссылка]
19. Осокин, Д. Двухмерная оценка позы нескольких человек в режиме реального времени на ЦП: облегченный OpenPose. arXiv 2018, arXiv:1811.12004.
20. Цао З.; Идальго, Г.; Саймон, Т.; Вэй, ЮВ; Шейх, Ю. OpenPose: двухмерная оценка позы нескольких человек в реальном времени с использованием сходства частей Поля. IEEE Транс. Паттерн Анал. Мах. Интел. 2021, 43, 172–186. [Перекрестная ссылка] [ПабМед]
21. Ховард А.; Сэндлер, М.; Чу, Г.; Чен, Л.; Чен, Б.; Тан, М.; Ван, В.; Чжу, Ю.; Панг, Р.; Васудеван, В.; и другие. Поиск МобилНетВ3. arXiv 2019, arXiv:1905.02244.
22. Сицилиано, Б.; Скьявикко, Л.; Виллани, Л.; Ориоло, Дж. Робототехника – моделирование, планирование и контроль; Спрингер: Лондон, Великобритания, 2009 г.
23. Редмон, Дж.; Диввала, С.; Гиршик, Р.; Фархади, А. Вы смотрите только один раз: унифицированное обнаружение объектов в реальном времени. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Лас-Вегас, Невада, США, 27–30 июня 2016 г.; стр. 779–788.
24. Брукс, Дж. Коко, аннотатор. 2019. Доступно онлайн: https://github.com/jsbroks/coco-annotator/. (по состоянию на 23 июля 2024 г.).

Machine Translated by Google

25. Рончи, М.Р.; Перона, П. Бенчмаркинг и диагностика ошибок при многоэкземплярной оценке позы. arXiv 2017, arXiv:1707.05388.
26. Хартекслабс; Лин, Т. LabelImg. 2015. Доступно онлайн: https://github.com/heartexlabs/labelImg. (по состоянию на 23 июля 2024 г.).

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Улучшенное управление активным подавлением помех для постоянного Магнитный синхронный двигатель

Живэй Хуан [1], Юаньхао Чэн 1,2,*, Сы Чэнь 1,2, Сюйхуэй Чжан [3], Цзявэй Сян 1,2 и Суньань Ван [4]

[1] Колледж машиностроения и электротехники Университета Вэньчжоу, Вэньчжоу 325060, Китай; 22461440036@stu.wzu.edu.cn (ZH); 20200316@wzu.edu.cn (SC); jwxiang@wzu.edu.cn (JX)
[2] Институт и интеллектуальное производство Пинъян, Университет Вэньчжоу, Вэньчжоу 325035, Китай
3 Ключевая лаборатория Шэньси интеллектуального мониторинга горного электромеханического оборудования, Сианьский университет науки и технологий, Сиань 710064, Китай;
[4] zhangxh@xust.edu.cn Школа машиностроения, Сианьский университет Цзяотун, Сиань 710049, Китай; пиланг@xjtu.edu.cn
* Переписка: yhcheng@wzu.edu.cn

Аннотация: Для улучшения характеристик управления синхронным двигателем с постоянными магнитами (СДПМ) в условиях внешних возмущений предложен улучшенный алгоритм управления активным подавлением возмущений (IADRC) . Поскольку нелинейная функция в традиционном алгоритме ADRC недостаточно гладкая в точках останова, что напрямую влияет на производительность управления, предлагается инновационная нелинейная функция для эффективного улучшения сходимости и стабильности. На этой основе построен предлагаемый IADRC , и результаты сравнительного моделирования с ADRC и другими IADRC показывают, что достигается более высокая скорость ответа, более высокая точность и более высокая надежность.

Ключевые слова: улучшенный ADRC; нелинейная функция; ПМСМ; контроль скорости

1. Введение

Благодаря преимуществам высокой удельной мощности, простой конструкции, простоты обслуживания и удобного регулирования скорости синхронные двигатели с постоянными магнитами (СДПМ) широко используются в транспортных средствах на новой энергии, станках с числовым программным управлением (ЧПУ) и другом высокотехнологичном оборудовании [1]. ,2]. Технология пропорционально-интегрально-дифференциального (ПИД) управления, которая широко используется в различных отраслях промышленности, управляет традиционной системой регулирования скорости PMSM. Хотя ПИД-регулятор имеет преимущества простой структуры и простоты реализации [3,4], для достижения точного управления необходимо создать точную математическую модель управляемого объекта. Однако при моделировании СДСМ структура двигателя часто упрощается. Это упрощение оказалось неадекватным, когда СДСМ работает на низких скоростях, что приводит к заметному ухудшению характеристик управления, как подчеркнуто в работе. [5]. Более того, шум и помехи не могут быть эффективно подавлены с помощью ПИД, поэтому ученые предложили множество улучшенных систем ПИД [6,7]. За счет оптимизации параметров регулировки можно значительно улучшить характеристики управления , однако помехозащищенность, скорость реагирования и точность все еще необходимо повысить.

Современная теория управления предлагает новое решение для повышения производительности систем управления. Одновременно с развитием технологии управления СДСМ были предложены различные эффективные методы управления. Согласно Реф. [8] нелинейный ПИД- регулятор повышает эффективность управления за счет адаптивного преобразования сигнала ошибки и динамической настройки параметров ПИД, что позволяет контроллеру эффективно подавлять шум во входном сигнале и решать проблему компромисса между скоростью реакции и перерегулированием, свойственным обычному ПИД-регулятору. контроллеры. Кроме того, для повышения надежности системы также используется нечеткий ПИД-регулятор [9]. Кроме того, в последние годы управление скользящим режимом (SMC), несомненно, стало важнейшим достижением в области техники управления двигателями [10,11]. Это может повысить эффективность управления системами PMSM, демонстрируя при этом быструю сходимость.

check for updates

Цитирование: Хуан, З.; Ченг, Ю.; Чен, С.; Чжан, Х.; Сян, Дж.; Ван, С. Улучшенное управление активным подавлением помех для синхронного двигателя с постоянными магнитами. Электроника 2024, 13, 3023. https://doi.org/10.3390/electronics13153023 .

Академический редактор: Фабио Корти

Поступила: 25 июня 2024 г.
Пересмотрено: 25 июля 2024 г.
Принято: 30 июля 2024 г.
Опубликовано: 31 июля 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

Machine Translated by Google

и сильная устойчивость к помехам. Однако из-за особенностей вибраций структуры переменных режима скольжения вопрос о том, как эффективно подавить это вибрирование, стал важной темой. Робастное управление также представляет собой класс методов, направленных на неопределенные системы, и были изучены законы робастного управления, основанные на парадигмах H и μ-синтезе, которые обеспечивают хорошее подавление возмущений и устойчивость к изменениям параметров [12]. Хотя эти алгоритмы эффективно улучшают эффект управления системой PMSM, им по-прежнему трудно достичь превосходных характеристик управления при устранении неопределенностей в моделях и борьбе с внешними помехами.

Управление активным подавлением возмущений (ADRC) — это стратегия нелинейного управления, предложенная Цзинцин Ханем, которая позволяет эффективно справляться с неопределенностями и внешними возмущениями [13]. Полезность этого подхода отражается в его превосходной антиинтерференционной способности, сильной адаптивности к изменению и неопределенности параметров системы, высокой точности управления, быстрой скорости реакции, простой структуре и легкости реализации [14], и он привлек широкое внимание в во многих технических областях [15–17]. ADRC состоит в основном из трех компонентов: отслеживающего дифференциатора (TD), наблюдателя расширенного состояния (ESO), а также закона управления нелинейными ошибками состояния с обратной связью (NLSEF). TD может сглаживать входной сигнал, уменьшать влияние шума и предоставлять дифференциальную информацию о сигнале, чтобы помочь контроллеру лучше реагировать на изменения в системе. ESO является основным компонентом ADRC, который используется для оценки переменных состояния и общих нарушений системы. Он имеет возможность наблюдать и смягчать неопределенности и внешние возмущения в режиме реального времени, тем самым повышая надежность системы и устойчивость к помехам. Кроме того, NLSEF генерирует команды управления на основе наблюдаемой ошибки состояния, а также информации о возмущениях для получения точного управления контролируемым объектом. Стратегия управления с нелинейной обратной связью, обеспечиваемая NLSEF, может значительно улучшить динамический отклик системы . ADRC унаследовал суть традиционного ПИД-регулятора и может эффективно управлять системами с нелинейностью и неопределенностью, не полагаясь на точные модели [18,19]. Например, на основе ADRC и скользящего режима управления Fang et al. [20] предложили новый интегрированный метод проектирования контуров скорости и положения и реализовали управление скоростью и положением СДСМ. Несмотря на достижение минимального перерегулирования системы, сложность настройки параметров остается проблемой. Обращаясь к этому, Li et al. [21] представили ADRC со скользящим режимом , заменив ESO на наблюдатель нелинейных возмущений, чтобы упростить настройку параметров. Основываясь на своих исследованиях, Ge et al. [22] использовали алгоритм оптимизации роя частиц для улучшения ADRC для управления ориентацией космического корабля. Вместо того, чтобы зависеть от эмпирического выбора, параметры теперь определяются посредством математической оптимизации . Ссылка. [23] предложили нечеткий ADRC дробного порядка, который смягчает ограничения традиционных методов управления (такие как низкая точность и медленный отклик) и используется для управления многочисленными шарнирными движениями манипулятора. Рамлави и Чидан [24] изучили подход к отслеживанию управления линейным активным подавлением помех для повышения производительности и надежности отслеживания контроллера и решили проблему неопределенности модели и нарушений окружающей среды одноколесного робота. Опираясь на исследования Партовибахша и Лю [25], Го и Чжао [26] расширили возможности отслеживания мобильных роботов с помощью компенсационного механизма ADRC. Lv [27] предлагает метод нечеткого автоматического подавления возмущений (Fuzzy-ADRC) для трехфазного четырехплечевого инвертора для подавления пульсаций крутящего момента двигателя в сложных условиях эксплуатации. Ван [28] в ответ на плохую производительность управления, вызванную фиксированными параметрами ADRC для PMSM без подшипника, предлагает метод динамической настройки параметров для ADRC, основанный на генетическом алгоритме и нейронной сети обратного распространения ошибки. Фанг [29] предложил метод ADR

Хотя технология ADRC эффективно улучшает эффект управления нелинейными и неопределенными системами, все еще существуют ограничения, такие как недостаточность эффективных методологий установки параметров, отсутствие возможности оценки быстро меняющихся во времени возмущений и функция падения , принятая в традиционных ADRC, не быть достаточно плавным в точках переключения между нелинейным участком и линейным участком [30,31]. Учитывая проблемы, с которыми в настоящее время сталкиваются системы управления PMSM с контроллерами ADRC, цель данной стат

Machine Translated by Google

вносят следующий вклад: предлагается полиномиальная нелинейная функция в сочетании с косинусной функцией для улучшения функции fal , что эффективно улучшает гладкость нелинейной функции и повышает надежность ADRC. Результаты исследования показывают , что в отличие от обычного ADRC и других улучшенных ADRC (IADRC), предлагаемый здесь IADRC значительно увеличивает скорость реагирования, точность и надежность системы .

Организационная структура данного документа изложена ниже. В разделе 2 представлена математическая модель PMSM. В разделе 3 представлена стратегия улучшения ADRC и построен контроллер IADRC. В разделе 4 представлена конструкция системы управления скоростью PMSM с использованием предлагаемого IADRC. В разделе 5 проводятся симуляционный эксперимент и анализ, а также проверяются хорошие характеристики управления предлагаемого IADRC для систем PMSM.

2. Математическое моделирование СДСМ.

Структура ротора и расположение постоянных магнитов СДСМ [32] показаны на рисунке 1. По конструктивным характеристикам СДСМ можно разделить на два типа: поверхностного монтажа и встроенные. СДСМ поверхностного монтажа также называют СДСМ с неявнополюсным полюсом. Поскольку его проницаемость очень близка к проницаемости вакуума, изменение между сопротивлением и индуктивностью очень мало, что обеспечивает хорошие характеристики неявнополюсного ротора. Это также приводит к тому, что магнитное поле постоянного магнита имеет примерно синусоидальное распределение, и двигатель имеет лучшую производительность. Для встроенного PMSM сопротивление прямой оси намного выше, чем у квадратурной оси, поэтому индуктивность прямой оси намного ниже, чем у квадратурной оси, а ротор имеет явнополюсные характеристики, поэтому также называется PMSM с явно выраженным полюсом. Встроенный PMSM реализует бездатчиковое управление посредством явнополюсного эффекта и увеличивает плотность мощности за счет крутящего момента эл Однако конструкция ротора сложна, коэффициент рассеяния магнитного потока велик, а стоимость изготовления высока.

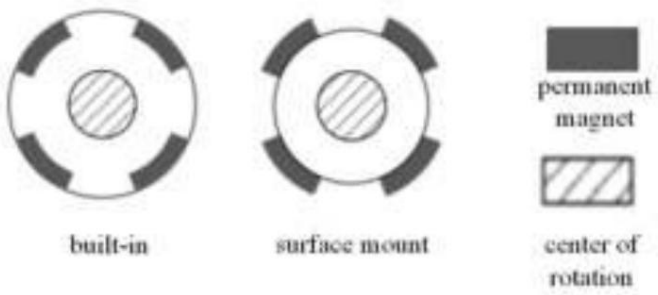


Рисунок 1. Принципиальная схема конструкции ротора СДСМ.

В связи с разной структурой СДСМ режимы их работы и формирования их электромагнитных моментов также будут разными. Следовательно, для достижения лучшего эффекта управления необходимо сформулировать множество различных схем управления в соответствии с внутренней структурой СДСМ. Учитывая, что СДПМ поверхностного монтажа имеет низкую стоимость и одинаковую индуктивность оси dq, проще составить уравнение электромагнитного момента. Поэтому в данном исследовании выбран СДСМ поверхностного монтажа.

Для анализа математической модели трехфазного PMSM необходимо сделать следующие предположения: сделаны для удобства анализа:

(1) Пренебречь сопротивлением сердечников статора и ротора, не принимая во внимание потери из-за вихревые токи или гистерезис.

(2) Материал постоянного магнита непроводящий, и его проницаемость эквивалентна к воздуху.

(3) Ротор не имеет демпфирующих обмоток.

Machine Translated by Google

(4) Магнитное поле возбуждения от постоянного магнита и магнитное поле реакции якоря от трехфазной обмотки синусоидально распределены по воздушному зазору.

(5) Форма волны индуцированной электродвижущей силы в фазной обмотке имеет синусоидальную форму. шаблон.

Полагая, что вышеупомянутые условия выполняются, согласно электромагнитной индукции и закону напряжения Кирхгофа уравнение напряжения трехфазного СДСМ в естественной системе координат можно выразить следующим образом:

$$\begin{aligned} &= dt\ d\psi b \frac{d\psi a}{} + Rsia\ ua \\ ub &= \frac{}{dt} + Pcuc \\ uc &= dt \frac{d\psi c}{} + Rsic \end{aligned} \quad (1)$$

Уравнение напряжения статора СДСМ в двухфазной вращающейся системе координат может быть следующим: полученные с помощью преобразования Кларка и преобразования Парка [33]:

$$\begin{aligned} d\psi q \quad & \frac{d\psi d}{} + Rsid \quad \omega e\psi q\ ud = dt \\ & \frac{}{} + Rsiq + \omega e\psi d\ uq = dt \end{aligned} \quad (2)$$

Уравнение цепи магнетизма статора:

$$\begin{aligned} &\psi d = Ldid + \psi f\ \psi q = \\ &Lqiq \end{aligned} \quad (3)$$

Объединение уравнения (3) в (2) дает:

$$\begin{aligned} d\psi q \quad & \frac{d\psi d}{} + Rsid \quad \omega eLqiq\ ud = dt \\ & \frac{}{} + Rsiq + \omega e(Ldid + \psi f)\ uq = dt \end{aligned} \quad (4)$$

где ua, ub, uc — трехфазные напряжения обмотки статора, за единицу измерения — B; Rs — сопротивление статора, единицей измерения является Ом ; ia, ib, ic — трехфазный ток обмотки статора, единица измерения — A; ψa, ψb, ψc — поток обмотки статора, за единицу измерения принят Wb; ψf — потокосцепление постоянного магнита, единица измерения — Wb; Ld, Lq – индуктивные составляющие индуктивности статора по оси dq соответственно, с единицей измерения H; ωe представляет собой электрическую угловую скорость двигателя, единицей измерения является рад/с.

Путем анализа компонентов крутящего момента сопротивления и динамического крутящего момента выводится уравнение механического движения двигателя:

$$\frac{d\omega m}{Дж} = Te \quad B\omega m \quad T\ dt \quad (5)$$

где J — момент инерции, T — момент нагрузки, а B — коэффициент вязкого трения.

С помощью преобразования Кларка и преобразования Парка уравнение электромагнитного момента получается ниже:

$$Te = \frac{3}{2} pn\ \psi f\ iq + (Ld \quad Lq)idiq \quad (6)$$

где Te – электромагнитный момент; 3/2 – коэффициент при применении принципа равноамплитудного преобразования; Pn относится к общему количеству пар полюсов двигателя.

Преобразованное уравнение механического движения:

$$\frac{d\omega m}{Дж\ дт} \text{ "=" } \frac{3}{2} pn\ \psi f\ iq + (Ld \quad Lq)idiq \quad B\omega m \quad T \quad (7)$$

3. Разработка улучшенного контроллера ADRC.

Традиционный подход ADRC рассматривает неопределенность, внутренние и внешние нарушения системы как общие нарушения системы. И это может достичь реального

Machine Translated by Google

точная по времени оценка и компенсация этого возмущения с использованием ESO и NLSEF. Этот подход не опирается на точную модель объекта и значительно расширяет возможности защиты от помех. Однако функция fal традиционного ADRC недостаточно плавна в точках переключения между нелинейным и линейным участками, в результате чего эффект управления не является идеальным. Поэтому в этой статье предлагается улучшенный контроллер ADRC, основанный на улучшенной функции падения .

3.1. Улучшенный дизайн функции Fal

Функция fal — нелинейная функция, которая играет жизненно важную роль в ADRC. Это не только позволяет ADRC лучше оценивать и компенсировать помехи в системе, тем самым улучшая характеристики защиты от помех, но также позволяет уменьшить усиление, когда ошибка велика, и увеличить, когда ошибка мала. Выгодно повысить скорость реакции системы на большие ошибки и точность управления на небольшие ошибки, одновременно повышая надежность и адаптируемость системы.

В традиционном ADRC fal(e, α, δ) как важнейшая нелинейная функция может эффективно оценивать как внутренние, так и внешние возмущения системы, тем самым генерируя выходной сигнал ADRC. Форма функции fal показана в уравнении (8):

$$f\,al(e, \alpha, \beta) = \begin{cases} |e|^{\alpha} \text{знак}(e), & |e| \quad \delta, \\ \dfrac{e}{\delta^{1-\alpha}}, & |e| < \delta \end{cases} \tag{8}$$

При δ = 0,001, когда α принимает разные значения, кривая функции fal показана на рисунке 2.

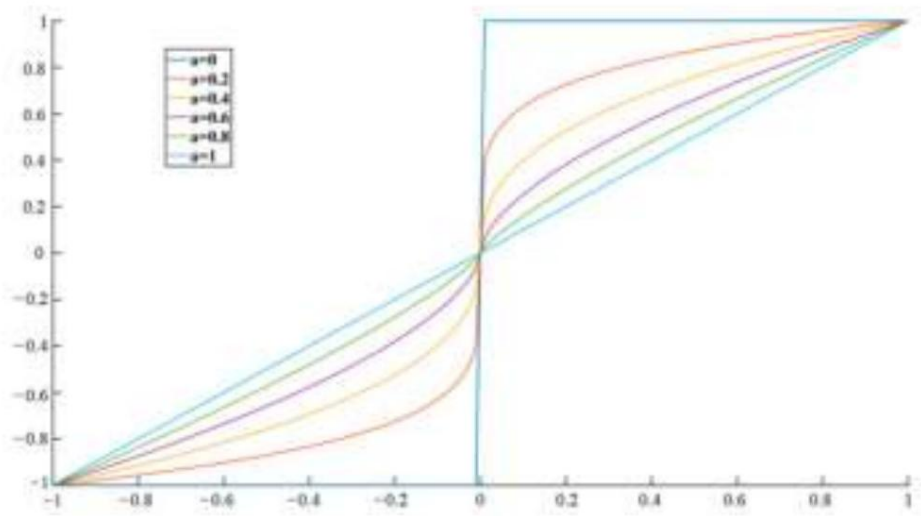


Рис. 2. Кривые функции fal(e, α, δ) при различных значениях α .

С точки зрения конвергенции терминальный аттрактор $|e|^{\alpha}$ знак(e) сходится за конечное время , особенно подходит для управления вблизи начальной стадии (|e| < 1), что оказывает усиливающее влияние на ошибку, но не подходит для управления вдали от исходной стадии, и сходимость будет медленнее, чем при классическом линейном управлении. Линейный член сходится экспоненциально, что подходит для управления вдали от начала координат (|e| > 1).
Однако в функции fal использование терминального аттрактора и линейного члена поменяно местами, что приводит к сходимости функции fal в целом за неопределенное время, независимо от того, насколько мала начальная ошибка e(0), поэтому линейный член может быть улучшен для повышения способности сходимости.

Из уравнения (8) очевидно, что значения производной в точках излома различаются после применения производной функции fal . Из рисунка 2 также видно, что кривые функции fal(e, α, δ) являются непрерывными, но не гладкими, когда e = |δ|, и происходит внезапное изменение, которое приведет к ухудшению управляемости системы. .

Machine Translated by Google

На основе поставленных задач в данной работе предлагается полиномиальная нелинейная функция Ifal(e, α, δ) , объединяющая линейные и тригонометрические функции. Причины принятия комбинации тригонометрических функций и полиномов заключаются в следующем: (1) Тригонометрические функции и полиномы имеют явные математические формы, что делает с ними легко справиться аналитически и теоретически.

(2) Правильно разработанные тригонометрические функции и полиномы могут уменьшить перерегулирование системы и повысить скорость отклика и стабильность системы.

(3) Хотя добавленная тригонометрическая функция делает Ifal более сложным, чем исходная функция fal , эффективность вычислений может быть повышена с помощью справочной таблицы и методов интерполяции, а лучший эффект управления может быть достигнут при условии обеспечения реального времени. производительность и более низкие требования к оборудованию.

Выражение функции показано ниже:

$$I\,f\,al(e,\alpha,\beta)=\begin{cases}\dfrac{|e|^{\alpha}\,знак(e)\,k1}{\delta^{\alpha-3}e\cos(e-\delta)\,k1}+\dfrac{\delta^{\frac{\alpha}{e}-2}\,k2\cos(e-\delta)}{}+k3\delta+\alpha\delta\alpha-1\,e, & |e|<\delta k2\end{cases} \tag{9}$$

Идея построения функции заключается в следующем: учитывая, что основная проблема исходной функции fal состоит в том, что сегменты недостаточно гладкие, исходная линейная функция заменяется полиномиальной функцией, объединенной с тригонометрической функцией и линейной функцией при |e | δ в этой статье. Затем, согласно условию непрерывности функции fal на отрезках и равенства значений производной функции, получают связь между тремя постоянными коэффициентами при k1, k2 и k3

Исходя из изложенных представлений, можно знать, что при e = |δ| значения функции по обе стороны от функции Ифал равны, а после взятия производной функции Ифал значения производных с обеих сторон также равны .

С учетом вышеизложенных концепций устанавливается следующее уравнение:

$$\begin{aligned}I\,f\,al(e,\alpha,\beta)&=I\,f\,al+(e,\alpha,\beta)\\ I\,f\,al'(e,\alpha,\beta)&=I\,f\,al'+(e,\alpha,\beta)\end{aligned} \tag{10}$$

Производное выражение функции Ifal после расчета имеет следующий вид:

$$I\,f\,al'(e,\alpha,\beta)=\begin{cases}\alpha\delta(\alpha-1) & \\ \delta^{\alpha-3}e(k1-1)+\delta^{\alpha-2}e(k2-1)+\alpha\delta(\alpha-1), & |e|>\delta\end{cases} \tag{11}$$

Объединив уравнения (10) и (11), мы можем получить уравнение (12).

$$\frac{\delta^{\alpha-3}ek1\cos(e-\delta)}{k1}+\frac{\delta^{\frac{\alpha}{e}-2}k2\cos(e-\delta)+}{}=\frac{\delta^{\alpha-3}e(k1-1)-\delta^{\alpha-2}e_{\alpha\delta(\alpha-1)}}{k3\delta+\alpha\delta\alpha-1\,e=|e|\,k2\,(k2-1)+\alpha\delta(\alpha-1)}\,знак(e)^{\alpha} \tag{12}$$

Когда |e| = δ, уравнение (3) упрощается до следующего вида:

$$\frac{\delta\,k1+\alpha-3}{k1}+\frac{\delta\,k2+\alpha-2}{\delta}+k3\delta+\alpha\delta=\delta\,k2\,\delta\,k1+\alpha-3\,{}^{\alpha}\quad k2+\alpha-2+\alpha\delta\alpha-1=\alpha\delta\alpha-1 \tag{13}$$

Затем выводится соотношение для k1, k2 и k3 : k2 = k1 1, k3 = k1 (k1 1) , которое вводится в функцию Ifal для получения $\frac{k1\,2\quad k1+1}{\text{новой нелинейной}}$ α, функции:

$$I\,f\,al(e,\alpha,\delta)=\begin{cases}\dfrac{|e|^{\alpha}\,знак(e)}{\delta^{\alpha-3}ek1\cos(e-\delta)\,k1}+\dfrac{\delta^{\frac{\alpha}{e}-2}\,k1-1\cos(e-\delta)\,\delta\,k1+\alpha-3}{}+\alpha\delta\alpha-1\,e, & |e|>\delta\\ k1\,e,1|e|\quad\delta+(1-\alpha)\delta\,\alpha+k1-1 & \end{cases} \tag{14}$$

Кривые функции fal до и после улучшения нарисованы в MATLAB 2020a . из MathWorks и показаны на рисунке 3.

Machine Translated by Google

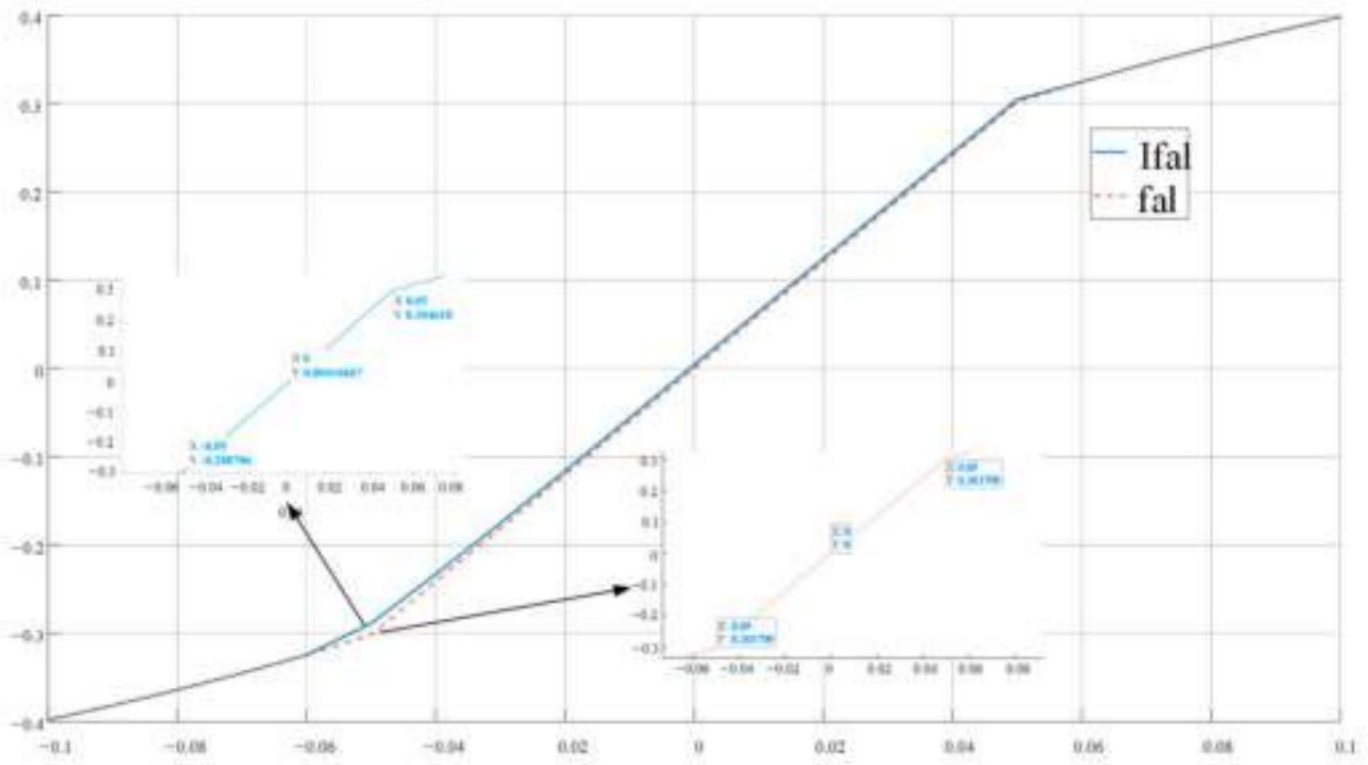


Рис. 3. Кривые функции fal(e, α, δ) до и после улучшения.

Из графика видно, что кривая функции Ifal(e, α, δ) более гладкая, чем кривая функции fal . Функция Ифала непрерывна и выводима в тех точках, где |e| = δ. В диапазоне (-0,06, 0,06) выходной сигнал функции Ifal выше, чем у функции fal , то есть функция Ifal сходится быстрее и более плавно.

3.2. Улучшенный дизайн ESO (IESO)

Более того, IESO служит ядром IADRC. Он не только обеспечивает оценку состояния системы PMSM, но также классифицирует все типы как внутренних, так и внешних возмущений, в дополнение к неопределенностям, охватывающим всю совокупность возмущений системы PMSM , а также оценивает и компенсирует общее возмущение в реальном времени. Подобно ESO, IESO полагается исключительно на входные и выходные данные системы и не требует точной математической модели возмущений. Такой подход эффективно повышает производительность и надежность системы управления. Формулировка IESO следующая:

$$\begin{array}{l} e2 = z21 \quad n = bu(t) \\ \dot{z}_{21} + z22 \quad k31 \, I\,fal(e2, \alpha2, \delta2) = \quad k32 \, I\,fal(e2, \alpha2, \delta2) \\ \dot{z}_{22} \end{array} \tag{15}$$

3.3. Улучшенный дизайн NLSEF (INLSEF)

NLSEF реализует эффективный контроль нелинейности, неопределенности и внешних возмущений в ADRC и повышает эффективность и надежность управления системы. Функция fal является ключевым нелинейным элементом NLSEF для обработки сигналов ошибок. Он может адаптивно управлять усилением в зависимости от величины ошибки, уменьшая коэффициент усиления в случаях существенной ошибки и увеличивая коэффициент усиления, когда ошибка минимальна, тем самым повышая точность и надежность механизма управления. Выражение INLSEF следующее:

$$\begin{array}{l} e1 = z11 \quad z21 \, z22 + \\ и\,(т) = - \quad \frac{k2}{I\,fal(e1, \alpha1, \delta1)\, b} \end{array} \tag{16}$$

Machine Translated by Google

где z21 представляет собой наблюдение динамической реакции контроллера, а z21 обозначает наблюдение возмущения, b — коэффициент компенсации, а k2 — коэффициент усиления регулятора.

3.4. Проектирование системы управления скоростью PMSM на основе предложения IADRC

На рисунке 4 изображена принципиальная схема системы управления скоростью PMSM, предложенной IADRC. Контроллер косвенно управляет скоростью и крутящим моментом PMSM, управляя выходным сигналом инвертора. В системе используется двойное управление с обратной связью, при котором токовая петля работает под управлением PI, а петля скорости управляется согласно предложению IADRC.

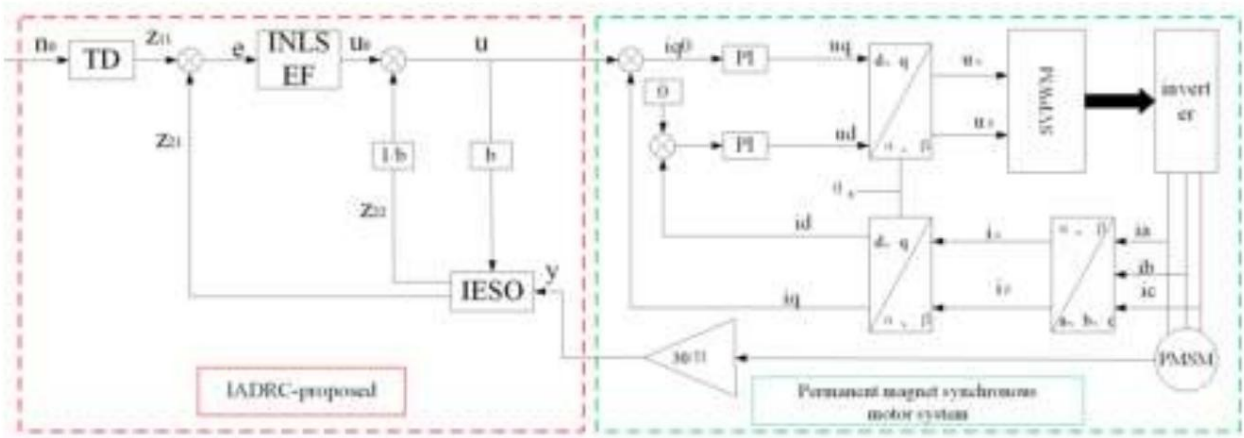


Рисунок 4. Схема, показывающая систему контроля скорости, предложенную PMSM IADRC.

4. Моделирование и экспериментальная оценка

Чтобы проверить влияние алгоритма, предложенного IADRC, на управление скоростью PMSM, традиционные ADRC, SMC, нечеткий PID и IADRC в Ref. В качестве алгоритмов сравнения приняты работы [34] . Модели системы регулирования скорости PMSM на основе этих пяти алгоритмов управления строятся в MATLAB/Simulink, после чего проводятся сравнительные эксперименты . Технические характеристики PMSM в имитационных моделях:
изложено в таблице 1.

Таблица 1. Технические характеристики СДСМ.

Технические характеристики	Ценить
Фазовое сопротивление статора R (Ом)	2,785
Индуктивность статора по прямой оси LD (Гн)	8,5 × 10–3 8,5
Поперечная индуктивность статора LQ (Гн)	× 10–3
Потокосцепление постоянного магнита ψf (Вб)	0,175 1
Инерция J (кг·м2/рад)	× 10–3
Пары полюсов, P	4
Коэффициент трения B	1 × 10 4

Методология, включенная в эту рукопись, изложена следующим образом: ожидаемая скорость PMSM составляет 1500 об/мин. Чтобы проверить надежность алгоритма, предложенного IADRC , сначала добавляется момент нагрузки. Когда фактическая скорость достигает ожидаемой скорости и становится стабильной, к выходному валу PMSM внезапно прикладывается момент нагрузки 2 Н·м за 0,1 с, чтобы сравнить кривую реакции на холостом ходу и противоинтерференционную способность PMSM под воздействием. из вышеперечисленных пяти алгоритмов управления. Общее время эксперимента составляет 0,15

Настройки параметров в алгоритме, предложенном IADRC, следующие. Согласно конкретным описаниям трех компонентов ADRC в разделе 3, два параметра α, δ в TD и NLSEF устанавливаются на основе опыта: α1 = 0,4, δ1 = 0,09, α2 = 0,4, δ Два параметра в ESO определяются посредством непрерывного моделирования и настройки: α3 = 0,25, δ3 = 0,04. Коэффициент скорости k1 в TD, коэффициент усиления регулятора k2 и коэффициент компенсации b NLSEF, полученные из математической модели и параметров PMSM.

Machine Translated by Google

системы принимаются равными k1 = 5355, k2 = 5000, b = 1030,66 соответственно. Аналогично, калибровка приросты продукции в ЭСО принимаются равными k31 = 8500, k32 = 500 000. Ключевые параметры в контроллер показаны в таблице 2.

Таблица 2. Значения основных параметров контроллера.

Компонент	Технические характеристики	Ценить
ТД	Коэффициент отслеживания α1	0,4
	Коэффициент фильтра δ1	0,09
	Коэффициент скорости k1	5355
НЛСЕФ	Коэффициент отслеживания α2	0,4
	Коэффициент фильтра δ2	0,09
	Коэффициент усиления регулятора k2	5000
	Компенсационный коэффициент b	1030,66
ESO	Фактор отслеживания α3	0,25
	Коэффициент фильтра δ3	0,04
	Калибровочные коэффициенты выхода k31	8500
	Калибровочные коэффициенты выхода k32	500 000

В соответствии с указанными выше настройками параметров три алгоритма применяются к Регулятор скорости PMSM. Результаты эксперимента показаны ниже:

На рисунках 5 и 6 видно, что на начальном этапе запуска PMSM IADRC предложенный в данной статье имеет лучший эффект среди трех алгоритмов. Оно не только не имеет перерегулирование, но также достигает и стабилизируется на ожидаемой скорости 0,027 с. Максимум Значение отклонения ошибки составляет 31,28 об/мин, что составляет около 2,09% от ожидаемой скорости. IADRC в исх. [34] требуется 0,045 с, чтобы достичь и стабилизировать ожидаемую скорость, а Перерегулирование составляет 61,64 об/мин, что составляет около 4,11% от ожидаемой скорости. И время для традиционное ADRC для достижения и стабилизации ожидаемой скорости составляет 0,085 с, а его превышение также является самой большой и составляет 119,04 об/мин, что составляет около 7,89% от ожидаемой скорости. СМК требуется 0,075 с для достижения и стабилизации ожидаемой скорости с максимальным отклонением 423,71 об/мин, что составляет примерно 28,25% от ожидаемой скорости. Нечеткий ПИД принимает 0,043 с для достижения и стабилизации ожидаемой скорости с максимальным отклонением ошибки 30,27 об/мин, что составляет около 2,02% от ожидаемой скорости.

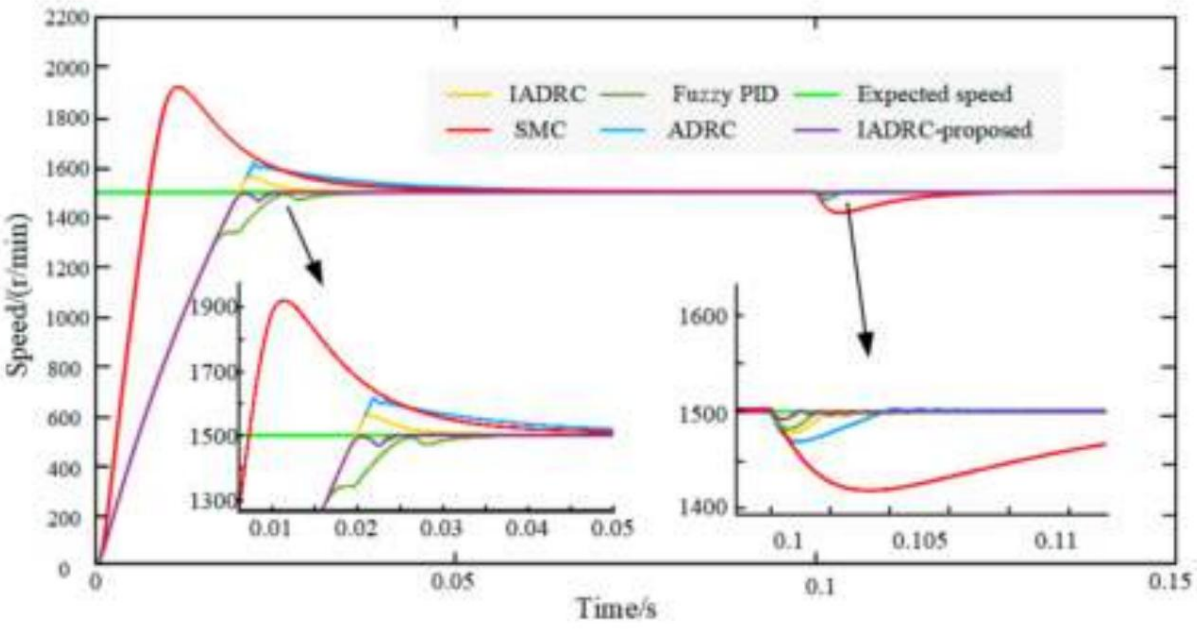


Рисунок 5. Кривые скоростного отклика 5 алгоритмов без гауссовского шума.

Machine Translated by Google

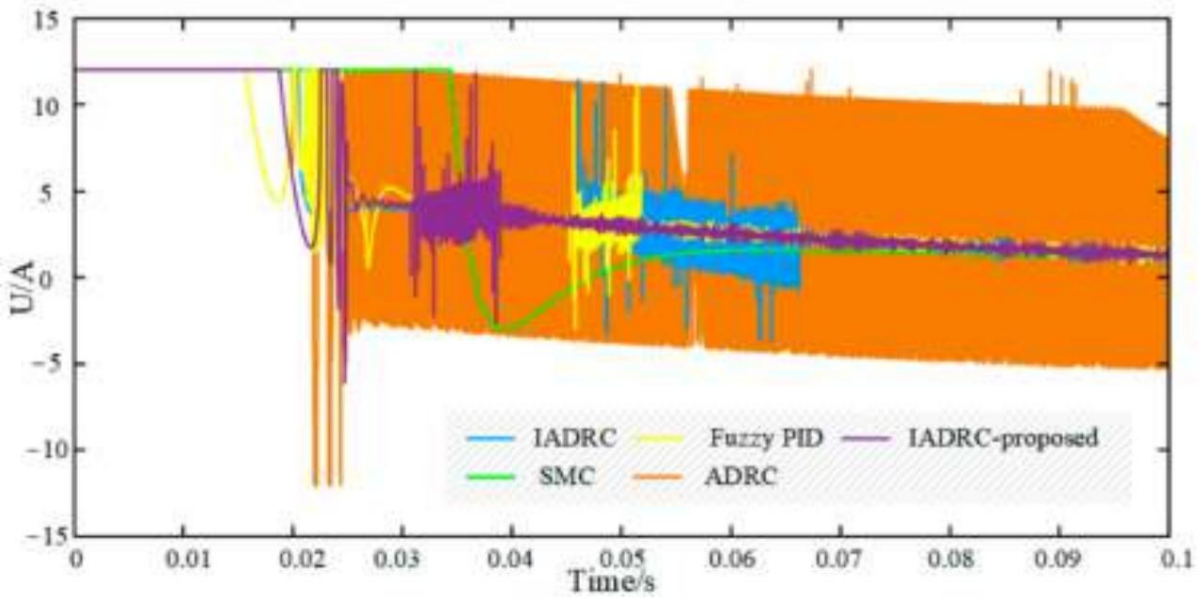


Рисунок 6. Входные данные управления объектом для 5 алгоритмов без гауссовского шума.

Когда скорость PMSM стабилизируется на ожидаемом уровне, момент ступенчатой нагрузки прикладывается через 0,1 с. Алгоритм IADRC, предложенный в этой статье, практически не затрагивается, а кривая скорости колеблется лишь незначительно. Максимальная погрешность скорости составляет 7,39 об/мин, что составляет около 0,49% от ожидаемой скорости, а затем быстро (0,004 с) возвращается к ожидаемому значению. Кривая скорости алгоритма IADRC в Ref. [34] выявили максимальную ошибку скорости 20,73 об/мин, что составляет около 1,38% от ожидаемой скорости, и вернулись к ожидаемому значению через 0,006 с. А кривая скорости традиционного алгоритма ADRC дает максимальную ошибку скорости 29,85 об/мин, что составляет около 1,99% от ожидаемой скорости, и возвращается к ожидаемому значению через 0,034 с. Кривая скорости алгоритма SMC дает максимальную ошибку скорости 81,98 об/мин, что составляет примерно 5,47% от ожидаемой скорости, и возвращается к ожидаемому значению через 0,028 с. Кривая скорости нечеткого ПИД-алгоритма дает максимальную ошибку скорости 18,11 об/мин, что составляет около 1,21% от ожидаемой скорости, и возвращается к ожидаемому значению через 0,07 с.

Различные неопределенные факторы влияют на систему PMSM в практической работе, например, шум окружающей среды и изменения сопротивления, вызванные непрерывной работой двигателя или изменениями температуры окружающей среды. Чтобы проверить надежность алгоритма, предложенного IADRC, при различных изменениях параметров и неопределенностях, в систему добавлены эксперименты по моделированию шумовых помех и изменения сопротивления двигателя.

Чтобы проверить надежность алгоритма, предложенного IADRC, в эту статью также включены сравнительные эксперименты с входным сигналом управления объектом U в условиях гауссовского шума, сравнивающие его с другими алгоритмами. В начале добавляется гауссовский шум, и для сигнала устанавливается среднее значение 1 и дисперсия 200. Другие параметры двигателя показаны в таблицах 1 и 2. Результаты моделирования показаны на рисунках 7 и 8.

Сравнивая рисунки 5–8, очевидно, что кривые скоростного отклика и выход системы U существенно не изменяются до и после добавления гауссовского шума для всех пяти алгоритмов. Кроме того, кривые скоростного отклика и выход системы U , управляемые алгоритмом, предложенным IADRC, имеют наименьшее время колебаний и наименьшую амплитуду колебаний, что подтверждает стабильность алгоритма, предложенного IADRC.

Для дальнейшей проверки влияния изменений сопротивления двигателя на предложенные IADRC значения сопротивления двигателя установлены на 2,5875, 2,875 и 3,1625 соответственно. Остальные параметры соответствуют таблице 1. Результаты эксперимента показаны на рисунке 9.

Machine Translated by Google

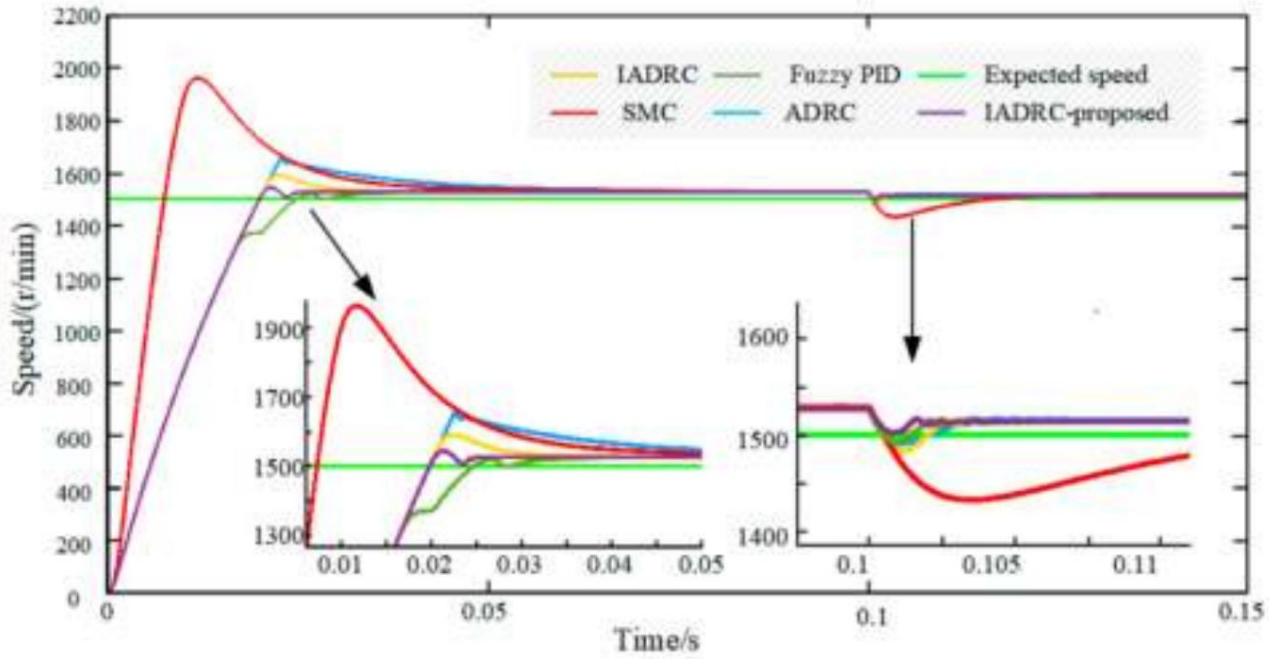


Рисунок 7. Кривые скоростного отклика 5 алгоритмов при гауссовском шуме.

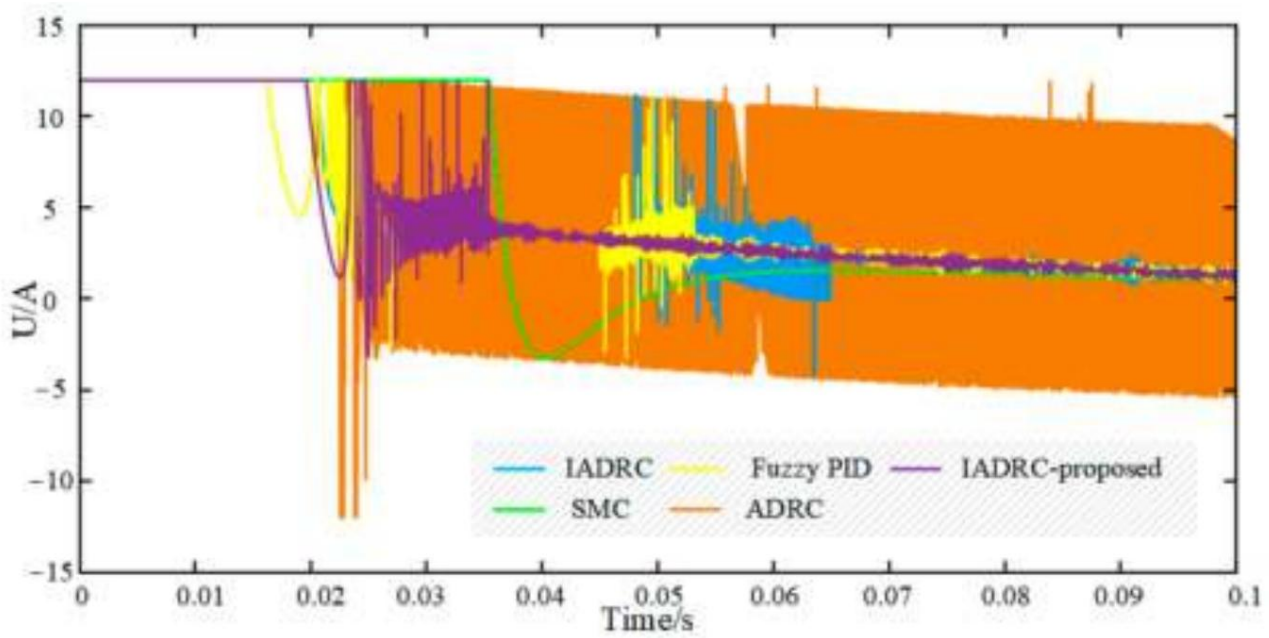


Рисунок 8. Входные данные управления объектом для 5 алгоритмов в условиях гауссовского шума.

Из рисунка 9 очевидно, что, хотя перерегулирование и время достижения стабильности изменяются, изменения незначительны. Это указывает на то, что после изменения внешних условий эффект управления все еще очень хороший, что подтверждает устойчивость алгоритма, предложенного IADRC .

Таким образом, фактически доказано, что представленный нами IADRC демонстрирует повышенную оперативность, более высокую точность управления, более высокую надежность и способность защиты от помех. Конкретные характеристики таковы: наш собственный алгоритм IADRC на 40,00% лучше, чем IADRC в Ref. [34], с точки зрения быстродействия на 68,24% лучше, чем традиционный ADRC, на 64,00% лучше, чем SMC, и на 37,21% лучше, чем нечеткий PID. С точки зрения точности он на 49,25% лучше, чем IADRC в работе. [34], на 73,72% лучше, чем традиционный ADRC, на 92,62% лучше, чем SMC, и на 3,35% хуже, чем нечеткий PID. Более того, когда момент ступенчатой нагрузки

Machine Translated by Google

При применении в состоянии стабильной работы алгоритм IADRC, предложенный в этой статье, на 33,33% лучше, чем алгоритм IADRC в работе. [34], с точки зрения быстродействия на 88,24% лучше, чем традиционный ADRC, на 85,71% лучше, чем SMC, и на 94,29% лучше, чем нечеткий PID. С точки зрения точности он на 64,35% лучше, чем IADRC в работе. [34], на 75,24% лучше, чем традиционный ADRC, на 90,99% лучше, чем SMC, и на 59,19% лучше, чем нечеткий PID. Кроме того, экспериментальные тесты с добавлением гауссовского шума и изменением сопротивления двигателя подтверждают, что предложенный алгоритм IADRC имеет хорошую надежность.

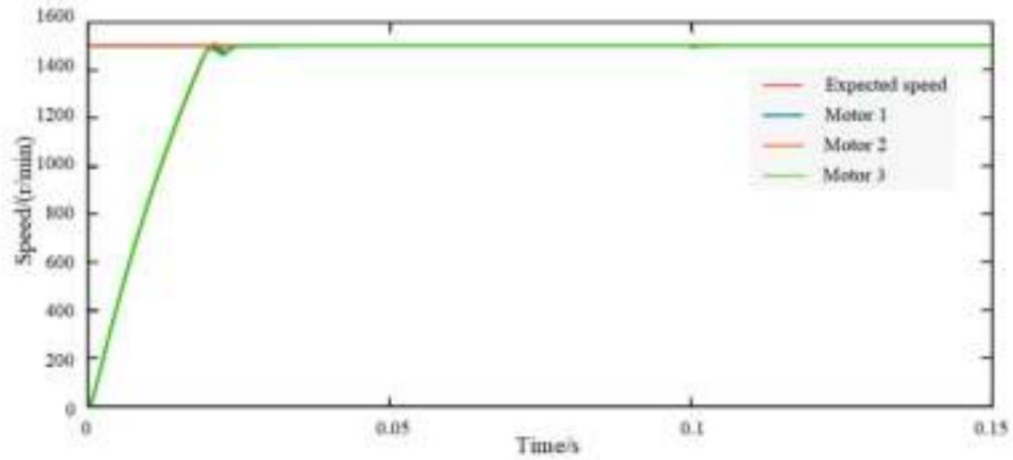


Рисунок 9. Кривые скорости двигателя при различных сопротивлениях.

5. Выводы

В этой статье поясняется усовершенствованная стратегия IADRC, основанная на функции падения , и достигается лучшая эффективность управления для PMSM по сравнению с традиционным ADRC, IADRC в Ref. [34], SMC и нечеткий ПИД. Результаты моделирования показывают, что по сравнению с ADRC, IADRC в Ref. [34], SMC и нечеткого PID, IADRC, предложенный в этой статье, имеет более высокую скорость отклика. Время достижения и стабилизации ожидаемой скорости на 68,24% короче, чем у ADRC, и на 40,00% меньше, чем у IADRC в ссылке. [34], на 64,00% короче, чем у SMC и на 37,21% короче, чем у нечеткого PID. При этом предложенный в статье IADRC не имеет перерегулирования, а максимальное снижение ошибки составляет 73,72% по сравнению с ADRC и 49,25% по сравнению с IADRC в работе. [34], 92,62% по сравнению с SMC и -3,35% по сравнению с нечетким PID. В случае внезапного возмущения, когда скорость PMSM стабилизируется на ожидаемой скорости, время достижения и стабилизации ожидаемой скорости на 88,24% короче, чем у ADRC, на 33,33% меньше, чем у IADRC в ссылке. [34], на 85,71% короче, чем у SMC и на 94,29% короче, чем у нечеткого PID. Кроме того, максимальная ошибка IADRC, предложенная в этой статье, на 75,24% ниже, чем у ADRC, и на 64,35% ниже, чем у IADRC в Ref. [34], на 90,99% ниже, чем у SMC и на 59,19% ниже, чем у нечеткого PID. Кроме того, хорошая надежность также проверяется путем добавления гауссова шума и изменения сопротивления двигателя.

Наша будущая работа будет сосредоточена на применении алгоритма, предложенного IADRC, к физическому управлению СДСМ и дальнейшему совершенствованию алгоритма в соответствии с фактической ситуацией управления.

Вклад автора: Написание — подготовка оригинального черновика, ZH; написание — обзор и редактирование, YC; ресурсы, CK; надзор, XZ, JX и SW. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Исследование финансировалось Национальным фондом естественных наук Китая (грант № 52205316) и Китайским фондом естественных наук провинции Чжэцзян (грант № LTGN24E050002). А APC финансировался крупным научно-техническим инновационным проектом Китая в Вэньчжоу в рамках гранта номер ZN2022002.

Machine Translated by Google

Заявление о доступности данных: оригинальные материалы, представленные в исследовании, включены в статью, дальнейшие запросы можно направить соответствующему автору.

Благодарности: Авторы благодарны редакторам и анонимным рецензентам за содержательные комментарии и предложения.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Болоньяни, С.; Тубиана, Л.; Зильотто, М. Расширенная настройка фильтра Калмана в бездатчиковых приводах СИММ. IEEE Транс. Индийское прил. 2003, 39, 1741–1747. [Перекрестная ссылка]
2. Уоллмарк, О.; Харнефорс, Л.; Карлсон О. Алгоритмы управления отказоустойчивым приводом СИММ. IEEE Транс. Индийский Электрон. 2007, 54, 1973–1980. [Перекрестная ссылка]
3. Гашти А.; Акбаримайд А. Проектирование противозавершающего ПИ-регулятора для ЛПЧ нелинейной энергосистемы, совмещенной с ДСТ АЭС и звеном постоянного тока. Электр. англ. 2020, 102, 793–809. [Перекрестная ссылка]
4. Эрруисси, Р.; Аль-Дурра, А.; Муйин, С.М. Экспериментальная проверка нового ПИ-регулятора скорости для приводов двигателей переменного тока с улучшенными переходными характеристиками. IEEE Транс. Система управления. Технол. 2018, 26, 1414–1421. [Перекрестная ссылка]
5. Заки, М.; Хатер, М.; Ясин, Х.; Шокралла С. Оценка очень низких и нулевых скоростей бездатчиковых асинхронных электроприводов. Электр. Система питания Рез. 2010, 80, 143–151. [Перекрестная ссылка]
6. Ши, Р. Улучшение алгоритма прогнозирующего управления на основе ПИД-регулятора нечеткого дробного порядка. Дж. Интелл. Сист. 2023, 32, 876. [CrossRef]
7. Тан, В.; Хан, В.; Сюй, Дж. ПИД в пространстве состояний: недостающее звено между классическим и современным управлением. IEEE Доступ 2022, 10, 116540–116553. [Перекрестная ссылка]
8. Аталик, Г.; Сентурк, С. Интуиционистские нечеткие контрольные диаграммы, основанные на интуиционистском методе нечеткого ранжирования для TIFN. Мягкий компьютер. 2022, 26, 11403–11407. [Перекрестная ссылка]
9. Ту, WW; Донг, Дж. К. Устойчивое управление скользящим режимом для класса нелинейных систем с помощью двухслойной схемы скользящего режима. Дж. Франкл. Инст. англ. Прил. Математика. 2023, 360, 10227–10250. [Перекрестная ссылка]
10. Чжан, GC; Ли, XF; Ся, ЮК; Он, С. Адаптивное управление скользящим режимом для двумерной нелинейной модели Форназини-Марчезини с учетом квантования и выпадения пакетов. Межд. Дж. Сист. наук. 2021, 52, 3001–3012. [Перекрестная ссылка]
11. Чжао Дж.; Чжао, Т.; Лю, Н. Активное управление подавлением возмущений дробного порядка с нечеткой самонастройкой для прецизионной стабилизированной платформы. Энтропия 2022, 24, 1681. [CrossRef] [ПабМед]
12. Сян, CD; Петерсен, ИК; Донг, Д.Ю. Когерентное робастное H - управление неопределенными линейными квантовыми системами с прямыми и косвенными муфты. Дж. Франкл. Инст. англ. Прил. Математика. 2023, 360, 13845–13869. [Перекрестная ссылка]
13. Хан Дж. От ПИД-регулятора к активному управлению подавлением возмущений. Пер. Индийский Электрон. 2009, 56, 900–906. [Перекрестная ссылка]
14. Лю, Джей Джей; Чен, Х.; Ван, Л.; Ю, М.Ю. Активное подавление помех для усовершенствованной глубинной модели АНПА. Прил. Мех. Матер. 2014, 687, 157–162. [Перекрестная ссылка]
15. Мэн, Ю.; Лю, Б.; Ван, Л. Управление скоростью PMSM на основе оптимизированного контроллера ADRC. Математика. Пробл. англ. 2019, 2019, 18. [Перекрестная ссылка]
16. Сюн, С.; Се, Х.; Сонг, К.; Чжан, Г. Метод отслеживания скорости для автономного вождения через ADRC с расширенным наблюдателем состояния. Прил. наук. 2019, 9, 3339. [CrossRef]
17. Ван, Дж.; Лю, Х.; Юань, Дж.; Шен, Ю.; Чжан, Х.; Ван, Х.; Чжэн Ю. Управление движением автономного подводного аппарата на основе Дробное исчисление. Активное подавление возмущений. Дж. Мар. англ. 2021, 9, 1306. [CrossRef]
18. Ян, З.; Цянь, К.; Солнце, Х.; Ван, К. Расширенная стратегия линейного ADRC для бездатчикового управления IPMSM с учетом перекрестных Факторы связи. IEEE Транс. Трансп. Электриф. 2023, 9, 4437–4446. [Перекрестная ссылка]
19. Цзо, Ю. Ф.; Чжан, Дж.; Лю, К.; Чжан Т. Интегрированная конструкция сервосистем синхронных двигателей с постоянными магнитами на основе активных управление подавлением помех. Пер. Китай Электротех. Соц. 2016, 31, 51–58. [Перекрестная ссылка]
20. Фанг, С.; Мэн, Дж.; Мэн, Ю.; Ван, Ю.; Хуанг, Д. Активное управление током подавления возмущений в дискретном времени в двигателе с постоянными магнитами Низкая скорость с использованием режима резонансного скольжения. IEEE Транс. Трансп. Электриф. 2023, 9, 4783–4794. [Перекрестная ссылка]
21. Ли, З.; Чжан, З.; Ван, Дж.; Ван, С.; Чен, Х.; Сан, Х. Система управления ADRC PMLSM на основе нового наблюдателя со скользящим режимом несингулярного терминала . Энергии 2022, 15, 3720. [CrossRef]
22. Ге, Х.; Сан, К. Оптимальное управление космическим кораблем с развертываемыми солнечными батареями с использованием алгоритма оптимизации роя частиц. наук. Китайские технологии. наук. 2011, 54, 1107–1112. [Перекрестная ссылка]
23. Ламрауи, ХК; Цидан З. Управление отслеживанием скорости мобильного робота одноколесного типа на базе LADRC. В материалах 3-й Международной конференции IEEE по науке управления и системной инженерии (ICCSSE) 2017 г., Пекин, Китай, 17–19 августа 2017 г.; стр. 200–204. [Перекрестная ссылка]
24. Партовибахш, М.; Лю, Г. Оценка коэффициента скольжения и управление колесным мобильным роботом на различных поверхностях. В материалах Международной конференции IEEE 2015 г. по кибертехнологиям в автоматизации, управлении и интеллектуальных системах (CYBER), Шэньян, Китай, 8–12 июня 2015 г.; стр. 566–571. [Перекрестная ссылка]
25. Дин, Л.; Гао, Х.-Б.; Дэн, Z.-Q.; Ли, З.; Ся, К.-Р.; Дуань, Г.-Р. Следование по траектории колесных планетарных роботов-исследователей передвижение по деформируемой пересеченной местности. наук. World J. 2014, 2014, 793526. [CrossRef] [ПабМед]
26. Фэн, Х.; Го, Б.-З. Активное управление подавлением возмущений: старые и новые результаты. Анну. Рев. Контроль 2017, 44, 238–248. [Перекрестная ссылка]

Machine Translated by Google

27. Льв, С.; Ван, Б.; Чен, Дж.; Чжан, Р.; Донг, Х.; Ван, С. Исследование метода подавления пульсаций крутящего момента для управления нечетким активным подавлением возмущений для синхронного двигателя с постоянными магнитами. Электроника 2024, 13, 1280. [CrossRef]
28. Ван, Х.; Чжу, Х. Управление активным подавлением возмущений в безподшипниковом синхронном двигателе с постоянными магнитами на основе генетического алгоритма и метода динамической настройки параметров нейронной сети. Электроника 2023, 12, 1455. [CrossRef]
29. Фанг, К.; Чжоу, Ю.; Ма, С.; Чжан, К.; Ван, Ю.; Хуанфу, Х. Технология сервоуправления электромеханическими приводами на основе Активное управление подавлением помех. Электроника 2023, 12, 1934. [CrossRef]
30. Вс, К.; Лю, М.; Лю, К.; Фэн, Х.; Ву, Х. Метод управления промышленным квадрокоптером БПЛА, основанный на нечеткой адаптивной линейно-активной системе. Управление подавлением помех. Электроника 2021, 10, 376. [CrossRef]
31. Штумбергер, Б.; Штумбергер, Г.; Хаджиселимович, М.; Хамлер, А.; Горикан, В.; Есеник, М.; Трлеп, М. Сравнение крутящего момента трехфазных синхронных двигателей с постоянными магнитами с различным расположением постоянных магнитов. Дж. Магн. Магн. Матер. 2007, 316, e261–e264. [Перекрестная ссылка]
32. Чжан, XG; Фанг, С.Дж.; Чжан Х. Прогнозируемое управление с прогнозированием на основе сильной устойчивой модели с компенсацией текущих ошибок для Системы привода ПМСМ. IEEE Транс. Индийский Электрон. 2024, 1–10. [Перекрестная ссылка]
33. Витор, АЛО; Скалассара, PR; Гёдтель, А.; Эндо, В. Паттерны, основанные на преобразованиях Кларка и Парка вейвлет-коэффициентов для Классификация неисправностей электрических машин. J. Управление автомат. Электр. Сист. 2023, 34, 230–245. [Перекрестная ссылка]
34. Сюй, П. Исследование системы управления скоростью PMSM горного автотранспорта на основе нечеткой самоиммунитета. Магистерская диссертация, Аньхой Технологический университет, Аньхой, Китай, 2022 г.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Исследование и смягчение компромисса между производительностью и справедливостью с помощью выборки по защищенным категориям

Гидеон Попула и Джон Шеппард *

Школа вычислительной техники Джанфорте, Государственный университет Монтаны, Бозман, MT
59717, США; gideon.popoola@student.montana.edu
* Адрес для переписки: john.sheppard@montana.edu

Аннотация: Алгоритмы машинного обучения стали обычным явлением в повседневном принятии решений, а системы помощи при принятии решений широко распространены в нашей повседневной жизни. Следовательно, в последние годы растет количество исследований по предотвращению и смягчению потенциальной предвзятости и несправедливости прогнозов, сделанных этими алгоритмами . В большинстве исследований по вопросам справедливости и смягчения предвзятости в машинном обучении каждая защищенная переменная часто рассматривается отдельно, но на самом деле один человек может принадлежать к нескольким защищенным категориям. Таким образом, в этой работе было рассмотрено объединение набора защищенных переменных и создание новых столбцов, которые разделяют эти защищенные переменные на множес
Эти новые подкатегории, как правило, крайне несбалансированы, поэтому к смягчению предвзятости подходили как к проблеме несбалансированной классификации. В частности, были разработаны и исследованы четыре новых метода индивидуальной выборки для выборки этих новых подкатегорий. Эти новые методы выборки называются передискретизацией по защищенной категории, пропорциональной выборкой по защищенной категории, методом синтетической выборки меньшинства по защищенной категории (PC-SMOTE) и адаптивной синтетической выборкой по защищенной категории (PC-ADASYN). Эти методы выборки модифицируют существующий метод выборки, ориентируя выборку на новые подкатегории, а не на метку класса. Влияние этих стратегий выборки затем оценивалось на основе классической эффективности и справедливости в условиях классификации.
Эффективность классификации измерялась с использованием точности и F1 на основе обучающих одномерных деревьев решений, а справедливость измерялась с использованием уравненных нечетных разностей и статистической четности. Чтобы оценить влияние справедливости на производительность, эти показатели сравнивались с глубиной дерева решений. Результаты показывают, что предложенные методы смогли определить оптимальные точки, в результате чего справедливость была увеличена без снижения производительности, что смягчило любой потенциальный компромисс между производительностью и справедливостью.

Ключевые слова: справедливость; защищенные категории; машинное обучение; выборка

check for updates

Цитирование: Попула, Г.; Шеппард, Дж.
Расследование и смягчение последствий
Компромисс между производительностью и справедливостью через
Выборка по защищенной категории.
Электроника 2024, 13, 3024. https://doi.org/10.3390/
electronics13153024 .

Академические редакторы: Нюша Шафиабади и
Цзяньлун Чжоу

Поступила: 11 июня 2024 г.
Пересмотрено: 23 июля 2024 г.
Принято: 24 июля 2024 г.
Опубликовано: 31 июля 2024 г.

Копирайт: © 2024 авторов.
Лицензиат MDPI, Базель, Швейцария.
Эта статья находится в открытом доступе.
распространяется на условиях и
условия Creative Commons
Лицензия с указанием авторства (CC BY)
(https://creativecommons.org/licenses/by/
4.0/).

1. Введение

Поскольку алгоритмы машинного обучения (МО) все больше доминируют в системах принятия решений и системах помощи в принятии решений, их широкое распространение в различных секторах поднимает насущные вопросы справедливости и прозрачности их прогнозов [1]. Потенциал этих алгоритмов увековечивать или усугублять существующие социальные предубеждения побудил провести значительный объем исследований по изучению и смягчению алгоритмической несправедливости. Это очень важно, поскольку решения, на которые влияют эти алгоритмы, оказывают глубокое влияние на людей, влияя на результаты в самых разных областях: от финансов и занятости до уголовного правосудия и здравоохранения [2].

Источник несправедливости и предвзятости в сфере ОД многогранен [3]. В частности, возможно, что несправедливость возникает непосредственно из самих алгоритмов МО из-за возможного несовпадения основного индуктивного смещения алгоритмов по отношению к целевой концепции и распределению данных. Это называется алгоритмической предвзятостью. Альтернативная проблема заключается в потенциальной предвзятости, присутствующей в данных, используемых для обучения моделей, где, как прямой результат типичного допущения «независимости и одинакового распределения» (IID), используемого в больш

Machine Translated by Google

методов, результатом обучения является распространение смещения в прогнозах таким образом, чтобы они соответствовали смещению в самих базовых данных. Именно эта последняя ситуация и является предметом нашей работы.

1.1. Предвзятость и несправедливость в машинном обучении

Предвзятость — это предвзятое, несправедливое или неравное обращение с человеком или группой, основанное на конкретных характеристиках, часто называемых чувствительными или защищенными характеристиками [4]. Примеры этих защищенных функций включают возраст, расу, инвалидность, пол и пол [5]. Предвзятость в сфере ОД можно грубо разделить на несопоставимое обращение (прямая несправедливость) и воздействие (косвенная несправедливость) [6]. Прямая несправедливость возникает, когда защищенные функции явно используются при принятии решений. Косвенная несправедливость сегодня становится все более распространенной. Этот тип несправедливости не использует явно защищенные атрибуты; вместо этого это происходит, когда использование переменных, связанных с этими атрибутами, приводит к значительно различным результатам для защищенных групп. Эти другие переменные известны как прокси-функции. Примеры реальной предвзятости включают историческую американскую практику «красной линии», когда жителям почтовых индексов, населенных преимущественно меньшинствами, отказывали в выдаче ипотечного кредита, гендерную предвзятость при приеме на работу в Amazon, расовую предвзятость в отношении дозаторов мыла Google и т. д

Хотя эти инструменты помощи при принятии решений помогают автоматизировать процесс принятия решений, эти инструменты могут привести к несправедливому обращению с отдельными лицами или группами, как прямо, так и косвенно [8]. Несправедливость может возникнуть в нескольких областях моделирования, например, в наборе обучающих данных. Это может произойти, когда набор обучающих данных не обеспечивает достоверного представления защищенных категорий, поэтому становится трудно определить «основную истину». Например , рассмотрим набор данных компании, в которой определенная группа населения исторически сталкивалась с дискриминацией. В частности, предположим, что сотрудницы-женщины в этой компании не получили продвижения по службе так же, как их коллеги-мужчины, которые, напротив, добились карьерного роста, несмотря на то, что обе группы работают на одном и том же уровне. В этой ситуации истинная ценность вклада сотрудниц – основная истина – не видна. В результате алгоритм МО, обученный на этих данных, скорее всего, обнаружит и учтет эту предвзятость, тем самым закрепляя существующие предрассудки. Это может привести к тому, что алгоритм будет принимать дискриминационные решения, например, рекомендовать кандидатов-мужчин на работу или продвижение по службе чаще, чем кандидатов-женщин с такой же или более высокой квалификацией.

Другая область, где может возникнуть несправедливость, — это сам алгоритм МО [9]. Алгоритмы МО по-прежнему могут принимать дискриминационные решения, даже если они обучены на объективном наборе данных, где «основная истина» представлена точно. Такая ситуация возникает, когда ошибки системы непропорционально сильно влияют на представителей определенной группы или меньшинства. Например, рассмотрим алгоритм обнаружения рака молочной железы, который демонстрирует значительно более высокий уровень ложноотрицательных результатов у чернокожих по сравнению с белыми, что означает, что он не может чаще выявлять рак молочной железы у чернокожих пациентов, чем у белых пациентов. Если этот алгоритм будет использоваться для информирования о рекомендациях по лечению, он будет ошибочно рекомендовать не лечить большее количество чернокожих, чем белых, что приведет к расовому неравенству в результатах здравоохранения. Это подчеркивает острую необходимость гарантировать, чтобы алгоритмы работали одинаково во всех группах с точки зрения обучающих данных и того, как их ошибки влияют на различные группы населения. Результаты предыдущей литературы сообщают о нескольких случаях применения алгоритмов, приводящих к несправедливому обращению, например, «красной черте» и расовому профилированию [10], ипотечной дискриминации [11], трудоустройству и подбору персонала [12].

Хотя значительные усилия были направлены на устранение систематической ошибки в прогнозах ОД [13,14], большая часть существующих исследований была сосредоточена на смягчении систематической ошибки для отдельных защищенных атрибутов в отдельности [15]. Например, на наборе данных с двумя защищенными атрибутами, расой и полом, большинство существующих подходов могут изучить либо справедливую модель, включающую расу, либо справедливую модель, включающую пол, но не справедливую модель, включающую и расу, и пол [7]. Однако реальные личности не единичны; они сложны и многогранны: люди часто одновременно принадлежат к нескольким охраняемым группам [16]. Например, человек может подвергаться дискриминации одновременно по нескольким защищенным признакам, таким как возраст, раса и пол. Эта интерсекциональность может привести к сложным формам предвзятости и

Machine Translated by Google

дискриминации, которые не могут быть адекватно решены мерами по обеспечению справедливости с использованием одной переменной. Поэтому крайне важно разработать методологии, которые целостно рассматривают многомерный характер личной идентичности. Этот проект направлен на устранение этого разрыва путем рассмотрения комбинаций защищенных категорий, тем самым синтезируя эти защищенные категории в комплексные многокатегорийные группы, а также направлен на более эффективное решение многоуровневых сложностей предвзятости с использованием новых методов выборки защищенных категорий, тем самым признавая и обращаясь к многогранному характеру личных идентичностей и потенциальных предубеждений

Работа, представленная в этой статье, мотивирована проблемой использования алгоритмов МО для принятия решений в социально чувствительных областях, таких как оценка кредита, найм или оценка ипотеки, работая с этой ситуацией, когда человек может принадлежать к нескольким защищенным категориям. Учитывая помеченный набор обучающих данных, содержащий две или более защищенные функции, предлагаемый метод объединяет эти защищенные атрибуты, а затем разбивает их на новые мультикатегории. Эти новые категории, вероятно, будут крайне несбалансированными, и их необходимо сбалансировать, чтобы повысить справедливость прогнозов наших алгоритмов машинного обучения. Популярные методы выборки, такие как избыточная выборка [17], метод синтетической избыточной выборки меньшинства (SMOTE) [18], адаптивная синтетическая выборка (ADASYN) [19] и т. д., осуществляют выборку данных по меткам классов, что не соответствует цели наше исследование выборки по новым мультикатегориям. Следовательно, предлагается новый класс модификаций этих методов выборки, который может осуществлять выборку по новой категории, а не по меткам классов. Этот новый класс модифицированной выборки называется выборкой по защищенной категории. Предлагаемые в результате методы выборки по защищенным категориям используются для выборки и балансировки новых категорий перед выполнением классификации. Новизна данной работы двоякая. Во-первых, предлагаемый подход объединяет защищенные категории для формирования новых мультикатегорий, которые имитируют то, как личность человека выглядит в реальном мире. Во-вторых, это модификация существующих методов выборки, чтобы они соответствовали выборке этих новых категорий, чтобы гарантировать, что все новые категории имеют одинаков

Только в демонстрационных целях в качестве алгоритма классификации было выбрано одномерное дерево решений. Цель состоит в том, чтобы продемонстрировать влияние различных методов выборки на производительность, ожидая, что аналогичные тенденции будут проявляться независимо от основного метода обучения. Предложенный метод выборки сравнивался с базовым уровнем (несемплированные данные) с использованием точности и F1 в качестве показателей эффективности классификации, а также уравненных различий шансов и статистической паритета в качестве показателей справедливости. Кроме того, было проведено несколько анализов, чтобы показать, как максимальная глубина дерева решений влияет на точность и справедливость.

1.2. Вопрос исследования

Исходя из эмпирических наблюдений о том, что иногда существует компромисс между справедливостью и эффективностью ОД [20], это исследование пытается ответить на несколько вопросов, например, как можно смягчить этот компромисс. В частности, мы пытаемся ответить, может ли метод выборки по защищенным категориям для достижения справедливости смягчить этот компромисс. Кроме того, можем ли мы разработать методологическую основу, которая эффективно смягчает предвзятости по этим объединенным защищенным переменным, не ставя под угрозу точность прогнозирования моделей МО? Наконец, мы планируем ответить на вопрос о том, как глубина дерева решений влияет на показатели точности и справедливости, исследуя тем самым взаимосвязь между уровнем соответствия (недооснащение через переоснащение) и справедливостью.

1.3. Гипотеза

Мы предполагаем, что, используя сложные методы выборки по защищенным категориям, разработанные для этих недавно сформулированных многокатегорийных групп, мы можем значительно повысить справедливость модели с точки зрения уравненных различий в шансах, не снижая при этом производительность классификации с точки зрения точности и F1. Кроме того, мы исследуем хрупкий баланс между справедливостью и точностью, предполагая, что можно определить стратегические точки, где справедливость может быть максимизирована без пагубного воздействия на производительность.

Это исследование бросает вызов существующим утверждениям о существовании компромиссов в вопросах справедливости и ОД.

Machine Translated by Google

прогноз. Это закладывает основу для будущих исследований многомерной природы идентичности и дискриминации в автоматизированных системах принятия решений.

1.4. Взносы

Широкая проблема справедливости в машинном обучении важна тем, что распространенность систем искусственного интеллекта и машинного обучения сегодня оказывает серьезное влияние на жизнь и средства к существованию люде
Несмотря на то, что внимание к добросовестному отмыванию денег существенно возросло, по-прежнему существует потребность в методах продвижения добросовестного отмывания денег без негативного влияния на эффективность ОД. Основываясь на углубленном обзоре литературы и указанной выше потребности в такого рода работе, представленные здесь методы вносят

следующий вклад: 1. Общепринятое предположение о том, что существует неотъемлемый компромисс между справедливостью и производительностью (т. е. точностью).) в машинном обучении подвергается сомнению с помощью доказательств , подтверждающих эту проблему. В частности, результаты этой статьи показывают, что такой компромисс можно смягчить, предполагая, что любой компромисс, скорее всего, связан с тем, как управляются данные.

2. Представлены четыре новых метода предварительной обработки данных выборки, основанные на применении стратегии многокатегорийной выборки с использованием данных, собранных в защищенных категориях. Методы исходят из предположения и соответствующей гипотезы о том, что балансировка данных на основе этих многокатегорийных свойств может повысить справедливость, не оказывая негативного влияния на производительность модели машинного обучения.
3. Результаты экспериментов представлены с использованием трех наборов данных, тщательно изученных сообществом справедливого ОД. Эксперименты включают сравнение с традиционными методами обучения без повторной выборки, чтобы продемонстрировать относительные эффекты предлагаемых методов. Результаты показывают, что два из предложенных методов, метод синтетической меньшинственной выборки защищенной категории (PC-SMOTE) и адаптивная синтетическая выборка защищенной категории (PC-ADASYN), особенно эффективны для повышения как справедливости, так и производительности.
4. Подробный анализ потенциального воздействия недостаточного и переоснащения на справедливость представлен путем изучения различных уровней модели дерева решений с использованием и без использования предлагаемых методов выборки. Результаты демонстрируют способность предложенных методов определять идеальный уровень дерева, на котором справедливость и точность максимальны.

Благодаря вышеуказанному вкладу эта работа представляет собой значительный шаг вперед в решении проблем справедливости в машинном обучении. Ключевой вывод из представленных здесь методов и результатов заключается в том, что справедливость можно обеспечить без ущерба для производительности модели.

1,5. Организация

Этот документ организован следующим образом. В разделе 2 представлены подробное объяснение справедливости и обсуждение нескольких технических показателей справедливости. Затем в разделе 3 описывается предыдущая литература, посвященная стратегиям смягчения предвзятости. В разделе 4 мы описываем предлагаемые нами методы выборки, набор данных и подход к настройке гиперпараметров. В разделе 5 мы представляем результаты нескольких экспериментов вместе с проверкой статистических гипотез в качестве средства проверки этих результатов. В разделе 6 обсуждаются экспериментальные результаты и анализируется, как каждый алгоритм работает с каждым набором данных и каждой метрикой. Также представлены дополнительные результаты о влиянии глубины дерева на справедливость и точность . В разделе 7 представлены ограничения этой работы и соответствующие направления будущей работы, а в разделе 8 представлен ряд выводов.

2. Предыстория В

этом исследовании рассматривается вопрос справедливости при прогнозировании результата y Y на основе набора характеристики x X Rd и некоторые дополнительные защищенные атрибуты s S , факторов, таких как раса, Rp , пол и пол. Например, при прогнозировании кредита x представляет финансовую историю заявителя, s — это его раса и пол, о которых он сам сообщил, а y — одобрен ли его кредит.

Machine Translated by Google

или отказано. Модель прогнозирования считается справедливой, если ее ошибки равномерно распределены между защищенными группами, такими как представители разных рас или полов. Предсказания класса на основе обучающих данных D обозначаются как Yˆ D := h(x, s) для некоторого h : X × S Y из класса H. Защищенные атрибуты s S в нашем исследовании предполагаются двоичными со специальным значение n , обозначающее непривилегированную группу. Например, S может означать расу, а n – «небелый»; следовательно, бинарная природа S равна {w, n}, где w представляет белых заявителей, которые являются привилегированной группой, а n представляет небелых заявителей, которые являются непривилегированной группой. Определение можно далее обобщить на небинарные случаи.

Дискриминация в помеченных наборах данных может быть определена как набор данных D, Икс, набор функций и защищенный набор атрибутов S со значением домена {w, n}. Дискриминация в D по отношению к группе S = n , обозначенной как diss=n(D), определяется как

$$\text{diss=n(D)} = \frac{|\{x \quad D : x(s) = w, h(x) = +\}|}{|\{x \quad D : x(s) = w\}|} \quad \frac{|\{x \quad D : x(s) = n, h(x) = +\}|}{\{x \quad D : x(s) = n\}|}$$

Приведенное выше определение можно перевести на разницу в вероятности нахождения заявителя в положительном классе для каждого домена защищенных атрибутов {w, n}. Наше исследование расширяет приведенное выше определение, рассматривая набор данных D, который содержит два или более защищенных атрибута.

Используются два популярных показателя справедливости. Первый — это уравненная нечетная разница (EOD), которая измеряет, насколько дискриминативным или справедливым является наш прогноз. EOD утверждает, что двоичный классификатор yˆ является справедливым, если его уровень ложноотрицательных результатов (FNR) и уровень истинного положительного результата (TPR) равны во всей области S [21]. FPR и TPR относительно защищенного атрибута s S со значением n можно определить как

$$TPRn(\hat{y}) = P(\hat{y} = 1 \,|\, y = 1, S = n)$$

$$FPRn(\hat{y}) = P(\hat{y} = 1 \,|\, y = 0, S = n)$$

Затем EOD определяется математически как разница между TPR и FPR в разных группах защищенного атрибута. То есть,

$$EOD = TPRn(\hat{y}) \quad FPRn(\hat{y}).$$

У справедливого классификатора EOD равен 0, а у несправедливого классификатора EOD равен 1. Хотя достижение полностью справедливого классификатора на практике практически невозможно, это исследование направлено на улучшение EOD без снижения точности. Тогда для EOD

$$FPRn(\hat{y}) = P(\hat{y} = 1 \,|\, y = 0, S = n) = TPRn(\hat{y}) = P(\hat{y} = 1 \,|\, y = 1, S = n)$$

и

$$FPRw(\hat{y}) = P(\hat{y} = 1 \,|\, y = 0, S = w) = TPRw(\hat{y}) = P(\hat{y} = 1 \,|\, y = 0, S = w)$$

Чтобы распространить приведенное выше определение EOD на нашу мультикатегорию, EOD рассчитывается для каждого столбца, затем макросреднее EOD представляется как окончательное EOD. Второй показатель, используемый для измерения справедливости прогнозирования МО, — это статистическая паритетность (СП). С.П. определяет справедливость как равную вероятность быть классифицированным как позитивное [22]. Это можно интерпретировать как то, что каждая группа в защищенном атрибуте имеет одинаковую вероятность быть классифицированной с положительным результатом.

$$P(\hat{y} = 1 \,|S = w) = P(\hat{y} = 1 \,|S = n)$$

3. Обзор литературы

Алгоритмы МО, которые все чаще используются для принятия решений в таких важных приложениях, как рецидивизм, кредитный скоринг, решения о кредите и т. д., изначально можно было бы считать справедливыми и свободными от присущей им предвзятости. Однако в действительности они могут унаследовать любую предвзятость или дискриминацию, присутствующую в данных, на которых они обучаются, как отмечает Берт [23]. Более того, простого удаления защищенных переменных из набора данных недостаточно для борьбы с косвенной дискриминацией и

Machine Translated by Google

на самом деле может скрыть это. Это признание усилило потребность в более совершенных инструментах, сделав выявление и предотвращение дискриминации важной областью исследований, как подчеркивается в [24–27].

Предвзятость в машинном обучении — быстрорастущая тема в исследовательском сообществе машинного обучения. Предвзятость в модели ОД может привести к несправедливому обращению с людьми, принадлежащими к определенным защищенным группам. В последнее время лидеры отрасли начали уделять все больше внимания предвзятости в моделях и программном обеспечении ML. Институт инженеров по электротехнике и электронике (IEEE) [28], Microsoft [29] и Европейский Союз [30] недавно опубликовали принципы справедливого поведения ИИ. Эти организации заявили, что модели машинного обучения должны быть справедливыми в реальных приложениях. Стратегии уменьшения смещения включают в себя изменение одного или нескольких из следующих элементов , чтобы гарантировать, что прогнозы, сделанные алгоритмом ML, менее предвзяты: (a) обучающие данные, (b) алгоритм ML и (c) сами последующие прогнозы. Они, соответственно, подразделяются на подходы предварительной обработки [31], внутренней обработки [32] и постобработки [21].

Во-первых, данные обучения могут быть предварительно обработаны, чтобы снизить несправедливость или предвзятость перед обучением модели. Камиран и Колдерс [6] предлагают выборку или повторное взвешивание данных для нейтрализации дискриминации. Этот подход может скорректировать представление или важность определенных точек данных, чтобы отдать предпочтение (или уменьшить преимущество) одного класса над другим. Другой метод предполагает непосредственное изменение отдельных записей данных, чтобы уменьшить дискриминацию, как показано в [33]. Например, этот подход предполагает изменение значений в наборе данных, чтобы уменьшить заметные предвзятости в отношении определенных групп. Кроме того, концепция t-близости, введенная Sondeck et al. [34], применяется для контроля дискриминации в работе [35]. Использование t-близости гарантирует, что распределение конфиденциальных атрибутов в любой заданной группе будет близко к распределению атрибута во всем наборе данных, тем самым сохраняя конфиденциальность и предотвращая дискриминацию на основе конфиденциальных атрибутов. Общей чертой этих подходов является балансирование контроля дискриминации с полезностью обработанных данных, то есть минимизация систематической ошибки без значительного ущерба для точности, репрезентативности и общей полезности данных для прогнозного моделирования или анализа. Этот баланс необходим для обеспечения того, чтобы усилия по обеспечению справедливости непреднамеренно не снижали качество или применимость данных.

В целом метод предварительной обработки можно разделить на три категории: (1) модификация данных, (2) удаление данных и (3) повторная выборка данных. Методы первой категории направлены на изменение значений точек обучающих данных (включая значения защищенных атрибутов, значения классов и значения функций), чтобы снизить систематическую ошибку в наборе данных. Примером такого метода является массирование данных, предложенное в [15]. Их подход ранжирует данные обучения, а данные, близкие к границе принятия решения, как в привилегированных, так и в непривилегированных группах переворачиваются. В качестве альтернативы Fahse et al. предложил оптимизированный метод предварительной обработки, который изучает вероятностное преобразование, которое редактирует классы и признаки с индивидуальным искажением и групповой справедливостью. [23]. В [36] исходные значения атрибутов заменяются значениями, выбранными независимо от метки класса, чтобы обучить модель, примерно достигая уравненных шансов. Аналогичным образом, Пэн и др. [37] заменяют значения защищенных атрибутов значениями, предсказанными на основе других атрибутов, аналогично вменению данных.

Методы второй категории направлены на обучение справедливой модели путем удаления определенных функций из обучающего набора. Примером этого метода является подавление данных, предложенное Dhar et al. [38]. В их статье защищенные атрибуты и функции, которые тесно связаны с защищенными атрибутами, также известные как прокси-атрибуты, удаляются из набора данных для обучения справедливой модели.

Методы третьей категории направлены на обучение справедливой модели либо путем корректировки весов выборки, либо путем избыточной выборки набора данных. Например, Красанакис и др. [39] предложили метод повторного взвешивания, который итеративно адаптирует веса обучающей выборки с помощью теоретически обоснованной модели для смягчения компромисса между предвзятостью и точностью. В [40] Чакраборти и др. предложил FairSMOTE в качестве метода избыточной выборки обучающих точек из групп меньшинств с искусственными точками данных на основе метода синтетической избыточной выборки меньшинства (SMOTE) [18] для достижения сбалансированного распределения классов. Кроме того, Ян и др. [41] предложили передискретизировать данные обучения групп меньшинств с помощью искусственных точек данных для достижения сбаланси

Machine Translated by Google

ции. В отличие от FairSMOTE, авторы сосредоточились на сценариях, в которых защищенные атрибуты неизвестны, и применили метод кластеризации для идентификации различных демографических групп.

В обработку входят методы, которые изменяют способ обучения модели ML в качестве средства уменьшения систематической ошибки. В [42] был предложен подход состязательного устранения смещений. Этот подход изучает классификатор для повышения точности и справедливости прогноза за счет включения переменной для заинтересованной группы путем одновременного изучения предиктора и противника. Это приводит к созданию беспристрастного классификатора, поскольку прогнозы не содержат никакой информации о групповой дискриминации, которую мог бы использовать противник. В качестве альтернативы в [43] был предложен алгоритм, который принимает метрику справедливости как часть функции потерь и возвращает модель, обученную для этой метрики справедливости . Камишима и др. [22] предложили метод регуляризации , который включал штрафной член в функцию потерь классификатора для получения несмещенного прогноза. Зафар и др. [44] разработали новый метод взвешивания, с помощью которого они настраивают вес выборки для каждого обучающего набора данных для достижения конкретной цели справедливости, например, уравнивания шансов на проверочных данных. В последнее время к уменьшению предвзятости стали относиться как к проблеме ограниченной оптимизации путем добавления ограничения справедливости и оптимизации потерь, чтобы они соответствовали этому ограничению [45,46]. Кроме того, в некоторых работах нейронные сети модифицируются с использованием исключения для удаления нейронов, принадлежащих к защищенным атрибутам [47].

Методы постобработки уменьшают систематическую ошибку после подбора модели машинного обучения и включают такие подходы , как калибровка, оптимизация ограничений и определение порога преобразования [6]. Такие методы предлагают алгоритм, который дает благоприятные результаты непривилегированным группам и неблагоприятные результаты благоприятным группам в пределах заданного доверительного интервала вокруг границы решения с наибольшей неопределенностью. Например, один из подходов изменяет пиковые пороги классификатора, чтобы получить заданную цель равных возможностей или уравненных шансов. Еще один подход предполагает случайное преобразование классов определенных предсказаний в разные классы [48].

Несколько новых исследований [49,50] объединили предварительную, внутреннюю или постобработку для формирования ансамблевого метода. Например, Бхаскаруни и др. [50] сочетают передискретизацию класса, защищенного от дисбаланса, со сдвигом границы решения и метода постобработки для решения проблемы несправедливости.

Исследователи углубились в различные концепции дискриминации и справедливости в рамках алгоритмического принятия решений. Например, несопоставимое воздействие (ранее называемое косвенной справедливостью) измеряется посредством статистического паритета и групповой справедливости, как обсуждалось Бхаскаруни и др. [50]. С другой стороны, концепция индивидуальной справедливости, также представленная Бхаскаруни и др., подчеркивает, что к одинаковым людям следует относиться одинаково, независимо от их групповой принадлежности. Этот подход фокусируется на справедливости на индивидуальном уровне, гарантируя, что решения принимаются на основе соответствующих качеств, а не групповых стереотипов или предубеждений.

Как подчеркнули Чжан и Нил [16], ключевой целью классификаторов и других прогностических моделей является достижение одинакового уровня ошибок в разных группах. Аналогичным образом, обеспечение калибровки или отсутствие предсказывающей предвзятости в прогнозах, как обсуждалось Hardt et al. [21] имеет решающее значение. Однако противоречие между этими понятиями — калибровкой и равными коэффициентами ошибок — исследуется Дворком и др. [51] и Pleiss et al. [52], что указывает на то, что одновременное удовлетворение обоих требований может оказаться сложной задачей. Карими-Хагиги и Кастильо [53] представляют соответствующую работу, исследующую сложности, присущие достижению алгоритмической справедливости. Фридлер и др. [54] далее исследует компромиссы при соблюдении различных определений алгоритмической справедливости, особенно с точки зрения общественной безопасности. Учитывая, что наша работа сосредоточена на предварительной обработке, а не на моделировании, такие соображения, как сбалансированная частота ошибок и предсказательная погрешность, становятся менее применимыми напрямую.

Судя по нашему обзору различных методов предварительной обработки, выяснилось, что не проводилось никаких работ по моделированию справедливости для двух или более защищенных атрибутов одновременно. Кроме того, метод выборки, использованный в предыдущей работе, был ориентирован только на выборку на основе меток классов, а не защищенных категорий. Следовательно, в этой статье особое внимание уделяется предварительной обработке, поскольку она представляет собой наиболее адаптируемый аспект конвейера обработки данных [55]. Предварительная обработка отличается тем, что она не зависит от выбора алгоритма моделирования и может быть легко выполнен

Machine Translated by Google

включены в механизмы выпуска и публикации данных. Эта независимость и гибкость делают предварительную обработку критически важной для обеспечения качества и справедливости данных перед выполнением любого аналитического или прогнозного моделирования. Наконец, мы сосредоточимся на новых методах выборки, которые выбирают защищенную категорию при обучении данных для построения справедливой модели.

4. Методология.

Целью нашей работы является изучение методов выборки для повышения справедливости в машинном обучении без соответствующего снижения эффективности прогнозирования, тем самым смягчая компромисс между справедливостью и производительностью . В результате предлагаются четыре новых метода выборки, ориентированных на достижение этой цели. Эти методы выборки решают проблему несбалансированного класса, возникающую из-за новой мультикатегории, созданной из-за комбинации защищенных категорий. Необходимы специальные методы выборки, поскольку существующие методы отбирают данные на основе классов меньшинства и большинства, но для снижения справедливости выборка из новых мультикатегорий осуществляется на равной основе. Это, в свою очередь, требует изменения существующих методов выборки для выборки данных на основе этих новых категорий. Это приводит к появлению четырех новых методов выборки: передискретизация по защищенной категории, пропорциональная выборка по защищенной категории, SMOTE по защищенной категории (PC- SMOTE) и ADASYN по защищенной категории (PC-ADASYN).

4.1. Передискретизация защищенных

категорий. При передискретизации защищенных категорий первым шагом является объединение защищенных категорий в наборе данных и кодирование этой комбинации для создания нашей новой мультикатегории. Например, в наборе данных о доходах взрослых возраст (молодой и взрослый), раса (белый и другие) и пол (мужской и женский) объединяются для создания восьми новых категорий: ADULTWHITEMALE, ADULTOTHERSMALE, YOUNGWHITEMALE, YOUNGOTHERSMAIL, ADULTWHITEFEMALE, ADULTWHITEMALE, ADULTWHITEMALE, YOUNGWHITEMALE, YOUNGOTHERSMAIL, ADULTWHITEFEMALE, ADULTOTHERSFEMALE, YOUNGWHITEFEMALE и Эти новые категории имеют разные размеры выборки, и цель нашей передискретизации защищенных категорий — сбалансировать эту новую категорию так, чтобы размер выборки каждой из новых категорий соответствовал размеру категории с наибольшим размером выборки. Чтобы избежать утечки данных, набор данных разделяется на обучающий и тестовый, при этом передискретизация применяется только к обучающим данным, а затем выполняется тестирование на несемплированном тестовом наборе.

Псевдокод в алгоритме 1 подробно показывает наш метод передискретизации защищенной категории. В алгоритме в качестве базовой использовалась самая большая категория, поскольку это категория с наибольшим размером выборки. В результате процесса выборки получаются новые обучающие данные со сбалансированным размером выборки по новой категории. Алгоритм работает путем выборки остальных защищенных категорий в соответствии с размером выборки базового уровня. Эта выборка выполняется путем многократного повторения категорий вместе с их метками классов.

4.2. Пропорциональная выборка защищенной категории

Метод пропорциональной выборки по защищенной категории представляет собой обобщение передискретизации по защищенной категории, поскольку процесс начинается с установки целевого размера выборки (который является настраиваемым гиперпараметром, а не просто размером наибольшей мультикатегории), обозначаемого как targetSamples. Это соответствует желаемому количеству экземпляров, необходимых для каждой категории. Эта цель обеспечивает единообразие по всем категориям, снижая риск смещения модели в сторону более частых категорий. Типичный результат применения этого метода заключается в том, что некоторые категории, которые имеют больше выборок, чем целевые выборки, будут иметь недостаточную выборку, в то время как другие будут иметь избыточную выборку, чтобы получить равную их долю в наборе обучающих данных. Псевдокод в алгоритме 2 показывает шаг за шагом метод пропорциональной выборки защищенной категории.

Machine Translated by Google

Алгоритм 1: передискретизация защищенной категории

```
baselineCount  сумма записей в «Largest_Category» Xtrain
totalCount  количество записей в Xtrain
baselineProportion  baselineCount/totalCount
BalanceData  инициализировать пустой набор данных
категории  список имен столбцов в Xtrain, начинающихся с «combined_category_»
для каждой категории в категориях делайте
    CategoryData  выбрать записи в Xtrain , где категория = 1
    CategoryData  объединить CategoryData с соответствующими метками из ytrain
    CategoryCount  количество записей в категорииData
    targetCount  целая часть totalCount × baselineProportion
    если CategoryCount < targetCount , то
        sampledData  образец targetCount из категорииData с заменой
        сбалансированные данные  добавить sampledData к сбалансированным данным
    еще
        сбалансированные данные  добавить категориюДанные к сбалансированнымДанным
конец, если
конец
вернуть сбалансированные данные
```

Алгоритм 2. Пропорциональная выборка защищенной категории.

```
целевые выборки  5000
sampledBalanced  инициализировать пустой набор данных
для каждого столбца в new_categories.columns выполните
    CategoryRows  выбрать строки в new_categories , где столбец = 1
sampledRows  образец targetSamples записей из категорииRows с заменой
    для каждого столбца в oneHotEncodedBalanced.columns выполните
        sampledRows[col]  0
конец для
    sampledRows[столбец]  1
    sampledBalanced  добавить sampledRows к sampledBalanced
конец
возврат выборки сбалансированный
```

4.3. Защищенная категория SMOTE

Метод передискретизации синтетического меньшинства защищенной категории (PC-SMOTE)

Метод выборки — это более сложный процесс, направленный на имитацию SMOTE, но модифицированный для выборка наших новых категорий, а не меток классов. При таком подходе первым шагом было измените SMOTE, чтобы использовать фиксированное количество соседей и случайным образом выбирать одного соседа для интерполяция, а не усреднение всех из них. Псевдокод в алгоритме 3 показывает процедура ПК-SMOTE. Поскольку цель метода состоит в том, чтобы использовать его для новых выборка категорий, она не касается непосредственно генерации меток классов. Следовательно, новый необходима функция, которая может генерировать новую метку класса для синтетических данных. Для этого новый определена функция, которая генерирует метки классов на основе количества новых синтетических данных генерируется и предварительно выбранное соотношение баланса между двумя классами. Алгоритм 4 показывает, как наша новая функция генерирует метки для наших синтетических образцов. Алгоритм сначала определяет количество образцов для каждого класса на основе соотношения баланса и генерирует образец необходимо для каждого класса. Метки классов перетасовываются, чтобы предотвратить алгоритмическую предвзятость в созданные классы.

Эти два алгоритма объединяются вместе для формирования PC-SMOTE, как показано в алгоритме 5. В подходе к достижению баланса между несколькими категориями каждая отдельная категория повторяется. так, чтобы было идентифицировано подмножество данных, связанных с этой категорией. Номер синтетических образцов, необходимых для достижения заранее определенного максимального размера по каждой категории, затем рассчитано. Если требуются дополнительные образцы, данные генерируются с помощью PC-SMOTE.

Machine Translated by Google

который интерполирует между существующими точками данных и их ближайшими соседями. Одновременно, сбалансированное распределение меток синтетических классов создается с указанным соотношением баланса с помощью с использованием алгоритма 4. Эти синтетические функции и метки затем включаются в подмножество обучения для каждой категории. Процесс повторяется для всех категорий, в результате чего сбалансированный набор данных. Гиперпараметрами в этом алгоритме 5 являются количество соседей. и соотношение баланса.

Алгоритм 3. Пользовательский метод синтетической передискретизации меньшинства (SMOTE)

```
процедура CUSTOMSMOTE(данные, n_samples)
    синтетические образцы  нулевая матрица размера (n_samples× количество столбцов в данных)
3: nn  NearestNeighbors(n_neighbors = 7).fit(данные)
    соседи  nn.kneighbors(data, return_distance = False)
    для i  от 1 до n_samples do
        sampleIdx  случайное целое число от 0 до (количество строк в данных — 1)
        nnIdx  случайный выбор из соседей[sampleIdx, 1 :]
        di ff  данные [nnIdx]  данные [sampleIdx]
        вес  случайное число из равномерного распределения от 0 до 1
        синтетические образцы[i]  данные[sampleIdx] + вес × разница ff
конец для
    вернуть синтетические образцы
завершение процедуры
```

Алгоритм 4. Генерация сбалансированных синтетических меток

```
процедура GENBALSYNTHLABELS(n_samplesNeeded, BalanceRatio)
    nClass1  int(n_samplesNeeded × BalanceRatio)
    nClass0  n_samplesNeeded — nClass1
    синтетические метки  [0] × nClass0 + [1] × nClass1
    SHUFFLE(syntheticLabels) возвращает                    Случайным образом перетасуйте метки.
    синтетическиеLabels
завершение процедуры
```

Алгоритм 5 SMOTE защищенной категории

```
balancedDataList  инициализировать пустой список
для каждой категории в категориях делаем
    CategorySubset  выбрать строки в train_data st 'combined_category' == категория
    функции  удалить «класс», «комбинированная_категория» из категорииПодмножество
    nSamples  max_size — количество строк в категорииSubset
    если nSamples > 0, то
        синтетические функции  PCSmote( функции, nSamples)
        синтетическиеLabels  GenBalSynthLabels(nSamples, BalanceRatio)
        синтетическиеОсобенности[' class' ]  синтетические метки
        синтетическиеОсобенности[' комбинированная_категория' ]  категория
        категорияПодмножествоСбалансированное  объединить категорииПодмножество и синтетическиеОсобенности
    еще
        категорияПодмножествоСбалансированный  категорияПодмножество
конец, если
добавить категориюSubsetBalanced в BalanceDataList.
конец
сбалансированные данные  добавить сбалансированный список данных и сбросить индекс
```

4.4. Защищенная категория ADASYN

Метод ADASYN защищенной категории имитирует адаптивное синтетическое меньшинство (ADASYN). выборки, но немного изменен для достижения нашей цели выборки по защищенным категориям. Наш алгоритм PC-ADASYN показан в алгоритме 6. Он расширяет ADASYN, фокусируясь на категории

Machine Translated by Google

плотность, а не классовый дисбаланс. В частности, эта функция работает путем поиска ближайшего соседей данных, а затем вычисление плотности категории каждой точки данных в ее пределах. ближайшее соседство. Он взвешивает эти плотности в обратном порядке, чтобы отдать приоритет категориям меньшинств. что повышает вероятность создания синтетических выборок из недостаточно представленных категорий. синтетические образцы создаются путем интерполяции между выбранными точками данных и их соседями, аналогично SMOTE, но использует случайный вес для изменения интерполяции, что обеспечивает разнообразный синтетический набор данных. Такой подход помогает устранить дисбаланс на уровне категорий и обогащает дисперсию набора данных, потенциально повышая надежность и справедливость моделей ML. обучен на этих данных. Поскольку этот метод выборки также генерирует новые выборки путем интерполяции, Алгоритм 4 используется для генерации меток классов для новых синтетических образцов.

Алгоритм 6 PC-ADASYN для балансировки по категориям

```
процедура PCADASYNCATEGORIES (данные, метки, n_samplesNeeded, n_neighbours)
    n_neighbours  n_neighbours + 1 3: nn                    Включая саму точку данных
NearestNeighbors(n_neighbours).fit(данные)
    расстояния, индексы  nn.kneighbors(данные)
    плотности  нулевой массив длины (данные)
    для i  0 до длины (данные) - 1 сделать
        текущая_категория  метки[i]
        соседние_индексы  индексы[i][1 :]                   Пропустить самоуказатель
        плотности[i]  СУММА(метки[neighbor_indices] == текущая_категория)
конец для
    веса  1/(плотности + 1) веса  веса/                     Добавьте 1, чтобы предотвратить деление на ноль.
    SUM(веса) синтетические образцы  пустой                 Нормализовать веса
    список
    sampleIndices  случайный выбор с заменой по длине (данным) с использованием весов
    для каждого idx в sampleIndices выполните
        baseIdx  idx
        Idx соседа  RANDOMCHOICE(index[baseIdx][1 :])
        di ff  данные [идентификатор соседа]  данные [baseIdx]
        синтетический образец  data[baseIdx] + RANDOM() × d ff
        добавить синтетический образец к синтетическим образцам
конец
    возвращаемый массив (синтетические образцы)
завершение процедуры
```

Алгоритмы 4 и 6 объединяются для формирования нашей выборки ADASYN защищенной категории. как показано в алгоритме 7. Для каждой категории данные, соответствующие этой категории, изолируются, и вычисляется дефицит размера относительно самой крупной категории. Если дополнительные необходимы образцы, применяется метод PC-ADASYN, генерирующий синтетические характеристики, которые соблюдайте характеристики распределения категории. Эти функции затем дополняются с синтетически сгенерированными метками, поддерживающими заранее заданное соотношение классов. процесс не только исправляет дисбаланс категорий, но и обогащает набор данных, потенциально повышение точности прогнозирования и справедливости моделей, обученных на этих данных.

4.5. Настройка набора данных и гиперпараметров

Для проверки четырех методов отбора проб был необходим классификатор, позволяющий оценить влияние на справедливость и производительность. В конечном счете, тип классификатора не имеет прямого значения, поскольку Цель состоит в том, чтобы смягчить компромисс между справедливостью и производительностью, а не найти лучший классификатор. Поэтому мы решили использовать одномерные деревья решений на основе CART [56] из-за их устойчивость к шуму и недостающим данным. Конкретная реализация, которую мы выбрали, была взято из библиотеки sklearn (версия 1.5.1) [57]. Кроме того, деревья решений позволяют нам контролировать силу соответствия, устанавливая глубину изученного дерева. Это позволяет нам сравнить справедливость и производительность на разных уровнях соответствия.

Machine Translated by Google

Процесс начинается с объединения защищенных категорий в каждом наборе данных, применения горячего кодирования для создания новых многокатегорийных объектов, а затем выполнения кодирования меток. Затем наборы данных были разделены с использованием стратифицированной 10-кратной перекрестной проверки, чтобы обеспечить репрезентативное распределение классов в каждом сгибе. Для каждого сгиба обучение проводилось на выборочных данных с использованием ранее описанных методов, а классификация тестировалась на соответствующих несемплированных тестовых наборах. Согласованность в обучении модели поддерживалась за счет применения одинаковой глубины дерева для всех методов выборки, а предоставленные результаты представляют собой средние значения 10-кратных прогонов с соответствующими доверительными интервалами.

Алгоритм 7 ADASYN защищенной категории

```
1: BalanceDataList инициализировать пустой список 2:
для каждой категории в категориях do
3: CategorySubset выбрать из new_data3 st 'combined_category' == функции категории удалить
4: 'class', 'combined_category' из категорииПодмножество CategoryLabels извлечь
5: 'combined_category' из категорииSubset nSamplesNeeded max_size минус
6: количество строк в категорииПодмножество, если nSamplesNeeded > 0, тогда
7: синтетические функции
8: PCadasyn( features, CategoryLabels, nSamplesNeeded) синтетическиеLabels
9: GenBalSynthLabels(nSamplesNeeded, BalanceRatio) синтетические функции[ ′ class ]
10: синтетические метки синтетические характеристики
11: [ ′ комбинированная_категория ] категорияSubsetBa копье
12: объединить категориюSubset с синтетическими функциями еще категорияSubsetBalanced
13:
14: CategorySubset 15: завершить, если 16: добавить

категориюSubsetBalanced к сбалансированномуDataList 17: завершить
для 18:
сбалансированныеДанные объединить сбалансированныйDataList и сбросить индекс
```

Для нашего анализа из репозитория UCI [58] были выбраны три набора данных : набор данных о доходах взрослых [59], набор данных German Credit [60] и набор данных «Профилирование управления исправительными правонарушителями для альтернативных санкций» (COMPAS). Набор данных о доходах взрослых предназначен для прогнозирования того, зарабатывает ли человек более 50 000 долларов США, и включает восемь категориальных и четыре числовых атрибута с защищенными переменными, соответствующими возрасту (молодой или взрослый), полу (мужской или женский) и расе (белый или другие). Набор данных о доходах взрослого населения был передан в дар UCI в 1996 году. Набор данных German Credit, используемый для прогнозирования кредитоспособности, включает 20 категориальных и два числовых атрибута с защищенными переменными возраста и пола. Набор немецких кредитных данных был передан в дар UCI в 1994 году. Набор данных COMPAS [61], который оценивает уровень рецидивизма в Соединенных Штатах, включает шесть категориальных и шесть числовых характеристик с защищенными переменными возраста, расы и пола. Набор данных был опубликован в 2018 году. Эти наборы данных были выбраны потому, что они представляют собой самые современные наборы данных для измерения предвзятости и дискриминации и широко используются в других исследованиях алгоритмической предвзятости и справедливости (см. Раздел 3). Кроме того, наборы данных имеют разные размеры: от маленьких до больших, что делает их подходящими для тестирования наших методов

Настройка гиперпараметров проводилась с использованием поиска по сетке [62] для изучения широкого диапазона параметров, дополненного визуальными оценками для определения оптимальных настроек, которые уравновешивают разницу уравненных шансов (EOD) и точность. Для набора данных о доходах взрослых оптимальные гиперпараметры включали максимальную глубину дерева 3, а для PC-SMOTE и PC-ADASYN — настройку ближайшего соседа 5 со сбалансированным коэффициентом 0,34. Эти параметры были столь же эффективны для набора данных German Credit. Для набора данных COMPAS максимальная глубина дерева 2 была оптимальной для всех методов выборки. PC-SMOTE и PC-ADASYN были настроены на настройку ближайшего соседа 3 и сбалансированное соотношение 0,60.

Machine Translated by Google

5. Результаты

Четыре стратегии выборки были применены к трем наборам данных, описанным выше, и оценили их влияние, используя простой одномерный классификатор дерева решений. Результаты в В таблицах 1–3 показаны заметные различия в эффективности моделей при пяти стратегиях выборки: без выборки, передискретизация, пропорциональная выборка, PC-SMOTE и PC-ADASYN включены наши три набора данных. Каждый метод оценивался по точности, макросу F1, Equalized Разница шансов (EOD) и статистическая паритетность (SP). Результаты были измерены с точностью и макрос F1, поскольку эти две метрики являются наиболее популярными метриками классификации. Также, ограничение показателя двумя делает результаты сопоставимыми для статистического анализа.

Чтобы убедиться в статистической значимости результатов каждого метода, мы использовали метод Фридмана. тест, непараметрическая альтернатива однофакторному дисперсионному анализу с повторными измерениями. На обнаружив значимые результаты теста Фридмана, мы приступили к апостериорному исследованию Неменьи. тест. Этот тест используется для оценки парных сравнений между методами установления какие методы статистически отличаются друг от друга. Преимущество теста Неменьи в этом параметре, поскольку он учитывает множественные сравнения, не предполагая нормального распределения, тем самым предоставляя надежный способ понять конкретные парные различия.

Для набора данных о доходах взрослого населения (таблица 1) метод без выборки дал точность 0,82, устанавливая высокий базовый уровень для сравнения. Однако оно продемонстрировало несколько предвзятое отношение прогнозирование с EOD 0,36 и минимальным несоответствием в показателях прогнозирования (SP = 0,02). В напротив, передискретизация сохранила ту же точность, но немного снизила макро F1. до 0,65, что указывает на потенциальные проблемы переобучения и ухудшение справедливости (EOD = 0,66) и увеличивающееся неравенство в показателях прогнозирования (SP = 0,09). Пропорциональная выборка уменьшена точность до 0,79, но улучшил макрос F1 до 0,79, что предполагает лучший баланс между точность и отзыв. Однако это значительно увеличило SP до 0,71, что указывает на существенную неравенство в показателях положительных прогнозов, что вызывает опасения по поводу справедливости модели. два специальных подхода для SMOTE и ADASYN были разработаны специально для улучшения по этим показателям. PC-SMOTE показал умеренную производительность с точностью 0,81 и улучшенный EOD 0,25, что предполагает повышенную справедливость по сравнению с базовой передискретизацией. и метод без выборки. Тем не менее, он по-прежнему зафиксировал более низкий макрос F1 (0,63), что указывает на проблемы искажения при генерировании синтетических данных. PC-ADASYN оказался наиболее сбалансированный подход, сохраняющий высокую точность (0,82) и лучшую обработку класса дисбалансы, с умеренным улучшением справедливости (EOD = 0,28) и контролируемым увеличение разницы в скорости прогнозирования (SP = 0,09). В целом, базовая точность статистически значительно лучше, чем пропорциональная выборка, хотя она не является статистически значимой, поскольку по сравнению с другими методами отбора проб. Для макроса F1 пропорциональная выборка статистически значительно лучше, чем другие методы выборки. Для EOD: PC-ADASYN статистически лучше, чем другие методы выборки, тогда как метод SP без выборки статистически значительно лучше, чем другие методы выборки.

Таблица 1. Результаты методов выборки по набору данных о доходах взрослых с доверительными интервалами 95%.

Выборка Метод	Точность	Макро Ф1	ОВП	СП
Без выборки	0,82 ± 0,00	0,66 ± 0,01	0,36 ± 0,18	0,02 ± 0,00
Избыточная выборка	0,82 ± 0,02	0,65 ± 0,06	0,66 ± 0,2	0,09 ± 0,02
Пропорциональная выборка	0,79 ± 0,05	0,79 ± 0,06	0,46 ± 0,12	0,71 ± 0,10
ПК-SMOTE	0,81 ± 0,05	0,63 ± 0,07	0,25 ± 0,21	0,07 ± 0,00
ПК-АДАСИН	0,82 ± 0,04	0,64 ± 0,07	0,28 ± 0,19	0,09 ± 0,02

Результаты экспериментов с набором данных German Credit также демонстрируют различное влияние. каждой стратегии выборки, особенно в отношении справедливости и точности, как показано в таблице 2. Без выборки базовая модель достигла точности 0,72, но продемонстрировала значительные смещение в прогнозе с EOD 0,88, что указывает на существенное несоответствие ошибок. ставки между группами в защищенных атрибутах. Реализация передискретизации поддерживается точность при улучшении макроса F1 до 0,68 и заметном снижении EOD до 0,37, хотя и на

Machine Translated by Google

стоимость увеличения SP до 0,35, что подчеркивает потенциальный компромисс между различными вариантами справедливости
меры. Пропорциональная выборка немного снизила точность до 0,68, но добилась наилучшего результата.
F1-оценка 0,69. Это также снизило EOD до 0,32, что позволяет предположить, что это эффективно уравновешивает прогноз.
качество со справедливостью. PC-SMOTE показывает улучшение точности до 0,73, но ниже
макро F1 0,55; справедливость модели показывает огромное улучшение по сравнению с базовым уровнем с
EOD 0,15 и умеренный SP 0,1. PC-ADASYN показывает такую же точность, как и
базовый уровень 0,72, хотя самый низкий макро-F1 равен 0,48, что предполагает потенциальный компромисс.
в точности и отзыве. Тем не менее, модель демонстрирует наилучшие результаты прогнозирования справедливости при
EOD 0,13 и SP 0,06. В целом, результат показывает, что точность отсутствия выборки
не является статистически значимым для других методов выборки, кроме пропорциональной выборки
в то время как для EOD результаты всех методов выборки статистически значимы в
по сравнению с безвыборочным методом.

Таблица 2. Результаты методов выборки на основе набора данных German Credit с доверительным интервалом 95%.

Выборка Метод	Точность	Макро Ф1	ОВП	СП
Без выборки	0,72 ± 0,03	0,62 ± 0,06	0,88 ± 0,10	0,07 ± 0,01
Избыточная	0,72 ± 0,05	0,68 ± 0,04	0,37 ± 0,16	0,35 ± 0,11
выборка Проп.	0,69 ± 0,04	0,69 ± 0,07	0,32 ± 0,26	0,29 ± 0,10
Выборка PC-SMOTE	0,73 ± 0,05	0,55 ± 0,09	0,15 ± 0,09	0,1 ± 0,02
ПК-АДАСИН	0,72 ± 0,03	0,48 ± 0,11	0,13 ± 0,02	0,06 ± 0,00

В Таблице 3 показаны результаты нашего опыта работы с набором данных COMPAS. Эти результаты
выявить значительные различия в производительности модели при различных стратегиях выборки.
Базовый подход без выборки позволил достичь точности и макропоказателя F1, равного
0,89, но показал более высокие различия в показателях справедливости с уравненной разницей шансов.
(EOD) 0,39 и статистического паритета (SP) 0,29. Это подчеркивает потенциальные предубеждения, которые
нескорректированные модели могут относиться к защищенным группам. Применение передискретизации
немного улучшилась точность до 0,90, но также значительно улучшилась справедливость, уменьшив EOD до
0,26. Это предполагает эффективность сокращения различий в результатах без ущерба для
СП. И наоборот, пропорциональная выборка, одновременно повышая точность и макро F1 до 0,90 и
0,91, соответственно, также достигло EOD 0,26, что улучшило его по сравнению с базовым уровнем, в то время как
также зафиксирован более высокий SP, равный 0,36, что указывает на потенциальное увеличение неравенства положительных
результаты в группах. PC-SMOTE и PC-ADASYN с одинаковыми показателями точности,
макросу F1 и SP удалось сохранить улучшение справедливости с EOD 0,30,
хотя эти методы также увеличили SP до 0,47. В целом результаты показывают, что оба
точность и EOD наших методов отбора проб статистически значительно лучше, чем
безвыборочный метод.

Таблица 3. Результаты методов выборки по набору данных COMPAS с доверительными интервалами 95%.

Выборка Метод	Точность	Макро Ф1	ОВП	СП
Без выборки	0,89 ± 0,04	0,89 ± 0,04	0,39 ± 0,15	0,29 ± 0,17
Oversample	0,90 ± 0,03	0,90 ± 0,05	0,26 ± 0,14	0,25 ± 0,07
Prop. Sample PC-	0,90 ± 0,05	0,91 ± 0,04	0,26 ± 0,12	0,36 ± 0,10
SMOTE	0,91 ± 0,02	0,91 ± 0,02	0,30 ± 0,11	0,47 ± 0,19
ПК-АДАСИН	0,91 ± 0,02	0,91 ± 0,03	0,30 ± 0,13	0,47 ± 0,21

6. Обсуждение

Результаты экспериментов с тремя наборами данных подтверждают, что защищенная категория
выборка может заметно повысить справедливость модели, часто без существенного ущерба для точности
прогнозирования. В некоторых случаях также отмечалось улучшение точности и макрос F1.
продемонстрировано. Ориентируясь на результаты набора данных о доходах взрослых, PC-SMOTE и PC-ADASYN

Machine Translated by Google

продемонстрировали заметные улучшения в EOD и сохранили умеренный уровень SP. Эффективность этих методов во многом объясняется их сложными методами интерполяции. Например, визуальный анализ деревьев решений, созданных без выборки и с помощью PC-ADASYN, позволяет получить четкое представление о контрастах. Примеры одной репрезентативной складки показаны на рисунках 1 и 2 соответственно. Дерево решений, полученное без выборки, выбрало свой корень с признаком, тесно связанным с защищенными атрибутами, действуя таким образом как прокси-атрибут. Это привело к выраженной ошибке прогнозиро
И наоборот, дерево решений, обученное на данных, сгенерированных с помощью PC-ADASYN, начиналось с функции , которая очень хорошо обобщала прогнозы и смягчала предвзятость, о чем свидетельствует заметное повышение справедливости модели и более высокая примесь Джини, что указывает на более чистое начальное разделение.

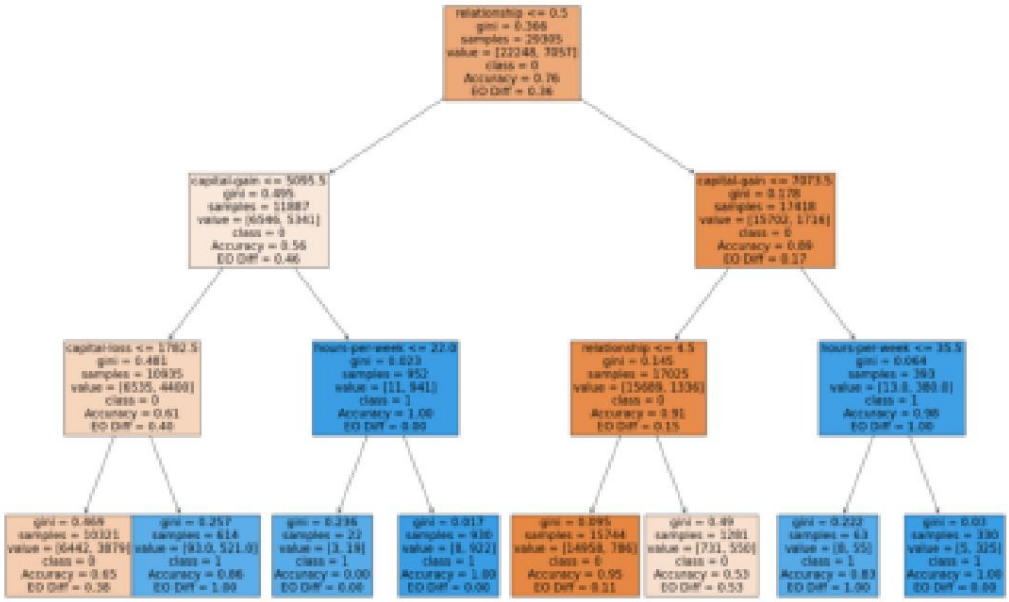


Рисунок 1. Пример дерева решений, обученного на основе доходов взрослого населения без выборки.

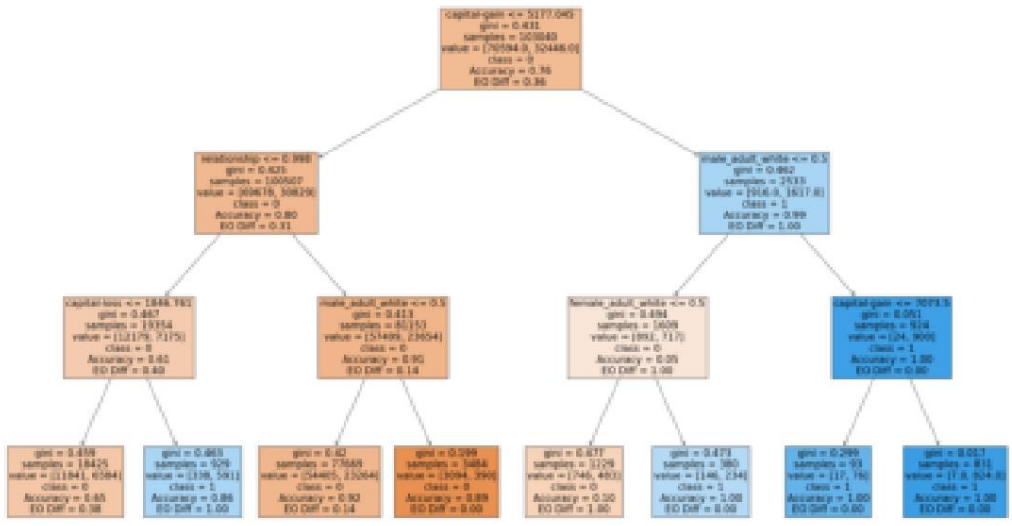


Рисунок 2. Дерево решений PC-ADASYN по доходу взрослого населения.

Machine Translated by Google

6.1. Сравнение справедливости и производительности

При сравнении передискретизации и пропорциональной выборки подходы этих методов к увеличению размера выборки путем дублирования существующих строк данных оказались простыми и не привели к существенному улучшению EOD. Этот результат имеет смысл, поскольку эти методы имеют тенденцию воспроизводить существующие предубеждения, которые потенциально могут усугубить проблемы справедливости, а не облегчить их. Это становится очевидным, в частности, при рассмотрении набора данных о доходах взрослых, где классы крайне несбалансированы. Этим наивным стратегиям репликации не хватает возможностей интерполяции PC-SMOTE и PC-ADASYN для корректировки выборок вблизи границ решения, что имеет решающее значение для уменьшения систематической ошибки в наборе данных. Напротив, стратегии интерполяции, используемые SMOTE и ADASYN, расширяют набор данных и повышают его разнообразие. Это особенно эффективно для выборок вблизи границ решения, где небольшие сдвиги в характеристиках могут существенно повлиять на справедливость прогнозов. Интерполируя между выборками, SMOTE и ADASYN эффективно перемещают эти граничные выборки к более справедливым областям пространства признаков, тем самым напрямую противодействуя и уменьшая смещение более эффективно, чем методы, которые увеличивают объем выборки без изменения структуры данных.

Функция генерации классов (алгоритм 4) также помогает увеличить общее распределение классов. Результаты наших PC-SMOTE и PC-ADASYN по доходам взрослых также демонстрируют превосходство над результатами, полученными в [41], где была получена точность 0,59 и SP 0,17. Кроме того, результаты PC-SMOTE и PC-ADASYN показывают превосходство над результатами, полученными в [40], где EOD 0,89 и немного лучшая точность 0,84.

Изучая результаты набора данных German Credit, мы наблюдали тенденцию, аналогичную той, которая была отмечена в наборе данных о доходах взрослых: метод без выборки имеет очень высокую погрешность в отношении EOD. Низкое значение SP, равное 0,07, указывает на минимальную разницу в показателях положительных прогнозов между группами, но это само по себе не является хорошим способом измерения справедливости, поскольку в предпочитаемой группе больше образцов, чем в неблагоприятной. Это приводит к искажению расчета SP, поскольку учитываются только положительные решения в каждой группе, на которые влияет размер выборки. Одним из выводов является важность использования нескольких показателей справедливости для всестороннего анализа влияния модели на все заинтересованные стороны. При передискретизации мы увидели улучшение EOD при аналогичном SP; это показывает, что увеличение количества выборок для каждого из защищенных атрибутов мультикатегории повышает справедливость в отношении EOD. Кроме того, обновленный SP отражает то, как будет выглядеть наличие более равного количества образцов для каждой мультикатегории, в отличие от базового варианта, где в предпочтительной группе в пять раз больше образцов, чем в неблагоприятной группе. Точность пропорциональной выборки падает, поскольку базовое количество выборок, выбранных после настройки гиперпараметра, было недостаточным для того, чтобы модель могла обобщить набор тестов без выборки, что привело к переобучению. Переобучение было подтверждено с учетом точности обучения. Интересно, что модель не меняет точность на полноту, как другие модели, и это дает пропорциональную выборку самого высокого макроса F1.

Что касается справедливости, мы обнаружили увеличение EOD по сравнению с базовым уровнем и избыточной выборкой. Это происходит потому, что каждая мультикатегория представлена в одних и тех же базовых количествах. Это может повысить справедливость модели, поскольку теперь модель имеет более широкое представление о категориях и делает более точные прогнозы и, в конечном итоге, более справедливые решения. PC-SMOTE и PC-ADASYN играют ключевую роль в значительном уменьшении систематической ошибки в прогнозах Такая согласованность подтвердила надежность этих методов в различных наборах данных. Примечательно , что ни один из методов не ставит под угрозу точность, в то время как оба повышают справедливость, иллюстрируя их эффективность в решении компромиссов, обычно связанных с прогнозным моделированием.

Эти результаты демонстрируют сильные возможности интерполяции, присущие PC-SMOTE и PC-ADASYN. Эти методы эффективно перераспределяют выборки в пространстве признаков, особенно перемещая образцы из малообеспеченных регионов от отрицательных к более положительным границам принятия решений. Такие корректировки имеют решающее значение для смягчения предвзятых результатов и обеспечения справедливости в автоматизированных процессах принятия решений. Макрос F1 в обеих моделях снижается по сравнению с базовым уровнем, поскольку для того, чтобы модель лучше работала при обобщении, требуется большее количество выборок , что этот набор данных не поддерживает. В частности, набор данных содержит 700 образцов для класса 0 и 300 образцов для класса 1, что означает, что набор тестов

Machine Translated by Google

имеет только 30 образцов для класса 1. Из-за такого небольшого количества образцов обе модели обменивают полноту на точность в классе 1. Примечательно, что мы наблюдали низкую полноту для класса 1, что в конечном итоге приводит к низкому показателю f1 для класса 1, и поскольку макрос F1 усредняет два показателя f1 и обрабатывает их одинаково, это влияет на производительность обеих моделей в макросе F1. В целом эти две модели дают более справедливую модель с хорошей точностью по сравнению с базовым и двумя другими методами выборки.

Оценка набора данных COMPAS еще раз подтверждает эффективность наших методов выборки . Выявленные отчетливые закономерности совпадают с теми, которые наблюдаются в наборах данных о доходах взрослых и немецких кредитах, что подчеркивает надежность наших выводов. Примечательно, что методы передискретизации и пропорциональной выборки продемонстрировали существенные улучшения в выравненной разнице шансов (EOD) и точности, в то время как передискретизация также заметно улучшает SP. Это улучшение, вероятно, связано с уникальным составом и балансом набора данных COMPAS, в отличие от других наборов данных, в которых классы несбалансированы. Успех передискретизации и пропорциональной выборки в этом контексте можно объяснить сбалансированным характером набора данных, который позволяет многократное дублирование существующих строк (методы выборки, используемые этими методами) для улучшения набора данных без внесения значительного перекоса в сторону какой-либо конкретной выборки. сорт. Этот метод эффективно дополняет представление всех классов и защищенных атрибутов в сбалансированной форме, что делает эти методы особенно эффективными для наборов данных, где области объектов в равной степени вносят вклад в прогнозы и где исходные распределения классов не страдают от серьезного дисбаланса . Это можно дополнительно проверить с помощью их макроса F1, поскольку ни одна из моделей не меняет точность на полноту. Улучшение SP при передискретизации можно объяснить большим количеством выборок при передискретизации по сравнению с пропорциональной выборкой. Что касается PC-SMOTE и PC-ADASYN, эти алгоритмы демонстрируют улучшение по сравнению с базовым уровнем как по точности, так и по EOD. Эти тенденции соответствуют предыдущим результатам. Примечательно, что это приводит к значительному падению SP, которое можно объяснить нашей новой меткой, созданной для того, чтобы сместить набор данных в сторону отрицательного класса. Эти результаты показывают сложность оптимизации для двух или более показателей справедливости одновременно и то, как эта оптимизация может влиять друг на друга.

6.2. Влияние глубины дерева на справедливость и

точность. В этом исследовании также изучалось влияние глубины дерева решений на производительность модели , в частности изучалось, как изменения глубины дерева влияют на точность и EOD. Понимание эффекта глубины имеет решающее значение, поскольку оно дает представление о последствиях, варьирующихся от недостаточного до переобучения, и помогает определить оптимальный уровень сложности, при котором и точность, и справедливость максимальны. Первоначально дереву решений разрешалось без ограничений расти на всю глубину, которая в среднем составляла около 30 ветвей. Затем дерево осмотрели визуально, чтобы определить максимальную глубину, исключая нерасщепляющиеся ветки. Для систематического анализа влияния глубины дерева максимальная глубина деревьев могла варьироваться от 1 до 30. Каждый предел глубины оценивался с использованием десятикратной перекрестной проверки, чтобы обеспечить надежность и обобщаемость результатов.

Для каждой конфигурации глубины дерева на тестовом наборе измерялись точность и EOD . Кроме того, для десятикратных показателей были рассчитаны 95% доверительные интервалы . Этот статистический анализ выявил глубину, на которой дерево решений уравновешивало компромисс между точностью и справедливостью, а также учитывало лежащий в основе компромисс между статистической погрешностью и дисперсией. Таким образом, удалось определить «золотую середину» — деликатную точку, в которой дерево решений сохраняет высокую точность прогнозирования без ущерба для справедливости, эффективно противодействуя часто упоминаемому компромиссу, представленному в предыдущей
На рисунках 3–7 показаны графики точности и EOD в зависимости от максимальной глубины для каждого из пяти методов выборки в наборе данных о доходах взрослых.

Machine Translated by Google

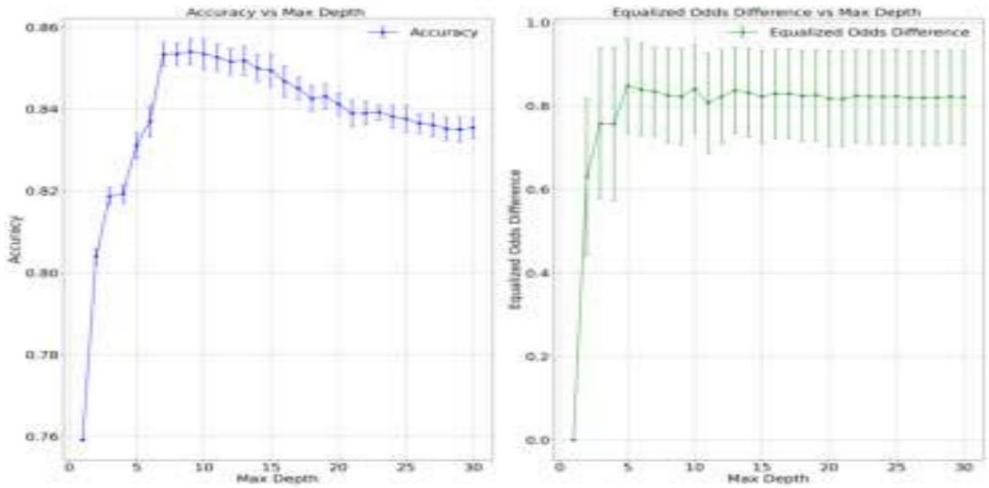


Рисунок 3. Графики доходов взрослых без выборки, показывающие точность и EOD с 95% доверительными интервалами в зависимости от максимальной глубины дерева в диапазоне от 1 до 30.

Судя по результатам, показанным на рисунках 3 и 4, наблюдается заметное первоначальное увеличение точности по мере увеличения максимальной глубины как для методов без выборки, так и для методов с избыточной выборкой. Однако оба метода демонстрируют снижение точности начиная с глубины 10, что указывает на начало переобучения. Соответственно, ЭОД резко уменьшается с увеличением глубины примерно до глубины 10, после чего стабилизируется. Эта закономерность указывает на то, что, хотя более глубокие деревья изначально улучшают справедливость, они в конечном итоге достигают порога, за которым дальнейшего улучшения не наблюдается. Напоминая, что целью справедливости было минимизировать EOD, ключевым наблюдением является то, что установка максимальной глубины от трех до пяти обеспечивает оптимальный баланс между достижением высокой точности и

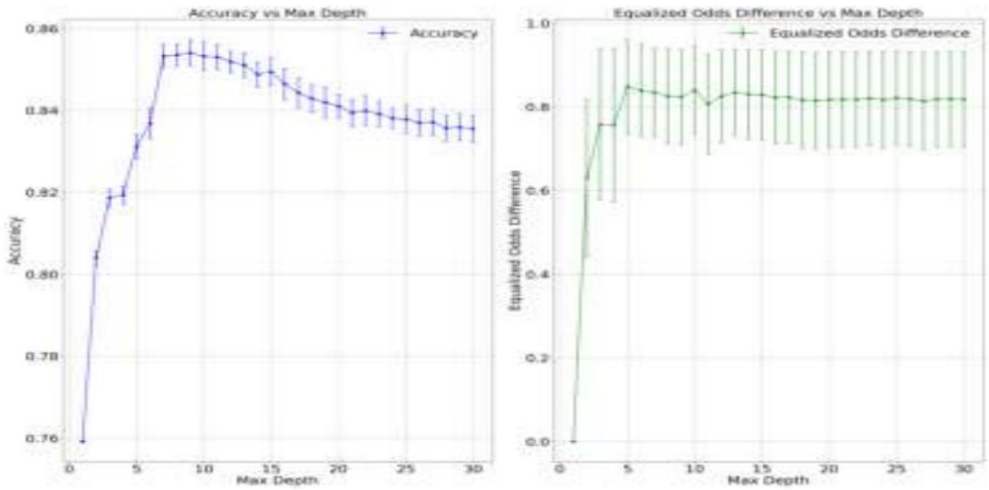


Рисунок 4. Графики доходов взрослых с использованием передискретизации, показывающие точность и EOD с 95% доверительными интервалами в зависимости от максимальной глубины дерева в диапазоне от 1 до 30.

При рассмотрении результатов, показанных на рисунке 5, метод пропорциональной выборки постоянно увеличивает точность с глубиной дерева, достигая максимума на глубине около 26. И наоборот, EOD сначала увеличивается, а затем уменьшается и стабилизируется на глубине около 15. Наблюдаются широкие доверительные интервалы. в метрике EOD предполагают значительную вариабельность результатов справедливости. Этот вывод подчеркивает важность выбора глубины, которая сводит к минимуму изменчивость справедливости при максимальной точности.

Machine Translated by Google

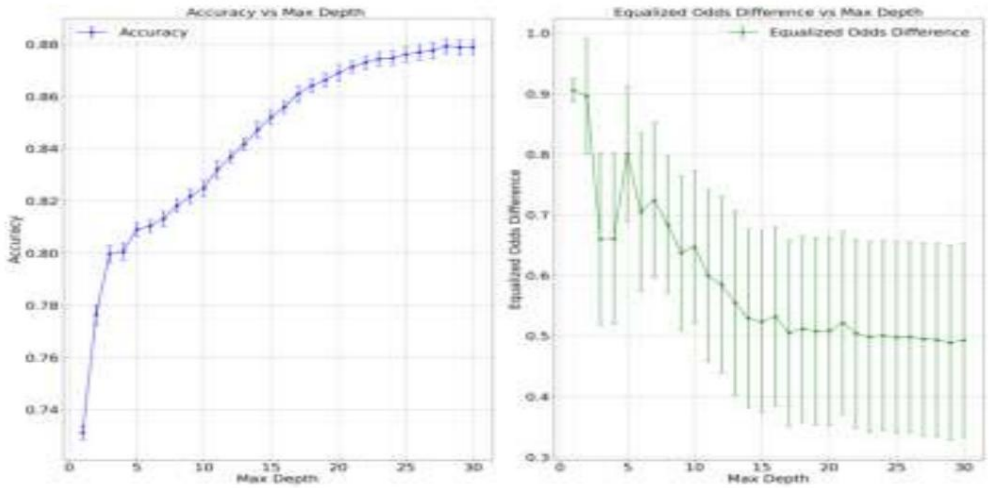


Рисунок 5. Графики доходов взрослых с использованием пропорциональной выборки, показывающие точность и EOD с 95% доверительными интервалами в зависимости от максимальной глубины дерева в диапазоне от 1 до 30.

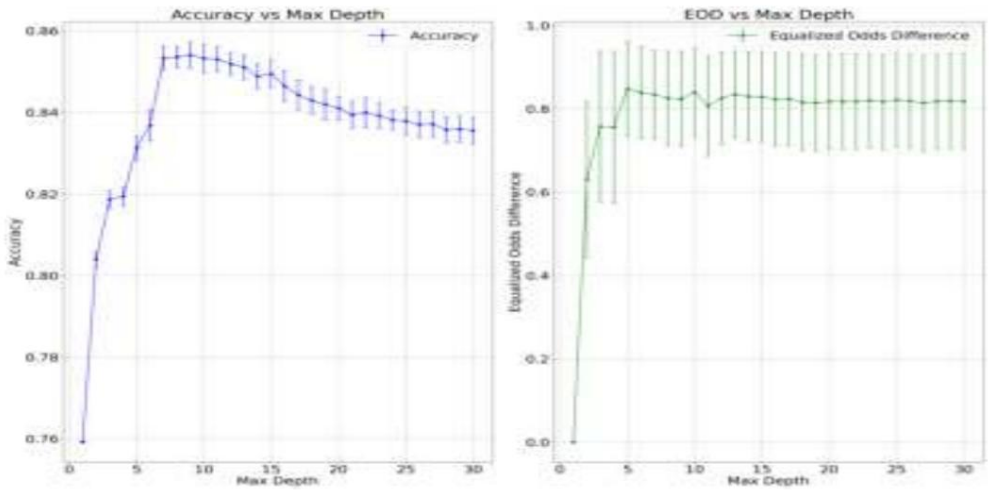


Рисунок 6. Графики доходов взрослых с использованием PC-SMOTE, показывающие точность и EOD с 95% доверительными интервалами в зависимости от максимальной глубины дерева в диапазоне от 1 до 30.

На рисунках 6 и 7, представляющих результаты с использованием PC-SMOTE и PC-ADASYN, соответственно , наблюдаются небольшие тенденции к снижению точности, которые ненадолго улучшаются перед новым снижением – закономерность, указывающая на переобучение на больших глубинах. Показатели EOD для этих методов показывают начальную стабильность на более низких глубинах, всплеск на средних глубинах и снижение, что предполагает сложное взаимодействие между генерацией синтетических образцов и определением границ принятия решений . Учитывая эти наблюдения, для наших экспериментов была выбрана максимальная глубина 3, поскольку она представляет собой «золотую середину», где оптимизируются как точность, так и EO

Учитывая эти результаты, один из выводов состоит в том, чтобы бросить вызов часто предполагаемому компромиссу между точностью и справедливостью, продемонстрировав, что наш метод PC-ADASYN последовательно превосходит базовые показатели для всех трех наборов данных с точки зрения как точности, так и справедливости. Этот вывод важен, поскольку предполагает повышение справедливости модели без ущерба для точности с помощью соответствующих методов выборки и возможной настройки модели. Однако наш анализ также выявляет сценарии, в которых корректировка сложности модели, в частности максимальной глубины деревьев решений, может повысить точность за счет справедливости, о чем свидетельствует увеличение разницы уравненных шансов (EOD). Однако ожидается, что сочетание методов выборки с методами внутренней обработки, такими как регуляризация на основе справедливости, может компенсировать эти эффекты. Эти решения подчеркивают дискреционную власть исследователей в балансировании показателей эффективности модели в зависимости от конкретных целей и ограничений их исслед

Количество выборки и временная сложность аналогичны любому другому методу выборки. Как

Machine Translated by Google

Увеличение количества выборки приводит к увеличению временной сложности, но в целом методы выборки имеют ту же временную сложность, что и лежащие в их основе алгоритмы, поскольку они имеют ту же функциональность.

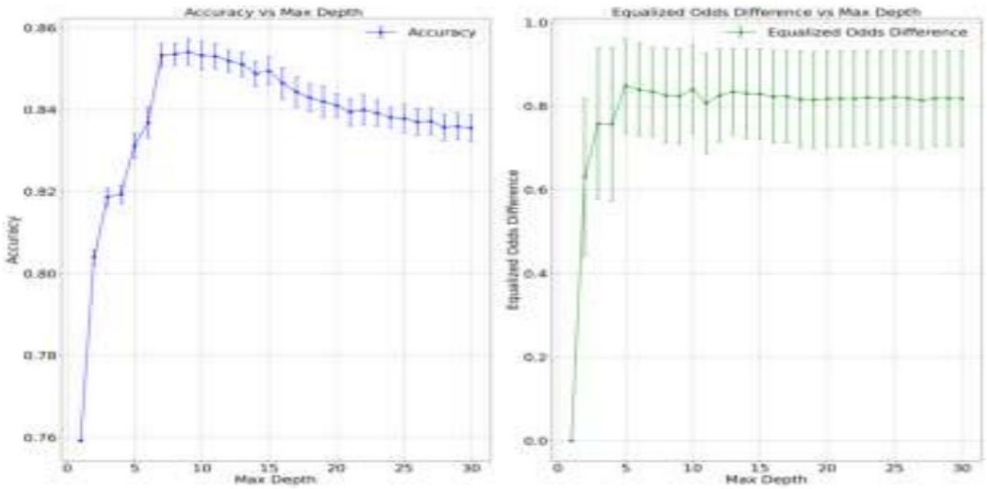


Рисунок 7. Графики доходов взрослых с использованием PC-ADASYN, показывающие точность и EOD с 95% доверительными интервалами относительно максимальной глубины дерева в диапазоне от 1 до 30.

Более того, наши результаты подчеркивают сложность одновременной оптимизации нескольких показателей справедливости. Например, усилия по улучшению статистической четности (SP) путем предпочтения более позитивных прогнозов для каждой защищенной группы в наборе данных COMPAS привели к непреднамеренному сокращению негативных прогнозов. Этот сдвиг отрицательно повлиял на показатель ложноположительных результатов (FPR), компонент EOD, тем самым ухудшив показатель EOD по мере улучшения SP. Этот феномен иллюстрирует присущие математические противоречия между показателями справедливости, когда оптимизация одного может пагубно повлиять на другой. Набор данных COMPAS с его почти сбалансированным распределением по классам представляет собой конкретный пример того, как характеристики набора данных могут влиять на поведение показателей справедливости. Оптимизация SP в этом контексте подразумевает искаженное измерение справедливости, особенно там, где существуют внутренние различия между группами в защищенных атрибутах. Это подтверждается литературными данными, указывающими на то, что SP может неадекватно учитывать групповые различия, что потенциально приводит к ошибочным выводам о справедливости модели [63].

7. Ограничения и будущая работа

Сама природа этого исследования такова, что невозможно рассмотреть все вопросы, связанные со справедливостью и так называемым компромиссом между справедливостью и эффективностью. Таким образом, в работе, представленной здесь, существуют ограничения. Несмотря на это, мы надеемся и намерены, что работа, о которой здесь сообщается, предложит дополнительные направления исследований в этой важной области.

Одним из ограничений этого исследования является то, что наш метод выборки не был специально разработан для оптимизации произвольных показателей справедливости. Другими словами, поскольку часто существуют неизбежные компромиссы между доступным набором показателей справедливости, было принято решение сосредоточиться на подходе, который не зависит от показателей, признавая, что результаты могли отличаться для других показателей. Это также одна из причин, почему мы наблюдали разное поведение между EOD и SP.

Кроме того, признается, что, хотя лежащий в основе метод ОД не должен иметь отношения к предлагаемому методу, он фактически не тестировался. Поэтому в будущем это исследование будет расширено за счет рассмотрения влияния других алгоритмов машинного обучения, таких как логистическая регрессия, нечеткий ID3, K-ближайший сосед и ансамблевые методы, такие как случайные леса или деревья с градиентным усилением, для оценки обобщаемости наших новые методы отбор Целью такого исследования будет проверка того, что наши методы независимы от используемого алгоритма ML. Кроме того, это поможет проверить, являются ли наблюдаемые улучшения в справедливости и точности специфичными для модели или могут применяться повсеместно.

Machine Translated by Google

Кроме того, признается, что рассматривались только три отдельных набора данных — наборы данных , которые широко изучались в этой области. Это вызывает обеспокоенность тем, что методы адаптируются к этим данным, а не решают более широкую проблему справедливости в ОД. Для решения этой проблемы планируются эксперименты с более крупными и разнообразными наборами данных, которые позволят глубже понять масштабируемость и надежность наших методов. Еще одной областью будущей работы является усовершенствование нашего подхода к выборке по нескольким категориям путем включения более детальных подразделений защищенных категорий, что потенциально может выявить более тонкие предубеждения и обеспечить более детальное понимание справедливости.

Наконец, следует признать, что альтернативные методы были предложены для уменьшения предвзятости , и эти методы вообще не изучались в этой работе. Будущая работа повлечет за собой сравнения с другими стратегиями выборки. Будет проведено более прямое сравнение предложенных методов с методами первичной и постобработки. Например , включение методов внутренней обработки, таких как регуляризация [22], или методов постобработки , таких как рандомизированный пороговый оптимизатор [64], будет изучено как возможные средства для дальнейшего улучшения как справедливости, так и производительности.

8. Выводы

В этом исследовании была исследована проблема систематической ошибки прогнозов МО и был разработан метод, основанный на объединении защищенных переменных в новую мультикатегорию. В частности, основное внимание уделялось предложенному в литературе вопросу о компромиссе между предвзятостью и производительностью и поиске метода смягчения этого компромисса. Предлагаемый новый многокатегорийный подход отражает многогранную идентичность людей, признавая сложное взаимодействие атрибутов, которые определяют сценарии реального мира. Учитывая присущий этой мультикатегории дисбаланс, были разработаны четыре метода выборки, адаптированные к этим сложным категориям, а не к традиционным меткам классов. Для применения базового классификатора были обучены деревья решений, и эффективность этих методов оценивалась с использованием трех наборов данных, которые часто используются в исследованиях справедливости. Эффективность методов сравнивалась с базовыми методами без выборки с использованием точности, макроса F1, разницы уравненных шансов (EOD) и статистической четности (SP) в качестве показателей оценки.

Результаты экспериментов показывают, что два недавно разработанных метода выборки — PC-SMOTE и PC-ADASYN — успешно повышают справедливость без ущерба для точности. Примечательно, что в некоторых случаях эти методы также повышали точность, тем самым предоставляя доказательства, опровергающие популярные утверждения о компромиссе между справедливостью и производительностью. Дальнейший анализ влияния максимальной глубины дерева на производительность модели показал, что, хотя увеличение глубины изначально повышает точность, в конечном итоге это приводит к ее снижению. И наоборот , увеличение глубины отрицательно влияет на справедливость, подчеркивая проблему баланса сложности и справедливости. Однако были определены оптимальные глубины дерева, которые одновременно повышают точность и EOD, подчеркивая возможность достижения справедливости без ущерба для производительности.

В конечном счете, эти результаты бросают вызов преобладающим представлениям о неявном компромиссе между производительностью и справедливостью в исследованиях по снижению предвзятости, предполагая, что тщательно разработанные стратегии смягчения предвзятости способны обойти этот компромисс. Наш подход создает новый прецедент для разработки более справедливых алгоритмов прогнозирования путем переопределения того, как защищенные атрибуты используются при обучении модели.

Вклад автора: концептуализация, GP и JS; формальный анализ, GP и JS; расследование, врач общей практики; методология, врач общей практики; программное обеспечение, ГП; супервизия, JS; валидация, GP и JS; визуализация, ГП; письмо — первоначальный вариант, ГП; написание — рецензирование и редактирование, JS. Все авторы прочитали опубликованную версию рукописи и согласились с ней.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: наборы данных и код, реализованные в этой исследовательской работе, были загружены на https://github.com/horlahsunbo/New-folder. (по состоянию на 10 июня 2024 г.).

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Machine Translated by Google

Рекомендации

1. Майна, И.В.; Белтон, ТД; Гинзберг, С.; Сингх, А.; Джонсон, Т.Дж. Десятилетие изучения скрытых расовых/этнических предубеждений в здравоохранении. поставщики, использующие тест неявной ассоциации. Соц. наук. Мед. 2018, 199, 219–229. [Перекрестная ссылка]
2. Салими, Б.; Родригес, Л.; Хау, Б.; Сучу, Д. Причинное восстановление базы данных для обеспечения алгоритмической справедливости. В материалах Международной конференции по управлению данными 2019 г., Амстердам, Нидерланды, 30 июня – 5 июля 2019 г.; стр. 793–810.
3. Кордзаде, Н.; Гасемагаи, М. Алгоритмическая предвзятость: обзор, синтез и направления будущих исследований. Евро. Ж. Инф. Сист. 2022, 31, 388–409. [Перекрестная ссылка]
4. Песах, Д.; Шмуэли, Э. Обзор справедливости в машинном обучении. АКМ Компьютер. Выж. 2022, 55, 1–44. [Перекрестная ссылка]
5. Агаи, С.; Азизи, МJ; Ваянос, П. Изучение оптимальных и справедливых деревьев решений для недискриминационного принятия решений. Учеб. ААААИ Конф. Артиф. Интел. 2019, 33, 1418–1426. [Перекрестная ссылка]
6. Камиран Ф.; Колдерс, Т. Методы предварительной обработки данных для классификации без дискриминации. Знать. Инф. Сист. 2012, 33, 1–33. [Перекрестная ссылка]
7. Кальмон, Ф.; Вэй, Д.; Винзамури, Б.; Рамамурти, КНЦ; Варшни, КР Оптимизированная предварительная обработка для предотвращения дискриминации. В Proceedings of the Advances in Neural Information Processing Systems, Лонг-Бич, Калифорния, США, 4–9 декабря 2017 г.
8. Шахбази, Н.; Лин, Ю.; Асуде, А.; Джагадиш, Х.В. Смещение представления данных: обзор методов идентификации и разрешения. АКМ Компьютер. Выж. 2023, 55, 1–39. [Перекрестная ссылка]
9. Чен, З.; Чжан, Дж. М.; Сарро, Ф.; Харман, М. Комплексное эмпирическое исследование методов уменьшения предвзятости в машинном обучении. классификаторы. АКМ Транс. Программное обеспечение англ. Методол. 2023, 32, 1–30. [Перекрестная ссылка]
10. Пержинский А.; Берг, Калифорния; Томас, К.; Чембалли, А.; Смит, Т.; Шик, С.; Ганзлер, Д.; Сегал, А.Р. Расовая дискриминация и экономические факторы в изменении границ районов Огайо. Буа преподобный Soc. наук. Рез. Гонка 2023, 20, 293–309. [Перекрестная ссылка]
11. Стейл, Дж. П.; Олбрайт, Л.; Раг, Дж.С.; Мэсси, Д.С. Социальная структура ипотечной дискриминации. Дом. Стад. 2018, 33, 759–776. [Перекрестная ссылка]
12. Сальгадо, Дж. Ф.; Москосо, С.; Гарсиа-Искьердо, Алабама; Андерсон, Н.Р. Формирование инклюзивных рабочих мест посредством социального диалога; Springer: Берлин/Гейдельберг, Германия, 2017.
13. Ливи, С.; Мини, Г.; Уэйд, К.; Грин, Д. Смягчение гендерной предвзятости в наборах данных машинного обучения. В материалах о предвзятости и социальных аспектах поиска и рекомендаций: первый международный семинар, BIAS 2020, Лиссабон, Португалия, 14 апреля 2020 г.; стр. 12–26.
14. Хорт, М.; Чен, З.; Чжан, Дж. М.; Харман, М.; Сарро, Ф. Уменьшение смещения для классификаторов машинного обучения: комплексный обзор. ACM J. Ответственные вычисления. 2023, 1, 1–52. [Перекрестная ссылка]
15. Фазе, Т.; Хубер, В.; Ван Гиффен, Б. Управление предвзятостью в проектах машинного обучения. В «Инновациях через информационные системы: Том II: Сборник последних исследований по технологическим проблемам»; Шпрингер: Берлин/Гейдельберг, Германия, 2021 г.; стр. 94–109.
16. Чжан З.; Нил, Д.Б. Выявление значительной предсказательной ошибки в классификаторах. arXiv 2017, arXiv:1611.08292. [Перекрестная ссылка]
17. Чжэн, З.; Кай, Ю.; Ли, Ю. Метод передискретизации для несбалансированной классификации. Вычислить. Поставить в известность. 2015, 34, 1017–1037.
18. Чавла, Невада; Бойер, КВ; Холл, Лоу; Кегельмейер, WP SMOTE: Техника чрезмерной выборки синтетического меньшинства. Дж. Артиф. Интел. Рез. 2002, 16, 321–357. [Перекрестная ссылка]
19. Он, Х.; Бай, Ю.; Гарсия, Э.А.; Ли, С. АДАСИН: Адаптивный синтетический подход к выборке для несбалансированного обучения. В материалах Международной совместной конференции IEEE по нейронным сетям 2008 г. (Всемирный конгресс IEEE по вычислительному интеллекту), Гонконг, Китай, 1–8 июня 2008 г.; стр. 1322–1328.
20. Янссен, П.; Садовский, Б.М. Предвзятость в алгоритмах: о компромиссе между точностью и справедливостью. В материалах 23-й конференции Международного общества электросвязи, проводимой раз в два года, Гетеборг, Швеция, 21–23 июня 2021 г.
21. Хардт, М.; Прайс, Э.; Сребро, Н. Равенство возможностей в контролируемом обучении. В материалах о достижениях в области нейронных технологий Системы обработки информации, Барселона, Испания, 5–10 декабря 2016 г.
22. Камишима, Т.; Акахо, С.; Асох, Х.; Сакума, Дж. Классификатор, учитывающий справедливость, с регуляризатором для устранения предрассудков. В материалах машинного обучения и обнаружения знаний в базах данных: Европейская конференция, ECML PKDD 2012, Бристоль, Великобритания, 24–28 сентября 2012 г.; стр. 35–50.
23. Колдерс, Т.; Жлиобайте И. Почему объективные вычислительные процессы могут привести к дискриминационным процедурам принятия решений. Дискриминация и конфиденциальность в информационном обществе: интеллектуальный анализ данных и профилирование в больших базах данных; Springer: Берлин/Гейдельберг, Германия, 2013 г.; стр. 23–33.
24. Берт А. Как бороться с дискриминацией в ИИ. Гарвардский бизнес-обзор. 2020 г. Доступно онлайн: https://hbr.org/2020/08/how-to-fight-discrimination-in-ai . (по состоянию на 12 июля 2024 г.).
25. Зиглер А.; Адмуссен, В. Обнаружение расовой дискриминации со стороны полиции. Северо-Западный университет. Закон ред. 2021 г., 115, 987–1054.
26. Грабович П.; Перелло, Н.; Мишра А. Как обучать модели, не распространяющие дискриминацию? Блог Equate и Machine Learning , Массачусетский университет, Амхерст. 2022. Доступно онлайн: https://groups.cs.umass.edu/equate-ml/2022/04/07/how-to-train-models-that-do-not-propagate-discrimination/ . (по состоянию на 12 июля 2024 г.).
27. Хани, Ф.; Лян, П. От дискриминации в машинном обучении к дискриминации в законодательстве, Часть 1: Неравномерное обращение. Блог Стэнфордской лаборатории искусственного интеллекта . 2022. Доступно онлайн: https://ai.stanford.edu/blog/discrimination_in_ML_and_law/. (по состоянию на 12 июля 2024 г.).
28. Шахриари, К.; Шахриари, М. Обзор стандарта IEEE — Этически согласованный дизайн: видение приоритета благополучия человека с помощью искусственного интеллекта и автономных систем. В материалах Канадской международной конференции по гуманитарным технологиям IEEE (IHTC), Торонто, Онтарио, Канада, 21–22 июля 2017 г.; стр. 197–201.

Machine Translated by Google

29. Европейская комиссия; Генеральный директорат сетей связи, контента и технологий. Рекомендации по этике для надежного ИИ. Офис публикаций. 2019. Доступно онлайн: https://op.europa.eu/en/publication-detail/-/publication/d39885. 69-0434-11ea-8c1f-01aa75ed71a1/language-en/format-PDF/source-337437547 (по состоянию на 10 июня 2024 г.)
30. Эберс, М.; Хох, ВРС; Розенкранц, Ф.; Рушемайер, Х.; Штайнрёттер, Б. Предложение Европейской комиссии по Закону об искусственном интеллекте – критическая оценка членов Общества юристов по робототехнике и искусственному интеллекту (RAILS). J 2021, 4, 589–603. [Перекрестная ссылка]
31. Хаджян, С.; Доминго-Феррер, Дж. Методика предотвращения прямой и косвенной дискриминации при интеллектуальном анализе данных. IEEE Транс. Знать. Данные англ. 2012, 25, 1445–1459. [Перекрестная ссылка]
32. Фиш, Б.; Кун, Дж.; Лелкес, А.Д. Подход, основанный на доверии, позволяющий сбалансировать справедливость и точность. В материалах SIAM 2016 г. Международная конференция по интеллектуальному анализу данных, Майами, Флорида, США, 5–7 мая 2016 г.; стр. 144–152.
33. Хаджян С. Одновременное предотвращение дискриминации и защита конфиденциальности при публикации и майнинге данных. arXiv 2013, arXiv:1306.6805. [Перекрестная ссылка]
34. Сондек, ЛП; Лоран, М.; Фрей, В. Метрика уровня семантической дискриминации для измерения конфиденциальности, которая ставит под сомнение преимущество t-близости перед l-разнообразием. В материалах 14-й Международной конференции по безопасности и криптографии, Мадрид, Испания, 24–26 июля 2017 г.; Том 6, стр. 285–294.
35. Руджери, С. Использование анонимности t-близости для контроля недискриминации. Пер. Частные данные. 2014, 2, 99–129.
36. Романо, Ю.; Бейтс, С.; Кандес, Э. Достижение равных шансов путем повторной выборки чувствительных атрибутов. В Proceedings of the Advance in Neural Information Processing Systems, онлайн, 6–12 декабря 2020 г.; стр. 361–371.
37. Пэн, К.; Чакраборти, Дж.; Мензис, Т. Фэйрмаск: Повышение справедливости за счет ребалансировки защищенных атрибутов на основе моделей. IEEE Транс. Программное обеспечение англ. 2022, 49, 2426–2439. [Перекрестная ссылка]
38. Дхар, П.; Глисон, Дж.; Рой, А.; Кастильо, CD; Челлаппа, Р. ПАСС: Система подавления защищенных атрибутов для уменьшения предвзятости при распознавании лиц. В материалах Международной конференции IEEE/CVF по компьютерному зрению, Монреаль, Квебек, Канада, 10–17 октября 2021 г.; стр. 15087–15096.
39. Красанакис Э.; Спиромитрос-Сиуфис, Э.; Пападопулос, С.; Компациарис, Ю. Адаптивное чувствительное повторное взвешивание для смягчения предвзятости в классификации с учетом принципа справедливости. В материалах конференции World Wide Web, Лион, Франция, 23–27 апреля 2018 г.; стр. 853–862.
40. Ранчич, С.; Радованович, С.; Делибашич, Б. Исследование методов передискретизации для справедливых моделей машинного обучения. В материалах систем поддержки принятия решений XI: Системы поддержки принятия решений, аналитика и технологии в ответ на глобальное кризисное управление: 7-я Международная конференция по технологиям систем поддержки принятия решений, ICDSST 2021, Лафборо, Великобритания, 26–28 мая 2021 г.; Шпрингер: Берлин/Гейдельберг, Германия, 2021 г.; стр. 110–123.
41. Ян, С.; те Као, Х.; Феррара, Э. Балансировка справедливых классов: повышение справедливости модели без учета чувствительных атрибутов. В материалах 29-й Международной конференции ACM по управлению информацией и знаниями, виртуальное мероприятие, Ирландия, 19–23 октября 2010 г.; стр. 1715–1724.
42. ХуЧжан, Б.; Лемуан, Б.; Митчелл, М. Смягчение нежелательных предубеждений с помощью состязательного обучения. В материалах конференции AAAI/ACM по искусственному интеллекту, этике и обществу, Новый Орлеан, Луизиана, США, 2–3 февраля 2018 г.; стр. 335–340.
43. Селис, Л.Э.; Хуанг, Л.; Кесвани, В.; Вишно, Н.К. Классификация с ограничениями справедливости: метаалгоритм с доказуемыми гарантиями. В материалах конференции по вопросам справедливости, подотчетности и прозрачности, Атланта, Джорджия, США, 29–31 января 2019 г.; стр. 319–328.
44. Зафар, М.Б.; Валера И.; Родригес, МГ; Гуммади, К.П. Справедливость за пределами разного обращения и разного воздействия: классификация обучения без разного жестокого обращения. В материалах 26-й Международной конференции по Всемирной паутине, Перт, Австралия, 3–7 апреля 2017 г.; стр. 1171–1180.
45. Агарвал, А.; Бейгельцимер, А.; Дудик, М.; Лэнгфорд, Дж.; Уоллах, Х. Подход к справедливой классификации с помощью сокращений. В материалах Международная конференция по машинному обучению, Стокгольм, Швеция, 10–15 июля 2018 г.; стр. 60–69.
46. Лоуи, А.; Бахарлуи, С.; Паван, Р.; Разавиян, М.; Бейрами, А. Схема стохастической оптимизации для справедливой минимизации рисков. Пер. Мах. Учиться. Рез. 2022. [Перекрестная ссылка]
47. Спинелли И.; Скардапане, С.; Хусейн, А.; Унчини, А. Фэйрдроп: Смещенное выпадение ребер для повышения справедливости в обучении графическому представлению . IEEE Транс. Артиф. Интел. 2021, 3, 344–354. [Перекрестная ссылка]
48. Хорт, М.; Чжан, Дж. М.; Сарро, Ф.; Харман, М. Файреа: Подход к изменению поведения модели для сравнительного анализа методов смягчения предвзятости. В материалах 29-й совместной встречи ACM по Европейской конференции по разработке программного обеспечения и симпозиума по основам программной инженерии, Афины, Греция, 23–28 августа 2012 г.; стр. 994–1006.
49. Бхаскаруни, Д.; Хм.; Лан, К. Повышение справедливости прогнозирования с помощью модельного ансамбля. В материалах 31-й Международной конференции IEEE по инструментам с искусственным интеллектом (ICTAI), Портленд, Орегон, США, 4–6 ноября 2019 г.; стр. 1810–1814.
50. Иосифидис В.; Фетаху, Б.; Нтуци, Э. Фаэ: Структура ансамбля, учитывающая справедливость. В материалах Международной конференции IEEE по большим данным (Big Data), Лос-Анджелес, Калифорния, США, 9–12 декабря 2019 г.; стр. 1375–1380.
51. Дворк, К.; Хардт, М.; Питасси, Т.; Рейнгольд, О.; Земель, Р. Справедливость через осознанность. В материалах 3-го Нововведения в Конференция по теоретической информатике, Кембридж, Массачусетс, США, 8–10 января 2012 г.; стр. 214–226.
52. Плейсс, Г.; Рагхаван, М.; Ву, Ф.; Кляйнберг, Дж.; Вайнбергер, К.К. О справедливости и калибровке. В материалах о достижениях в Нейронные системы обработки информации, Лонг-Бич, Калифорния, США, 4–8 декабря 2017 г.
53. Карими-Хагиги, М.; Кастильо, К. Улучшение инструмента прогнозирования рецидивов с помощью машинного обучения: эффективность и алгоритмическая справедливость. В материалах восемнадцатой Международной конференции по искусственному интеллекту и праву, Сан-Паулу, Бразилия, 21–25 июня 2017 г.; стр. 210–214.

Machine Translated by Google

54. Фридлер, С.А.; Шайдеггер, К.; Венкатасубраманиан, С. (Не)возможность справедливости: разные системы ценностей требуют разных механизмов для справедливого принятия решений. Коммун. ACM 2021, 64, 136–143. [Перекрестная ссылка]
55. Гарсия, В.; Санчес, Х.С.; Моллинеда, Р.А. Об эффективности методов предварительной обработки при работе с различными уровнями классового дисбаланса. Информационная система. 2012, 25, 13–21. [Перекрестная ссылка]
56. Брейман Л. Деревья классификации и регрессии; Рутледж: Лондон, Великобритания, 2017 г.
57. Педрегоса, Ф.; Варокво, Г.; Грамфорт, А.; Мишель, В.; Тирион, Б.; Гризель, О.; Блондель, М.; Преттенхофер, П.; Вайс, Р.; Дюбур, В.; и другие. Scikit-learn: машинное обучение на Python. Дж. Мах. Учиться. Рез. 2011, 12, 2825–2830.
58. Келли, М.; Лонгджон, Р.; Ноттингем, К. Репозиторий машинного обучения UCI. 2024 г. Доступно онлайн: https://archive.ics.uci.edu. (по состоянию на 10 июня 2024 г.).
59. Беккер, Б.; Кохави, Р. Взрослый. Репозиторий машинного обучения UCI. 1996. Доступно онлайн: https://archive.ics.uci.edu/dataset/2/. взрослый (по состоянию на 10 июня 2024 г.). [Перекрестная ссылка]
60. Хофманн, Х.; Statlog (кредитные данные Германии). Репозиторий машинного обучения UCI. 1994. Доступно в Интернете: https://archive.ics.uci.edu/dataset/144/statlog+немецкий+кредит+данные (по состоянию на 10 июня 2024 г.). [Перекрестная ссылка]
61. Дрессел, Дж.; Фарид, Х. Точность, справедливость и пределы прогнозирования рецидивизма. наук. Адв. 2018, 4, eaao5580. [Перекрестная ссылка]
62. Хуан, Ц.; Мао, Дж.; Лю, Ю. Улучшенный алгоритм поиска по сетке для оптимизации параметров СВР. В материалах 14-й Международной конференции IEEE по коммуникационным технологиям 2012 г., Чэнду, Китай, 9–11 ноября 2012 г.; стр. 1022–1026.
63. Катон, С.; Хаас, К. Справедливость в машинном обучении: опрос. АКМ Компьютер. Выж. 2023, 56, 1–38. [Перекрестная ссылка]
64. Алабдулмохсин И.; Лючич, М. Почти оптимальный алгоритм устранения смещения обученных моделей машинного обучения. В материалах 35-я конференция по нейронным системам обработки информации, онлайн, 6–14 декабря 2021 г.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Безопасная обработка и распространение данных, управляемых конфиденциально Межпланетная файловая система с использованием доказательств с нулевым разглашением

Кёхей Сибано 1,*, Кенсуке Ито [1], Чанхи Хан [2], Цз Тат Чу [2], Ватару Одзаки [2] и Генто Моги [1]

[1] Департамент управления технологиями для инноваций, Школа инженерии, Токийский университет, Токио 113-8656,
[2] Япония Callisto Inc., Токио 171-0022,
Япония * Адрес для переписки: shibano@tmi.tu-tokyo.ac.jp

Аннотация: В этом исследовании предлагается новый метод совместного использования данных, который использует частную межпланетную файловую систему — децентрализованную систему хранения, работающую в закрытой сети, — для распространения данных внешним объектам, обеспечивая при этом проверку их подлинности. Среди двух режимов работы IPFS, публичного и частного, в данном исследовании основное внимание уделяется методу использования частной IPFS. Частная IPFS не открыта для широкой публики; хотя предложенный метод и создает риск фальсификации данных при передаче данных внешним сторонам, он обеспечивает подлинность полученных данных. В частности, этот метод применяет тип доказательства с нулевым разглашением, а именно протокол zk-SNARK Groth16, чтобы гарантировать, что данные соответствуют идентификатору контента в частной IPFS. Более того, имя получателя встраивается в распространяемые данные, чтобы предотвратить несанкционированное вторичное распространение. Эксперименты подтвердили эффективность предложенного метода при размере данных изображения до 120× В будущих исследованиях предложенный метод будет применяться к более крупным и разнообразным типам данных.

Ключевые слова: ИПФС; доказательство с нулевым разглашением; цирк; зк-СНАРКИ; частная IPFS; распространение данных; обработка данных ; безопасность данных

check for updates

Цитирование: Шибано, К.; Ито, К.; Хан, К.; Чу, ТТ; Одзаки, В.; Моги, Г. Безопасная обработка и распространение данных, управляемых в частной межпланетной файловой системе, с использованием нулевого знания Доказательства. Электроника 2024, 13, 3025. https://doi.org/10.3390/электроника13153025

Академический редактор: Ария Гангопадьяй

Поступила: 13 июня 2024 г.
Пересмотрено: 14 июля 2024 г.
Принято: 20 июля 2024 г.
Опубликовано: 31 июля 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Децентрализованные системы надежны, поскольку у них нет единой точки отказа; поэтому они широко применяются в корпоративных секторах, включая криптовалюту, управление цепочками поставок, финансовые услуги и цифровую идентификацию. Для хранения данных большого размера, таких как изображения , этим системам требуются функции хранения, которые по своей сути децентрализованы. Блокчейн, обычно используемый совместно, обычно обрабатывает данные меньшего размера, такие как истории транзакций, и работает как база данных реестра. Межпланетная файловая система (IPFS) — это известная децентрализованная система хранения, которая хранит данные на нескольких узлах для повышения доступности данных. IPFS имеет два варианта: общедоступная IPFS, в которой данные могут храниться любым пользователем с неограниченным доступом, и частная IPFS, в которой создается закрытая сеть, доступная только внутри определенных организаций или групп, что обеспечивает повышенную конфиденциальность и безопасность.

При хранении данных в IPFS крайне важно понимать различия между общедоступной и частной IPFS. Публичная IPFS позволяет любому получить доступ к данным, тогда как частная IPFS доступна только в определенных организациях или группах, что повышает конфиденциальность и безопасность. При хранении конфиденциальной информации, например конфиденциальных данных, в общедоступной IPFS применение соответствующей схемы шифрования имеет жизненно важное значение для обеспечения защиты данных. Напротив, частная IPFS обеспечивает более высокий уровень безопасности хранения данных, поскольку доступна только внутри закрытой сети.

В частности, для таких организаций, как корпорации или медицинские учреждения, хранение данных в общедоступной IPFS, несмотря на использование надежных технологий шифрования, несет в себе неизбежные риски. Более того, в таких случаях постоянно существует вероятность утечки данных из-за операционных ошибок ; хотя данные зашифрованы, они доступны всему миру, что делает их уязвимыми для атак методом перебора и других угроз безопасности.

Machine Translated by Google

Что касается доступности, общедоступная IPFS позволяет обычным пользователям напрямую получать доступ и извлекать данные. данные. Однако в частных IPFS данные должны быть получены от членов организации или группа, составляющая сеть. Если во время этого процесса данные будут подделаны, пользователи могут неспособен его обнаружить. Поэтому доверяя посредникам, ответственным за обработку данных передача в таких случаях становится обязательной. Для решения вышеупомянутой проблемы доверия был разработан новый здесь предлагается метод распространения данных, хранящихся в частной IPFS, внешним объектам. при этом проверяя его подлинность. Протокол Groth16 [1] zk-SNARKs, тип ZKP применяется к данным, хранящимся в частной IPFS, для обеспечения их подлинности. Более того, информация получателя встраивается в распространяемые данные, чтобы предотвратить несанкционированное вторичное распространение. Предлагаемый метод обмена данными важен потому что он адаптирован к частному случаю IPFS.

Различия в нескольких аспектах, включая безопасность и доступность, при хранении данные в общедоступных IPFS и частных IPFS в домене предприятия суммируются в Таблица 1. В этом исследовании предлагаются решения угроз, связанных с частными IPFS.

Таблица 1. Сравнение общедоступной и частной IPFS в корпоративной сфере.

	Публичная IPFS	Частная IPFS
Модель доверия	Доверчивый	Требуется доверие к операционной группе
Ограничения доступа	Доступен любому	Доступно только внутри операционной группы
Риск утечки данных	Постоянный риск утечки из-за ошибки пользователя	Низкий риск утечки в закрытой сети
Обращение с конфиденциальной информацией Информация	Требуется правильное шифрование	Данные, хранящиеся в IPFS, сами по себе не требуют шифрования высокого уровня; есть точка доверия при передаче данных пользователям
Риск атаки методом грубой силы	Всегда присутствует	Низкий
Метод получения данных	Прямой доступ пользователей	Данные, полученные от членов организации или группы
Угрозы	Требуется шифрование, предотвращающее дешифрование неавторизованными пользователями.	Существует риск фальсификации при передаче данных пользователям.

Оставшаяся часть статьи организована следующим образом. В разделе 2 представлены соответствующие предыдущие исследовать. В разделе 3 описаны фундаментальные технологии, а именно ZKP и zk-SNARK. В разделе 4 описывается структура предлагаемого метода, а в разделе 5 излагаются возможные применения этого метода. В разделе 6 представлена реализация этого метода, а в разделе 7 обсуждаются эксперименты, проведенные для проверки эффективности выполнение. В разделе 8 представлено обсуждение экспериментальных результатов, а в разделе 9 представлены выводы документа и обзор будущих задач.

2. Сопутствующие исследования

Существующие децентрализованные системы используют IPFS, особенно в сочетании с блокчейном. технологии. Кумар и др. [2] предложили метод безопасного управления медицинскими данными путем интеграция IPFS с блокчейном. Азбег и др. [3] специально предложил систему, которая управлял и хранил медицинские данные с использованием частной IPFS и разрешенного блокчейна использование повторного шифрования прокси-сервера для обеспечения безопасного дешифрования назначенными врачами. Когда врач получает некоторые данные пациента, он/она получает повторно зашифрованные данные через больницу. Хоссан и др. [4] также предложили систему безопасной записи информации для совместного использования поездок. сервисы с использованием IPFS и частного блокчейна.

Сосредоточив внимание на контроле над распространением данных, управляемых IPFS, Lin et al. [5] предложил систему защиты частных данных с использованием улучшенной IPFS в сочетании с блокчейн. Эта система записывала метаданные файлов и получала разрешения на блокчейн. предоставление пользователям возможности контролировать общий доступ к файлам. Кроме того, в системе реализованы эффективные функции управления с использованием смарт-контрактов, что повышает безопасность и управление данными.

Machine Translated by Google

гибкость. Батта и др. [6] разработали систему, которая использовала многостороннюю аутентификацию (MPA), повторное шифрование прокси и смарт-контракты на блокчейне для децентрализованного контроля доступа к зашифрованным данным, хранящимся в IPFS. Хуанг и др. [7] представили доверенный прокси-сервер IPFS для реализации контроля доступа и управления групповыми ключами для зашифрованных данных, хранящихся в IPFS. Сан и др. [8] предложили систему, которая позволяла только лицам с соответствующими атрибутами расшифровывать зашифрованные данные, хранящиеся в IPFS, с использованием системы шифрования на основе атрибутов политики зашифрованного текста, что способствовало эффективному управлению медицинской информацией. Канг и др. [9] обеспечили распространение данных, управляемых с использованием частной IPFS и частной цепочки блоков, внешним пользователям с использованием именованной сети передачи данных (NDN). Более того, Уддин и др. [10] предложили систему обмена файлами , которая использовала IPFS и технологию инфраструктуры открытых ключей (PKI), не требуя доверенной третьей стороны.

В нескольких исследованиях ZKP использовался для распространения данных. Например, Ли и др. [11] предложили систему управления трафиком, сохраняющую конфиденциальность, которая сочетает в себе неинтерактивные доказательства диапазона с нулевым разглашением и блокчейн. Прототип, использующий Hyperledger Fabric и Hyperledger Ursa, отвечал требованиям конфиденциальности данных для управления трафиком в режиме ре

В этом исследовании предлагается метод соответствующей обработки и распространения данных, управляемых в частной IPFS, пользователям за пределами сети, тем самым предлагая другой подход, чем те, которые предлагались в предыдущих исследованиях. В некоторых исследованиях перешифрование прокси-сервера было принято в качестве подходящего метода хранения и распространения данных в IPFS [3,6]. Используя этот метод, распределенные данные можно повторно зашифровать для расшифровки с помощью закрытого ключа получателя. Более того, при хранении данных в IPFS запись хеш-значения предварительно зашифрованных данных в блокчейн позволяет получателям проверять правильность полученных данных после расшифровки. Однако этот метод не может обрабатывать случаи обработки данных, например встраивания имени получателя в расшифрованные данные, как в этом исследовании.

3. Доказательство с нулевым

разглашением ZKP — это криптографический протокол, который позволяет доказывающему доказать верификатору достоверность предложения, не раскрывая никакой дополнительной информации, кроме достоверности предложения. Предложение этого исследования заключается в том, что данные, предоставляемые внешнему объекту, генерируются на основе заданного CID. Наша цель — позволить члену частной IPFS (проверяющему) доказать это утверждение внешнему лицу (проверяющему как получателю данных) без раскрытия какой-либо другой важной информации (например, прав доступа IPFS и ключей шифрования).

ZKP, в частности протокол Groth16 zk-SNARK, используемый в этом исследовании, начинается с доверенной настройки, в которой обе стороны устанавливают общедоступные параметры, которые имеют решающее значение для безопасного создания и проверки доказательств. В схеме ZKP мы сначала генерируем схему, описывающую процесс, для которого предназначено доказательство. Схема включает в себя условия, которые необходимо проверить, например наличие CID. С помощью схемы ZKP создаются криптографические ключи, в частности ключ подтверждения и ключ проверки. Эти ключи имеют решающее значение для создания доказательства схемы и ее проверки. Используя ключ доказательства и входные данные, средство доказательства генерирует доказательство, которое показывает соответствие выходных данных условиям, указанным в схеме. Затем верификатор использует ключ проверки для проверки доказательства и выходных данных. Если доказательство действительно, оно подтверждает целостность данных без раскрытия какой-либо основной информации.

В этом исследовании обработка Groth16 выполняется с использованием circom [12,13] и snarkjs [14]. Схема процесса представлена на рисунке 1.

zk-SNARKs используется из-за его неинтерактивного характера и эффективности, что особенно выгодно для систем, использующих смарт-контракты, из-за низких вычислительных затрат для проверки доказательств. Кроме того, zk-SNARKs известен своими высокими вычислительными требованиями и потребностью в расширенных характеристиках ПК. Например, в одном из типичных приложений zk-SNARKs, Zcash [15], процесс генерации доказательства занимает более полуминуты для одной анонимной транзакции [16].

Machine Translated by Google

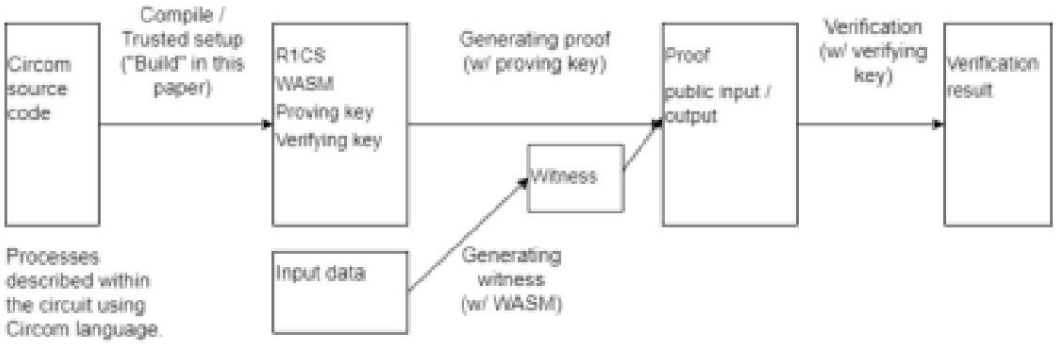


Рисунок 1. Доказательство с нулевым разглашением с использованием circom.

4. Предлагаемая

система На рисунке 2 представлен обзор предлагаемой системы.

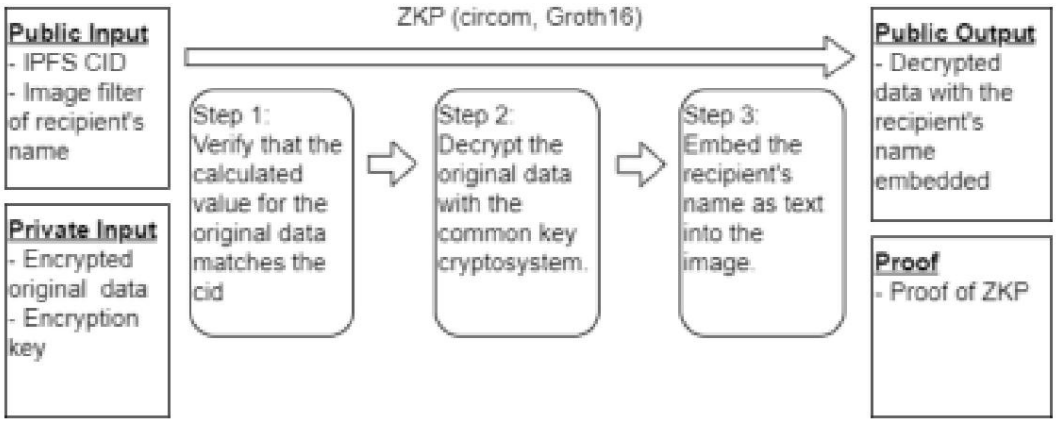


Рисунок 2. Технологическая схема ZKP: пример обработки данных изображения.

Здесь мы делаем следующие предположения: •

Частная IPFS управляется ограниченным числом участников. • Данные

хранятся в IPFS в зашифрованном формате с использованием криптографии с симметричным

ключом. • Ключ шифрования принадлежит исключительно одному лицу среди участников, т.е.

администратор.

Предлагаемая система облегчает создание доказательства ZKP через схему участником, владеющим ключом шифрования (эквивалентным администратору). Входные и выходные данные (кроме доказательств) этого процесса следующие: • Публичные

входные данные: –

CID исходных зашифрованных данных; –

Фильтр для встраивания имени получателя в данные; • Частный ввод:

– Зашифрованные

исходные данные; – Ключ

шифрования; • Выход:

– Расшифрованные данные со встроенным именем получателя.

Обратите внимание, что реализация, оптимизированная для ZK (раздел 6.1), добавляет больше информации к общедоступным данным. В частности, внутренний процесс ZKP включает в себя

следующие шаги: 1. Вычисление CID исходных зашифрованных данных для проверки соответствия с кодом.

Теред CID.

Machine Translated by Google

2. Расшифровка исходных данных с использованием криптографии с симметричным ключом.
3. Встраивание имени получателя в необработанные данные расшифрованного изображения на основе фильтра.
представлено в публичном доступе.

В предлагаемой системе мы не сосредотачиваемся исключительно на данных изображения, но используем их в качестве примеров данных для проверки применимости предлагаемой схемы. Фильтры используются для повышения эффективности обработки внутри схемы. Поскольку встраивание данных имени внутри схемы требует больших вычислительных ресурсов, значительная часть обработки изображения заранее выполняется вне схемы и создается фильтр. Используя общедоступные данные и доказательства, получатель может убедиться, что расшифрованные данные (т. е. выходные данные) со встроенным их именем созданы из исходных данных (содержащихся в частных входных данных), управляемых с помощью CID. Получатель может проверить наличие CID хотя бы у одного участника сети, чтобы подтвердить наличие CID в частной IPFS.

Таким образом, вышеупомянутый процесс позволяет получателю проверить полученное данные, выполнив следующие задачи:

- убедитесь, что полученные данные были сгенерированы из данных, управляемых с помощью CID частная IPFS,
- подтвердите, используя доказательство того, что весь процесс был правильно проведен без непосредственного знания зашифрованных данных или ключа шифрования,
- и • убедитесь, что CID существует конкретно в частной IPFS, запросив хотя бы одного члена сети.

Новизна предлагаемой системы заключается в том, что она позволяет проверять подлинность данных, доверяя хотя бы одному участнику сети, даже если получатель не контролирует ключ шифрования. (Для получателя естественно доверять хотя бы одному члену конкретной многочленной системы. Если ни одному из участников нельзя доверять, то будет незначительный стимул для получения данных, управляемых этой сетью).

5. Возможные применения В

медицинской отрасли данные о диагнозах пациентов обрабатываются в нескольких медицинских учреждениях. Используя предлагаемую систему, пациенты могут проверить, действительно ли получаемые ими данные обрабатываются в частной IPFS для обеспечения их подлинности. Здесь также предлагается универсальная платформа для исследований и разработок машинного обучения с медицинскими изображениями [17]. На этой платформе анонимные медицинские изображения управляются частным IPFS, которым управляет группа медицинских учреждений. Система позволяет исследователям машинного обучения, внешним по отношению к сети, проверять, действительно ли данные изображения управляются в частной IPFS. Более того, встраивая информацию об исследователях машинного обучения в данные изображения, медицинские учреждения могут снизить риск вторичного распространения.

Если приложение не ограничивается вставкой имени получателя, потенциальные возможности применения предлагаемой системы могут быть дополнительно расширены. Например, рассмотрим сценарий, в котором конкретная компания устанавливает частную IPFS для обмена конфиденциальными документами между компаниями своей группы. Если данные сотрудников включены, то сокрытие личных данных и их распространение внешним организациям позволяет этим организациям подтвердить связь сотрудников с компанией, обеспечивая при этом защиту их конфиденциальности.

Кроме того, предположим, что университет настроил частную IPFS, чтобы разрешить доступ к данным об успеваемости студентов только академическому персоналу. В этом случае учащиеся могут убедиться, что полученные ими данные о производительности действительно управляются в частной IPFS.

Таким образом, предлагаемая система поддерживает гибридный вариант — распределение внутренних данных по конкретные внешние объекты по мере необходимости, что распространено в реальных условиях.

6. Реализация Предлагаемая

система была реализована для обработки данных изображения с использованием circom, известного инструмента, специализирующегося на построении схем zk-SNARKs. Circcom позволяет описывать вычислительные процессы внутри схемы, используя свой уникальный язык.

Machine Translated by Google

исполняемый файл, созданный после компиляции, можно вызвать через библиотеку JavaScript snarkjs. Такое расположение позволяет описывать процессы схемы в circom, а внешняя обработка и проверка корректности схемы выполняются с использованием JavaScript. Для зк-CHAPKов использована схема Groth16; он известен своей относительно более высокой скоростью выполнения, чем другие схема зк-CHAPKC.

В частности, мы работали над двумя типами реализаций данных изображений: стандартным реализация с использованием общих криптографических методов и реализация, оптимизированная для ZK, с использованием криптографических методов, дружественных к ZK, для сокращения времени вычисления схема. Эти реализации использовались для сравнения требуемого времени вычислений. Схемы ЗКП требуют значительно большого времени вычислений, даже для вычислений, которые могут легко обрабатываться компьютерным программным обеспечением (это особенно заметно при работе с данные изображения). Поэтому эффективность вычислений имеет решающее значение для практического использования.

6.1. Стандартная реализация

В разделе 4 описан ввод данных в схему. Для упрощения внутрисхемного при обработке исходные зашифрованные данные были отформатированы как совместимые с данными растрового изображения. со стандартами OS/2. Первые 54 байта данных изображения хранят такую информацию, как ширина, высота и глубина цвета изображений [18]. Глубина цвета составляет 8 бит, и каждый цвет компоненту в RGB выделяется один байт, в результате чего на пиксель приходится 3 байта.

Первоначально система проверяет, являются ли зашифрованные данные, введенные как частные входные данные, соответствует CID, предоставленному в качестве общедоступного ввода. Если они не совпадают, система означает ошибка, и данные изображения, выводимые на выходе, представляют собой последовательность байтов, в которой все значения имеют значение 0x00. CID служит важнейшим механизмом для уникальной идентификации файлов и эффективного получение данных из IPFS. CID имеет две версии: V0 и V1 [19]. При этом более гибким использовалась версия CID V1. CID включает в себя хеш соответствующих данных, обеспечивая различные данные будут иметь разные CID. Обычно CID V1 рассчитывается с использованием хеша SHA256. функция, а стандартная реализация использует SHA256 для вычисления CID.

Структура данных CID V1 показана в таблице 2.

Таблица 2. Структура данных CID V1.

Позиция байта	Описание	Ценность в реализации
Первый байт	Многобазовый	0x01
Второй байт	префикс версии CID	0x55: необработанные данные
Третий байт	Идентификатор хеш-функции	0x12: SHA-256
Четвертый байт	Длина хеша	0x20: 32 байта
Начиная с пятого байта	Хэш-значение	Хэш-значение SHA-256 (32 байта)

Кодирование CID осуществляется с использованием Base32. Base32 кодирует последовательность байтов. построенный на основе этой структуры для генерации CID.

Внутри схемы введенное значение CID декодируется из Base32 и система проверяет, соответствует ли извлеченное значение хеша хешу SHA256, вычисленному из зашифрованные данные.

Впоследствии зашифрованные данные расшифровываются. В качестве шифрования используется AES-CTR. алгоритм, который является разновидностью криптографии с симметричным ключом. Шифрование AES-CTR и использовалась расшифровка в circom-chacha20 [20] . Для расшифровки шифрования AES-CTR: Требуются ключ шифрования и одноразовый номер, используемые во время шифрования. Они вводятся в схема как 256-битный ключ и 128-битный одноразовый номер соответственно в качестве частных входов. Более того, AES-CTR обрабатывает объемы данных, кратные 16. Следовательно, если длина изображения данные перед шифрованием не кратны 16, в конец данных добавляются нули (0x00) чтобы привести его в соответствие с этим требованием.

Наконец, к расшифрованным данным применяется фильтр для встраивания имени получателя. Реализация встраивания текста непосредственно в схему может существенно увеличить вычислительную нагрузку; поэтому вне схемы создается фильтр, выполняющий значительную

Machine Translated by Google

часть обработки изображения заранее. Для текста, представляющего имя получателя, используется шрифт Misaki [21]. Затем фильтр используется для оптимизации обработки внутри схемы. Фильтр представляет собой список чисел, в котором значения от 0 до 255 используются для изменения цвета каждого пикселя, если он отличается от цвета пикселей исходного изображения; более того, значение 300 указывает, что цвет должен оставаться таким же, как в исходном изображении. Этот фильтр представляет позицию на изображении, где должно быть вставлено имя получателя.

Внутри схемы указанные цвета пикселей в расшифрованных данных BMP изменяются на основе этого фильтра.

6.2. ZK-оптимизированная реализация

Реализация, оптимизированная для ZK, заменяет хеш-функцию, технологию шифрования и внутрисхемную обработку на стандартную реализацию шифрования, совместимую с ZK. Эта реализация повышает эффективность вычислений и не оценивает разницу в скорости вычислений между ZK-дружественным шифрованием и обычным шифрованием. Поэтому внутрисхемная обработка также была модифицирована.

Хэш Посейдона [22] использовался в качестве хэш-функции для вычисления CID. Примечательно, что использование хеша Poseidon для CID официально не поддерживается; поэтому он был разработан специально для данного исследования. Хотя SHA256 обычно используется в общих вычислениях, он требует значительного времени вычислений в схемах ZKP. Хэш Poseidon реализован в circom и JavaScript (circomlib [23] и circomlibjs [24] соответственно). Он вычисляется над конечным полем простого порядка и может принимать до 16 входных переменных. Используемый порядок меньше максимального значения в 32 байта, но больше максимального значения в 31 байт. Это указывает на то, что каждый из 16 входов должен содержать данные, не превышающие этот порядок. В этой реализации данные, предназначенные для вычисления хеш-функции, делятся на сегменты по 31 байт в качестве входных значений. Если деление превышает 16 сегментов, хеш Посейдона рассчитывается для первых 16 сегментов. Этот результат добавляется к следующим 15 сегментам данных для последующего ввода хэша Poseidon. Процесс повторяется до тех пор, пока все входные данные не будут использованы для вычисления хэша. В вычислительном отношении, если окончательный ввод не завершает 16 сегментов, недостающие входные данные обнуляются, чтобы гарантировать, что вычисление всегда включает 16 вх

При генерации CID из хэш-значения Poseidon последовательность байтов должна соответствовать структуре данных CID V1 и быть закодирована в формате Base32. Однако для дальнейшего сокращения времени вычислений в этой реализации отсутствует кодировка Base32 и напрямую используется хэш-значение Poseidon вместо CID. Разделение входных данных на 16 сегментов внутри схемы требует больших вычислительных ресурсов; следовательно, это деление выполняется вне схемы и подается на вход. В этом случае последовательность байтов зашифрованных данных и список значений для расчета хэша Poseidon предоставляются в качестве общедоступных входных данных, что позволяет проверить, что оба набора данных представляют одну и ту же информацию. Получатели могут подтвердить, что данные, вычисляемые для хэша Poseidon, и данные, расшифровываемые в схеме, идентичны, путем взаимного преобразования и проверки этих двух значений. В этом случае, поскольку пользователи могут получить расшифрованные данные, существует опасность утечки пароля в результате атак методом перебора или других средств.

В качестве технологии шифрования мы использовали шифрование Poseidon [25] вместо шифрования AES-CTR. Шифрование Poseidon, реализованное в circom и TypeScript (poseidon-cryption-circom2 [26]), предполагает получение открытого ключа получателя, генерацию общего ключа и обеспечение безопасного шифрования и дешифрования обеими сторонами. Однако в этом случае общий ключ генерируется напрямую и используется для шифрования и дешифрования. Схема снабжена двумя значениями, представляющими координаты эллиптической кривой, и значением nonce в качестве частных входных данных для шифрования. При этом фильтр реализован так же, как и в стандартной реализации.

7. Оценка

Мы создали пример программы на основе вышеупомянутых реализаций, которая использует circom для описания схемы и использует snarkjs для выполнения схемы и проверки.

Machine Translated by Google

доказательства. В качестве криптографических библиотек circom-chacha20 [20], circomlib [23], circomlibjs [24], и Poseidon-encryption-circom2 [26] .

Для каждой схемы были реализованы стандартная и оптимизированная для ZK реализации. и время их вычислений сравнивалось во время выполнения. Белые растровые изображения были целевыми изображениями, а эксперименты проводились с буквой «A» в качестве встроенный персонаж. Поскольку встраивание любого количества символов не меняет обработку с помощью фильтра встраивание одного символа позволяло сравнивать время вычислений. Кроме того, мы варьировали размеры изображений, чтобы измерить время выполнения каждой схемы. Использовались размеры 10×10, 15×15, 30×15, 30×30, 60×30, 60×60, 120×60, 120×120. и 180 × 120 пикселей. Средой выполнения была Windows 11 с Ryzen 9 3950X. ЦП и 128 ГБ ОЗУ работают под управлением Ubuntu 22.04 в среде WSL2.

На рисунке 3 показан пример изображения, сгенерированного схемой, специально для размера 60 × 30 пикселей. размер с использованием реализации, оптимизированной для ZK. Представлены результаты для каждого размера изображения. в таблице 3, где нелинейные ограничения указывают количество нелинейных ограничений в время сборки — это время, необходимое для компиляции схемы и вывода схемы, а также доказательства gen time — это время, необходимое для генерации доказательств с помощью схемы. Поскольку стандартная реализация использует шифрование AES-CTR, данные с номером 0x00 добавляются в конце, чтобы гарантировать, что размер ввода кратен 16.

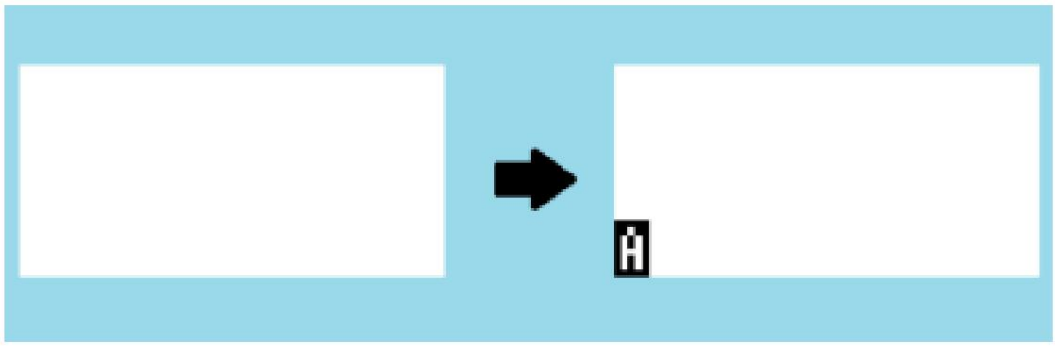


Рисунок 3. Изображение, сгенерированное схемой для размера изображения 60 × 30 с помощью ZK-оптимизированной реализации.

Таблица 3. Сравнение времени выполнения схемы.

	Пиксель	Размер изображения [Байт]	Нелинейный Ограничения	Время сборки [PC]	Доказательство ген. Время [мс]
Стандартный					
	10 × 10	384	558 341	668 220	14 377
	15 × 15	784	1 095 292	1 316 666	25 545
	30 × 15	1440	1 980 688	1 696 370	35 161
	30 × 30	2816	3 867 989	3 657 578	65 785
	60 × 30	5456	7 453 300	7 864 583	126 262
ZK-оптимизированный					
	10 × 10	376	35 407	128 979	3076
	15 × 15	775	72 725	173 210	4032
	30 × 15	1435	134 663	275 630	5900
	30 × 30	2815	263 450	470 872	9733
	60 × 30	5455	509 375	850 612	17 571
	60 × 60	10855	1 013 483	1 257 418	29 044
	120 × 120	43255	4 036 913	6 754 928	96 580

В стандартной и ZK-оптимизированной реализациях для 60×60 пикселей и 180×120 пикселей. размера изображения, в системе закончилась память, и вычисление не удалось завершить. В реализации, оптимизированной с помощью ZK, количество нелинейных ограничений было уменьшено до примерно на одну десятую меньше, чем в стандартной реализации для того же размера изображения.

Machine Translated by Google

Это сократило время сборки и проверки. Однако максимально управляемый размер изображения по-прежнему составлял всего лишь 120 × 120 пикселей, что значительно мало для практических приложений.

8. Обсуждение

Несмотря на то, что во время реализации ZK-оптимизации использовались криптографические технологии, совместимые с ZK, а внутрисхемные процессы были оптимизированы, максимально управляемый размер изображения составлял примерно 120 × 120 пикселей. Это ограничивает практическую полезность изображениями значительно малого размера. Однако исследования, направленные на повышение производительности ZKP, продолжаются, и будущие технологические достижения могут позволить обрабатывать изображения большего размера. Например, Чжан и др. [27] добились десятикратного ускорения zk-SNARK с использованием ASIC. Ма и др. [16] аналогичным образом использовали графические процессоры для ускорения времени создания доказательства, добившись производительности до 48,1 раз выше по сравнению с традиционными методами. Кроме того, предложены методы упрощения вычислительных процессов, например метод «сворачивания». Этот метод сжимает доказываемые предложения [28]. По мере постепенного изучения возможностей повышения скорости потребление памяти, вероятно, также будет оптимизировано. Это потенциально позволит в будущем обрабатывать изображения большего размера.

Кроме того, мы обнаружили, что наш метод предложения может обрабатывать данные размером около 10 КБ. Хотя прямое применение нашего предложения к реалистичным данным изображения (от нескольких МБ до десятков МБ) является сложной задачей, разделение данных на фрагменты путем изменения шифрования и встроенных строк может сделать приложение осуществимым.

Более того, в наших реализациях требуется значение, основанное на размере исходных данных, подлежащих обработке (шифрованию), в качестве аргумента во время генерации схемы. Поэтому для каждых данных необходимо генерировать схему. Время генерации схемы (время сборки) значительно увеличивается с увеличением размера данных изображения; например, даже в реализации, оптимизированной с помощью ZK, для создания схемы для изображения размером 120 × 120 требуется более 112 минут (6 754 928 мс). Однако после генерации схемы время генерации доказательства в тех же условиях невелико, примерно 97 с (96 580 мс). Другими словами, после создания схемы генерация доказательства не требует много времени. Этот факт не создает никаких практических проблем в случаях , когда одно и то же изображение распространяется среди разных людей.

В реализациях, дружественных к ZK, зашифрованные данные вводятся как общедоступные данные. Обработка ключей шифрования изображений требует тщательного рассмотрения. Данные, управляемые в частной IPFS, зашифрованы. Однако в случае утечки ключей шифрования зашифрованные данные могут быть расшифрованы. Таким образом, конкретные пользователи, управляющие частной IPFS, должны стать администраторами, чтобы тщательно управлять ключами, или в той же сети можно создать блокчейн типа консорциума для надлежащего установления и управления правами доступа.

9. Выводы

Здесь был предложен новый метод для распространения данных, хранящихся в частной IPFS, внешним объектам, обеспечивая при этом возможность проверки их подлинности. В этом методе применяется тип ZKP, zk- SNARKs, для проверки CID данных и внедрения имени получателя. Этот подход позволяет внешним объектам проверять, что полученные данные созданы на основе исходных данных в частной IPFS, не требуя таких подробностей, как права доступа IPFS и ключи шифрования.

Для повышения вычислительной эффективности предлагаемого метода были реализованы стандартная реализация с использованием традиционных криптографических методов и ZK-оптимизированная реализация с использованием криптографических схем, дружественных к ZK . Эксперименты с образцом программы подтвердили эффективность предложенного метода при размере данных изображения до 120×120 пикселей.

Этот предлагаемый метод расширяет диапазон использования децентрализованных систем хранения до гибридного случая — при необходимости распределяя внутренние данные по конкретным внешним объектам. Это исследование открывает новый способ обмена конфиденциальной информацией между различными секторами внутри группы и за ее пределами. Однако для широкой практической применимости предлагаемого метода к более крупным и разнообразным типам данных, таким как изображения и видео, скорость обработки должна

Machine Translated by Google

должны быть улучшены и необходимо использовать методы разделения данных, которые находятся в рамках наших будущих исследований.

Вклад авторов: концептуализация, KS, CH, TTC и WO; письмо – подготовка оригинального черновика , KC; написание — рецензирование и редактирование, KC и KИ; надзор, генеральный директор; администрация проекта, K.C. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Это исследование было проведено в рамках совместного исследовательского проекта Токийского университета и Callisto Inc. при финансовой поддержке Callisto Inc. Эта работа была поддержана кафедрой инноваций в области блокчейна и Центром будущих наук и технологий Мохаммеда бен Салмана для Саудовской Аравии. -Японское видение 2030 (MbSC2030) в Токийском университете.

Заявление о доступности данных. Исходный код, использованный для моделирования, доступен на GitHub. https://github.com/blockchaininnovation/circom_image_processing (по состоянию на 21 марта 2024 г.).

Конфликты интересов: авторы Чанги Хан, Цз Тат Чу и Ватару Одзаки работали в компании Callisto Inc. Остальные авторы заявляют, что исследование проводилось в отсутствие каких-либо коммерческих или финансовых отношений, которые могли бы быть истолкованы как потенциальный конфликт интересов. .

Рекомендации

1. Грот Дж. О размере неинтерактивных аргументов на основе пар. В Proceedings of the Advances in Cryptology – EUROCRYPT 2016: 35-я ежегодная международная конференция по теории и применению криптографических методов, Вена, Австрия, 8–12 мая 2016 г.; Слушания, часть II 35; Springer: Берлин/Гейдельберг, Германия, 2016 г.; стр. 305–326.

2. Кумар, С.; Бхарти, АК; Амин, Р. Децентрализованное безопасное хранение медицинских записей с использованием блокчейна и IPFS: сравнительный анализ анализ с будущими направлениями. Безопасность. Прив. 2021, 4, e162. [Перекрестная ссылка]

3. Азбег, К.; Учетто, О.; Андалусси, SJ BlockMedCare: система здравоохранения на основе Интернета вещей, блокчейна и IPFS для данных. безопасность управления. Египет. Поставить в известность. Дж. 2022, 23, 329–343. [Перекрестная ссылка]

4. Хоссан, М.С.; Хатун, М.Л.; Рахман, С.; Рено, С.; Ахмед, М. Обеспечение безопасности службы совместного использования поездок с использованием IPFS и Hyperledger на основе частного блокчейна. В материалах 24-й Международной конференции по компьютерным и информационным технологиям (ICCIT) 2021 г., Дакка, Бангладеш, 18–20 декабря 2021 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2021 г.; стр. 1–6.

5. Лин, Ю.; Чжан, К. Метод защиты частных данных в IPFS. В материалах 24-й Международной конференции IEEE 2021 года по совместной компьютерной работе в дизайне (CSCWD), Далянь, Китай, 5–7 мая 2021 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2021 г.; стр. 404–409.

6. Батта, А.А.; Мадин, ММ; Альзааби, Х.; Якуб, И.; Салах, К.; Джаяраман, Р. Многосторонняя авторизация на основе блокчейна для доступа к зашифрованным данным IPFS. IEEE Access 2020, 8, 196813–196825. [Перекрестная ссылка]

7. Хуанг, Х.С.; Чанг, Т.С.; Ву, Дж. Я. Безопасная система обмена файлами, основанная на IPFS и блокчейне. В материалах 2-й Международной конференции по электронике, Сингапур, 8–10 июля 2020 г.; стр. 96–100.

8. Сан, Дж.; Яо, Х.; Ван, С.; Ву, Ю. Схема безопасного хранения и доступа к электронным медицинским записям на основе блокчейна в IPFS. IEEE Access 2020, 8, 59389–59401. [Перекрестная ссылка]

9. Канг, П.; Ян, В.; Чжэн, Дж. Метод совместного использования частного файлового хранилища Blockchain на основе IPFS. Датчики 2022, 22, 5100. [CrossRef] [ПабМед]

10. Уддин, Миннесота; Хаснат, АНМА; Насрин, С.; Алам, MS; Юсуф, М.А. Безопасная система обмена файлами с использованием технологий блокчейн, ipfs и pki. В материалах 5-й Международной конференции по электрическим информационным и коммуникационным технологиям (EICT) 2021 г., Кхулна, Бангладеш, 17–19 декабря 2021 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2021 г.; стр. 1–5.

11. Ли, В.; Го, Х.; Неджад, М.; Шен, СС Управление трафиком с сохранением конфиденциальности: подход, вдохновленный блокчейном и доказательством с нулевым разглашением . IEEE Access 2020, 8, 181733–181743. [Перекрестная ссылка]

12. Беллес-Муньос, М.; Изабель, М.; Муньос-Тапия, JL; Рубио, А.; Бэйлина, Дж. Сирком: Язык описания схем для построения приложения с нулевым разглашением. IEEE Транс. Надежная безопасность. Вычислить. 2022, 20, 4733–4751. [Перекрестная ссылка]

13. Официальный сайт Circom. Доступно онлайн: https://iden3.io/circom. (по состоянию на 24 марта 2024 г.).

14. Репозиторий Snarkjs на Github. Доступно онлайн: https://github.com/iden3/snarkjs . (по состоянию на 4 июня 2024 г.).

15. ZCash. Доступно онлайн: https://z.cash/ (по состоянию на 12 июля 2024 г.).

16. Ма, В.; Сюн, К.; Ши, Х.; Макс.; Джин, Х.; Куанг, Х.; Гао, М.; Чжан, Ю.; Шен, Х.; Ху, В. Гзкп: Система доказательства с нулевым разглашением, ускоренная на графическом процессоре. В материалах 28-й Международной конференции ACM по архитектурной поддержке языков программирования и операционных систем, том 2, Ванкувер, Британская Колумбия, Канада, 25–29 марта 2023 г.; стр. 340–353.

17. Хан, К.; Шибано, К.; Одзаки, В.; Осаки, К.; Харагути, Т.; Хирахара, Д.; Кимура, С.; Кобаяши, Ю.; Моги, Г. Универсальная платформа для исследований и разработок искусственного интеллекта в области медицинской визуализации, включающая сбор, отбор, аннотирование и предварительную обработку данных. В Proceedings of the Medical Imaging 2024: Imaging Informatics for Healthcare, Research and Applications, Сан-Диего, Калифорния, США, 18–23 февраля 2024 г.; SPIE: Беллингем, Вашингтон, США, 2024 г.; Том 12931, стр. 311–315.

18. Миано, Дж. Форматы файлов сжатых изображений: Jpeg, png, gif, xbm, bmp; Addison-Wesley Professional: Бостон, Массачусетс, США, 1999 г.

Machine Translated by Google

19. Идентификаторы контента (CID). Доступно онлайн: https://docs.ipfs.tech/concepts/content-addressing/#cids-are-not-file-hashes. (по состоянию на 24 марта 2024 г.).
20. circom-chacha20 Репозиторий Github. Доступно онлайн: https://github.com/reclaimprotocol/circom-chacha20. (доступ 24 март 2024 г.).
21. Японский шрифт «Misaki Font» размером 8 × 8 точек. Доступно онлайн: https://littlelimit.net/misaki.htm. (по состоянию на 11 июня 2024 г.). (На японском языке)
22. Грасси, Л.; Ховратович Д.; Рехбергер, К.; Рой, А.; Шофнеггер, М. Посейдон: Новая хэш-функция для систем доказательства с {нулевым разглашением} . В материалах 30-го симпозиума по безопасности USENIX (USENIX Security 21), Ванкувер, Британская Колумбия, Канада, 11–13 августа 2021 г.; стр. 519–535.
23. Репозиторий Circcomlib на Github. Доступно онлайн: https://github.com/iden3/circomlib . (по состоянию на 21 марта 2024 г.).
24. Репозиторий Circcomlibjs на Github. Доступно онлайн: https://github.com/iden3/circomlibjs . (по состоянию на 21 марта 2024 г.).
25. Ховратович, Д. Шифрование с помощью Посейдона. 2019. Доступно онлайн: https://drive.google.com/file/d/1EVrP3DzoGbmzkRmYn.yEDciQcXVU7GlOd/просмотр (по состоянию на 19 июля 2024 г.).
26. Репозиторий Poseidon-Encryption-Circom2 на Github. Доступно онлайн: https://github.com/Shigoto-dev19/poseidon-encryption-c. ircom2 (по состоянию на 21 марта 2024 г.).
27. Чжан Ю.; Ван, С.; Чжан, Х.; Донг, Дж.; Мао, Х.; Лонг, Ф.; Ван, К.; Чжоу, Д.; Гао, М.; Сан, Г. Пипецк: Ускорение доказательства с нулевым разглашением с помощью конвейерной архитектуры. В материалах 48-го ежегодного международного симпозиума по компьютерной архитектуре (ISCA) ACM/IEEE 2021 г., Валенсия, Испания, 14–18 июня 2021 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2021 г.; стр. 416–428.
28. Котапалли, А.; Сетти, С.; Циалла, И. Нова: Рекурсивные аргументы с нулевым разглашением из схем свертывания. В материалах ежегодной международной конференции по криптологии, Санта-Барбара, Калифорния, США, 15–18 августа 2022 г.; Шпрингер: Чам, Швейцария, 2022 г.; стр. 359–388.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Исследование и смягчение компромисса между производительностью и справедливостью с помощью выборки по защищенным категориям

Гидеон Попула и Джон Шеппард *

Школа вычислительной техники Джанфорте, Государственный университет Монтаны, Бозман, MT
59717, США; gideon.popoola@student.montana.edu
* Адрес для переписки: john.sheppard@montana.edu

Аннотация: Алгоритмы машинного обучения стали обычным явлением в повседневном принятии решений, а системы помощи при принятии решений широко распространены в нашей повседневной жизни. Следовательно, в последние годы растет количество исследований по предотвращению и смягчению потенциальной предвзятости и несправедливости прогнозов, сделанных этими алгоритмами . В большинстве исследований по вопросам справедливости и смягчения предвзятости в машинном обучении каждая защищенная переменная часто рассматривается отдельно, но на самом деле один человек может принадлежать к нескольким защищенным категориям. Таким образом, в этой работе было рассмотрено объединение набора защищенных переменных и создание новых столбцов, которые разделяют эти защищенные переменные на множес
Эти новые подкатегории, как правило, крайне несбалансированы, поэтому к смягчению предвзятости подходили как к проблеме несбалансированной классификации. В частности, были разработаны и исследованы четыре новых метода индивидуальной выборки для выборки этих новых подкатегорий. Эти новые методы выборки называются передискретизацией по защищенной категории, пропорциональной выборкой по защищенной категории, методом синтетической выборки меньшинства по защищенной категории (PC-SMOTE) и адаптивной синтетической выборкой по защищенной категории (PC-ADASYN). Эти методы выборки модифицируют существующий метод выборки, ориентируя выборку на новые подкатегории, а не на метку класса. Влияние этих стратегий выборки затем оценивалось на основе классической эффективности и справедливости в условиях классификации.
Эффективность классификации измерялась с использованием точности и F1 на основе обучающих одномерных деревьев решений, а справедливость измерялась с использованием уравненных нечетных разностей и статистической четности. Чтобы оценить влияние справедливости на производительность, эти показатели сравнивались с глубиной дерева решений. Результаты показывают, что предложенные методы смогли определить оптимальные точки, в результате чего справедливость была увеличена без снижения производительности, что смягчило любой потенциальный компромисс между производительностью и справедливостью.

Ключевые слова: справедливость; защищенные категории; машинное обучение; выборка

check for updates

Цитирование: Попула, Г.; Шеппард, Дж. Расследование и смягчение последствий Компромисс между производительностью и справедливостью через Выборка по защищенной категории. Электроника 2024, 13, 3024. https://doi.org/10.3390/electronics13153024 .

Академические редакторы: Нюша Шафиабади и Цзяньлун Чжоу

Поступила: 11 июня 2024 г.
Пересмотрено: 23 июля 2024 г.
Принято: 24 июля 2024 г.
Опубликовано: 31 июля 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Поскольку алгоритмы машинного обучения (МО) все больше доминируют в системах принятия решений и системах помощи в принятии решений, их широкое распространение в различных секторах поднимает насущные вопросы справедливости и прозрачности их прогнозов [1]. Потенциал этих алгоритмов увековечивать или усугублять существующие социальные предубеждения побудил провести значительный объем исследований по изучению и смягчению алгоритмической несправедливости. Это очень важно, поскольку решения, на которые влияют эти алгоритмы, оказывают глубокое влияние на людей, влияя на результаты в самых разных областях: от финансов и занятости до уголовного правосудия и здравоохранения [2].

Источник несправедливости и предвзятости в сфере ОД многогранен [3]. В частности, возможно, что несправедливость возникает непосредственно из самих алгоритмов МО из-за возможного несовпадения основного индуктивного смещения алгоритмов по отношению к целевой концепции и распределению данных. Это называется алгоритмической предвзятостью. Альтернативная проблема заключается в потенциальной предвзятости, присутствующей в данных, используемых для обучения моделей, где, как прямой результат типичного допущения «независимости и одинакового распределения» (IID), используемого в больш

Machine Translated by Google

методов, результатом обучения является распространение смещения в прогнозах таким образом, чтобы они соответствовали смещению в самих базовых данных. Именно эта последняя ситуация и является предметом нашей работы.

1.1. Предвзятость и несправедливость в машинном обучении

Предвзятость — это предвзятое, несправедливое или неравное обращение с человеком или группой, основанное на конкретных характеристиках, часто называемых чувствительными или защищенными характеристиками [4]. Примеры этих защищенных функций включают возраст, расу, инвалидность, пол и пол [5]. Предвзятость в сфере ОД можно грубо разделить на несопоставимое обращение (прямая несправедливость) и воздействие (косвенная несправедливость) [6]. Прямая несправедливость возникает, когда защищенные функции явно используются при принятии решений. Косвенная несправедливость сегодня становится все более распространенной. Этот тип несправедливости не использует явно защищенные атрибуты; вместо этого это происходит, когда использование переменных, связанных с этими атрибутами, приводит к значительно различным результатам для защищенных групп. Эти другие переменные известны как прокси-функции. Примеры реальной предвзятости включают историческую американскую практику «красной линии», когда жителям почтовых индексов, населенных преимущественно меньшинствами, отказывали в выдаче ипотечного кредита, гендерную предвзятость при приеме на работу в Amazon, расовую предвзятость в отношении дозаторов мыла Google и т. д

Хотя эти инструменты помощи при принятии решений помогают автоматизировать процесс принятия решений, эти инструменты могут привести к несправедливому обращению с отдельными лицами или группами, как прямо, так и косвенно [8]. Несправедливость может возникнуть в нескольких областях моделирования, например, в наборе обучающих данных. Это может произойти, когда набор обучающих данных не обеспечивает достоверного представления защищенных категорий, поэтому становится трудно определить «основную истину». Например , рассмотрим набор данных компании, в которой определенная группа населения исторически сталкивалась с дискриминацией. В частности, предположим, что сотрудницы-женщины в этой компании не получили продвижения по службе так же, как их коллеги-мужчины, которые, напротив, добились карьерного роста, несмотря на то, что обе группы работают на одном и том же уровне. В этой ситуации истинная ценность вклада сотрудниц – основная истина – не видна. В результате алгоритм МО, обученный на этих данных, скорее всего, обнаружит и учтет эту предвзятость, тем самым закрепляя существующие предрассудки. Это может привести к тому, что алгоритм будет принимать дискриминационные решения, например, рекомендовать кандидатов-мужчин на работу или продвижение по службе чаще, чем кандидатов-женщин с такой же или более высокой квалификацией.

Другая область, где может возникнуть несправедливость, — это сам алгоритм МО [9]. Алгоритмы МО по-прежнему могут принимать дискриминационные решения, даже если они обучены на объективном наборе данных, где «основная истина» представлена точно. Такая ситуация возникает, когда ошибки системы непропорционально сильно влияют на представителей определенной группы или меньшинства. Например, рассмотрим алгоритм обнаружения рака молочной железы, который демонстрирует значительно более высокий уровень ложноотрицательных результатов у чернокожих по сравнению с белыми, что означает, что он не может чаще выявлять рак молочной железы у чернокожих пациентов, чем у белых пациентов. Если этот алгоритм будет использоваться для информирования о рекомендациях по лечению, он будет ошибочно рекомендовать не лечить большее количество чернокожих, чем белых, что приведет к расовому неравенству в результатах здравоохранения. Это подчеркивает острую необходимость гарантировать, чтобы алгоритмы работали одинаково во всех группах с точки зрения обучающих данных и того, как их ошибки влияют на различные группы населения. Результаты предыдущей литературы сообщают о нескольких случаях применения алгоритмов, приводящих к несправедливому обращению, например, «красной черте» и расовому профилированию [10], ипотечной дискриминации [11], трудоустройству и подбору персонала [12].

Хотя значительные усилия были направлены на устранение систематической ошибки в прогнозах ОД [13,14], большая часть существующих исследований была сосредоточена на смягчении систематической ошибки для отдельных защищенных атрибутов в отдельности [15]. Например, на наборе данных с двумя защищенными атрибутами, расой и полом, большинство существующих подходов могут изучить либо справедливую модель, включающую расу, либо справедливую модель, включающую пол, но не справедливую модель, включающую и расу, и пол [7]. Однако реальные личности не единичны; они сложны и многогранны: люди часто одновременно принадлежат к нескольким охраняемым группам [16]. Например, человек может подвергаться дискриминации одновременно по нескольким защищенным признакам, таким как возраст, раса и пол. Эта интерсекциональность может привести к сложным формам предвзятости и

Machine Translated by Google

дискриминации, которые не могут быть адекватно решены мерами по обеспечению справедливости с использованием одной переменной. Поэтому крайне важно разработать методологии, которые целостно рассматривают многомерный характер личной идентичности. Этот проект направлен на устранение этого разрыва путем рассмотрения комбинаций защищенных категорий, тем самым синтезируя эти защищенные категории в комплексные многокатегорийные группы, а также направлен на более эффективное решение многоуровневых сложностей предвзятости с использованием новых методов выборки защищенных категорий, тем самым признавая и обращаясь к многогранному характеру личных идентичностей и потенциальных предубеждений

Работа, представленная в этой статье, мотивирована проблемой использования алгоритмов МО для принятия решений в социально чувствительных областях, таких как оценка кредита, найм или оценка ипотеки, работая с этой ситуацией, когда человек может принадлежать к нескольким защищенным категориям. Учитывая помеченный набор обучающих данных, содержащий две или более защищенные функции, предлагаемый метод объединяет эти защищенные атрибуты, а затем разбивает их на новые мультикатегории. Эти новые категории, вероятно, будут крайне несбалансированными, и их необходимо сбалансировать, чтобы повысить справедливость прогнозов наших алгоритмов машинного обучения. Популярные методы выборки, такие как избыточная выборка [17], метод синтетической избыточной выборки меньшинства (SMOTE) [18], адаптивная синтетическая выборка (ADASYN) [19] и т. д., осуществляют выборку данных по меткам классов, что не соответствует цели наше исследование выборки по новым мультикатегориям. Следовательно, предлагается новый класс модификаций этих методов выборки, который может осуществлять выборку по новой категории, а не по меткам классов. Этот новый класс модифицированной выборки называется выборкой по защищенной категории. Предлагаемые в результате методы выборки по защищенным категориям используются для выборки и балансировки новых категорий перед выполнением классификации. Новизна данной работы двоякая. Во-первых, предлагаемый подход объединяет защищенные категории для формирования новых мультикатегорий, которые имитируют то, как личность человека выглядит в реальном мире. Во-вторых, это модификация существующих методов выборки, чтобы они соответствовали выборке этих новых категорий, чтобы гарантировать, что все новые категории имеют одинаков

Только в демонстрационных целях в качестве алгоритма классификации было выбрано одномерное дерево решений. Цель состоит в том, чтобы продемонстрировать влияние различных методов выборки на производительность, ожидая, что аналогичные тенденции будут проявляться независимо от основного метода обучения. Предложенный метод выборки сравнивался с базовым уровнем (несемплированные данные) с использованием точности и F1 в качестве показателей эффективности классификации, а также уравненных различий шансов и статистической паритета в качестве показателей справедливости. Кроме того, было проведено несколько анализов, чтобы показать, как максимальная глубина дерева решений влияет на точность и справедливость.

1.2. Вопрос исследования

Исходя из эмпирических наблюдений о том, что иногда существует компромисс между справедливостью и эффективностью ОД [20], это исследование пытается ответить на несколько вопросов, например, как можно смягчить этот компромисс. В частности, мы пытаемся ответить, может ли метод выборки по защищенным категориям для достижения справедливости смягчить этот компромисс. Кроме того, можем ли мы разработать методологическую основу, которая эффективно смягчает предвзятости по этим объединенным защищенным переменным, не ставя под угрозу точность прогнозирования моделей МО? Наконец, мы планируем ответить на вопрос о том, как глубина дерева решений влияет на показатели точности и справедливости, исследуя тем самым взаимосвязь между уровнем соответствия (недооснащение через переоснащение) и справедливостью.

1.3. Гипотеза

Мы предполагаем, что, используя сложные методы выборки по защищенным категориям, разработанные для этих недавно сформулированных многокатегорийных групп, мы можем значительно повысить справедливость модели с точки зрения уравненных различий в шансах, не снижая при этом производительность классификации с точки зрения точности и F1. Кроме того, мы исследуем хрупкий баланс между справедливостью и точностью, предполагая, что можно определить стратегические точки, где справедливость может быть максимизирована без пагубного воздействия на производительность.

Это исследование бросает вызов существующим утверждениям о существовании компромиссов в вопросах справедливости и ОД.

Machine Translated by Google

прогноз. Это закладывает основу для будущих исследований многомерной природы идентичности и дискриминации в автоматизированных системах принятия решений.

1.4. Взносы

Широкая проблема справедливости в машинном обучении важна тем, что распространенность систем искусственного интеллекта и машинного обучения сегодня оказывает серьезное влияние на жизнь и средства к существованию люде
Несмотря на то, что внимание к добросовестному отмыванию денег существенно возросло, по-прежнему существует потребность в методах продвижения добросовестного отмывания денег без негативного влияния на эффективность ОД. Основываясь на углубленном обзоре литературы и указанной выше потребности в такого рода работе, представленные здесь методы вносят

следующий вклад: 1. Общепринятое предположение о том, что существует неотъемлемый компромисс между справедливостью и производительностью (т. е. точностью).) в машинном обучении подвергается сомнению с помощью доказательств , подтверждающих эту проблему. В частности, результаты этой статьи показывают, что такой компромисс можно смягчить, предполагая, что любой компромисс, скорее всего, связан с тем, как управляются данные.

2. Представлены четыре новых метода предварительной обработки данных выборки, основанные на применении стратегии многокатегорийной выборки с использованием данных, собранных в защищенных категориях. Методы исходят из предположения и соответствующей гипотезы о том, что балансировка данных на основе этих многокатегорийных свойств может повысить справедливость, не оказывая негативного влияния на производительность модели машинного обучения.
3. Результаты экспериментов представлены с использованием трех наборов данных, тщательно изученных сообществом справедливого ОД. Эксперименты включают сравнение с традиционными методами обучения без повторной выборки, чтобы продемонстрировать относительные эффекты предлагаемых методов. Результаты показывают, что два из предложенных методов, метод синтетической меньшинственной выборки защищенной категории (PC-SMOTE) и адаптивная синтетическая выборка защищенной категории (PC-ADASYN), особенно эффективны для повышения как справедливости, так и производительности.
4. Подробный анализ потенциального воздействия недостаточного и переоснащения на справедливость представлен путем изучения различных уровней модели дерева решений с использованием и без использования предлагаемых методов выборки. Результаты демонстрируют способность предложенных методов определять идеальный уровень дерева, на котором справедливость и точность максимальны.

Благодаря вышеуказанному вкладу эта работа представляет собой значительный шаг вперед в решении проблем справедливости в машинном обучении. Ключевой вывод из представленных здесь методов и результатов заключается в том, что справедливость можно обеспечить без ущерба для производительности модели.

1,5. Организация

Этот документ организован следующим образом. В разделе 2 представлены подробное объяснение справедливости и обсуждение нескольких технических показателей справедливости. Затем в разделе 3 описывается предыдущая литература, посвященная стратегиям смягчения предвзятости. В разделе 4 мы описываем предлагаемые нами методы выборки, набор данных и подход к настройке гиперпараметров. В разделе 5 мы представляем результаты нескольких экспериментов вместе с проверкой статистических гипотез в качестве средства проверки этих результатов. В разделе 6 обсуждаются экспериментальные результаты и анализируется, как каждый алгоритм работает с каждым набором данных и каждой метрикой. Также представлены дополнительные результаты о влиянии глубины дерева на справедливость и точность . В разделе 7 представлены ограничения этой работы и соответствующие направления будущей работы, а в разделе 8 представлен ряд выводов.

2. Предыстория В

этом исследовании рассматривается вопрос справедливости при прогнозировании результата y Y на основе набора характеристики x X Rd и некоторые дополнительные защищенные атрибуты s S , факторов, таких как раса, Rp , пол и пол. Например, при прогнозировании кредита x представляет финансовую историю заявителя, s — это его раса и пол, о которых он сам сообщил, а y — одобрен ли его кредит.

Machine Translated by Google

или отказано. Модель прогнозирования считается справедливой, если ее ошибки равномерно распределены между защищенными группами, такими как представители разных рас или полов. Предсказания класса на основе обучающих данных D обозначаются как Yˆ D := h(x, s) для некоторого h : X × S Y из класса H. Защищенные атрибуты s S в нашем исследовании предполагаются двоичными со специальным значение n , обозначающее непривилегированную группу. Например, S может означать расу, а n – «небелый»; следовательно, бинарная природа S равна {w, n}, где w представляет белых заявителей, которые являются привилегированной группой, а n представляет небелых заявителей, которые являются непривилегированной группой. Определение можно далее обобщить на небинарные случаи.

Дискриминация в помеченных наборах данных может быть определена как набор данных D, Икс, набор функций и защищенный набор атрибутов S со значением домена {w, n}. Дискриминация в D по отношению к группе S = n , обозначенной как diss=n(D), определяется как

$$\text{diss=n(D)} = \frac{|\{x \quad D : x(s) = w, h(x) = +\}|}{|\{x \quad D : x(s) = w\}|} \quad \frac{|\{x \quad D : x(s) = n, h(x) = +\}|}{\{x \quad D : x(s) = n\}|}$$

Приведенное выше определение можно перевести на разницу в вероятности нахождения заявителя в положительном классе для каждого домена защищенных атрибутов {w, n}. Наше исследование расширяет приведенное выше определение, рассматривая набор данных D, который содержит два или более защищенных атрибута.

Используются два популярных показателя справедливости. Первый — это уравненная нечетная разница (EOD), которая измеряет, насколько дискриминативным или справедливым является наш прогноз. EOD утверждает, что двоичный классификатор yˆ является справедливым, если его уровень ложноотрицательных результатов (FNR) и уровень истинного положительного результата (TPR) равны во всей области S [21]. FPR и TPR относительно защищенного атрибута s S со значением n можно определить как

$$TPRn(\hat{y}) = P(\hat{y} = 1 \,|\, y = 1, S = n)$$

$$FPRn(\hat{y}) = P(\hat{y} = 1 \,|\, y = 0, S = n)$$

Затем EOD определяется математически как разница между TPR и FPR в разных группах защищенного атрибута. То есть,

$$EOD = TPRn(\hat{y}) \quad FPRn(\hat{y}).$$

У справедливого классификатора EOD равен 0, а у несправедливого классификатора EOD равен 1. Хотя достижение полностью справедливого классификатора на практике практически невозможно, это исследование направлено на улучшение EOD без снижения точности. Тогда для EOD

$$FPRn(\hat{y}) = P(\hat{y} = 1 \,|\, y = 0, S = n) = TPRn(\hat{y}) = P(\hat{y} = 1 \,|\, y = 1, S = n)$$

и

$$FPRw(\hat{y}) = P(\hat{y} = 1 \,|\, y = 0, S = w) = TPRw(\hat{y}) = P(\hat{y} = 1 \,|\, y = 0, S = w)$$

Чтобы распространить приведенное выше определение EOD на нашу мультикатегорию, EOD рассчитывается для каждого столбца, затем макросреднее EOD представляется как окончательное EOD. Второй показатель, используемый для измерения справедливости прогнозирования МО, — это статистическая паритетность (СП). С.П. определяет справедливость как равную вероятность быть классифицированным как позитивное [22]. Это можно интерпретировать как то, что каждая группа в защищенном атрибуте имеет одинаковую вероятность быть классифицированной с положительным результатом.

$$P(\hat{y} = 1 \,|S = w) = P(\hat{y} = 1 \,|S = n)$$

3. Обзор литературы

Алгоритмы МО, которые все чаще используются для принятия решений в таких важных приложениях, как рецидивизм, кредитный скоринг, решения о кредите и т. д., изначально можно было бы считать справедливыми и свободными от присущей им предвзятости. Однако в действительности они могут унаследовать любую предвзятость или дискриминацию, присутствующую в данных, на которых они обучаются, как отмечает Берт [23]. Более того, простого удаления защищенных переменных из набора данных недостаточно для борьбы с косвенной дискриминацией и

Machine Translated by Google

на самом деле может скрыть это. Это признание усилило потребность в более совершенных инструментах, сделав выявление и предотвращение дискриминации важной областью исследований, как подчеркивается в [24–27].

Предвзятость в машинном обучении — быстрорастущая тема в исследовательском сообществе машинного обучения. Предвзятость в модели ОД может привести к несправедливому обращению с людьми, принадлежащими к определенным защищенным группам. В последнее время лидеры отрасли начали уделять все больше внимания предвзятости в моделях и программном обеспечении ML. Институт инженеров по электротехнике и электронике (IEEE) [28], Microsoft [29] и Европейский Союз [30] недавно опубликовали принципы справедливого поведения ИИ. Эти организации заявили, что модели машинного обучения должны быть справедливыми в реальных приложениях. Стратегии уменьшения смещения включают в себя изменение одного или нескольких из следующих элементов , чтобы гарантировать, что прогнозы, сделанные алгоритмом ML, менее предвзяты: (a) обучающие данные, (b) алгоритм ML и (c) сами последующие прогнозы. Они, соответственно, подразделяются на подходы предварительной обработки [31], внутренней обработки [32] и постобработки [21].

Во-первых, данные обучения могут быть предварительно обработаны, чтобы снизить несправедливость или предвзятость перед обучением модели. Камиран и Колдерс [6] предлагают выборку или повторное взвешивание данных для нейтрализации дискриминации. Этот подход может скорректировать представление или важность определенных точек данных, чтобы отдать предпочтение (или уменьшить преимущество) одного класса над другим. Другой метод предполагает непосредственное изменение отдельных записей данных, чтобы уменьшить дискриминацию, как показано в [33]. Например, этот подход предполагает изменение значений в наборе данных, чтобы уменьшить заметные предвзятости в отношении определенных групп. Кроме того, концепция t-близости, введенная Sondeck et al. [34], применяется для контроля дискриминации в работе [35]. Использование t-близости гарантирует, что распределение конфиденциальных атрибутов в любой заданной группе будет близко к распределению атрибута во всем наборе данных, тем самым сохраняя конфиденциальность и предотвращая дискриминацию на основе конфиденциальных атрибутов. Общей чертой этих подходов является балансирование контроля дискриминации с полезностью обработанных данных, то есть минимизация систематической ошибки без значительного ущерба для точности, репрезентативности и общей полезности данных для прогнозного моделирования или анализа. Этот баланс необходим для обеспечения того, чтобы усилия по обеспечению справедливости непреднамеренно не снижали качество или применимость данных.

В целом метод предварительной обработки можно разделить на три категории: (1) модификация данных, (2) удаление данных и (3) повторная выборка данных. Методы первой категории направлены на изменение значений точек обучающих данных (включая значения защищенных атрибутов, значения классов и значения функций), чтобы снизить систематическую ошибку в наборе данных. Примером такого метода является массирование данных, предложенное в [15]. Их подход ранжирует данные обучения, а данные, близкие к границе принятия решения, как в привилегированных, так и в непривилегированных группах переворачиваются. В качестве альтернативы Fahse et al. предложил оптимизированный метод предварительной обработки, который изучает вероятностное преобразование, которое редактирует классы и признаки с индивидуальным искажением и групповой справедливостью. [23]. В [36] исходные значения атрибутов заменяются значениями, выбранными независимо от метки класса, чтобы обучить модель, примерно достигая уравненных шансов. Аналогичным образом, Пэн и др. [37] заменяют значения защищенных атрибутов значениями, предсказанными на основе других атрибутов, аналогично вменению данных.

Методы второй категории направлены на обучение справедливой модели путем удаления определенных функций из обучающего набора. Примером этого метода является подавление данных, предложенное Dhar et al. [38]. В их статье защищенные атрибуты и функции, которые тесно связаны с защищенными атрибутами, также известные как прокси-атрибуты, удаляются из набора данных для обучения справедливой модели.

Методы третьей категории направлены на обучение справедливой модели либо путем корректировки весов выборки, либо путем избыточной выборки набора данных. Например, Красанакис и др. [39] предложили метод повторного взвешивания, который итеративно адаптирует веса обучающей выборки с помощью теоретически обоснованной модели для смягчения компромисса между предвзятостью и точностью. В [40] Чакраборти и др. предложил FairSMOTE в качестве метода избыточной выборки обучающих точек из групп меньшинств с искусственными точками данных на основе метода синтетической избыточной выборки меньшинства (SMOTE) [18] для достижения сбалансированного распределения классов. Кроме того, Ян и др. [41] предложили передискретизировать данные обучения групп меньшинств с помощью искусственных точек данных для достижения сбаланси

Machine Translated by Google

ции. В отличие от FairSMOTE, авторы сосредоточились на сценариях, в которых защищенные атрибуты неизвестны, и применили метод кластеризации для идентификации различных демографических групп.

В обработку входят методы, которые изменяют способ обучения модели ML в качестве средства уменьшения систематической ошибки. В [42] был предложен подход состязательного устранения смещений. Этот подход изучает классификатор для повышения точности и справедливости прогноза за счет включения переменной для заинтересованной группы путем одновременного изучения предиктора и противника. Это приводит к созданию беспристрастного классификатора, поскольку прогнозы не содержат никакой информации о групповой дискриминации, которую мог бы использовать противник. В качестве альтернативы в [43] был предложен алгоритм, который принимает метрику справедливости как часть функции потерь и возвращает модель, обученную для этой метрики справедливости . Камишима и др. [22] предложили метод регуляризации , который включал штрафной член в функцию потерь классификатора для получения несмещенного прогноза. Зафар и др. [44] разработали новый метод взвешивания, с помощью которого они настраивают вес выборки для каждого обучающего набора данных для достижения конкретной цели справедливости, например, уравнивания шансов на проверочных данных. В последнее время к уменьшению предвзятости стали относиться как к проблеме ограниченной оптимизации путем добавления ограничения справедливости и оптимизации потерь, чтобы они соответствовали этому ограничению [45,46]. Кроме того, в некоторых работах нейронные сети модифицируются с использованием исключения для удаления нейронов, принадлежащих к защищенным атрибутам [47].

Методы постобработки уменьшают систематическую ошибку после подбора модели машинного обучения и включают такие подходы , как калибровка, оптимизация ограничений и определение порога преобразования [6]. Такие методы предлагают алгоритм, который дает благоприятные результаты непривилегированным группам и неблагоприятные результаты благоприятным группам в пределах заданного доверительного интервала вокруг границы решения с наибольшей неопределенностью. Например, один из подходов изменяет пиковые пороги классификатора, чтобы получить заданную цель равных возможностей или уравненных шансов. Еще один подход предполагает случайное преобразование классов определенных предсказаний в разные классы [48].

Несколько новых исследований [49,50] объединили предварительную, внутреннюю или постобработку для формирования ансамблевого метода. Например, Бхаскаруни и др. [50] сочетают передискретизацию класса, защищенного от дисбаланса, со сдвигом границы решения и метода постобработки для решения проблемы несправедливости.

Исследователи углубились в различные концепции дискриминации и справедливости в рамках алгоритмического принятия решений. Например, несопоставимое воздействие (ранее называемое косвенной справедливостью) измеряется посредством статистического паритета и групповой справедливости, как обсуждалось Бхаскаруни и др. [50]. С другой стороны, концепция индивидуальной справедливости, также представленная Бхаскаруни и др., подчеркивает, что к одинаковым людям следует относиться одинаково, независимо от их групповой принадлежности. Этот подход фокусируется на справедливости на индивидуальном уровне, гарантируя, что решения принимаются на основе соответствующих качеств, а не групповых стереотипов или предубеждений.

Как подчеркнули Чжан и Нил [16], ключевой целью классификаторов и других прогностических моделей является достижение одинакового уровня ошибок в разных группах. Аналогичным образом, обеспечение калибровки или отсутствие предсказывающей предвзятости в прогнозах, как обсуждалось Hardt et al. [21] имеет решающее значение. Однако противоречие между этими понятиями — калибровкой и равными коэффициентами ошибок — исследуется Дворком и др. [51] и Pleiss et al. [52], что указывает на то, что одновременное удовлетворение обоих требований может оказаться сложной задачей. Карими-Хагиги и Кастильо [53] представляют соответствующую работу, исследующую сложности, присущие достижению алгоритмической справедливости. Фридлер и др. [54] далее исследует компромиссы при соблюдении различных определений алгоритмической справедливости, особенно с точки зрения общественной безопасности. Учитывая, что наша работа сосредоточена на предварительной обработке, а не на моделировании, такие соображения, как сбалансированная частота ошибок и предсказательная погрешность, становятся менее применимыми напрямую.

Судя по нашему обзору различных методов предварительной обработки, выяснилось, что не проводилось никаких работ по моделированию справедливости для двух или более защищенных атрибутов одновременно. Кроме того, метод выборки, использованный в предыдущей работе, был ориентирован только на выборку на основе меток классов, а не защищенных категорий. Следовательно, в этой статье особое внимание уделяется предварительной обработке, поскольку она представляет собой наиболее адаптируемый аспект конвейера обработки данных [55]. Предварительная обработка отличается тем, что она не зависит от выбора алгоритма моделирования и может быть легко выполнен

Machine Translated by Google

включены в механизмы выпуска и публикации данных. Эта независимость и гибкость делают предварительную обработку критически важной для обеспечения качества и справедливости данных перед выполнением любого аналитического или прогнозного моделирования. Наконец, мы сосредоточимся на новых методах выборки, которые выбирают защищенную категорию при обучении данных для построения справедливой модели.

4. Методология.

Целью нашей работы является изучение методов выборки для повышения справедливости в машинном обучении без соответствующего снижения эффективности прогнозирования, тем самым смягчая компромисс между справедливостью и производительностью . В результате предлагаются четыре новых метода выборки, ориентированных на достижение этой цели. Эти методы выборки решают проблему несбалансированного класса, возникающую из-за новой мультикатегории, созданной из-за комбинации защищенных категорий. Необходимы специальные методы выборки, поскольку существующие методы отбирают данные на основе классов меньшинства и большинства, но для снижения справедливости выборка из новых мультикатегорий осуществляется на равной основе. Это, в свою очередь, требует изменения существующих методов выборки для выборки данных на основе этих новых категорий. Это приводит к появлению четырех новых методов выборки: передискретизация по защищенной категории, пропорциональная выборка по защищенной категории, SMOTE по защищенной категории (PC- SMOTE) и ADASYN по защищенной категории (PC-ADASYN).

4.1. Передискретизация защищенных

категорий. При передискретизации защищенных категорий первым шагом является объединение защищенных категорий в наборе данных и кодирование этой комбинации для создания нашей новой мультикатегории. Например, в наборе данных о доходах взрослых возраст (молодой и взрослый), раса (белый и другие) и пол (мужской и женский) объединяются для создания восьми новых категорий: ADULTWHITEMALE, ADULTOTHERSMALE, YOUNGWHITEMALE, YOUNGOTHERSMAIL, ADULTWHITEFEMALE, ADULTWHITEMALE, ADULTWHITEMALE, YOUNGWHITEMALE, YOUNGOTHERSMAIL, ADULTWHITEFEMALE, ADULTOTHERSFEMALE, YOUNGWHITEFEMALE и Эти новые категории имеют разные размеры выборки, и цель нашей передискретизации защищенных категорий — сбалансировать эту новую категорию так, чтобы размер выборки каждой из новых категорий соответствовал размеру категории с наибольшим размером выборки. Чтобы избежать утечки данных, набор данных разделяется на обучающий и тестовый, при этом передискретизация применяется только к обучающим данным, а затем выполняется тестирование на несемплированном тестовом наборе.

Псевдокод в алгоритме 1 подробно показывает наш метод передискретизации защищенной категории. В алгоритме в качестве базовой использовалась самая большая категория, поскольку это категория с наибольшим размером выборки. В результате процесса выборки получаются новые обучающие данные со сбалансированным размером выборки по новой категории. Алгоритм работает путем выборки остальных защищенных категорий в соответствии с размером выборки базового уровня. Эта выборка выполняется путем многократного повторения категорий вместе с их метками классов.

4.2. Пропорциональная выборка защищенной категории

Метод пропорциональной выборки по защищенной категории представляет собой обобщение передискретизации по защищенной категории, поскольку процесс начинается с установки целевого размера выборки (который является настраиваемым гиперпараметром, а не просто размером наибольшей мультикатегории), обозначаемого как targetSamples. Это соответствует желаемому количеству экземпляров, необходимых для каждой категории. Эта цель обеспечивает единообразие по всем категориям, снижая риск смещения модели в сторону более частых категорий. Типичный результат применения этого метода заключается в том, что некоторые категории, которые имеют больше выборок, чем целевые выборки, будут иметь недостаточную выборку, в то время как другие будут иметь избыточную выборку, чтобы получить равную их долю в наборе обучающих данных. Псевдокод в алгоритме 2 показывает шаг за шагом метод пропорциональной выборки защищенной категории.

Machine Translated by Google

Алгоритм 1: передискретизация защищенной категории

```
baselineCount  сумма записей в «Largest_Category» Xtrain
totalCount  количество записей в Xtrain
baselineProportion  baselineCount/totalCount
BalanceData  инициализировать пустой набор данных
категории  список имен столбцов в Xtrain, начинающихся с «combined_category_»
для каждой категории в категориях делайте
    CategoryData  выбрать записи в Xtrain , где категория = 1
    CategoryData  объединить CategoryData с соответствующими метками из ytrain
    CategoryCount  количество записей в категорииData
    targetCount  целая часть totalCount × baselineProportion
    если CategoryCount < targetCount , то
        sampledData  образец targetCount из категорииData с заменой
        сбалансированные данные  добавить sampledData к сбалансированным данным
    еще
        сбалансированные данные  добавить категориюДанные к сбалансированнымДанным
конец, если
конец
вернуть сбалансированные данные
```

Алгоритм 2. Пропорциональная выборка защищенной категории.

```
целевые выборки  5000
sampledBalanced  инициализировать пустой набор данных
для каждого столбца в new_categories.columns выполните
    CategoryRows  выбрать строки в new_categories , где столбец = 1
sampledRows  образец targetSamples записей из категорииRows с заменой
    для каждого столбца в oneHotEncodedBalanced.columns выполните
        sampledRows[col]  0
конец для
    sampledRows[столбец]  1
    sampledBalanced  добавить sampledRows к sampledBalanced
конец
возврат выборки сбалансированный
```

4.3. Защищенная категория SMOTE

Метод передискретизации синтетического меньшинства защищенной категории (PC-SMOTE)

Метод выборки — это более сложный процесс, направленный на имитацию SMOTE, но модифицированный для выборка наших новых категорий, а не меток классов. При таком подходе первым шагом было измените SMOTE, чтобы использовать фиксированное количество соседей и случайным образом выбирать одного соседа для интерполяция, а не усреднение всех из них. Псевдокод в алгоритме 3 показывает процедура ПК-SMOTE. Поскольку цель метода состоит в том, чтобы использовать его для новых выборка категорий, она не касается непосредственно генерации меток классов. Следовательно, новый необходима функция, которая может генерировать новую метку класса для синтетических данных. Для этого новый определена функция, которая генерирует метки классов на основе количества новых синтетических данных генерируется и предварительно выбранное соотношение баланса между двумя классами. Алгоритм 4 показывает, как наша новая функция генерирует метки для наших синтетических образцов. Алгоритм сначала определяет количество образцов для каждого класса на основе соотношения баланса и генерирует образец необходимо для каждого класса. Метки классов перетасовываются, чтобы предотвратить алгоритмическую предвзятость в созданные классы.

Эти два алгоритма объединяются вместе для формирования PC-SMOTE, как показано в алгоритме 5. В подходе к достижению баланса между несколькими категориями каждая отдельная категория повторяется. так, чтобы было идентифицировано подмножество данных, связанных с этой категорией. Номер синтетических образцов, необходимых для достижения заранее определенного максимального размера по каждой категории, затем рассчитано. Если требуются дополнительные образцы, данные генерируются с помощью PC-SMOTE.

Machine Translated by Google

который интерполирует между существующими точками данных и их ближайшими соседями. Одновременно, сбалансированное распределение меток синтетических классов создается с указанным соотношением баланса с помощью с использованием алгоритма 4. Эти синтетические функции и метки затем включаются в подмножество обучения для каждой категории. Процесс повторяется для всех категорий, в результате чего сбалансированный набор данных. Гиперпараметрами в этом алгоритме 5 являются количество соседей. и соотношение баланса.

Алгоритм 3. Пользовательский метод синтетической передискретизации меньшинства (SMOTE)

```
процедура CUSTOMSMOTE(данные, n_samples)
    синтетические образцы  нулевая матрица размера (n_samples× количество столбцов в данных)
3: nn  NearestNeighbors(n_neighbors = 7).fit(данные)
    соседи  nn.kneighbors(data, return_distance = False)
    для i  от 1 до n_samples do
        sampleIdx  случайное целое число от 0 до (количество строк в данных — 1)
        nnIdx  случайный выбор из соседей[sampleIdx, 1 :]
        di ff  данные [nnIdx]  данные [sampleIdx]
        вес  случайное число из равномерного распределения от 0 до 1
        синтетические образцы[i]  данные[sampleIdx] + вес × разница ff
конец для
    вернуть синтетические образцы
завершение процедуры
```

Алгоритм 4. Генерация сбалансированных синтетических меток

```
процедура GENBALSYNTHLABELS(n_samplesNeeded, BalanceRatio)
    nClass1  int(n_samplesNeeded × BalanceRatio)
    nClass0  n_samplesNeeded — nClass1
    синтетические метки  [0] × nClass0 + [1] × nClass1
    SHUFFLE(syntheticLabels) возвращает                    Случайным образом перетасуйте метки.
    синтетическиеLabels
завершение процедуры
```

Алгоритм 5 SMOTE защищенной категории

```
balancedDataList  инициализировать пустой список
для каждой категории в категориях делаем
    CategorySubset  выбрать строки в train_data st 'combined_category' == категория
    функции  удалить «класс», «комбинированная_категория» из категорииПодмножество
    nSamples  max_size — количество строк в категорииSubset
    если nSamples > 0, то
        синтетические функции  PCSmote( функции, nSamples)
        синтетическиеLabels  GenBalSynthLabels(nSamples, BalanceRatio)
        синтетическиеОсобенности[' class' ]  синтетические метки
        синтетическиеОсобенности[' комбинированная_категория' ]  категория
        категорияПодмножествоСбалансированное  объединить категорииПодмножество и синтетическиеОсобенности
    еще
        категорияПодмножествоСбалансированный  категорияПодмножество
конец, если
добавить категориюSubsetBalanced в BalanceDataList.
конец
сбалансированные данные  добавить сбалансированный список данных и сбросить индекс
```

4.4. Защищенная категория ADASYN

Метод ADASYN защищенной категории имитирует адаптивное синтетическое меньшинство (ADASYN). выборки, но немного изменен для достижения нашей цели выборки по защищенным категориям. Наш алгоритм PC-ADASYN показан в алгоритме 6. Он расширяет ADASYN, фокусируясь на категории

Machine Translated by Google

плотность, а не классовый дисбаланс. В частности, эта функция работает путем поиска ближайшего соседей данных, а затем вычисление плотности категории каждой точки данных в ее пределах. ближайшее соседство. Он взвешивает эти плотности в обратном порядке, чтобы отдать приоритет категориям меньшинств. что повышает вероятность создания синтетических выборок из недостаточно представленных категорий. синтетические образцы создаются путем интерполяции между выбранными точками данных и их соседями, аналогично SMOTE, но использует случайный вес для изменения интерполяции, что обеспечивает разнообразный синтетический набор данных. Такой подход помогает устранить дисбаланс на уровне категорий и обогащает дисперсию набора данных, потенциально повышая надежность и справедливость моделей ML. обучен на этих данных. Поскольку этот метод выборки также генерирует новые выборки путем интерполяции, Алгоритм 4 используется для генерации меток классов для новых синтетических образцов.

Алгоритм 6 PC-ADASYN для балансировки по категориям

```
процедура PCADASYNCATEGORIES (данные, метки, n_samplesNeeded, n_neighbours)
    n_neighbours  n_neighbours + 1 3: nn                    Включая саму точку данных
NearestNeighbors(n_neighbours).fit(данные)
    расстояния, индексы  nn.kneighbors(данные)
    плотности  нулевой массив длины (данные)
    для i  0 до длины (данные) - 1 сделать
        текущая_категория  метки[i]
        соседние_индексы  индексы[i][1 :]                   Пропустить самоуказатель
        плотности[i]  СУММА(метки[neighbor_indices] == текущая_категория)
конец для
    веса  1/(плотности + 1) веса  веса/                     Добавьте 1, чтобы предотвратить деление на ноль.
    SUM(веса) синтетические образцы  пустой                 Нормализовать веса
    список
    sampleIndices  случайный выбор с заменой по длине (данным) с использованием весов
    для каждого idx в sampleIndices выполните
        baseIdx  idx
        Idx соседа  RANDOMCHOICE(index[baseIdx][1 :])
        di ff  данные [идентификатор соседа]  данные [baseIdx]
        синтетический образец  data[baseIdx] + RANDOM() × d ff
        добавить синтетический образец к синтетическим образцам
конец
    возвращаемый массив (синтетические образцы)
завершение процедуры
```

Алгоритмы 4 и 6 объединяются для формирования нашей выборки ADASYN защищенной категории. как показано в алгоритме 7. Для каждой категории данные, соответствующие этой категории, изолируются, и вычисляется дефицит размера относительно самой крупной категории. Если дополнительные необходимы образцы, применяется метод PC-ADASYN, генерирующий синтетические характеристики, которые соблюдайте характеристики распределения категории. Эти функции затем дополняются с синтетически сгенерированными метками, поддерживающими заранее заданное соотношение классов. процесс не только исправляет дисбаланс категорий, но и обогащает набор данных, потенциально повышение точности прогнозирования и справедливости моделей, обученных на этих данных.

4.5. Настройка набора данных и гиперпараметров

Для проверки четырех методов отбора проб был необходим классификатор, позволяющий оценить влияние на справедливость и производительность. В конечном счете, тип классификатора не имеет прямого значения, поскольку Цель состоит в том, чтобы смягчить компромисс между справедливостью и производительностью, а не найти лучший классификатор. Поэтому мы решили использовать одномерные деревья решений на основе CART [56] из-за их устойчивость к шуму и недостающим данным. Конкретная реализация, которую мы выбрали, была взято из библиотеки sklearn (версия 1.5.1) [57]. Кроме того, деревья решений позволяют нам контролировать силу соответствия, устанавливая глубину изученного дерева. Это позволяет нам сравнить справедливость и производительность на разных уровнях соответствия.

Machine Translated by Google

Процесс начинается с объединения защищенных категорий в каждом наборе данных, применения горячего кодирования для создания новых многокатегорийных объектов, а затем выполнения кодирования меток. Затем наборы данных были разделены с использованием стратифицированной 10-кратной перекрестной проверки, чтобы обеспечить репрезентативное распределение классов в каждом сгибе. Для каждого сгиба обучение проводилось на выборочных данных с использованием ранее описанных методов, а классификация тестировалась на соответствующих несемплированных тестовых наборах. Согласованность в обучении модели поддерживалась за счет применения одинаковой глубины дерева для всех методов выборки, а предоставленные результаты представляют собой средние значения 10-кратных прогонов с соответствующими доверительными интервалами.

Алгоритм 7 ADASYN защищенной категории

```
1: BalanceDataList инициализировать пустой список 2:
для каждой категории в категориях do
3: CategorySubset выбрать из new_data3 st 'combined_category' == функции категории удалить
4: 'class', 'combined_category' из категорииПодмножество CategoryLabels извлечь
5: 'combined_category' из категорииSubset nSamplesNeeded max_size минус
6: количество строк в категорииПодмножество, если nSamplesNeeded > 0, тогда
7: синтетические функции
8: PCadasyn( features, CategoryLabels, nSamplesNeeded) синтетическиеLabels
9: GenBalSynthLabels(nSamplesNeeded, BalanceRatio) синтетические функции[ ′ class ]
10: синтетические метки синтетические характеристики
11: [ ′ комбинированная_категория ] категорияSubsetBa копье
12: объединить категориюSubset с синтетическими функциями еще категорияSubsetBalanced
13:
14: CategorySubset 15: завершить, если 16: добавить

категориюSubsetBalanced к сбалансированномуDataList 17: завершить
для 18:
сбалансированныеДанные объединить сбалансированныйDataList и сбросить индекс
```

Для нашего анализа из репозитория UCI [58] были выбраны три набора данных : набор данных о доходах взрослых [59], набор данных German Credit [60] и набор данных «Профилирование управления исправительными правонарушителями для альтернативных санкций» (COMPAS). Набор данных о доходах взрослых предназначен для прогнозирования того, зарабатывает ли человек более 50 000 долларов США, и включает восемь категориальных и четыре числовых атрибута с защищенными переменными, соответствующими возрасту (молодой или взрослый), полу (мужской или женский) и расе (белый или другие). Набор данных о доходах взрослого населения был передан в дар UCI в 1996 году. Набор данных German Credit, используемый для прогнозирования кредитоспособности, включает 20 категориальных и два числовых атрибута с защищенными переменными возраста и пола. Набор немецких кредитных данных был передан в дар UCI в 1994 году. Набор данных COMPAS [61], который оценивает уровень рецидивизма в Соединенных Штатах, включает шесть категориальных и шесть числовых характеристик с защищенными переменными возраста, расы и пола. Набор данных был опубликован в 2018 году. Эти наборы данных были выбраны потому, что они представляют собой самые современные наборы данных для измерения предвзятости и дискриминации и широко используются в других исследованиях алгоритмической предвзятости и справедливости (см. Раздел 3). Кроме того, наборы данных имеют разные размеры: от маленьких до больших, что делает их подходящими для тестирования наших методов

Настройка гиперпараметров проводилась с использованием поиска по сетке [62] для изучения широкого диапазона параметров, дополненного визуальными оценками для определения оптимальных настроек, которые уравновешивают разницу уравненных шансов (EOD) и точность. Для набора данных о доходах взрослых оптимальные гиперпараметры включали максимальную глубину дерева 3, а для PC-SMOTE и PC-ADASYN — настройку ближайшего соседа 5 со сбалансированным коэффициентом 0,34. Эти параметры были столь же эффективны для набора данных German Credit. Для набора данных COMPAS максимальная глубина дерева 2 была оптимальной для всех методов выборки. PC-SMOTE и PC-ADASYN были настроены на настройку ближайшего соседа 3 и сбалансированное соотношение 0,60.

Machine Translated by Google

5. Результаты

Четыре стратегии выборки были применены к трем наборам данных, описанным выше, и оценили их влияние, используя простой одномерный классификатор дерева решений. Результаты в В таблицах 1–3 показаны заметные различия в эффективности моделей при пяти стратегиях выборки: без выборки, передискретизация, пропорциональная выборка, PC-SMOTE и PC-ADASYN включены наши три набора данных. Каждый метод оценивался по точности, макросу F1, Equalized Разница шансов (EOD) и статистическая паритетность (SP). Результаты были измерены с точностью и макрос F1, поскольку эти две метрики являются наиболее популярными метриками классификации. Также, ограничение показателя двумя делает результаты сопоставимыми для статистического анализа.

Чтобы убедиться в статистической значимости результатов каждого метода, мы использовали метод Фридмана. тест, непараметрическая альтернатива однофакторному дисперсионному анализу с повторными измерениями. На обнаружив значимые результаты теста Фридмана, мы приступили к апостериорному исследованию Неменьи. тест. Этот тест используется для оценки парных сравнений между методами установления какие методы статистически отличаются друг от друга. Преимущество теста Неменьи в этом параметре, поскольку он учитывает множественные сравнения, не предполагая нормального распределения, тем самым предоставляя надежный способ понять конкретные парные различия.

Для набора данных о доходах взрослого населения (таблица 1) метод без выборки дал точность 0,82, устанавливая высокий базовый уровень для сравнения. Однако оно продемонстрировало несколько предвзятое отношение прогнозирование с EOD 0,36 и минимальным несоответствием в показателях прогнозирования (SP = 0,02). В напротив, передискретизация сохранила ту же точность, но немного снизила макро F1. до 0,65, что указывает на потенциальные проблемы переобучения и ухудшение справедливости (EOD = 0,66) и увеличивающееся неравенство в показателях прогнозирования (SP = 0,09). Пропорциональная выборка уменьшена точность до 0,79, но улучшил макрос F1 до 0,79, что предполагает лучший баланс между точность и отзыв. Однако это значительно увеличило SP до 0,71, что указывает на существенную неравенство в показателях положительных прогнозов, что вызывает опасения по поводу справедливости модели. два специальных подхода для SMOTE и ADASYN были разработаны специально для улучшения по этим показателям. PC-SMOTE показал умеренную производительность с точностью 0,81 и улучшенный EOD 0,25, что предполагает повышенную справедливость по сравнению с базовой передискретизацией. и метод без выборки. Тем не менее, он по-прежнему зафиксировал более низкий макрос F1 (0,63), что указывает на проблемы искажения при генерировании синтетических данных. PC-ADASYN оказался наиболее сбалансированный подход, сохраняющий высокую точность (0,82) и лучшую обработку класса дисбалансы, с умеренным улучшением справедливости (EOD = 0,28) и контролируемым увеличение разницы в скорости прогнозирования (SP = 0,09). В целом, базовая точность статистически значительно лучше, чем пропорциональная выборка, хотя она не является статистически значимой, поскольку по сравнению с другими методами отбора проб. Для макроса F1 пропорциональная выборка статистически значительно лучше, чем другие методы выборки. Для EOD: PC-ADASYN статистически лучше, чем другие методы выборки, тогда как метод SP без выборки статистически значительно лучше, чем другие методы выборки.

Таблица 1. Результаты методов выборки по набору данных о доходах взрослых с доверительными интервалами 95%.

Выборка Метод	Точность	Макро Ф1	ОВП	СП
Без выборки	0,82 ± 0,00	0,66 ± 0,01	0,36 ± 0,18	0,02 ± 0,00
Избыточная выборка	0,82 ± 0,02	0,65 ± 0,06	0,66 ± 0,2	0,09 ± 0,02
Пропорциональная выборка	0,79 ± 0,05	0,79 ± 0,06	0,46 ± 0,12	0,71 ± 0,10
ПК-SMOTE	0,81 ± 0,05	0,63 ± 0,07	0,25 ± 0,21	0,07 ± 0,00
ПК-АДАСИН	0,82 ± 0,04	0,64 ± 0,07	0,28 ± 0,19	0,09 ± 0,02

Результаты экспериментов с набором данных German Credit также демонстрируют различное влияние. каждой стратегии выборки, особенно в отношении справедливости и точности, как показано в таблице 2. Без выборки базовая модель достигла точности 0,72, но продемонстрировала значительные смещение в прогнозе с EOD 0,88, что указывает на существенное несоответствие ошибок. ставки между группами в защищенных атрибутах. Реализация передискретизации поддерживается точность при улучшении макроса F1 до 0,68 и заметном снижении EOD до 0,37, хотя и на

Machine Translated by Google

стоимость увеличения SP до 0,35, что подчеркивает потенциальный компромисс между различными вариантами справедливости
меры. Пропорциональная выборка немного снизила точность до 0,68, но добилась наилучшего результата.
F1-оценка 0,69. Это также снизило EOD до 0,32, что позволяет предположить, что это эффективно уравновешивает прогноз.
качество со справедливостью. PC-SMOTE показывает улучшение точности до 0,73, но ниже
макро F1 0,55; справедливость модели показывает огромное улучшение по сравнению с базовым уровнем с
EOD 0,15 и умеренный SP 0,1. PC-ADASYN показывает такую же точность, как и
базовый уровень 0,72, хотя самый низкий макро-F1 равен 0,48, что предполагает потенциальный компромисс.
в точности и отзыве. Тем не менее, модель демонстрирует наилучшие результаты прогнозирования справедливости при
EOD 0,13 и SP 0,06. В целом, результат показывает, что точность отсутствия выборки
не является статистически значимым для других методов выборки, кроме пропорциональной выборки
в то время как для EOD результаты всех методов выборки статистически значимы в
по сравнению с безвыборочным методом.

Таблица 2. Результаты методов выборки на основе набора данных German Credit с доверительным интервалом 95%.

Выборка Метод	Точность	Макро Ф1	ОВП	СП
Без выборки	0,72 ± 0,03	0,62 ± 0,06	0,88 ± 0,10	0,07 ± 0,01
Избыточная	0,72 ± 0,05	0,68 ± 0,04	0,37 ± 0,16	0,35 ± 0,11
выборка Проп.	0,69 ± 0,04	0,69 ± 0,07	0,32 ± 0,26	0,29 ± 0,10
Выборка PC-SMOTE	0,73 ± 0,05	0,55 ± 0,09	0,15 ± 0,09	0,1 ± 0,02
ПК-АДАСИН	0,72 ± 0,03	0,48 ± 0,11	0,13 ± 0,02	0,06 ± 0,00

В Таблице 3 показаны результаты нашего опыта работы с набором данных COMPAS. Эти результаты
выявить значительные различия в производительности модели при различных стратегиях выборки.
Базовый подход без выборки позволил достичь точности и макропоказателя F1, равного
0,89, но показал более высокие различия в показателях справедливости с уравненной разницей шансов.
(EOD) 0,39 и статистического паритета (SP) 0,29. Это подчеркивает потенциальные предубеждения, которые
нескорректированные модели могут относиться к защищенным группам. Применение передискретизации
немного улучшилась точность до 0,90, но также значительно улучшилась справедливость, уменьшив EOD до
0,26. Это предполагает эффективность сокращения различий в результатах без ущерба для
СП. И наоборот, пропорциональная выборка, одновременно повышая точность и макро F1 до 0,90 и
0,91, соответственно, также достигло EOD 0,26, что улучшило его по сравнению с базовым уровнем, в то время как
также зафиксирован более высокий SP, равный 0,36, что указывает на потенциальное увеличение неравенства положительных
результаты в группах. PC-SMOTE и PC-ADASYN с одинаковыми показателями точности,
макросу F1 и SP удалось сохранить улучшение справедливости с EOD 0,30,
хотя эти методы также увеличили SP до 0,47. В целом результаты показывают, что оба
точность и EOD наших методов отбора проб статистически значительно лучше, чем
безвыборочный метод.

Таблица 3. Результаты методов выборки по набору данных COMPAS с доверительными интервалами 95%.

Выборка Метод	Точность	Макро Ф1	ОВП	СП
Без выборки	0,89 ± 0,04	0,89 ± 0,04	0,39 ± 0,15	0,29 ± 0,17
Oversample	0,90 ± 0,03	0,90 ± 0,05	0,26 ± 0,14	0,25 ± 0,07
Prop. Sample PC-	0,90 ± 0,05	0,91 ± 0,04	0,26 ± 0,12	0,36 ± 0,10
SMOTE	0,91 ± 0,02	0,91 ± 0,02	0,30 ± 0,11	0,47 ± 0,19
ПК-АДАСИН	0,91 ± 0,02	0,91 ± 0,03	0,30 ± 0,13	0,47 ± 0,21

6. Обсуждение

Результаты экспериментов с тремя наборами данных подтверждают, что защищенная категория
выборка может заметно повысить справедливость модели, часто без существенного ущерба для точности
прогнозирования. В некоторых случаях также отмечалось улучшение точности и макрос F1.
продемонстрировано. Ориентируясь на результаты набора данных о доходах взрослых, PC-SMOTE и PC-ADASYN

Machine Translated by Google

продемонстрировали заметные улучшения в EOD и сохранили умеренный уровень SP. Эффективность этих методов во многом объясняется их сложными методами интерполяции. Например, визуальный анализ деревьев решений, созданных без выборки и с помощью PC-ADASYN, позволяет получить четкое представление о контрастах. Примеры одной репрезентативной складки показаны на рисунках 1 и 2 соответственно. Дерево решений, полученное без выборки, выбрало свой корень с признаком, тесно связанным с защищенными атрибутами, действуя таким образом как прокси-атрибут. Это привело к выраженной ошибке прогнозиро
И наоборот, дерево решений, обученное на данных, сгенерированных с помощью PC-ADASYN, начиналось с функции , которая очень хорошо обобщала прогнозы и смягчала предвзятость, о чем свидетельствует заметное повышение справедливости модели и более высокая примесь Джини, что указывает на более чистое начальное разделение.

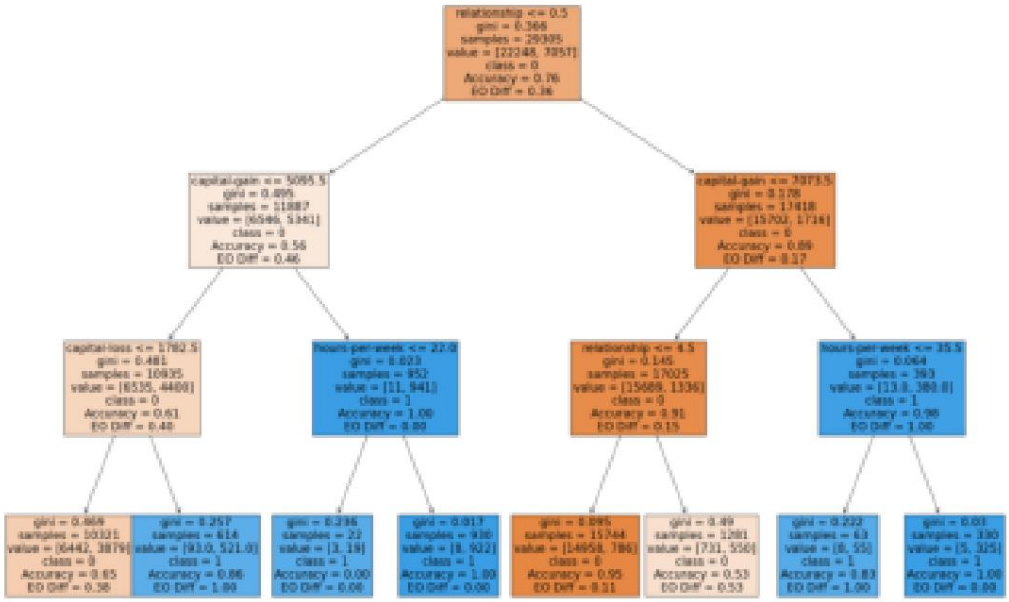


Рисунок 1. Пример дерева решений, обученного на основе доходов взрослого населения без выборки.

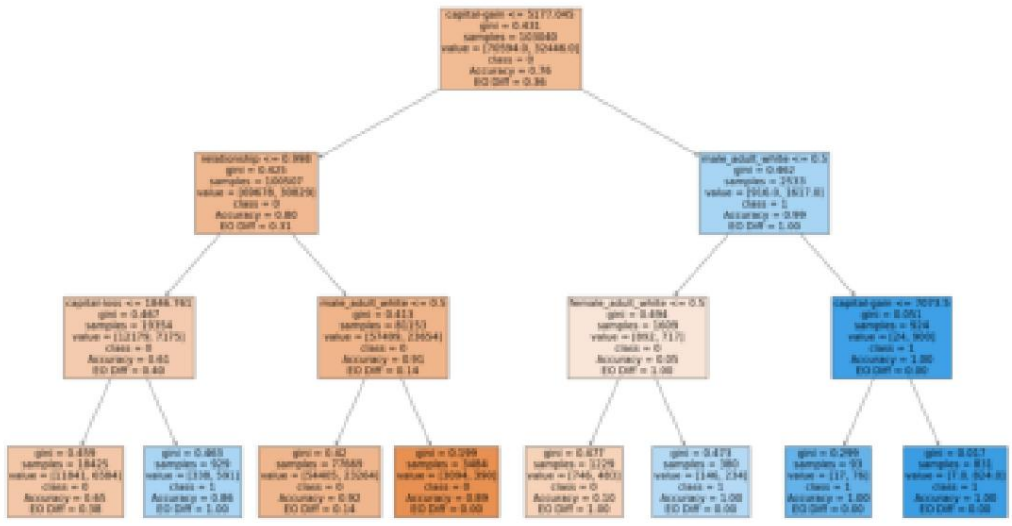


Рисунок 2. Дерево решений PC-ADASYN по доходу взрослого населения.

Machine Translated by Google

6.1. Сравнение справедливости и производительности

При сравнении передискретизации и пропорциональной выборки подходы этих методов к увеличению размера выборки путем дублирования существующих строк данных оказались простыми и не привели к существенному улучшению EOD. Этот результат имеет смысл, поскольку эти методы имеют тенденцию воспроизводить существующие предубеждения, которые потенциально могут усугубить проблемы справедливости, а не облегчить их. Это становится очевидным, в частности, при рассмотрении набора данных о доходах взрослых, где классы крайне несбалансированы. Этим наивным стратегиям репликации не хватает возможностей интерполяции PC-SMOTE и PC-ADASYN для корректировки выборок вблизи границ решения, что имеет решающее значение для уменьшения систематической ошибки в наборе данных. Напротив, стратегии интерполяции, используемые SMOTE и ADASYN, расширяют набор данных и повышают его разнообразие. Это особенно эффективно для выборок вблизи границ решения, где небольшие сдвиги в характеристиках могут существенно повлиять на справедливость прогнозов. Интерполируя между выборками, SMOTE и ADASYN эффективно перемещают эти граничные выборки к более справедливым областям пространства признаков, тем самым напрямую противодействуя и уменьшая смещение более эффективно, чем методы, которые увеличивают объем выборки без изменения структуры данных.

Функция генерации классов (алгоритм 4) также помогает увеличить общее распределение классов. Результаты наших PC-SMOTE и PC-ADASYN по доходам взрослых также демонстрируют превосходство над результатами, полученными в [41], где была получена точность 0,59 и SP 0,17. Кроме того, результаты PC-SMOTE и PC-ADASYN показывают превосходство над результатами, полученными в [40], где EOD 0,89 и немного лучшая точность 0,84.

Изучая результаты набора данных German Credit, мы наблюдали тенденцию, аналогичную той, которая была отмечена в наборе данных о доходах взрослых: метод без выборки имеет очень высокую погрешность в отношении EOD. Низкое значение SP, равное 0,07, указывает на минимальную разницу в показателях положительных прогнозов между группами, но это само по себе не является хорошим способом измерения справедливости, поскольку в предпочитаемой группе больше образцов, чем в неблагоприятной. Это приводит к искажению расчета SP, поскольку учитываются только положительные решения в каждой группе, на которые влияет размер выборки. Одним из выводов является важность использования нескольких показателей справедливости для всестороннего анализа влияния модели на все заинтересованные стороны. При передискретизации мы увидели улучшение EOD при аналогичном SP; это показывает, что увеличение количества выборок для каждого из защищенных атрибутов мультикатегории повышает справедливость в отношении EOD. Кроме того, обновленный SP отражает то, как будет выглядеть наличие более равного количества образцов для каждой мультикатегории, в отличие от базового варианта, где в предпочтительной группе в пять раз больше образцов, чем в неблагоприятной группе. Точность пропорциональной выборки падает, поскольку базовое количество выборок, выбранных после настройки гиперпараметра, было недостаточным для того, чтобы модель могла обобщить набор тестов без выборки, что привело к переобучению. Переобучение было подтверждено с учетом точности обучения. Интересно, что модель не меняет точность на полноту, как другие модели, и это дает пропорциональную выборку самого высокого макроса F1.

Что касается справедливости, мы обнаружили увеличение EOD по сравнению с базовым уровнем и избыточной выборкой. Это происходит потому, что каждая мультикатегория представлена в одних и тех же базовых количествах. Это может повысить справедливость модели, поскольку теперь модель имеет более широкое представление о категориях и делает более точные прогнозы и, в конечном итоге, более справедливые решения. PC-SMOTE и PC-ADASYN играют ключевую роль в значительном уменьшении систематической ошибки в прогнозах Такая согласованность подтвердила надежность этих методов в различных наборах данных. Примечательно , что ни один из методов не ставит под угрозу точность, в то время как оба повышают справедливость, иллюстрируя их эффективность в решении компромиссов, обычно связанных с прогнозным моделированием.

Эти результаты демонстрируют сильные возможности интерполяции, присущие PC-SMOTE и PC-ADASYN. Эти методы эффективно перераспределяют выборки в пространстве признаков, особенно перемещая образцы из малообеспеченных регионов от отрицательных к более положительным границам принятия решений. Такие корректировки имеют решающее значение для смягчения предвзятых результатов и обеспечения справедливости в автоматизированных процессах принятия решений. Макрос F1 в обеих моделях снижается по сравнению с базовым уровнем, поскольку для того, чтобы модель лучше работала при обобщении, требуется большее количество выборок , что этот набор данных не поддерживает. В частности, набор данных содержит 700 образцов для класса 0 и 300 образцов для класса 1, что означает, что набор тестов

Machine Translated by Google

имеет только 30 образцов для класса 1. Из-за такого небольшого количества образцов обе модели обменивают полноту на точность в классе 1. Примечательно, что мы наблюдали низкую полноту для класса 1, что в конечном итоге приводит к низкому показателю f1 для класса 1, и поскольку макрос F1 усредняет два показателя f1 и обрабатывает их одинаково, это влияет на производительность обеих моделей в макросе F1. В целом эти две модели дают более справедливую модель с хорошей точностью по сравнению с базовым и двумя другими методами выборки.

Оценка набора данных COMPAS еще раз подтверждает эффективность наших методов выборки . Выявленные отчетливые закономерности совпадают с теми, которые наблюдаются в наборах данных о доходах взрослых и немецких кредитах, что подчеркивает надежность наших выводов. Примечательно, что методы передискретизации и пропорциональной выборки продемонстрировали существенные улучшения в выравненной разнице шансов (EOD) и точности, в то время как передискретизация также заметно улучшает SP. Это улучшение, вероятно, связано с уникальным составом и балансом набора данных COMPAS, в отличие от других наборов данных, в которых классы несбалансированы. Успех передискретизации и пропорциональной выборки в этом контексте можно объяснить сбалансированным характером набора данных, который позволяет многократное дублирование существующих строк (методы выборки, используемые этими методами) для улучшения набора данных без внесения значительного перекоса в сторону какой-либо конкретной выборки. сорт. Этот метод эффективно дополняет представление всех классов и защищенных атрибутов в сбалансированной форме, что делает эти методы особенно эффективными для наборов данных, где области объектов в равной степени вносят вклад в прогнозы и где исходные распределения классов не страдают от серьезного дисбаланса . Это можно дополнительно проверить с помощью их макроса F1, поскольку ни одна из моделей не меняет точность на полноту. Улучшение SP при передискретизации можно объяснить большим количеством выборок при передискретизации по сравнению с пропорциональной выборкой. Что касается PC-SMOTE и PC-ADASYN, эти алгоритмы демонстрируют улучшение по сравнению с базовым уровнем как по точности, так и по EOD. Эти тенденции соответствуют предыдущим результатам. Примечательно, что это приводит к значительному падению SP, которое можно объяснить нашей новой меткой, созданной для того, чтобы сместить набор данных в сторону отрицательного класса. Эти результаты показывают сложность оптимизации для двух или более показателей справедливости одновременно и то, как эта оптимизация может влиять друг на друга.

6.2. Влияние глубины дерева на справедливость и

точность. В этом исследовании также изучалось влияние глубины дерева решений на производительность модели , в частности изучалось, как изменения глубины дерева влияют на точность и EOD. Понимание эффекта глубины имеет решающее значение, поскольку оно дает представление о последствиях, варьирующихся от недостаточного до переобучения, и помогает определить оптимальный уровень сложности, при котором и точность, и справедливость максимальны. Первоначально дереву решений разрешалось без ограничений расти на всю глубину, которая в среднем составляла около 30 ветвей. Затем дерево осмотрели визуально, чтобы определить максимальную глубину, исключая нерасщепляющиеся ветки. Для систематического анализа влияния глубины дерева максимальная глубина деревьев могла варьироваться от 1 до 30. Каждый предел глубины оценивался с использованием десятикратной перекрестной проверки, чтобы обеспечить надежность и обобщаемость результатов.

Для каждой конфигурации глубины дерева на тестовом наборе измерялись точность и EOD . Кроме того, для десятикратных показателей были рассчитаны 95% доверительные интервалы . Этот статистический анализ выявил глубину, на которой дерево решений уравновешивало компромисс между точностью и справедливостью, а также учитывало лежащий в основе компромисс между статистической погрешностью и дисперсией. Таким образом, удалось определить «золотую середину» — деликатную точку, в которой дерево решений сохраняет высокую точность прогнозирования без ущерба для справедливости, эффективно противодействуя часто упоминаемому компромиссу, представленному в предыдущей
На рисунках 3–7 показаны графики точности и EOD в зависимости от максимальной глубины для каждого из пяти методов выборки в наборе данных о доходах взрослых.

Machine Translated by Google

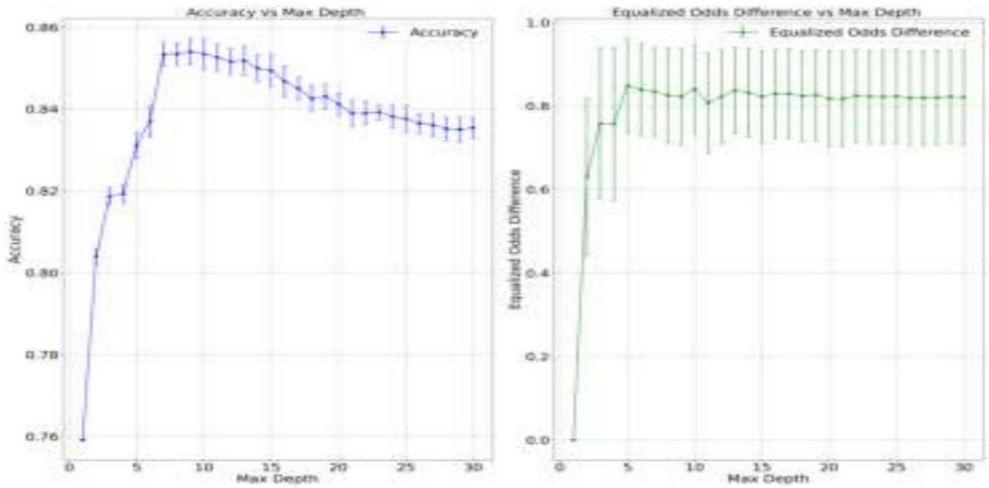


Рисунок 3. Графики доходов взрослых без выборки, показывающие точность и EOD с 95% доверительными интервалами в зависимости от максимальной глубины дерева в диапазоне от 1 до 30.

Судя по результатам, показанным на рисунках 3 и 4, наблюдается заметное первоначальное увеличение точности по мере увеличения максимальной глубины как для методов без выборки, так и для методов с избыточной выборкой. Однако оба метода демонстрируют снижение точности начиная с глубины 10, что указывает на начало переобучения. Соответственно, ЭОД резко уменьшается с увеличением глубины примерно до глубины 10, после чего стабилизируется. Эта закономерность указывает на то, что, хотя более глубокие деревья изначально улучшают справедливость, они в конечном итоге достигают порога, за которым дальнейшего улучшения не наблюдается. Напоминая, что целью справедливости было минимизировать EOD, ключевым наблюдением является то, что установка максимальной глубины от трех до пяти обеспечивает оптимальный баланс между достижением высокой точности и

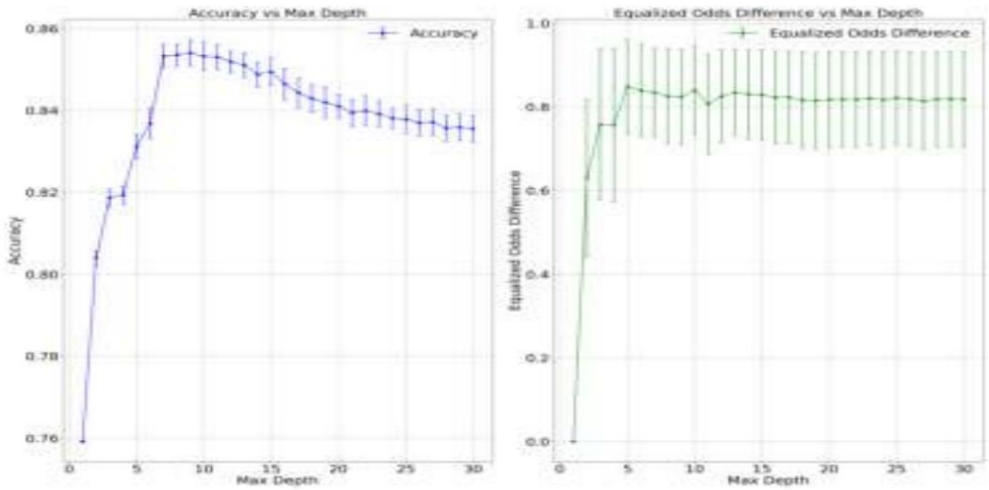


Рисунок 4. Графики доходов взрослых с использованием передискретизации, показывающие точность и EOD с 95% доверительными интервалами в зависимости от максимальной глубины дерева в диапазоне от 1 до 30.

При рассмотрении результатов, показанных на рисунке 5, метод пропорциональной выборки постоянно увеличивает точность с глубиной дерева, достигая максимума на глубине около 26. И наоборот, EOD сначала увеличивается, а затем уменьшается и стабилизируется на глубине около 15. Наблюдаются широкие доверительные интервалы. в метрике EOD предполагают значительную вариабельность результатов справедливости. Этот вывод подчеркивает важность выбора глубины, которая сводит к минимуму изменчивость справедливости при максимальной точности.

Machine Translated by Google

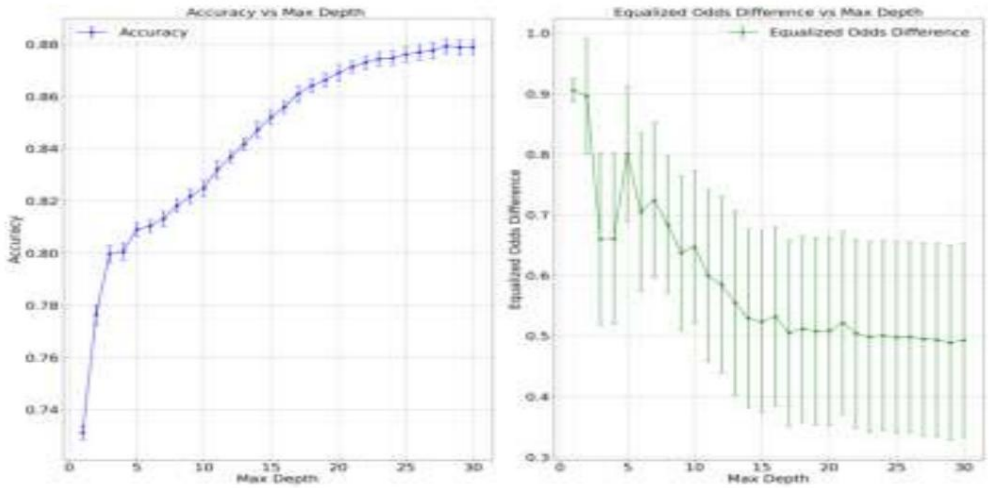


Рисунок 5. Графики доходов взрослых с использованием пропорциональной выборки, показывающие точность и EOD с 95% доверительными интервалами в зависимости от максимальной глубины дерева в диапазоне от 1 до 30.

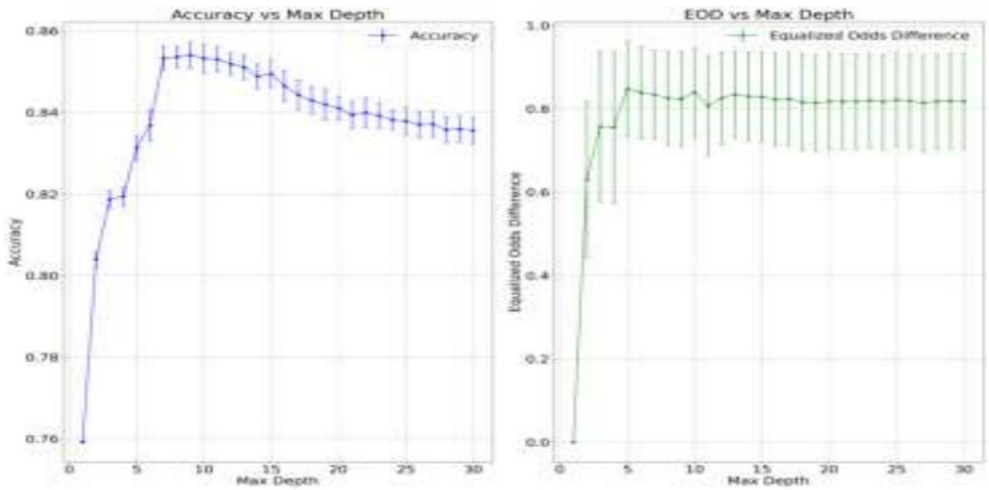


Рисунок 6. Графики доходов взрослых с использованием PC-SMOTE, показывающие точность и EOD с 95% доверительными интервалами в зависимости от максимальной глубины дерева в диапазоне от 1 до 30.

На рисунках 6 и 7, представляющих результаты с использованием PC-SMOTE и PC-ADASYN, соответственно , наблюдаются небольшие тенденции к снижению точности, которые ненадолго улучшаются перед новым снижением – закономерность, указывающая на переобучение на больших глубинах. Показатели EOD для этих методов показывают начальную стабильность на более низких глубинах, всплеск на средних глубинах и снижение, что предполагает сложное взаимодействие между генерацией синтетических образцов и определением границ принятия решений . Учитывая эти наблюдения, для наших экспериментов была выбрана максимальная глубина 3, поскольку она представляет собой «золотую середину», где оптимизируются как точность, так и EO

Учитывая эти результаты, один из выводов состоит в том, чтобы бросить вызов часто предполагаемому компромиссу между точностью и справедливостью, продемонстрировав, что наш метод PC-ADASYN последовательно превосходит базовые показатели для всех трех наборов данных с точки зрения как точности, так и справедливости. Этот вывод важен, поскольку предполагает повышение справедливости модели без ущерба для точности с помощью соответствующих методов выборки и возможной настройки модели. Однако наш анализ также выявляет сценарии, в которых корректировка сложности модели, в частности максимальной глубины деревьев решений, может повысить точность за счет справедливости, о чем свидетельствует увеличение разницы уравненных шансов (EOD). Однако ожидается, что сочетание методов выборки с методами внутренней обработки, такими как регуляризация на основе справедливости, может компенсировать эти эффекты. Эти решения подчеркивают дискреционную власть исследователей в балансировании показателей эффективности модели в зависимости от конкретных целей и ограничений их исслед

Количество выборки и временная сложность аналогичны любому другому методу выборки. Как

Machine Translated by Google

Увеличение количества выборки приводит к увеличению временной сложности, но в целом методы выборки имеют ту же временную сложность, что и лежащие в их основе алгоритмы, поскольку они имеют ту же функциональность.

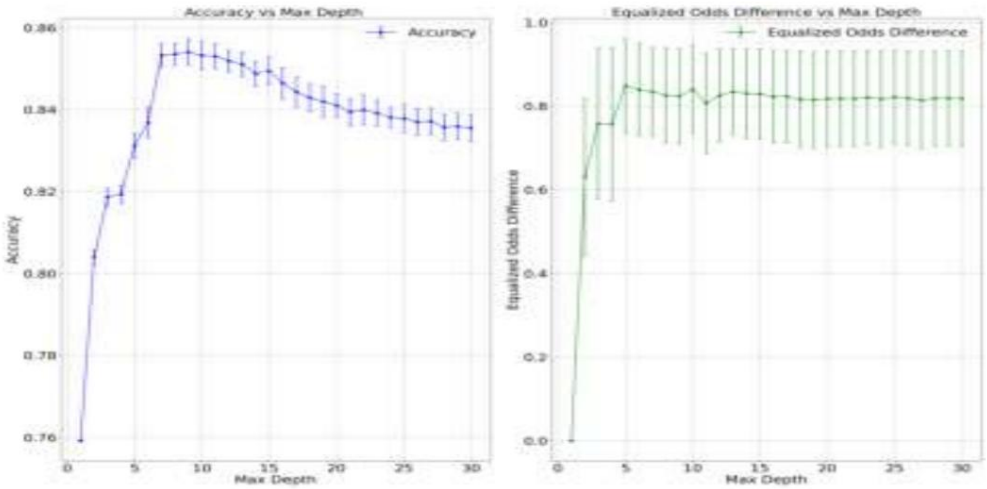


Рисунок 7. Графики доходов взрослых с использованием PC-ADASYN, показывающие точность и EOD с 95% доверительными интервалами относительно максимальной глубины дерева в диапазоне от 1 до 30.

Более того, наши результаты подчеркивают сложность одновременной оптимизации нескольких показателей справедливости. Например, усилия по улучшению статистической четности (SP) путем предпочтения более позитивных прогнозов для каждой защищенной группы в наборе данных COMPAS привели к непреднамеренному сокращению негативных прогнозов. Этот сдвиг отрицательно повлиял на показатель ложноположительных результатов (FPR), компонент EOD, тем самым ухудшив показатель EOD по мере улучшения SP. Этот феномен иллюстрирует присущие математические противоречия между показателями справедливости, когда оптимизация одного может пагубно повлиять на другой. Набор данных COMPAS с его почти сбалансированным распределением по классам представляет собой конкретный пример того, как характеристики набора данных могут влиять на поведение показателей справедливости. Оптимизация SP в этом контексте подразумевает искаженное измерение справедливости, особенно там, где существуют внутренние различия между группами в защищенных атрибутах. Это подтверждается литературными данными, указывающими на то, что SP может неадекватно учитывать групповые различия, что потенциально приводит к ошибочным выводам о справедливости модели [63].

7. Ограничения и будущая работа

Сама природа этого исследования такова, что невозможно рассмотреть все вопросы, связанные со справедливостью и так называемым компромиссом между справедливостью и эффективностью. Таким образом, в работе, представленной здесь, существуют ограничения. Несмотря на это, мы надеемся и намерены, что работа, о которой здесь сообщается, предложит дополнительные направления исследований в этой важной области.

Одним из ограничений этого исследования является то, что наш метод выборки не был специально разработан для оптимизации произвольных показателей справедливости. Другими словами, поскольку часто существуют неизбежные компромиссы между доступным набором показателей справедливости, было принято решение сосредоточиться на подходе, который не зависит от показателей, признавая, что результаты могли отличаться для других показателей. Это также одна из причин, почему мы наблюдали разное поведение между EOD и SP.

Кроме того, признается, что, хотя лежащий в основе метод ОД не должен иметь отношения к предлагаемому методу, он фактически не тестировался. Поэтому в будущем это исследование будет расширено за счет рассмотрения влияния других алгоритмов машинного обучения, таких как логистическая регрессия, нечеткий ID3, K-ближайший сосед и ансамблевые методы, такие как случайные леса или деревья с градиентным усилением, для оценки обобщаемости наших новые методы отбор Целью такого исследования будет проверка того, что наши методы независимы от используемого алгоритма ML. Кроме того, это поможет проверить, являются ли наблюдаемые улучшения в справедливости и точности специфичными для модели или могут применяться повсеместно.

Machine Translated by Google

Кроме того, признается, что рассматривались только три отдельных набора данных — наборы данных , которые широко изучались в этой области. Это вызывает обеспокоенность тем, что методы адаптируются к этим данным, а не решают более широкую проблему справедливости в ОД. Для решения этой проблемы планируются эксперименты с более крупными и разнообразными наборами данных, которые позволят глубже понять масштабируемость и надежность наших методов. Еще одной областью будущей работы является усовершенствование нашего подхода к выборке по нескольким категориям путем включения более детальных подразделений защищенных категорий, что потенциально может выявить более тонкие предубеждения и обеспечить более детальное понимание справедливости.

Наконец, следует признать, что альтернативные методы были предложены для уменьшения предвзятости , и эти методы вообще не изучались в этой работе. Будущая работа повлечет за собой сравнения с другими стратегиями выборки. Будет проведено более прямое сравнение предложенных методов с методами первичной и постобработки. Например , включение методов внутренней обработки, таких как регуляризация [22], или методов постобработки , таких как рандомизированный пороговый оптимизатор [64], будет изучено как возможные средства для дальнейшего улучшения как справедливости, так и производительности.

8. Выводы

В этом исследовании была исследована проблема систематической ошибки прогнозов МО и был разработан метод, основанный на объединении защищенных переменных в новую мультикатегорию. В частности, основное внимание уделялось предложенному в литературе вопросу о компромиссе между предвзятостью и производительностью и поиске метода смягчения этого компромисса. Предлагаемый новый многокатегорийный подход отражает многогранную идентичность людей, признавая сложное взаимодействие атрибутов, которые определяют сценарии реального мира. Учитывая присущий этой мультикатегории дисбаланс, были разработаны четыре метода выборки, адаптированные к этим сложным категориям, а не к традиционным меткам классов. Для применения базового классификатора были обучены деревья решений, и эффективность этих методов оценивалась с использованием трех наборов данных, которые часто используются в исследованиях справедливости. Эффективность методов сравнивалась с базовыми методами без выборки с использованием точности, макроса F1, разницы уравненных шансов (EOD) и статистической четности (SP) в качестве показателей оценки.

Результаты экспериментов показывают, что два недавно разработанных метода выборки — PC-SMOTE и PC-ADASYN — успешно повышают справедливость без ущерба для точности. Примечательно, что в некоторых случаях эти методы также повышали точность, тем самым предоставляя доказательства, опровергающие популярные утверждения о компромиссе между справедливостью и производительностью. Дальнейший анализ влияния максимальной глубины дерева на производительность модели показал, что, хотя увеличение глубины изначально повышает точность, в конечном итоге это приводит к ее снижению. И наоборот , увеличение глубины отрицательно влияет на справедливость, подчеркивая проблему баланса сложности и справедливости. Однако были определены оптимальные глубины дерева, которые одновременно повышают точность и EOD, подчеркивая возможность достижения справедливости без ущерба для производительности.

В конечном счете, эти результаты бросают вызов преобладающим представлениям о неявном компромиссе между производительностью и справедливостью в исследованиях по снижению предвзятости, предполагая, что тщательно разработанные стратегии смягчения предвзятости способны обойти этот компромисс. Наш подход создает новый прецедент для разработки более справедливых алгоритмов прогнозирования путем переопределения того, как защищенные атрибуты используются при обучении модели.

Вклад автора: концептуализация, GP и JS; формальный анализ, GP и JS; расследование, врач общей практики; методология, врач общей практики; программное обеспечение, ГП; супервизия, JS; валидация, GP и JS; визуализация, ГП; письмо — первоначальный вариант, ГП; написание — рецензирование и редактирование, JS. Все авторы прочитали опубликованную версию рукописи и согласились с ней.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: наборы данных и код, реализованные в этой исследовательской работе, были загружены на https://github.com/horlahsunbo/New-folder. (по состоянию на 10 июня 2024 г.).

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Machine Translated by Google

Рекомендации

1. Майна, И.В.; Белтон, ТД; Гинзберг, С.; Сингх, А.; Джонсон, Т.Дж. Десятилетие изучения скрытых расовых/этнических предубеждений в здравоохранении. поставщики, использующие тест неявной ассоциации. Соц. наук. Мед. 2018, 199, 219–229. [Перекрестная ссылка]
2. Салими, Б.; Родригес, Л.; Хау, Б.; Сучу, Д. Причинное восстановление базы данных для обеспечения алгоритмической справедливости. В материалах Международной конференции по управлению данными 2019 г., Амстердам, Нидерланды, 30 июня – 5 июля 2019 г.; стр. 793–810.
3. Кордзаде, Н.; Гасемагаи, М. Алгоритмическая предвзятость: обзор, синтез и направления будущих исследований. Евро. Ж. Инф. Сист. 2022, 31, 388–409. [Перекрестная ссылка]
4. Песах, Д.; Шмуэли, Э. Обзор справедливости в машинном обучении. АКМ Компьютер. Выж. 2022, 55, 1–44. [Перекрестная ссылка]
5. Агаи, С.; Азизи, MJ; Ваянос, П. Изучение оптимальных и справедливых деревьев решений для недискриминационного принятия решений. Учеб. ААААИ Конф. Артиф. Интел. 2019, 33, 1418–1426. [Перекрестная ссылка]
6. Камиран Ф.; Колдерс, Т. Методы предварительной обработки данных для классификации без дискриминации. Знать. Инф. Сист. 2012, 33, 1–33. [Перекрестная ссылка]
7. Кальмон, Ф.; Вэй, Д.; Винзамури, Б.; Рамамурти, КНЦ; Варшни, КР Оптимизированная предварительная обработка для предотвращения дискриминации. В Proceedings of the Advances in Neural Information Processing Systems, Лонг-Бич, Калифорния, США, 4–9 декабря 2017 г.
8. Шахбази, Н.; Лин, Ю.; Асуде, А.; Джагадиш, Х.В. Смещение представления данных: обзор методов идентификации и разрешения. АКМ Компьютер. Выж. 2023, 55, 1–39. [Перекрестная ссылка]
9. Чен, З.; Чжан, Дж. М.; Сарро, Ф.; Харман, М. Комплексное эмпирическое исследование методов уменьшения предвзятости в машинном обучении. классификаторы. АКМ Транс. Программное обеспечение англ. Методол. 2023, 32, 1–30. [Перекрестная ссылка]
10. Пержинский А.; Берг, Калифорния; Томас, К.; Чембалли, А.; Смит, Т.; Шик, С.; Ганзлер, Д.; Сегал, А.Р. Расовая дискриминация и экономические факторы в изменении границ районов Огайо. Буа преподобный Soc. наук. Рез. Гонка 2023, 20, 293–309. [Перекрестная ссылка]
11. Стейл, Дж. П.; Олбрайт, Л.; Раг, Дж.С.; Мэсси, Д.С. Социальная структура ипотечной дискриминации. Дом. Стад. 2018, 33, 759–776. [Перекрестная ссылка]
12. Сальгадо, Дж. Ф.; Москосо, С.; Гарсиа-Искьердо, Алабама; Андерсон, Н.Р. Формирование инклюзивных рабочих мест посредством социального диалога; Springer: Берлин/Гейдельберг, Германия, 2017.
13. Ливи, С.; Мини, Г.; Уэйд, К.; Грин, Д. Смягчение гендерной предвзятости в наборах данных машинного обучения. В материалах о предвзятости и социальных аспектах поиска и рекомендаций: первый международный семинар, BIAS 2020, Лиссабон, Португалия, 14 апреля 2020 г.; стр. 12–26.
14. Хорт, М.; Чен, З.; Чжан, Дж. М.; Харман, М.; Сарро, Ф. Уменьшение смещения для классификаторов машинного обучения: комплексный обзор. ACM J. Ответственные вычисления. 2023, 1, 1–52. [Перекрестная ссылка]
15. Фазе, Т.; Хубер, В.; Ван Гиффен, Б. Управление предвзятостью в проектах машинного обучения. В «Инновациях через информационные системы: Том II: Сборник последних исследований по технологическим проблемам»; Шпрингер: Берлин/Гейдельберг, Германия, 2021 г.; стр. 94–109.
16. Чжан З.; Нил, Д.Б. Выявление значительной предсказательной ошибки в классификаторах. arXiv 2017, arXiv:1611.08292. [Перекрестная ссылка]
17. Чжэн, З.; Кай, Ю.; Ли, Ю. Метод передискретизации для несбалансированной классификации. Вычислить. Поставить в известность. 2015, 34, 1017–1037.
18. Чавла, Невада; Бойер, КВ; Холл, Лоу; Кегельмейер, WP SMOTE: Техника чрезмерной выборки синтетического меньшинства. Дж. Артиф. Интел. Рез. 2002, 16, 321–357. [Перекрестная ссылка]
19. Он, Х.; Бай, Ю.; Гарсия, Э.А.; Ли, С. АДАСИН: Адаптивный синтетический подход к выборке для несбалансированного обучения. В материалах Международной совместной конференции IEEE по нейронным сетям 2008 г. (Всемирный конгресс IEEE по вычислительному интеллекту), Гонконг, Китай, 1–8 июня 2008 г.; стр. 1322–1328.
20. Янссен, П.; Садовский, Б.М. Предвзятость в алгоритмах: о компромиссе между точностью и справедливостью. В материалах 23-й конференции Международного общества электросвязи, проводимой раз в два года, Гетеборг, Швеция, 21–23 июня 2021 г.
21. Хардт, М.; Прайс, Э.; Сребро, Н. Равенство возможностей в контролируемом обучении. В материалах о достижениях в области нейронных технологий Системы обработки информации, Барселона, Испания, 5–10 декабря 2016 г.
22. Камишима, Т.; Акахо, С.; Асох, Х.; Сакума, Дж. Классификатор, учитывающий справедливость, с регуляризатором для устранения предрассудков. В материалах машинного обучения и обнаружения знаний в базах данных: Европейская конференция, ECML PKDD 2012, Бристоль, Великобритания, 24–28 сентября 2012 г.; стр. 35–50.
23. Колдерс, Т.; Жлиобайте И. Почему объективные вычислительные процессы могут привести к дискриминационным процедурам принятия решений. Дискриминация и конфиденциальность в информационном обществе: интеллектуальный анализ данных и профилирование в больших базах данных; Springer: Берлин/Гейдельберг, Германия, 2013 г.; стр. 23–33.
24. Берт А. Как бороться с дискриминацией в ИИ. Гарвардский бизнес-обзор. 2020 г. Доступно онлайн: https://hbr.org/2020/08/how-to-fight-discrimination-in-ai . (по состоянию на 12 июля 2024 г.).
25. Зиглер А.; Адмуссен, В. Обнаружение расовой дискриминации со стороны полиции. Северо-Западный университет. Закон ред. 2021 г., 115, 987–1054.
26. Грабович П.; Перелло, Н.; Мишра А. Как обучать модели, не распространяющие дискриминацию? Блог Equate и Machine Learning , Массачусетский университет, Амхерст. 2022. Доступно онлайн: https://groups.cs.umass.edu/equate-ml/2022/04/07/how-to-train-models-that-do-not-propagate-discrimination/ . (по состоянию на 12 июля 2024 г.).
27. Хани, Ф.; Лян, П. От дискриминации в машинном обучении к дискриминации в законодательстве, Часть 1: Неравномерное обращение. Блог Стэнфордской лаборатории искусственного интеллекта . 2022. Доступно онлайн: https://ai.stanford.edu/blog/discrimination_in_ML_and_law/. (по состоянию на 12 июля 2024 г.).
28. Шахриари, К.; Шахриари, М. Обзор стандарта IEEE — Этически согласованный дизайн: видение приоритета благополучия человека с помощью искусственного интеллекта и автономных систем. В материалах Канадской международной конференции по гуманитарным технологиям IEEE (IHTC), Торонто, Онтарио, Канада, 21–22 июля 2017 г.; стр. 197–201.

Machine Translated by Google

29. Европейская комиссия; Генеральный директорат сетей связи, контента и технологий. Рекомендации по этике для надежного ИИ. Офис публикаций. 2019. Доступно онлайн: https://op.europa.eu/en/publication-detail/-/publication/d39885. 69-0434-11ea-8c1f-01aa75ed71a1/language-en/format-PDF/source-337437547 (по состоянию на 10 июня 2024 г.)
30. Эберс, М.; Хох, ВРС; Розенкранц, Ф.; Рушемайер, Х.; Штайнрёттер, Б. Предложение Европейской комиссии по Закону об искусственном интеллекте – критическая оценка членов Общества юристов по робототехнике и искусственному интеллекту (RAILS). J 2021, 4, 589–603. [Перекрестная ссылка]
31. Хаджян, С.; Доминго-Феррер, Дж. Методика предотвращения прямой и косвенной дискриминации при интеллектуальном анализе данных. IEEE Транс. Знать. Данные англ. 2012, 25, 1445–1459. [Перекрестная ссылка]
32. Фиш, Б.; Кун, Дж.; Лелкес, А.Д. Подход, основанный на доверии, позволяющий сбалансировать справедливость и точность. В материалах SIAM 2016 г. Международная конференция по интеллектуальному анализу данных, Майами, Флорида, США, 5–7 мая 2016 г.; стр. 144–152.
33. Хаджян С. Одновременное предотвращение дискриминации и защита конфиденциальности при публикации и майнинге данных. arXiv 2013, arXiv:1306.6805. [Перекрестная ссылка]
34. Сондек, ЛП; Лоран, М.; Фрей, В. Метрика уровня семантической дискриминации для измерения конфиденциальности, которая ставит под сомнение преимущество t-близости перед l-разнообразием. В материалах 14-й Международной конференции по безопасности и криптографии, Мадрид, Испания, 24–26 июля 2017 г.; Том 6, стр. 285–294.
35. Руджери, С. Использование анонимности t-близости для контроля недискриминации. Пер. Частные данные. 2014, 2, 99–129.
36. Романо, Ю.; Бейтс, С.; Кандес, Э. Достижение равных шансов путем повторной выборки чувствительных атрибутов. В Proceedings of the Advance in Neural Information Processing Systems, онлайн, 6–12 декабря 2020 г.; стр. 361–371.
37. Пэн, К.; Чакраборти, Дж.; Мензис, Т. Фэйрмаск: Повышение справедливости за счет ребалансировки защищенных атрибутов на основе моделей. IEEE Транс. Программное обеспечение англ. 2022, 49, 2426–2439. [Перекрестная ссылка]
38. Дхар, П.; Глисон, Дж.; Рой, А.; Кастильо, CD; Челлаппа, Р. ПАСС: Система подавления защищенных атрибутов для уменьшения предвзятости при распознавании лиц. В материалах Международной конференции IEEE/CVF по компьютерному зрению, Монреаль, Квебек, Канада, 10–17 октября 2021 г.; стр. 15087–15096.
39. Красанакис Э.; Спиромитрос-Сиуфис, Э.; Пападопулос, С.; Компациарис, Ю. Адаптивное чувствительное перевзвешивание для смягчения предвзятости в классификации с учетом принципа справедливости. В материалах конференции World Wide Web, Лион, Франция, 23–27 апреля 2018 г.; стр. 853–862.
40. Ранчич, С.; Радованович, С.; Делибашич, Б. Исследование методов передискретизации для справедливых моделей машинного обучения. В материалах систем поддержки принятия решений XI: Системы поддержки принятия решений, аналитика и технологии в ответ на глобальное кризисное управление: 7-я Международная конференция по технологиям систем поддержки принятия решений, ICDSST 2021, Лафборо, Великобритания, 26–28 мая 2021 г.; Шпрингер: Берлин/Гейдельберг, Германия, 2021 г.; стр. 110–123.
41. Ян, С.; те Као, Х.; Феррара, Э. Балансировка справедливых классов: повышение справедливости модели без учета чувствительных атрибутов. В материалах 29-й Международной конференции ACM по управлению информацией и знаниями, виртуальное мероприятие, Ирландия, 19–23 октября 2010 г.; стр. 1715–1724.
42. ХуЧжан, Б.; Лемуан, Б.; Митчелл, М. Смягчение нежелательных предубеждений с помощью состязательного обучения. В материалах конференции AAAI/ACM по искусственному интеллекту, этике и обществу, Новый Орлеан, Луизиана, США, 2–3 февраля 2018 г.; стр. 335–340.
43. Селис, Л.Э.; Хуанг, Л.; Кесвани, В.; Вишно, Н.К. Классификация с ограничениями справедливости: метаалгоритм с доказуемыми гарантиями. В материалах конференции по вопросам справедливости, подотчетности и прозрачности, Атланта, Джорджия, США, 29–31 января 2019 г.; стр. 319–328.
44. Зафар, М.Б.; Валера И.; Родригес, МГ; Гуммади, К.П. Справедливость за пределами разного обращения и разного воздействия: классификация обучения без разного жестокого обращения. В материалах 26-й Международной конференции по Всемирной паутине, Перт, Австралия, 3–7 апреля 2017 г.; стр. 1171–1180.
45. Агарвал, А.; Бейгельцимер, А.; Дудик, М.; Лэнгфорд, Дж.; Уоллах, Х. Подход к справедливой классификации с помощью сокращений. В материалах Международная конференция по машинному обучению, Стокгольм, Швеция, 10–15 июля 2018 г.; стр. 60–69.
46. Лоуи, А.; Бахарлуи, С.; Паван, Р.; Разавиян, М.; Бейрами, А. Схема стохастической оптимизации для справедливой минимизации рисков. Пер. Мах. Учиться. Рез. 2022. [Перекрестная ссылка]
47. Спинелли И.; Скардапане, С.; Хусейн, А.; Унчини, А. Фэйрдроп: Смещенное выпадение ребер для повышения справедливости в обучении графическому представлению . IEEE Транс. Артиф. Интел. 2021, 3, 344–354. [Перекрестная ссылка]
48. Хорт, М.; Чжан, Дж. М.; Сарро, Ф.; Харман, М. Файреа: Подход к изменению поведения модели для сравнительного анализа методов смягчения предвзятости. В материалах 29-й совместной встречи ACM по Европейской конференции по разработке программного обеспечения и симпозиума по основам программной инженерии, Афины, Греция, 23–28 августа 2012 г.; стр. 994–1006.
49. Бхаскаруни, Д.; Хм.; Лан, К. Повышение справедливости прогнозирования с помощью модельного ансамбля. В материалах 31-й Международной конференции IEEE по инструментам с искусственным интеллектом (ICTAI), Портленд, Орегон, США, 4–6 ноября 2019 г.; стр. 1810–1814.
50. Иосифидис В.; Фетаху, Б.; Нтуци, Э. Фаэ: Структура ансамбля, учитывающая справедливость. В материалах Международной конференции IEEE по большим данным (Big Data), Лос-Анджелес, Калифорния, США, 9–12 декабря 2019 г.; стр. 1375–1380.
51. Дворк, К.; Хардт, М.; Питасси, Т.; Рейнгольд, О.; Земель, Р. Справедливость через осознанность. В материалах 3-го Нововведения в Конференция по теоретической информатике, Кембридж, Массачусетс, США, 8–10 января 2012 г.; стр. 214–226.
52. Плейсс, Г.; Рагхаван, М.; Ву, Ф.; Кляйнберг, Дж.; Вайнбергер, К.К. О справедливости и калибровке. В материалах о достижениях в Нейронные системы обработки информации, Лонг-Бич, Калифорния, США, 4–8 декабря 2017 г.
53. Карими-Хагиги, М.; Кастильо, К. Улучшение инструмента прогнозирования рецидивов с помощью машинного обучения: эффективность и алгоритмическая справедливость. В материалах восемнадцатой Международной конференции по искусственному интеллекту и праву, Сан-Паулу, Бразилия, 21–25 июня 2017 г.; стр. 210–214.

Machine Translated by Google

54. Фридлер, С.А.; Шайдеггер, К.; Венкатасубраманиан, С. (Не)возможность справедливости: разные системы ценностей требуют разных механизмов для справедливого принятия решений. Коммун. ACM 2021, 64, 136–143. [Перекрестная ссылка]
55. Гарсия, В.; Санчес, Х.С.; Моллинеда, Р.А. Об эффективности методов предварительной обработки при работе с различными уровнями классового дисбаланса. Информационная система. 2012, 25, 13–21. [Перекрестная ссылка]
56. Брейман Л. Деревья классификации и регрессии; Рутледж: Лондон, Великобритания, 2017 г.
57. Педрегоса, Ф.; Варокво, Г.; Грамфорт, А.; Мишель, В.; Тирион, Б.; Гризель, О.; Блондель, М.; Преттенхофер, П.; Вайс, Р.; Дюбур, В.; и другие. Scikit-learn: машинное обучение на Python. Дж. Мах. Учиться. Рез. 2011, 12, 2825–2830.
58. Келли, М.; Лонгджон, Р.; Ноттингем, К. Репозиторий машинного обучения UCI. 2024 г. Доступно онлайн: https://archive.ics.uci.edu. (по состоянию на 10 июня 2024 г.).
59. Беккер, Б.; Кохави, Р. Взрослый. Репозиторий машинного обучения UCI. 1996. Доступно онлайн: https://archive.ics.uci.edu/dataset/2/. взрослый (по состоянию на 10 июня 2024 г.). [Перекрестная ссылка]
60. Хофманн, Х.; Statlog (кредитные данные Германии). Репозиторий машинного обучения UCI. 1994. Доступно в Интернете: https://archive.ics.uci.edu/dataset/144/statlog+немецкий+кредит+данные (по состоянию на 10 июня 2024 г.). [Перекрестная ссылка]
61. Дрессел, Дж.; Фарид, Х. Точность, справедливость и пределы прогнозирования рецидивизма. наук. Адв. 2018, 4, eaao5580. [Перекрестная ссылка]
62. Хуан, Ц.; Мао, Дж.; Лю, Ю. Улучшенный алгоритм поиска по сетке для оптимизации параметров СВР. В материалах 14-й Международной конференции IEEE по коммуникационным технологиям 2012 г., Чэнду, Китай, 9–11 ноября 2012 г.; стр. 1022–1026.
63. Катон, С.; Хаас, К. Справедливость в машинном обучении: опрос. АКМ Компьютер. Выж. 2023, 56, 1–38. [Перекрестная ссылка]
64. Алабдулмохсин И.; Лючич, М. Почти оптимальный алгоритм устранения смещения обученных моделей машинного обучения. В материалах 35-я конференция по нейронным системам обработки информации, онлайн, 6–14 декабря 2021 г.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Безопасная обработка и распространение данных, управляемых конфиденциально Межпланетная файловая система с использованием доказательств с нулевым разглашением

Кёхей Сибано 1,*, Кенсуке Ито [1], Чанхи Хан [2], Цз Тат Чу [2], Ватару Одзаки [2] и Генто Моги [1]

[1] Департамент управления технологиями для инноваций, Школа инженерии, Токийский университет, Токио 113-8656,
[2] Япония Callisto Inc., Токио 171-0022,
Япония * Адрес для переписки: shibano@tmi.tu-tokyo.ac.jp

Аннотация: В этом исследовании предлагается новый метод совместного использования данных, который использует частную межпланетную файловую систему — децентрализованную систему хранения, работающую в закрытой сети, — для распространения данных внешним объектам, обеспечивая при этом проверку их подлинности. Среди двух режимов работы IPFS, публичного и частного, в данном исследовании основное внимание уделяется методу использования частной IPFS. Частная IPFS не открыта для широкой публики; хотя предложенный метод и создает риск фальсификации данных при передаче данных внешним сторонам, он обеспечивает подлинность полученных данных. В частности, этот метод применяет тип доказательства с нулевым разглашением, а именно протокол zk-SNARK Groth16, чтобы гарантировать, что данные соответствуют идентификатору контента в частной IPFS. Более того, имя получателя встраивается в распространяемые данные, чтобы предотвратить несанкционированное вторичное распространение. Эксперименты подтвердили эффективность предложенного метода при размере данных изображения до 120× В будущих исследованиях предложенный метод будет применяться к более крупным и разнообразным типам данных.

Ключевые слова: ИПФС; доказательство с нулевым разглашением; цирк; зк-СНАРКИ; частная IPFS; распространение данных; обработка данных ; безопасность данных

check for updates

Цитирование: Шибано, К.; Ито, К.; Хан, К.; Чу, ТТ; Одзаки, В.; Моги, Г. Безопасная обработка и распространение данных, управляемых в частной межпланетной файловой системе, с использованием нулевого знания Доказательства. Электроника 2024, 13, 3025. https://doi.org/10.3390/электроника13153025

Академический редактор: Ария Гангопадьяй

Поступила: 13 июня 2024 г.
Пересмотрено: 14 июля 2024 г.
Принято: 20 июля 2024 г.
Опубликовано: 31 июля 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Децентрализованные системы надежны, поскольку у них нет единой точки отказа; поэтому они широко применяются в корпоративных секторах, включая криптовалюту, управление цепочками поставок, финансовые услуги и цифровую идентификацию. Для хранения данных большого размера, таких как изображения , этим системам требуются функции хранения, которые по своей сути децентрализованы. Блокчейн, обычно используемый совместно, обычно обрабатывает данные меньшего размера, такие как истории транзакций, и работает как база данных реестра. Межпланетная файловая система (IPFS) — это известная децентрализованная система хранения, которая хранит данные на нескольких узлах для повышения доступности данных. IPFS имеет два варианта: общедоступная IPFS, в которой данные могут храниться любым пользователем с неограниченным доступом, и частная IPFS, в которой создается закрытая сеть, доступная только внутри определенных организаций или групп, что обеспечивает повышенную конфиденциальность и безопасность.

При хранении данных в IPFS крайне важно понимать различия между общедоступной и частной IPFS. Публичная IPFS позволяет любому получить доступ к данным, тогда как частная IPFS доступна только в определенных организациях или группах, что повышает конфиденциальность и безопасность. При хранении конфиденциальной информации, например конфиденциальных данных, в общедоступной IPFS применение соответствующей схемы шифрования имеет жизненно важное значение для обеспечения защиты данных. Напротив, частная IPFS обеспечивает более высокий уровень безопасности хранения данных, поскольку доступна только внутри закрытой сети.

В частности, для таких организаций, как корпорации или медицинские учреждения, хранение данных в общедоступной IPFS, несмотря на использование надежных технологий шифрования, несет в себе неизбежные риски. Более того, в таких случаях постоянно существует вероятность утечки данных из-за операционных ошибок ; хотя данные зашифрованы, они доступны всему миру, что делает их уязвимыми для атак методом перебора и других угроз безопасности.

Machine Translated by Google

Что касается доступности, общедоступная IPFS позволяет обычным пользователям напрямую получать доступ и извлекать данные. данные. Однако в частных IPFS данные должны быть получены от членов организации или группа, составляющая сеть. Если во время этого процесса данные будут подделаны, пользователи могут неспособен его обнаружить. Поэтому доверяя посредникам, ответственным за обработку данных передача в таких случаях становится обязательной. Для решения вышеупомянутой проблемы доверия был разработан новый здесь предлагается метод распространения данных, хранящихся в частной IPFS, внешним объектам. при этом проверяя его подлинность. Протокол Groth16 [1] zk-SNARKs, тип ZKP применяется к данным, хранящимся в частной IPFS, для обеспечения их подлинности. Более того, информация получателя встраивается в распространяемые данные, чтобы предотвратить несанкционированное вторичное распространение. Предлагаемый метод обмена данными важен потому что он адаптирован к частному случаю IPFS.

Различия в нескольких аспектах, включая безопасность и доступность, при хранении данные в общедоступных IPFS и частных IPFS в домене предприятия суммируются в Таблица 1. В этом исследовании предлагаются решения угроз, связанных с частными IPFS.

Таблица 1. Сравнение общедоступной и частной IPFS в корпоративной сфере.

	Публичная IPFS	Частная IPFS
Модель доверия	Доверчивый	Требуется доверие к операционной группе
Ограничения доступа	Доступен любому	Доступно только внутри операционной группы
Риск утечки данных	Постоянный риск утечки из-за ошибки пользователя	Низкий риск утечки в закрытой сети
Обращение с конфиденциальной информацией Информация	Требуется правильное шифрование	Данные, хранящиеся в IPFS, сами по себе не требуют шифрования высокого уровня; есть точка доверия при передаче данных пользователям
Риск атаки методом грубой силы	Всегда присутствует	Низкий
Метод получения данных	Прямой доступ пользователей	Данные, полученные от членов организации или группы
Угрозы	Требуется шифрование, предотвращающее дешифрование неавторизованными пользователями.	Существует риск фальсификации при передаче данных пользователям.

Оставшаяся часть статьи организована следующим образом. В разделе 2 представлены соответствующие предыдущие исследовать. В разделе 3 описаны фундаментальные технологии, а именно ZKP и zk-SNARK. В разделе 4 описывается структура предлагаемого метода, а в разделе 5 излагаются возможные применения этого метода. В разделе 6 представлена реализация этого метода, а в разделе 7 обсуждаются эксперименты, проведенные для проверки эффективности выполнение. В разделе 8 представлено обсуждение экспериментальных результатов, а в разделе 9 представлены выводы документа и обзор будущих задач.

2. Сопутствующие исследования

Существующие децентрализованные системы используют IPFS, особенно в сочетании с блокчейном. технологии. Кумар и др. [2] предложили метод безопасного управления медицинскими данными путем интеграция IPFS с блокчейном. Азбег и др. [3] специально предложил систему, которая управлял и хранил медицинские данные с использованием частной IPFS и разрешенного блокчейна использование повторного шифрования прокси-сервера для обеспечения безопасного дешифрования назначенными врачами. Когда врач получает некоторые данные пациента, он/она получает повторно зашифрованные данные через больницу. Хоссан и др. [4] также предложили систему безопасной записи информации для совместного использования поездок. сервисы с использованием IPFS и частного блокчейна.

Сосредоточив внимание на контроле над распространением данных, управляемых IPFS, Lin et al. [5] предложил систему защиты частных данных с использованием улучшенной IPFS в сочетании с блокчейн. Эта система записывала метаданные файлов и получала разрешения на блокчейн. предоставление пользователям возможности контролировать общий доступ к файлам. Кроме того, в системе реализованы эффективные функции управления с использованием смарт-контрактов, что повышает безопасность и управление данными.

Machine Translated by Google

гибкость. Батта и др. [6] разработали систему, которая использовала многостороннюю аутентификацию (MPA), повторное шифрование прокси и смарт-контракты на блокчейне для децентрализованного контроля доступа к зашифрованным данным, хранящимся в IPFS. Хуанг и др. [7] представили доверенный прокси-сервер IPFS для реализации контроля доступа и управления групповыми ключами для зашифрованных данных, хранящихся в IPFS. Сан и др. [8] предложили систему, которая позволяла только лицам с соответствующими атрибутами расшифровывать зашифрованные данные, хранящиеся в IPFS, с использованием системы шифрования на основе атрибутов политики зашифрованного текста, что способствовало эффективному управлению медицинской информацией. Канг и др. [9] обеспечили распространение данных, управляемых с использованием частной IPFS и частной цепочки блоков, внешним пользователям с использованием именованной сети передачи данных (NDN). Более того, Уддин и др. [10] предложили систему обмена файлами , которая использовала IPFS и технологию инфраструктуры открытых ключей (PKI), не требуя доверенной третьей стороны.

В нескольких исследованиях ZKP использовался для распространения данных. Например, Ли и др. [11] предложили систему управления трафиком, сохраняющую конфиденциальность, которая сочетает в себе неинтерактивные доказательства диапазона с нулевым разглашением и блокчейн. Прототип, использующий Hyperledger Fabric и Hyperledger Ursa, отвечал требованиям конфиденциальности данных для управления трафиком в режиме ре

В этом исследовании предлагается метод соответствующей обработки и распространения данных, управляемых в частной IPFS, пользователям за пределами сети, тем самым предлагая другой подход, чем те, которые предлагались в предыдущих исследованиях. В некоторых исследованиях перешифрование прокси-сервера было принято в качестве подходящего метода хранения и распространения данных в IPFS [3,6]. Используя этот метод, распределенные данные можно повторно зашифровать для расшифровки с помощью закрытого ключа получателя. Более того, при хранении данных в IPFS запись хеш-значения предварительно зашифрованных данных в блокчейн позволяет получателям проверять правильность полученных данных после расшифровки. Однако этот метод не может обрабатывать случаи обработки данных, например встраивания имени получателя в расшифрованные данные, как в этом исследовании.

3. Доказательство с нулевым

разглашением ZKP — это криптографический протокол, который позволяет доказывающему доказать верификатору достоверность предложения, не раскрывая никакой дополнительной информации, кроме достоверности предложения. Предложение этого исследования заключается в том, что данные, предоставляемые внешнему объекту, генерируются на основе заданного CID. Наша цель — позволить члену частной IPFS (проверяющему) доказать это утверждение внешнему лицу (проверяющему как получателю данных) без раскрытия какой-либо другой важной информации (например, прав доступа IPFS и ключей шифрования).

ZKP, в частности протокол Groth16 zk-SNARK, используемый в этом исследовании, начинается с доверенной настройки, в которой обе стороны устанавливают общедоступные параметры, которые имеют решающее значение для безопасного создания и проверки доказательств. В схеме ZKP мы сначала генерируем схему, описывающую процесс, для которого предназначено доказательство. Схема включает в себя условия, которые необходимо проверить, например наличие CID. С помощью схемы ZKP создаются криптографические ключи, в частности ключ подтверждения и ключ проверки. Эти ключи имеют решающее значение для создания доказательства схемы и ее проверки. Используя ключ доказательства и входные данные, средство доказательства генерирует доказательство, которое показывает соответствие выходных данных условиям, указанным в схеме. Затем верификатор использует ключ проверки для проверки доказательства и выходных данных. Если доказательство действительно, оно подтверждает целостность данных без раскрытия какой-либо основной информации.

В этом исследовании обработка Groth16 выполняется с использованием circom [12,13] и snarkjs [14]. Схема процесса представлена на рисунке 1.

zk-SNARKs используется из-за его неинтерактивного характера и эффективности, что особенно выгодно для систем, использующих смарт-контракты, из-за низких вычислительных затрат для проверки доказательств. Кроме того, zk-SNARKs известен своими высокими вычислительными требованиями и потребностью в расширенных характеристиках ПК. Например, в одном из типичных приложений zk-SNARKs, Zcash [15], процесс генерации доказательства занимает более полуминуты для одной анонимной транзакции [16].

Machine Translated by Google

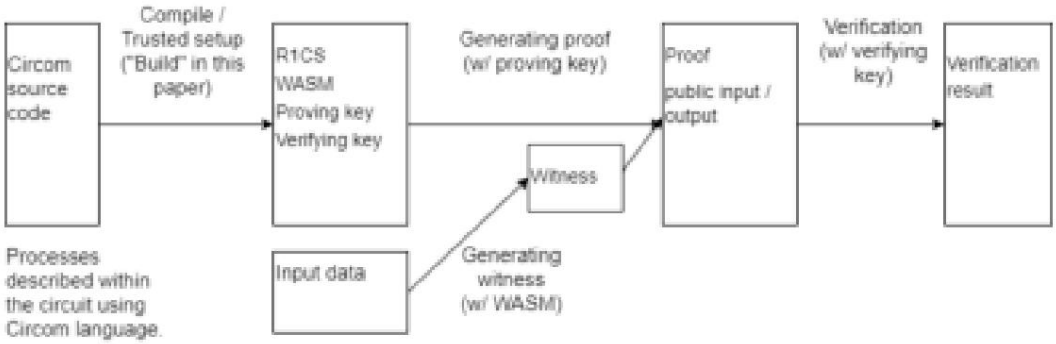


Рисунок 1. Доказательство с нулевым разглашением с использованием circom.

4. Предлагаемая

система На рисунке 2 представлен обзор предлагаемой системы.

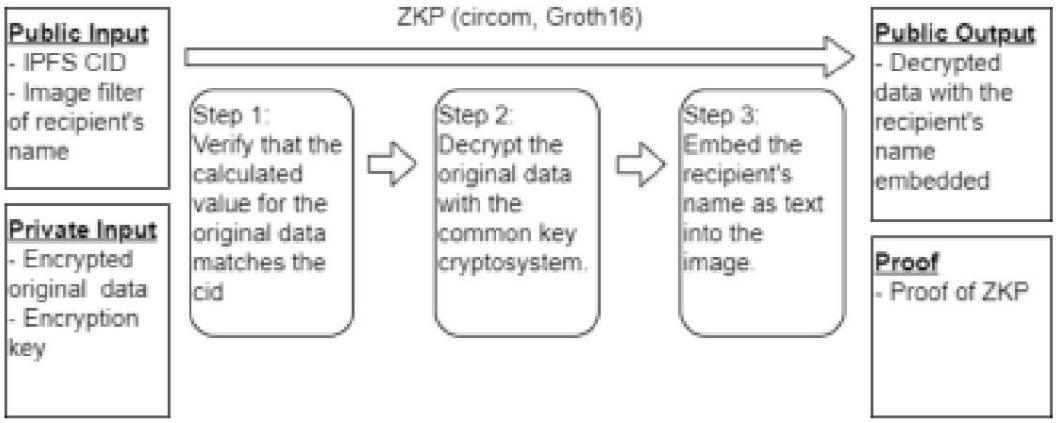


Рисунок 2. Технологическая схема ZKP: пример обработки данных изображения.

Здесь мы делаем следующие предположения: •

Частная IPFS управляется ограниченным числом участников. • Данные

хранятся в IPFS в зашифрованном формате с использованием криптографии с симметричным ключом. • Ключ шифрования принадлежит исключительно одному лицу среди участников, т.е. администратор.

Предлагаемая система облегчает создание доказательства ZKP через схему участником, владеющим ключом шифрования (эквивалентным администратору). Входные и выходные данные (кроме доказательств) этого процесса следующие: • Публичные

входные данные: –

CID исходных зашифрованных данных; –
Фильтр для встраивания имени получателя в данные; • Частный ввод:

– Зашифрованные

исходные данные; – Ключ
шифрования; • Выход:

– Расшифрованные данные со встроенным именем получателя.

Обратите внимание, что реализация, оптимизированная для ZK (раздел 6.1), добавляет больше информации к общедоступным данным. В частности, внутренний процесс ZKP включает в себя

следующие шаги: 1. Вычисление CID исходных зашифрованных данных для проверки соответствия с кодом. Теред CID.

Machine Translated by Google

2. Расшифровка исходных данных с использованием криптографии с симметричным ключом.
3. Встраивание имени получателя в необработанные данные расшифрованного изображения на основе фильтра.
представлено в публичном доступе.

В предлагаемой системе мы не сосредотачиваемся исключительно на данных изображения, но используем их в качестве примеров данных для проверки применимости предлагаемой схемы. Фильтры используются для повышения эффективности обработки внутри схемы. Поскольку встраивание данных имени внутри схемы требует больших вычислительных ресурсов, значительная часть обработки изображения заранее выполняется вне схемы и создается фильтр. Используя общедоступные данные и доказательства, получатель может убедиться, что расшифрованные данные (т. е. выходные данные) со встроенным их именем созданы из исходных данных (содержащихся в частных входных данных), управляемых с помощью CID. Получатель может проверить наличие CID хотя бы у одного участника сети, чтобы подтвердить наличие CID в частной IPFS.

Таким образом, вышеупомянутый процесс позволяет получателю проверить полученное данные, выполнив следующие задачи:

- убедитесь, что полученные данные были сгенерированы из данных, управляемых с помощью CID частная IPFS,
- подтвердите, используя доказательство того, что весь процесс был правильно проведен без непосредственного знания зашифрованных данных или ключа шифрования,
- и убедитесь, что CID существует конкретно в частной IPFS, запросив хотя бы одного члена сети.

Новизна предлагаемой системы заключается в том, что она позволяет проверять подлинность данных, доверяя хотя бы одному участнику сети, даже если получатель не контролирует ключ шифрования. (Для получателя естественно доверять хотя бы одному члену конкретной многочленной системы. Если ни одному из участников нельзя доверять, то будет незначительный стимул для получения данных, управляемых этой сетью).

5. Возможные применения В

медицинской отрасли данные о диагнозах пациентов обрабатываются в нескольких медицинских учреждениях. Используя предлагаемую систему, пациенты могут проверить, действительно ли получаемые ими данные обрабатываются в частной IPFS для обеспечения их подлинности. Здесь также предлагается универсальная платформа для исследований и разработок машинного обучения с медицинскими изображениями [17]. На этой платформе анонимные медицинские изображения управляются частным IPFS, которым управляет группа медицинских учреждений. Система позволяет исследователям машинного обучения, внешним по отношению к сети, проверять, действительно ли данные изображения управляются в частной IPFS. Более того, встраивая информацию об исследователях машинного обучения в данные изображения, медицинские учреждения могут снизить риск вторичного распространения.

Если приложение не ограничивается вставкой имени получателя, потенциальные возможности применения предлагаемой системы могут быть дополнительно расширены. Например, рассмотрим сценарий, в котором конкретная компания устанавливает частную IPFS для обмена конфиденциальными документами между компаниями своей группы. Если данные о сотрудниках включены, то сокрытие личных данных и их распространение внешним организациям позволяет этим организациям подтвердить связь сотрудников с компанией, обеспечивая при этом защиту их конфиденциальности.

Кроме того, предположим, что университет настроил частную IPFS, чтобы разрешить доступ к данным об успеваемости студентов только академическому персоналу. В этом случае учащиеся могут убедиться, что полученные ими данные о производительности действительно управляются в частной IPFS.

Таким образом, предлагаемая система поддерживает гибридный вариант — распределение внутренних данных по конкретные внешние объекты по мере необходимости, что распространено в реальных условиях.

6. Реализация Предлагаемая

система была реализована для обработки данных изображения с использованием circom, известного инструмента, специализирующегося на построении схем zk-SNARKs. Circcom позволяет описывать вычислительные процессы внутри схемы, используя свой уникальный язык.

Machine Translated by Google

исполняемый файл, созданный после компиляции, можно вызвать через библиотеку JavaScript snarkjs. Такое расположение позволяет описывать процессы схемы в circom, а внешняя обработка и проверка корректности схемы выполняются с использованием JavaScript. Для зк-CHAPKов использована схема Groth16; он известен своей относительно более высокой скоростью выполнения, чем другие схема зк-CHAPKC.

В частности, мы работали над двумя типами реализаций данных изображений: стандартным реализация с использованием общих криптографических методов и реализация, оптимизированная для ZK, с использованием криптографических методов, дружественных к ZK, для сокращения времени вычисления схема. Эти реализации использовались для сравнения требуемого времени вычислений. Схемы ЗКП требуют значительно большого времени вычислений, даже для вычислений, которые могут легко обрабатываться компьютерным программным обеспечением (это особенно заметно при работе с данные изображения). Поэтому эффективность вычислений имеет решающее значение для практического использования.

6.1. Стандартная реализация

В разделе 4 описан ввод данных в схему. Для упрощения внутрисхемного при обработке исходные зашифрованные данные были отформатированы как совместимые с данными растрового изображения. со стандартами OS/2. Первые 54 байта данных изображения хранят такую информацию, как ширина, высота и глубина цвета изображений [18]. Глубина цвета составляет 8 бит, и каждый цвет компоненту в RGB выделяется один байт, в результате чего на пиксель приходится 3 байта.

Первоначально система проверяет, являются ли зашифрованные данные, введенные как частные входные данные, соответствует CID, предоставленному в качестве общедоступного ввода. Если они не совпадают, система означает ошибка, и данные изображения, выводимые на выходе, представляют собой последовательность байтов, в которой все значения имеют значение 0x00. CID служит важнейшим механизмом для уникальной идентификации файлов и эффективного получение данных из IPFS. CID имеет две версии: V0 и V1 [19]. При этом более гибким использовалась версия CID V1. CID включает в себя хеш соответствующих данных, обеспечивая различные данные будут иметь разные CID. Обычно CID V1 рассчитывается с использованием хеша SHA256. функция, а стандартная реализация использует SHA256 для вычисления CID.

Структура данных CID V1 показана в таблице 2.

Таблица 2. Структура данных CID V1.

Позиция байта	Описание	Ценность в реализации
Первый байт	Многобазовый	0x01
Второй байт	префикс версии CID	0x55: необработанные данные
Третий байт	Идентификатор хеш-функции	0x12: SHA-256
Четвертый байт	Длина хеша	0x20: 32 байта
Начиная с пятого байта	Хэш-значение	Хэш-значение SHA-256 (32 байта)

Кодирование CID осуществляется с использованием Base32. Base32 кодирует последовательность байтов. построенный на основе этой структуры для генерации CID.

Внутри схемы введенное значение CID декодируется из Base32 и система проверяет, соответствует ли извлеченное значение хеша хешу SHA256, вычисленному из зашифрованные данные.

Впоследствии зашифрованные данные расшифровываются. В качестве шифрования используется AES-CTR. алгоритм, который является разновидностью криптографии с симметричным ключом. Шифрование AES-CTR и использовалась расшифровка в circom-chacha20 [20] . Для расшифровки шифрования AES-CTR: Требуются ключ шифрования и одноразовый номер, используемые во время шифрования. Они вводятся в схема как 256-битный ключ и 128-битный одноразовый номер соответственно в качестве частных входов. Более того, AES-CTR обрабатывает объемы данных, кратные 16. Следовательно, если длина изображения данные перед шифрованием не кратны 16, в конец данных добавляются нули (0x00) чтобы привести его в соответствие с этим требованием.

Наконец, к расшифрованным данным применяется фильтр для встраивания имени получателя. Реализация встраивания текста непосредственно в схему может существенно увеличить вычислительную нагрузку; поэтому вне схемы создается фильтр, выполняющий значительную

Machine Translated by Google

часть обработки изображения заранее. Для текста, представляющего имя получателя, используется шрифт Misaki [21]. Затем фильтр используется для оптимизации обработки внутри схемы. Фильтр представляет собой список чисел, в котором значения от 0 до 255 используются для изменения цвета каждого пикселя, если он отличается от цвета пикселей исходного изображения; более того, значение 300 указывает, что цвет должен оставаться таким же, как в исходном изображении. Этот фильтр представляет позицию на изображении, где должно быть вставлено имя получателя.

Внутри схемы указанные цвета пикселей в расшифрованных данных BMP изменяются на основе этого фильтра.

6.2. ZK-оптимизированная реализация

Реализация, оптимизированная для ZK, заменяет хеш-функцию, технологию шифрования и внутрисхемную обработку на стандартную реализацию шифрования, совместимую с ZK. Эта реализация повышает эффективность вычислений и не оценивает разницу в скорости вычислений между ZK-дружественным шифрованием и обычным шифрованием. Поэтому внутрисхемная обработка также была модифицирована.

Хэш Посейдона [22] использовался в качестве хэш-функции для вычисления CID. Примечательно, что использование хеша Poseidon для CID официально не поддерживается; поэтому он был разработан специально для данного исследования. Хотя SHA256 обычно используется в общих вычислениях, он требует значительного времени вычислений в схемах ZKP. Хэш Poseidon реализован в circom и JavaScript (circomlib [23] и circomlibjs [24] соответственно). Он вычисляется над конечным полем простого порядка и может принимать до 16 входных переменных. Используемый порядок меньше максимального значения в 32 байта, но больше максимального значения в 31 байт. Это указывает на то, что каждый из 16 входов должен содержать данные, не превышающие этот порядок. В этой реализации данные, предназначенные для вычисления хеш-функции, делятся на сегменты по 31 байт в качестве входных значений. Если деление превышает 16 сегментов, хеш Посейдона рассчитывается для первых 16 сегментов. Этот результат добавляется к следующим 15 сегментам данных для последующего ввода хэша Poseidon. Процесс повторяется до тех пор, пока все входные данные не будут использованы для вычисления хэша. В вычислительном отношении, если окончательный ввод не завершает 16 сегментов, недостающие входные данные обнуляются, чтобы гарантировать, что вычисление всегда включает 16 вх

При генерации CID из хэш-значения Poseidon последовательность байтов должна соответствовать структуре данных CID V1 и быть закодирована в формате Base32. Однако для дальнейшего сокращения времени вычислений в этой реализации отсутствует кодировка Base32 и напрямую используется хеш-значение Poseidon вместо CID. Разделение входных данных на 16 сегментов внутри схемы требует больших вычислительных ресурсов; следовательно, это деление выполняется вне схемы и подается на вход. В этом случае последовательность байтов зашифрованных данных и список значений для расчета хэша Poseidon предоставляются в качестве общедоступных входных данных, что позволяет проверить, что оба набора данных представляют одну и ту же информацию. Получатели могут подтвердить, что данные, вычисляемые для хэша Poseidon, и данные, расшифровываемые в схеме, идентичны, путем взаимного преобразования и проверки этих двух значений. В этом случае, поскольку пользователи могут получить расшифрованные данные, существует опасность утечки пароля в результате атак методом перебора или других способов.

В качестве технологии шифрования мы использовали шифрование Poseidon [25] вместо шифрования AES-CTR. Шифрование Poseidon, реализованное в circom и TypeScript (poseidon-cryption-circom2 [26]), предполагает получение открытого ключа получателя, генерацию общего ключа и обеспечение безопасного шифрования и дешифрования обеими сторонами. Однако в этом случае общий ключ генерируется напрямую и используется для шифрования и дешифрования. Схема снабжена двумя значениями, представляющими координаты эллиптической кривой, и значением nonce в качестве частных входных данных для шифрования. При этом фильтр реализован так же, как и в стандартной реализации.

7. Оценка

Мы создали пример программы на основе вышеупомянутых реализаций, которая использует circom для описания схемы и использует snarkjs для выполнения схемы и проверки.

Machine Translated by Google

доказательства. В качестве криптографических библиотек circom-chacha20 [20], circomlib [23], circomlibjs [24], и Poseidon-encryption-circom2 [26] .

Для каждой схемы были реализованы стандартная и оптимизированная для ZK реализации. и время их вычислений сравнивалось во время выполнения. Белые растровые изображения были целевыми изображениями, а эксперименты проводились с буквой «A» в качестве встроенный персонаж. Поскольку встраивание любого количества символов не меняет обработку с помощью фильтра встраивание одного символа позволяло сравнивать время вычислений. Кроме того, мы варьировали размеры изображений, чтобы измерить время выполнения каждой схемы. Использовались размеры 10×10, 15×15, 30×15, 30×30, 60×30, 60×60, 120×60, 120×120. и 180 × 120 пикселей. Средой выполнения была Windows 11 с Ryzen 9 3950X. ЦП и 128 ГБ ОЗУ работают под управлением Ubuntu 22.04 в среде WSL2.

На рисунке 3 показан пример изображения, сгенерированного схемой, специально для размера 60 × 30 пикселей. размер с использованием реализации, оптимизированной для ZK. Представлены результаты для каждого размера изображения. в таблице 3, где нелинейные ограничения указывают количество нелинейных ограничений в время сборки — это время, необходимое для компиляции схемы и вывода схемы, а также доказательства gen time — это время, необходимое для генерации доказательств с помощью схемы. Поскольку стандартная реализация использует шифрование AES-CTR, данные с номером 0x00 добавляются в конце, чтобы гарантировать, что размер ввода кратен 16.

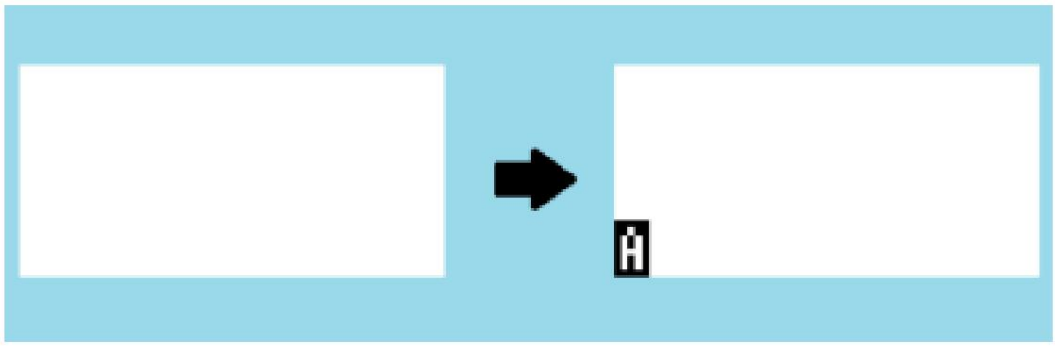


Рисунок 3. Изображение, сгенерированное схемой для размера изображения 60 × 30 с помощью ZK-оптимизированной реализации.

Таблица 3. Сравнение времени выполнения схемы.

	Пиксель	Размер изображения [Байт]	Нелинейный Ограничения	Время сборки [PC]	Доказательство ген. Время [мс]
Стандартный					
	10 × 10	384	558 341	668 220	14 377
	15 × 15	784	1 095 292	1 316 666	25 545
	30 × 15	1440	1 980 688	1 696 370	35 161
	30 × 30	2816	3 867 989	3 657 578	65 785
	60 × 30	5456	7 453 300	7 864 583	126 262
ZK-оптимизированный					
	10 × 10	376	35 407	128 979	3076
	15 × 15	775	72 725	173 210	4032
	30 × 15	1435	134 663	275 630	5900
	30 × 30	2815	263 450	470 872	9733
	60 × 30	5455	509 375	850 612	17 571
	60 × 60	10855	1 013 483	1 257 418	29 044
	120 × 120	43255	4 036 913	6 754 928	96 580

В стандартной и ZK-оптимизированной реализациях для 60×60 пикселей и 180×120 пикселей. размера изображения, в системе закончилась память, и вычисление не удалось завершить. В реализации, оптимизированной с помощью ZK, количество нелинейных ограничений было уменьшено до примерно на одну десятую меньше, чем в стандартной реализации для того же размера изображения.

Machine Translated by Google

Это сократило время сборки и проверки. Однако максимально управляемый размер изображения по-прежнему составлял всего лишь 120 × 120 пикселей, что значительно мало для практических приложений.

8. Обсуждение

Несмотря на то, что во время реализации ZK-оптимизации использовались криптографические технологии, совместимые с ZK, а внутрисхемные процессы были оптимизированы, максимально управляемый размер изображения составлял примерно 120 × 120 пикселей. Это ограничивает практическую полезность изображениями значительно малого размера. Однако исследования, направленные на повышение производительности ZKP, продолжаются, и будущие технологические достижения могут позволить обрабатывать изображения большего размера. Например, Чжан и др. [27] добились десятикратного ускорения zk-SNARK с использованием ASIC. Ма и др. [16] аналогичным образом использовали графические процессоры для ускорения времени создания доказательства, добившись производительности до 48,1 раз выше по сравнению с традиционными методами. Кроме того, предложены методы упрощения вычислительных процессов, например метод «сворачивания». Этот метод сжимает доказываемые предложения [28]. По мере постепенного изучения возможностей повышения скорости потребление памяти, вероятно, также будет оптимизировано. Это потенциально позволит в будущем обрабатывать изображения большего размера.

Кроме того, мы обнаружили, что наш метод предложения может обрабатывать данные размером около 10 КБ. Хотя прямое применение нашего предложения к реалистичным данным изображения (от нескольких МБ до десятков МБ) является сложной задачей, разделение данных на фрагменты путем изменения шифрования и встроенных строк может сделать приложение осуществимым.

Более того, в наших реализациях требуется значение, основанное на размере исходных данных, подлежащих обработке (шифрованию), в качестве аргумента во время генерации схемы. Поэтому для каждых данных необходимо генерировать схему. Время генерации схемы (время сборки) значительно увеличивается с увеличением размера данных изображения; например, даже в реализации, оптимизированной с помощью ZK, для создания схемы для изображения размером 120 × 120 требуется более 112 минут (6 754 928 мс). Однако после генерации схемы время генерации доказательства в тех же условиях невелико, примерно 97 с (96 580 мс). Другими словами, после создания схемы генерация доказательства не требует много времени. Этот факт не создает никаких практических проблем в случаях , когда одно и то же изображение распространяется среди разных людей.

В реализациях, дружественных к ZK, зашифрованные данные вводятся как общедоступные данные. Обработка ключей шифрования изображений требует тщательного рассмотрения. Данные, управляемые в частной IPFS, зашифрованы. Однако в случае утечки ключей шифрования зашифрованные данные могут быть расшифрованы. Таким образом, конкретные пользователи, управляющие частной IPFS, должны стать администраторами, чтобы тщательно управлять ключами, или в той же сети можно создать блокчейн типа консорциума для надлежащего установления и управления правами доступа.

9. Выводы

Здесь был предложен новый метод для распространения данных, хранящихся в частной IPFS, внешним объектам, обеспечивая при этом возможность проверки их подлинности. В этом методе применяется тип ZKP, zk- SNARKs, для проверки CID данных и внедрения имени получателя. Этот подход позволяет внешним объектам проверять, что полученные данные созданы на основе исходных данных в частной IPFS, не требуя таких подробностей, как права доступа IPFS и ключи шифрования.

Для повышения вычислительной эффективности предлагаемого метода были реализованы стандартная реализация с использованием традиционных криптографических методов и ZK-оптимизированная реализация с использованием криптографических схем, дружественных к ZK . Эксперименты с образцом программы подтвердили эффективность предложенного метода при размере данных изображения до 120×120 пикселей.

Этот предлагаемый метод расширяет диапазон использования децентрализованных систем хранения до гибридного случая — при необходимости распределяя внутренние данные по конкретным внешним объектам. Это исследование открывает новый способ обмена конфиденциальной информацией между различными секторами внутри группы и за ее пределами. Однако для широкой практической применимости предлагаемого метода к более крупным и разнообразным типам данных, таким как изображения и видео, скорость обработки должна

Machine Translated by Google

должны быть улучшены и необходимо использовать методы разделения данных, которые находятся в рамках наших будущих исследований.

Вклад авторов: концептуализация, KS, CH, TTC и WO; письмо – подготовка оригинального черновика , KC; написание — рецензирование и редактирование, KC и KИ; надзор, генеральный директор; администрация проекта, K.C. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Это исследование было проведено в рамках совместного исследовательского проекта Токийского университета и Callisto Inc. при финансовой поддержке Callisto Inc. Эта работа была поддержана кафедрой инноваций в области блокчейна и Центром будущих наук и технологий Мохаммеда бен Салмана для Саудовской Аравии. -Японское видение 2030 (MbSC2030) в Токийском университете.

Заявление о доступности данных. Исходный код, использованный для моделирования, доступен на GitHub. https://github.com/blockchaininnovation/circom_image_processing (по состоянию на 21 марта 2024 г.).

Конфликты интересов: авторы Чанги Хан, Цз Тат Чу и Ватару Одзаки работали в компании Callisto Inc. Остальные авторы заявляют, что исследование проводилось в отсутствие каких-либо коммерческих или финансовых отношений, которые могли бы быть истолкованы как потенциальный конфликт интересов. .

Рекомендации

1. Грот Дж. О размере неинтерактивных аргументов на основе пар. В Proceedings of the Advances in Cryptology – EUROCRYPT 2016: 35-я ежегодная международная конференция по теории и применению криптографических методов, Вена, Австрия, 8–12 мая 2016 г.; Слушания, часть II 35; Springer: Берлин/Гейдельберг, Германия, 2016 г.; стр. 305–326.

2. Кумар, С.; Бхарти, АК; Амин, Р. Децентрализованное безопасное хранение медицинских записей с использованием блокчейна и IPFS: сравнительный анализ анализ с будущими направлениями. Безопасность. Прив. 2021, 4, e162. [Перекрестная ссылка]

3. Азбег, К.; Учетто, О.; Андалусси, SJ BlockMedCare: система здравоохранения на основе Интернета вещей, блокчейна и IPFS для данных. безопасность управления. Египет. Поставить в известность. Дж. 2022, 23, 329–343. [Перекрестная ссылка]

4. Хоссан, М.С.; Хатун, М.Л.; Рахман, С.; Рено, С.; Ахмед, М. Обеспечение безопасности службы совместного использования поездок с использованием IPFS и Hyperledger на основе частного блокчейна. В материалах 24-й Международной конференции по компьютерным и информационным технологиям (ICCIT) 2021 г., Дакка, Бангладеш, 18–20 декабря 2021 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2021 г.; стр. 1–6.

5. Лин, Ю.; Чжан, К. Метод защиты частных данных в IPFS. В материалах 24-й Международной конференции IEEE 2021 года по совместной компьютерной работе в дизайне (CSCWD), Далянь, Китай, 5–7 мая 2021 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2021 г.; стр. 404–409.

6. Батта, А.А.; Мадин, ММ; Альзааби, Х.; Якуб, И.; Салах, К.; Джаяраман, Р. Многосторонняя авторизация на основе блокчейна для доступа к зашифрованным данным IPFS. IEEE Access 2020, 8, 196813–196825. [Перекрестная ссылка]

7. Хуанг, Х.С.; Чанг, Т.С.; Ву, Дж. Я. Безопасная система обмена файлами, основанная на IPFS и блокчейне. В материалах 2-й Международной конференции по электронике, Сингапур, 8–10 июля 2020 г.; стр. 96–100.

8. Сан, Дж.; Яо, Х.; Ван, С.; Ву, Ю. Схема безопасного хранения и доступа к электронным медицинским записям на основе блокчейна в IPFS. IEEE Access 2020, 8, 59389–59401. [Перекрестная ссылка]

9. Канг, П.; Ян, В.; Чжэн, Дж. Метод совместного использования частного файлового хранилища Blockchain на основе IPFS. Датчики 2022, 22, 5100. [CrossRef] [ПабМед]

10. Уддин, Миннесота; Хаснат, АНМА; Насрин, С.; Алам, MS; Юсуф, М.А. Безопасная система обмена файлами с использованием технологий блокчейн, ipfs и pki. В материалах 5-й Международной конференции по электрическим информационным и коммуникационным технологиям (EICT) 2021 г., Кхулна, Бангладеш, 17–19 декабря 2021 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2021 г.; стр. 1–5.

11. Ли, В.; Го, Х.; Неджад, М.; Шен, СС Управление трафиком с сохранением конфиденциальности: подход, вдохновленный блокчейном и доказательством с нулевым разглашением . IEEE Access 2020, 8, 181733–181743. [Перекрестная ссылка]

12. Беллес-Муньос, М.; Изабель, М.; Муньос-Тапия, JL; Рубио, А.; Бэйлина, Дж. Сирком: Язык описания схем для построения приложения с нулевым разглашением. IEEE Транс. Надежная безопасность. Вычислить. 2022, 20, 4733–4751. [Перекрестная ссылка]

13. Официальный сайт Circom. Доступно онлайн: https://iden3.io/circom. (по состоянию на 24 марта 2024 г.).

14. Репозиторий Snarkjs на Github. Доступно онлайн: https://github.com/iden3/snarkjs . (по состоянию на 4 июня 2024 г.).

15. ZCash. Доступно онлайн: https://z.cash/ (по состоянию на 12 июля 2024 г.).

16. Ма, В.; Сюн, К.; Ши, Х.; Макс.; Джин, Х.; Куанг, Х.; Гао, М.; Чжан, Ю.; Шен, Х.; Ху, В. Гзкп: Система доказательства с нулевым разглашением, ускоренная на графическом процессоре. В материалах 28-й Международной конференции ACM по архитектурной поддержке языков программирования и операционных систем, том 2, Ванкувер, Британская Колумбия, Канада, 25–29 марта 2023 г.; стр. 340–353.

17. Хан, К.; Шибано, К.; Одзаки, В.; Осаки, К.; Харагути, Т.; Хирахара, Д.; Кимура, С.; Кобаяши, Ю.; Моги, Г. Универсальная платформа для исследований и разработок искусственного интеллекта в области медицинской визуализации, включающая сбор, отбор, аннотирование и предварительную обработку данных. В Proceedings of the Medical Imaging 2024: Imaging Informatics for Healthcare, Research and Applications, Сан-Диего, Калифорния, США, 18–23 февраля 2024 г.; SPIE: Беллингем, Вашингтон, США, 2024 г.; Том 12931, стр. 311–315.

18. Миано, Дж. Форматы файлов сжатых изображений: Jpeg, png, gif, xbm, bmp; Addison-Wesley Professional: Бостон, Массачусетс, США, 1999 г.

Machine Translated by Google

19. Идентификаторы контента (CID). Доступно онлайн: https://docs.ipfs.tech/concepts/content-addressing/#cids-are-not-file-hashes. (по состоянию на 24 марта 2024 г.).
20. circom-chacha20 Репозиторий Github. Доступно онлайн: https://github.com/reclaimprotocol/circom-chacha20. (доступ 24 март 2024 г.).
21. Японский шрифт «Misaki Font» размером 8 × 8 точек. Доступно онлайн: https://littlelimit.net/misaki.htm. (по состоянию на 11 июня 2024 г.). (На японском языке)
22. Грасси, Л.; Ховратович Д.; Рехбергер, К.; Рой, А.; Шофнеггер, М. Посейдон: Новая хэш-функция для систем доказательства с {нулевым разглашением} . В материалах 30-го симпозиума по безопасности USENIX (USENIX Security 21), Ванкувер, Британская Колумбия, Канада, 11–13 августа 2021 г.; стр. 519–535.
23. Репозиторий Circcomlib на Github. Доступно онлайн: https://github.com/iden3/circomlib . (по состоянию на 21 марта 2024 г.).
24. Репозиторий Circcomlibjs на Github. Доступно онлайн: https://github.com/iden3/circomlibjs . (по состоянию на 21 марта 2024 г.).
25. Ховратович, Д. Шифрование с помощью Посейдона. 2019. Доступно онлайн: https://drive.google.com/file/d/1EVrP3DzoGbmzkRmYn.yEDciQcXVU7GlOd/просмотр (по состоянию на 19 июля 2024 г.).
26. Репозиторий Poseidon-Encryption-Circom2 на Github. Доступно онлайн: https://github.com/Shigoto-dev19/poseidon-encryption-c. ircom2 (по состоянию на 21 марта 2024 г.).
27. Чжан Ю.; Ван, С.; Чжан, Х.; Донг, Дж.; Мао, Х.; Лонг, Ф.; Ван, К.; Чжоу, Д.; Гао, М.; Сан, Г. Пипецк: Ускорение доказательства с нулевым разглашением с помощью конвейерной архитектуры. В материалах 48-го ежегодного международного симпозиума по компьютерной архитектуре (ISCA) ACM/IEEE 2021 г., Валенсия, Испания, 14–18 июня 2021 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2021 г.; стр. 416–428.
28. Котапалли, А.; Сетти, С.; Циалла, И. Нова: Рекурсивные аргументы с нулевым разглашением из схем свертывания. В материалах ежегодной международной конференции по криптологии, Санта-Барбара, Калифорния, США, 15–18 августа 2022 г.; Шпрингер: Чам, Швейцария, 2022 г.; стр. 359–388.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Автономный кооперативный навигационный подход для Несколько беспилотных наземных транспортных средств в переменной Коммуникационная среда

Сюдун Линь и Мэнсин Хуан *

Школа информатики и коммуникационной инженерии, Хайнаньский университет, Хайкоу 570228, Китай; linx424523604@163.com

* Переписка: huangmx09@163.com

Аннотация: Роботы помогают службам экстренного реагирования, удаленно собирая важную информацию. Использование нескольких совместных беспилотных наземных транспортных средств (UGV) для реагирования может сократить время реагирования, улучшить ситуационную осведомленность и минимизировать затраты. Надежная связь имеет решающее значение для нескольких UGV для реагирования на окружающую среду, поскольку нескольким роботам необходимо обмениваться информацией для совместной навигации и сбора данных. В этой работе мы исследуем политику управления для оптимальной связи между несколькими UGV и базовыми станциями (BS). Предлагается многоагентный алгоритм глубокого детерминированного градиента политики (MADDPG) для обновления политики управления для достижения максимального отношения сигнал/помеха. UGV связываются как со стационарными БС, так и с мобильными БС. Предлагаемая политика управления позволяет перемещаться по UGV и мобильным БС для оптимизации связи и мощности сигнала. Наконец, предлагается генетический алгоритм (ГА) для оптимизации гиперпараметров обучения на основе MADDPG. Результаты моделирования демонстрируют вычислительную эффективность и надежность алгоритма MADDPG на основе GA для управления несколькими

Ключевые слова: беспилотные наземные транспортные средства (БГА); генетический алгоритм (ГА); многоагентный глубокий детерминированный политический градиент (MADDPG); автономная навигация

check for updates

Цитирование: Лин, Х.; Хуанг, М. Ан Автономная кооперативная навигация Подход к множеству беспилотных Наземная техника в переменной Коммуникационная среда. Электроника 2024, 13, 3028. https://doi.org/10.3390/electronics13153028 .

Академические редакторы: Эрик Мэтсон и Молодой Им Чо

Поступила: 18 июня 2024 г.
Пересмотрено: 25 июля 2024 г.
Принято: 30 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария.
Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Сеть распределенных беспилотных наземных транспортных средств (БНА) и центрального контроллера известна как система управления несколькими БНГ [1]. Эта система обеспечивает автономное доминирование, автономную навигацию и автономное сотрудничество. Он может работать либо в пределах зоны ограниченного доступа, либо как часть более широкой транспортной системы. Системы управления несколькими UGV предлагают уникальный подход к навигации, который является высоконадежным, более экономичным и способствует экономии энергии. В последние годы острая потребность в навигационных системах с несколькими UGV вызвала растущее количество дискуссий в научных круга

Навигация UGV в коммуникационной среде была предметом исследований [6], а традиционные методы оптимизации дали хорошие результаты [7]. Для создания автономной навигационной системы Д. Чен с соавт. [8] разработал эвристический алгоритм Монте-Карло, основанный на дискретном преобразовании Хафа и локализации Монте-Карло, что обеспечивает низкую сложность обработки в реальном времени. В отличие от инноваций в алгоритмах , обеспечивающих надежную работу в неизвестных и загроможденных средах, HU Unlu et al. [9] создали надежный подход к инерциальной навигации с помощью машинного зрения, который может противостоять неопределенностям . В отличие от использования наглядных пособий X. Lyu et al. [10] были вдохновлены геометрической точкой зрения, и они разработали новую адаптивную схему технологии объединения навигационной информации с интегрированным коэффициентом совместного использования, которая имеет адаптивную навигацию в случае нелинейных систем и использует негауссово распределение. Эти традиционные методы оптимизации, упомянутые выше, легко реализовать. Однако эти методы должны быть представлены с предварительными условиями, что делает их подходящими только для статических сред. Более того, в действительности большинство сценариев

Machine Translated by Google

Следовательно, системы с несколькими UGV столкнутся с этими двумя проблемами при работе со сложными сценариями, и для эффективного их решения необходимо использование машинного обучения (ML) [12–14].

Существует веское обоснование для использования методов машинного обучения в навигации UGV, учитывая быстрый прогресс в области машинного обучения. Чтобы добиться улучшения характеристик дальности действия, H. Lee et al. [15] предоставили метод машинного обучения для расчета расстояния между БС и БПЛА, который позволяет осуществлять локализацию без какой-либо дополнительной инфраструктуры. Вместо того, чтобы полагаться на прямое определение дальности, HT Nguyen et al. разработала систему координации между беспилотными летательными аппаратами и UGV, обеспечивающую эффективную совместную навига Однако по мере усложнения среды моделирования эффективность предлагаемого решения быстро снижается. Для решения этой проблемы многообещающим выбором является использование алгоритмов обучения с подкреплением (RL). RL подчеркивает, как агенты могут найти лучшую политику для максимизации всех вознаграждений при взаимодействии с окружающей средой, что делает ее хорошо подходящей для исследования и адаптации к все более сложным средам [16].

В последнее время исследования были вызваны дискуссиями об использовании RL для решения проблемы совместной навигации нескольких UGV [17]. Чтобы избежать столкновений с препятствиями, X. Huang et al. [18] предложили инновационную систему навигации локального планирования маршрута UGV на основе глубокого RL, которая использует мультимодальное восприятие для облегчения изучения политики для создания гибких навигационных действий. В отличие от навигации с одним UGV, для улучшения средней спектральной эффективности S. Wu et al. [19] предложили технологию оптимизации траектории, основанную на совместном многоагентном глубоко детерминированном политическом градиенте (F-MADDPG), который наследует способность MADDPG совместно управлять несколькими UGV и использует совместное усреднение для устранения изоляции данных и ускорения конвергенции. Значительный прогресс был достигнут с помощью этих методов навигации UGV на основе RL. Однако они упускают из виду ограничения статических коммуникационных сред и проблемы конвергенции, возникающие из-за сложности среды.
Эти два элемента крайне важно учитывать при планировании совместной навигации в коммуникационной среде.

Учитывая ограничения совместной навигации по зоне покрытия связи для UGV, необходимо преодолеть три основные проблемы, такие как сложность одновременного управления UGV, различия в зоне покрытия связи и сложность среды совместного управления для UGV. Во-первых, традиционные методы управления, такие как Q-обучение [20], пропорционально-интегрально-дифференциальное (ПИД) управление [21] и глубокая Q-сеть [22] , часто дают неоптимальную производительность с точки зрения покрытия связи , когда нескольким агентам требуется одновременное управление. контроль. Во-вторых, учитывая изменчивость коммуникационной среды во время навигации нескольких UGV, часто встречаются области с плохой связью, что препятствует эффективному сотрудничеству между несколькими UGV.
Однако перспективное решение проблем многоагентного кооперативного управления предлагают многоагентные алгоритмы RL [23]. Эти алгоритмы направляют многоагентное сотрудничество посредством парадигмы централизованного обучения и децентрализованного выполнения (CTDE) [24].
Кроме того, в предлагаемом нами подходе мы вводим передвижную БС UGV, интегрированную с UGV, что позволяет динамически изменять стационарную среду связи. Это сотрудничество эффективно поддерживает навигационные задачи UGV. Однако возросшая сложность построенной среды может создать проблемы с эффективностью и сходимостью алгоритмов. К счастью, мы смягчаем трудности конвергенции за счет адаптивного обновления динамических гиперпараметров с использованием генетического алгоритма (ГА) [25]. К счастью, были проведены некоторые исследования по интеграции GA для настройки гиперпараметров в структуры RL.
А. Сегал и др. использовали ГА, чтобы найти воспроизведение ретроспективного опыта (HER), используемое в глубоком детерминированном политическом градиенте (DDPG) в задаче манипулирования роботом, чтобы помочь агенту ускорить обучение [26]. В отличие от изменения одного параметра, для задачи гибкого планирования цеха (FJSP) Chen R et al. предложил метод настройки параметров GA, основанный на Q-обучении, который изменяет несколько ключевых параметров Q-обучения для получения более высоких значений вознаграждения [27]. Однако этот подход, основанный на вознаграждениях, склонен впадать в локальную оптимальность. Более того, эти методы не подходят для сценариев, когда количество агентов увеличивается. Для решения этих проблем Алипур и др. предложил гиб

Machine Translated by Google

мультиагентная RL-эвристика для решения задачи коммивояжера. Таким образом, ГА с новым оператором пересечения действует как эвристика улучшения путешествия, а MARL действует как строительная эвристика [28]. Хотя этот подход позволяет избежать риска локальной оптимальности, он отказывается от процесса обучения MARL и использует его только как эвристику, вместо этого используя GA для обучения, что означает, что алгоритм не будет уделять слишком много внимания сотрудничество между интеллектуальными агентами. Лю и др. использовал децентрализованную частично наблюдаемую метод многоагентного планирования пути, основанный на эволюционном RL (MAPPER), для эффективного обучения стратегии местного планирования в смешанных динамических средах. Основываясь на многоагентном обучении с подкреплением, они использовали ГА для итеративного расширения первоначально обученного алгоритма. к более сложной модели. Хотя этот метод позволяет избежать снижения производительности в долгосрочных задачах, итеративный ГА не обязательно может хорошо адаптироваться к более сложным средам [29].

В наших исследованиях мы объединяем преимущества вышеупомянутых статей ГА и принимаем парадигма CTDE для проведения исследований в многоагентной среде RL. ГА назначает разные веса обновлениям алгоритма в зависимости от вклада перехода, что означает что мы уделяем больше внимания гиперпараметрам, которые в большей степени способствуют обновлению модели а не те, которые достигают более высоких значений вознаграждения. Это позволяет нам не впадать в локальная оптимальность при увеличении числа агентов.

Чтобы решить эти три проблемы и добиться совместной навигации в комплексе В различных средах предлагается новый метод навигации по покрытию связи с несколькими UGV , который основан на многоагентном глубоком детерминированном политическом градиенте с GA (GA- MADDPG). Ниже кратко изложены основные преимущества метода навигации по зоне покрытия связи с несколькими UGV:

- Комплексная мультиагентная модель объединена в совместную работу нескольких UGV. навигационной системы, а оптимальная координация работы нескольких БПЛА в зоне покрытия связи формулируется как многоагентный марковский процесс принятия решений в режиме реального времени. (МДП) модель. Все UGV настроены как независимые агенты с возможностями самоконтроля.
- Предлагается метод мультиагентной совместной навигации с расширенным покрытием связи . Благодаря внедрению мобильной базовой станции покрытие связи окружающая среда динамически изменяется. Результаты моделирования показывают, что этот метод эффективно улучшает качество связи во время навигации.
- Представлен адаптивный подход к гиперпараметрам на основе GA для оптимизации UGV. покрытие связи и навигация. Он присваивает веса гиперпараметрам по степени обновления алгоритма и делает выбор на основе размер веса при следующем выборе, который отличается от традиционной стратегии с фиксированными гиперпараметрами и может избежать локальных оптимумов.

Остальные части эссе организованы следующим образом. Моделирование систем связи и навигации с несколькими UGV подробно описано в разделе 2. Подробности представляемого нами метода RL изложены в разделе 3. Несколько экспериментальных сравнений в разделе 4 служат для проверки эффективности нашего подхода. В конце концов, мы обсуждаем будущее направления исследования и обобщить ключевые положения статьи в заключении раздела 5.

2. MDP для навигационного и коммуникационного покрытия для нескольких UGV в окружающей среде

Чтобы имитировать процесс принятия решений несколькими UGV в реальных системах, мы принимаем Модель MDP. С быстрым развитием мультиагентного RL MDP превратился в надежная модель принятия решений [30]. В этом исследовании мы создаем сложную среду с взаимодействующими тремя БПЛА и одной мобильной БС и включающей в себя различные препятствия. Кроме того, мы вводим понятие коммуникации, согласно которому общение зона покрытия определяется четырьмя фиксированными БС и одной мобильной БС в совокупности.

2.1. описание проблемы

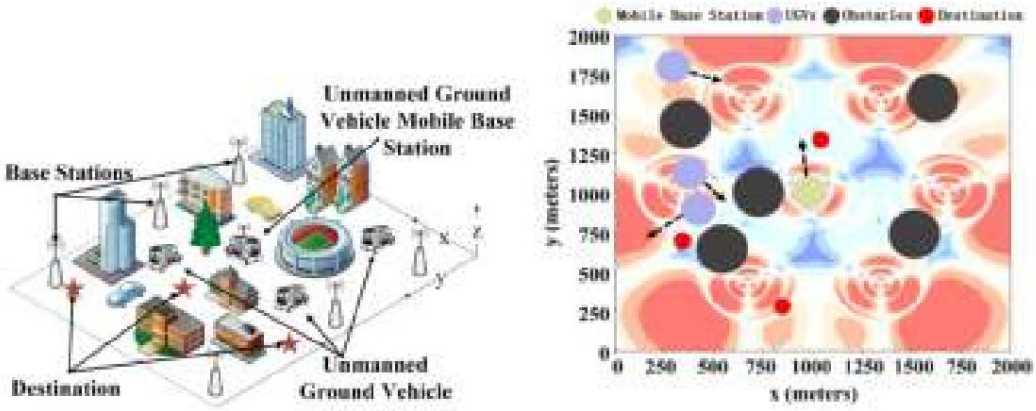
Основная цель нашей статьи — решение задач мультиагентной навигации в широкий спектр крупномасштабных, неизвестных и сложных сред как можно быстрее. Задача навигации требует, чтобы БПЛА могли взаимодействовать по различным характеристикам окружающей среды, с возможностью преодоления внешней экологической информации.

Machine Translated by Google

помех, а также с возможностью эффективного и автономного отслеживания целей в режиме реального времени. Более конкретно, степень покрытия связи устанавливается совместно как стационарными BS, так и мобильными BS. В пределах этой зоны покрытия связи мобильная БС и БПЛА осуществляют совместную навигацию. Мы создаем сценарий задачи с несколькими целями оптимизации. Целью UGV является успешное достижение пункта назначения, в то время как мобильной BS поручено динамически регулировать покрытие связи в режиме реального времени с целью оптимизации качества связи для UGV. UGV и мобильная БС осуществляют глобально оптимальную совместную навигацию для достижения своих общих целей.

2.2. Моделирование окружающей среды

Наша работа включает в себя моделирование реальной среды, в которой несколько UGV взаимодействуют для достижения целевой точки. Кроме того, эта среда включает в себя препятствия, которые препятствуют движению UGV, воспроизводя реальные сценарии. Мы используем среду мультиагентных частиц (MPE) [24] в качестве базовой среды для нашей вторичной разработки, как показано на рисунке 1. В этой среде мы используем M UGV (где M определяется как три), W мобильных BS (где W определяется как единица) и определенное количество препятствий. Целью UGV является совместное предотвращение столкновений и достижение соответствующих оптимальных целевых точек, принимая во внимание коммуникацию в глоба
UGV выбирают маршрут обхода препятствий с лучшей связью для координации своих действий.
движение к целевой точке (модель связи будет подробно рассмотрена в разделе 2.3). Задачей мобильной БС является улучшение связи для трех подвижных единиц путем регулирования зоны покрытия связи в глобальном состоянии, что показано на рисунке 1b.



(а) (б)

Рисунок 1. Схематическая диаграмма взаимодействия стаи UGV в среде с возможностью связи. (а) 3D городская среда. (б) Вид сверху на визуализируемую коммуникационную среду.

2.3. Моделирование покрытия связи

В нашем моделировании мы интегрировали связь в среду MPE и использовали ее в качестве критерия для оценки выполнения задачи. В этом подразделе мы представляем принятую нами модель канала связи, а также модель связи, на которую влияет движение мобильной БС, как показано на рисунке 1b. Область связи внутри среднего красного круга варьируется в зависимости от местоположения подвижной БС, как показано на рисунке 2. Обратите внимание, что на рисунках 2a–l представлены диаграммы, показывающие, как меняется среда связи при движении мобильной БС на этапе t. Мобильная БС первоначально расположена в центре на рисунке 2a и постепенно переходит в нижний правый угол, как показано на рисунке 2l. Такое перемещение мобильной БС вызвано наблюдением за схемой движения UGV. Следовательно, мобильная БС перемещается из центра в нижний правый угол, чтобы улучшить качество связи в этой области, тем самым расширяя красную зону покрытия, как показано на рисунке 2. И наоборот, перемещени

Machine Translated by Google

мобильная БС приводит к уменьшению зоны покрытия при улучшении качества связи в верхнем левом квадранте. Более того, чтобы подчеркнуть развивающееся качество общения и повысить ясность изменений в общении, мы обозначили различия между каждой диаграммой и ее предыдущим аналогом.

Мы построили в общей сложности M BS в среде, где M определяется как семь и включает одну мобильную BS. Прирост мощности сигнала, получаемый БПА от БС m (m 7), определяется как p – основной $\overset{M}{T}$. Впоследствии отношение сигнал/помеха (SIR) используется как критерий оценки связи БС. Этот критерий может быть выражен как:

$$SIRt \quad p \frac{\overset{\text{Это}}{\tau}}{m \not= It \cdot p} \quad (1)$$

где It {1, · · · , M} представляет собой БС, которые не связаны с УГВ на шаге t. Стоит отметить, что для простоты мы опустили влияние шума, поскольку хорошо известно, что производительность связи BS-UGV часто ограничивается помехами. Кроме того, при глобальном повторном использовании частот мы учли ситуацию наихудшего случая, в которой все эти несвязанные BS вносят вклад в состав помех в уравнении (1). В нашем исследовании мощность сигнала UGV на шаге t зависела главным образом от их относительного положения относительно БС, и pt можно записать как:

$$pt = P\bar{}\beta(qt)G(qt)\,\tilde{}ht \quad (2)$$

где P представляет мощность передачи BS, а β(qt) представляет крупномасштабное усиление канала; крупномасштабное усиление канала учитывает эффекты потерь на трассе и затухания теней. Это может быть выражено как:

$$\beta(qt) = \beta 0 \quad \frac{}{d0\ d(\text{кт})}^{\gamma} \quad (3)$$

где β0 — потери на трассе на опорном расстоянии d0, d(qt) — расстояние между UGV и БС, а γ — показатель степени потерь на трассе. G (qt) обозначает усиление антенны БС; усиление антенны BS учитывает направленное усиление UGV относительно антенны BS. Его можно представить диаграммой направленности антенны, которая обычно выражается как:

$$G(qt) = Gmax \cdot A(\theta t\ , \varphi t) \quad (4)$$

где Gmax — максимальное усиление антенны, а A(θt , φt) — функция усиления положения UGV относительно направления главного лепестка антенны. Эти параметры обычно зависят от местоположения UGV . Кроме того, случайная величина ˜ht используется для учета эффектов затухания. Важно отметить, что каждое UGV имеет независимый SIR на каждом этапе t, который используется для оценки производительности связи UGV в этот конкретный момент времени. Также следует отметить, что при инициализации сценария начальные положения всех базовых станций, включая подвижную базовую станцию, фиксированы, т.е. все они находятся в фиксированном положении, а затем три UGV и подвижная базовая станция. обучаются выполнять различные действия посредством стратегии, при этом на основе выбранного действия следующая позиция подвижной базовой станции определяется выбранным действием, а также исходным положением вместе взятым. На этом этапе значение qt фиксировано, поскольку qt связано только с переменной положения (x, y). Видно, что начальное положение мобильной БС задано заранее, тогда как последующее qt является переменной решения и определяется действием мобильной БС, целью которого является обеспечение лучшей среды связи для оставшихся UGV.

Machine Translated by Google

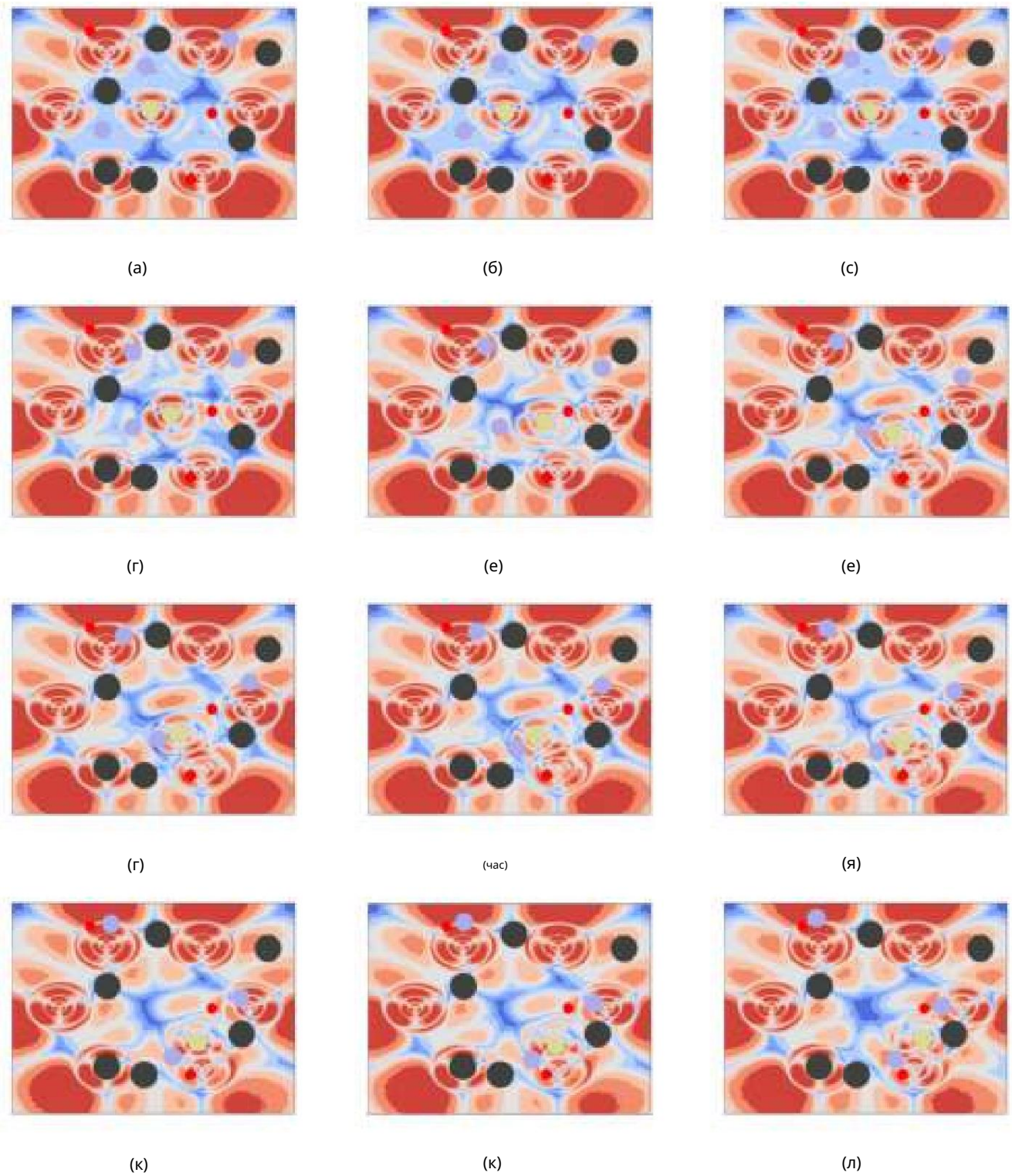


Рис. 2. Результаты изменения связи в среде по мере движения мобильной БС. красные области указывают на лучшую связь, тогда как синие области указывают на плохую связь. Коричневые кружки обозначают мобильную БС, синие кружки обозначают БПА, черные кружки обозначают препятствия, а красные круги обозначают целевые точки. Это движение направлено на повышение качества связи для УГВ.

Machine Translated by Google

2.4. Состояние и действие БПА. Состояние БПА

обозначается как s = (s1,s2, … ,sN). Для каждого UGV u состояние определяется как su = (sPu,sEu), где sPu = xu, yu, vxu, vyu, SIRu — комбинация положения (xu, yu), скорости (vxu, vyu) и SIRu. . Кроме того, sEu = xug, yug, x0, y0, vox, voy представляет данные о том, что UGV наблюдают другие UGV или препятствия. Термин su обозначает положение агента в системе координат. Однако во многих реальных ситуациях получить абсолютные местоположения может оказаться невозможным. Следовательно, агент и барьеры могут быть смоделированы в полярной системе координат движения. В нашей исходной формулировке sEu = xug, yug, x0, y0, vox, voy предназначено для представления наблюдаемых данных для каждого UGV u. Чтобы уточнить, (xug, yug) представляет собой расстояние от объекта g (включая UGV и все препятствия) до UGV u. A (x0, y0) представляет собой глобальную координату UGV u.

Посредством серии преобразований мы также можем получить глобальные позиции других объектов, наблюдаемых UGV u. Комбинация этих компонентов позволяет каждому UGV двигаться к своей цели, учитывая наличие и движение препятствий или других UGV.

Действие УГВ обозначается как a = (a1, a2, … , aN), которое определяется как совокупность индивидуальных действий каждого УГВ в многоагентной системе. В этой конкретной статье движение UGV упрощается за счет предположения о начальной скорости, равной 0, и постоянном ускорении, которое представлено формулой: vt = v0 + at, которое определяется как 2-мерный вектор.

2.5. Функция вознаграждения

Основная цель — найти оптимальную стратегию сотрудничества для конкретного состояния, чтобы совместно перемещаться на этапе t и следующем шаге t + 1 с улучшенной коммуникацией. На этапе t конкретное состояние обозначается как st . Награда за действие в может быть представлена как r(st , at). Следовательно, общая выгода от принятия политики π может быть выражена как:

$$R(\pi) = L \quad y \quad {}^{\tau} r(ст , \pi(ст)) \mid c0 = c \tag{5}$$

Наша цель – определить оптимальную стратегию, обозначаемую как π , что максимизирует общее вознаграждение при соблюдении всех заданных ограничений. Основной целью статьи является получение политики, которая дает максимально возможное вознаграждение, обозначаемое как R (π), среди всех возможных политик π.

Следует отметить, что принципы навигации мобильных БС аналогичны принципам навигации UGV. И UGV, и мобильная BS могут использовать схожие принципы планирования пути и обхода препятствий на основе их целевых позиций и текущих данных об окружающей среде. У UGV три задачи: во-первых, UGV достигают пункта назначения посредством совместной навигации. Во-вторых, UGV должны стараться избегать столкновений. В-третьих, UGV должны путешествовать в среде связи высокого качества. У мобильной БС всего две задачи: Одна – работать с БПЛА по корректировке зоны покрытия путем корректировки положения, обеспечивая таким образом движение БПА в качественной дальности связи. Один из них — максимально избегать столкновений, что аналогично цели UGV. Также важно отметить, что начальная позиция мобильной БС находится в самом центре сцены во всех настроенных нами сценариях и что она движется вместе с UGV, не упреждая их.

Таким образом, в этой модели обучения функция вознаграждения UGV в основном состоит из трех частей, основанных на теоретической основе. Во-первых, это связано с расстоянием между БПЛА и целевой точкой. Во-вторых, это связано с количеством столкновений, в том числе столкновений между БПЛА, столкновений БПЛА с мобильной БС, столкновений БПЛА с препятствиями. Наконец, это связано с SIR, полученным UGV на шаге t, который можно сформулировать как r(st , at).

$$r(st , at) = SIRt \quad D(БПА, цель) \quad coll \tag{6}$$

где SIRt представляет собой комплексное качество связи, полученное всеми UGV на каждом этапе t, а определение SIRt подробно представлено в уравнении (1).

D(UGVs, target) представляет собой сумму длин между всеми UGV и их соответствующими

Machine Translated by Google

пункты назначения на каждом шаге t; Следует отметить, что ни у одного UGV нет фиксированного пункта назначения, а это означает, что все UGV будут автономно определять пункт назначения, которого нужно достичь, на основе своих стратегий и наблюдений. Термин «col» представляет собой количество столкновений, произошедших между всеми UGV на каждом этапе t.

Формула расчета SIRt в уравнении (6) была представлена в разделе 2.3. Коммуникация напрямую влияет на функцию вознаграждения UGV, где более высокий уровень коммуникации приводит к большему вознаграждению. Следовательно, у UGV появляется стимул отдавать приоритет местам с лучшей связью, что побуждает их активно перемещаться в эти районы.

D(UGV, цель) рассчитывается по формуле:

$$D(\text{БПЛА, цель}) = (xu \quad xtarget) \qquad ^2 + (ю - уцель)^2 \tag{7}$$

где (xu, yu) содержит информацию о координатах для всех целевых точек, что указывает на то, что потери также уменьшаются по мере того, как расстояние между UGV и пунктом назначения становится меньше. Следовательно, меньшая потеря соответствует более высокой награде. По сути, UGV с большей вероятностью получат большую награду, когда они находятся ближе к целевой точке.

Термин «колл» может быть подтвержден как:

$$\text{кол} = \begin{matrix} 0, \text{ если } D(\text{UGV, кортеж}) > K \quad 1, \\ \text{если } D(\text{UGV, кортеж}) \quad K \end{matrix} \tag{8}$$

где D(UGVs, кортеж) представляет собой расстояния между UGV и различными объектами, такими как другие UGV, мобильная БС и препятствия в данном сценарии. Кроме того, константа «K» используется для оценки возможности столкновения. Если расстояние между любыми двумя объектами меньше значения K, регистрируется столкновение. Следовательно, используя этот подход, мультиагенты сотрудничают, чтобы минимизировать возникновение коллизий.

3. Навигация по многоагентному покрытию связи RL с GA. В этом разделе мы

описываем краткое изложение формулировки MDP для навигации по покрытию связи при сотрудничестве между мобильной BS и UGV.

Далее мы представляем алгоритм DDPG [31], который разработан для непрерывного пространства управления. Опираясь на эти основы, мы разрабатываем инновационный алгоритм RL под названием GA-MADPPG для решения проблем покрытия связи и навигации. Алгоритм GA-MADPPG состоит из двух основных компонентов. Во-первых, мы используем алгоритм MADPPG , который расширяет DDPG в соответствии с парадигмой CTDE. Это позволяет нам использовать преимущества MADPPG при работе с многоагентными системами и проблемами непрерывного управлен Во-вторых, мы интегрируем ГА в алгоритм MADPPG, позволяя обновлять гиперпараметры в реальном времени на основе функции потерь во время процесса обучения. Предлагаемая политика подчеркивает способность алгоритма GA-MADPPG динамически корректировать гиперпараметры на основе функции потерь. Объединив эти два компонента, GA-MADPPG стремится достичь эффективного покрытия связи и навигации в сложных средах.

3.1. Модель MDP

Многоагентная марковская игра, существенное расширение MDP в многоагентном сценарии, является предметом работы [32]. В этой игре теоретическое состояние N агентов представлено буквами s. В каждую эпоху t агенты отслеживают текущее состояние st и выбирают действие в . После этого состояние переходит в следующее состояние st + 1, и все агенты получают вознаграждение r(st , at).

Для оценки функций действие-значение и функций значения отображения состояния-действия, вычисление функции стоимости для стохастической политики влечет за собой:

$$V\pi(st) \models E \qquad \gamma \quad ^{\text{л}} \text{г}(\text{ст+л} , at+l) \mid \text{ул.} \tag{9}$$
$$\text{л=0}$$

Machine Translated by Google

г д е к о э ф ф и ц и е н т д и с к о н т и р о в а н и я р а в е н у [0, 1). А ф у н к ц и я д е й с т в и е -ц е н н о с т ь в ы ч и с л я е т с я с л е д у ю щ и м о б р а з о м:

$$Q\pi(st , at) = E \sum_{л=0}^{} л^{у} г(ст+л , at+l) \mid у л . , в \quad (10)$$

Из у ч е н и е и д е а л ь н о г о с т р а т е г и я, к о т о р а я о п т и м и з и р у е т о б щ у ю о ж и д а е м у ю п р и б ы л ь, я в л я е т с я ц е л ь ю π в с е х а г е н т о в.

$$\pi = \text{аргумент макс.}^{E} \sum_{т=0}^{} т^{у} г (ст , в) \quad (11)$$

3.2. Основы подхода DDPG

DDPG — э т о а л г о р и т м г л у б о к о г о д е т е р м и н и р о в а н н о г о п о л и т и ч е с к о г о г р а д и е н т а, р а з р а б о т а н н ы й д л я р е ш е н и я п р о б л е м у п р а в л е н и я н е п р е р ы в н ы м и д е й с т в и я м и. О н о с н о в а н н а г р а д и е н т а х п о л и т и к и и н а п р я м у ю к о р р е к т и р у е т п а р а м е т р ы п о л и т и к и θ д л я о п т и м и з а ц и и ц е л е в о й ф у н к ц и и.

$$J(\theta) = Es \quad p\pi|,a \quad \pi\theta \quad (12)$$

ч т о я в л я е т с я о с н о в н о й и д е е й DDPG, п о с к о л ь к у о н а п р е д п о л а г а е т и с п о л ь з о в а н и е п о л и т и ч е с к о г о г р а д и е н т а θ J(θ) н а к а ж д о м э т а п е. П о л и т и ч е с к и й г р а д и е н т м о ж н о в ы р а з и т ь с л е д у ю щ и м о б р а з о м:

$$\theta J(\theta) = Es \quad p\pi,a \quad \pi\theta [\quad \theta \log \pi\theta (a \mid s)Q \pi(s, a)] \quad (13)$$

г д е Qπ(s, a) = E R | Р а с п р е д е л е н и е . ^{т} т = c, a = a — ф у н к ц и я д е й с т в и е -ц е н н о с т ь, а р $^{\pi}$ э т о г о с у д а р с т в о

Д е т е р м и н и р о в а н н а я п о л и т и к а т а к ж е м о ж е т б ы т ь в к л ю ч е н а в с т р у к т у р у г р а д и е н т а п о л и т и к и и о б о з н а ч а е т с я к а к μθ : S A [1]. В ч а с т н о с т и, п р и о п р е д е л е н н ы х о б с т о я т е л ь с т в а х м ы м о ж е м з а п и с а т ь г р а д и е н т ц е л и J(θ) = Es р μ [R(s, a)] с л е д у ю щ и м о б р а з о м:

$$\theta J(\theta) = Es \quad D \quad \theta\mu\theta (a \mid s) \quad aQ \mu (s, a)| \; a=\mu\theta (c) \quad (14)$$

Т е о р е м а т р е б у е т, ч т о б ы п р о с т р а н с т в о д е й с т в и я а б ы л о н е п р е р ы в н ы м, п о с к о л ь к у о н о з а в и с и т о т aQμ (s, a).

Г л у б о к и е н е й р о н н ы е с е т и и с п о л ь з у ю т с я в м е т о д е DDPG, к о т о р ы й я в л я е т с я р а з н о в и д н о с т ь ю а л г о р и т м а д е т е р м и н и р о в а н н о г о г р а д и е н т а п о л и т и к и, д л я о ц е н к и п о л и т и к и μ и к р и т и к а Qμ. Э т о н е п о л и т и ч е с к и й п о д х о д, т о е с т ь о н у ч и т с я н а о п ы т е в о в р е м я о б у ч е н и я. П о м и м о о н л а й н -с е т и, DDPG т а к ж е и с п о л ь з у е т ц е л е в у ю с е т ь д л я с т а б и л и з а ц и и о б у ч е н и я. Ц е л е в а я с е т ь п е р и о д и ч е с к и п е р е с м а т р и в а е т с я, ч т о б ы с м я г ч и т ь п о с л е д с т в и я к о л е б а н и й п о л и т и к и в о в р е м я о б у ч е н и я.

3.3. Мультиагентный глубокий детерминированный градиент политики

П о л и т и к а DDPG д е м о н с т р и р у е т п р и с у щ у ю а г е н т у н а д е ж н о с т ь и в о з м о ж н о с т и о б о б щ е н и я, ч т о п р и в о д и т к м а к с и м а л ь н о й п р о и з в о д и т е л ь н о с т и [31]. Э т о п р е и м у щ е с т в о д е л а е т DDPG о с о б е н н о п о д х о д я щ и м д л я о б у ч е н и я в с л о ж н ы х о б с т о я т е л ь с т в а х, к о г д а п р и с у т с т в у ю т н е и з в е с т н ы е и в н е ш н и е п о м е х и. В с в е т е э т о г о м ы п р и н я л и п а р а д и г м у о б у ч е н и я н а в и г а ц и и п о з о н е п о к р ы т и я н а о с н о в е MADDPG. А г е н т в с р е д е а в т о н о м е н и н е м о ж е т в з а и м о д е й с т в о в а т ь с д р у г и м и а г е н т а м и, н о о н о щ у т и м. Н а к а ж д о м ш а г е t а г е н т н е м о ж е т н а б л ю д а т ь т е к у щ и е с х е м ы м о б и л ь н о с т и д р у г и х а г е н т о в. П р е и м у щ е с т в о CTDE з а к л ю ч а е т с я в т о м, ч т о о н у с т р а н я е т н е о б х о д и м о с т ь и с к а т ь к о м п р о м и с с ы м е ж д у а г е н т а м и, а ц е л ь о п т и м и з а ц и и — у в е л и ч и т ь о б щ и й д о х о д в с е х а г е н т о в [33].

$$G = \quad s\hat{}, a, p, г, o, и \quad (15)$$

г д е u п р е д с т а в л я е т и н д е к с к а ж д о г о а г е н т а, а s х р а н и т г л о б а л ь н ы е с т а т у с ы к а ж д о г о а г е н т а и л о к а л ь н ы е н а б л ю д е н и я. Ч л е н а п р е д с т а в л я е т а к т и в н о с т ь в с е х а г е н т о в, а в о з н а г р а ж д е н и е к а ж д о г о а г е н т а я в л я е т с я ч а с т ь ю т е н з о р а r. Ф у н к ц и я н а б л ю д е н и я о б о з н а ч а е т с я б у к в о й o, а р п р е д с т а в л я е т в е р о я т н о с т ь п е р е х о д а и з т е к у щ е г о с о с т о я н и я в с л е д у ю щ е е с о с т о я н и е .

Machine Translated by Google

Точ не е , в игре N аг е нтов и страте г ий, параме триз ованных θ = {θ1, . . . , θN}. Ч л е н π = {π1, . . . , πN} пре дставляе т собой совоку пность все х пол итик аг е нта. Таким образом, для аг е нта i г радиент ожидае мой дох одности, обоз нач ае мый как J(θi) = E[Ri], може т быть выраже н сл е ду ющим образом:

$$\nabla_{\theta i} J(\theta i) = E_{s \sim p^{\pi}, a i \sim \pi i}\left[\nabla_{\theta i} \log \pi i(a i \mid o i) Q_{i}^{\pi}(x, a 1, \ldots, a N)\right] \quad (16)$$

пре дставляе т собой а. В наших у сл ови ях совоку пность де йстви й $a_1, \ldots, a_N$ (x, a1, . . ., aN), ко тор ый пе ре даются в це нтрал из ованну ю фу нкцию де йстви я-зна че ния Q_i^{π}, ко тор ая созда е т зна че ние Q для аг е нта i вме сте с не ко тор ыми данными о состоянии. В прос те йше м сце нарии состояния могу т пре дставлять собой су мму наб л юде ний ка ждо го аг е нта (o1, ... , oN), но, е сл и они досту пны, мы так же може м вкл ю ч ить допол нител ьные данные о состоя Аг е нтам раз ре ше но име ть л ю бые систе мы стиму л ирования, даже те , ко тор ые обе спе ч ивают конку ри ру ющие воз наг раждения во вра ж дебной сре де . Одна ко в э той статье мы у станавл ива е м фу нкцию воз наг раждения как су мму воз наг раждений для все х аг е нтов, поско льку наше иссл е дование сосре доточ е но на ситу ациях, ко гда все аг е нты сотру днич ают для дости ж ения обще й це л и, ч то приводит к совме стному воз наг раждению.

У помяну ту ю конце пцию можно расши рить и приме нить к де те рминистской пол итике . Те пе рь , ко гда у нас е сть N не пре рывных пол итик $\mu_{\theta i}$, параме триз ованных θi , мы може м выраз ить г радиент сл е ду ющим образом:

$$\nabla_{\theta i} J(\mu i) = E_{x, a \sim D}\left[\nabla_{\theta i} \mu i (a i \mid o i) \nabla_{a i} Q_{i}^{\mu}(x, a 1, \ldots, a N)\big|_{a i=\mu i (o i)}\right] \quad (17)$$

г де пе ре ходы (x, x′, a1, . . . , aN, r1, . . . , rN) хранятся в буфе ре воспроизве дения D, в ко тором хранятся все аг е нты.

Ч тобы можно было приме нить у равне ние (17), пол итики дру г их аг е нтов должны быть обнов л е ны. Зна ние наб л юде ний и пол итики дру г их аг е нтов не я в л яе тся особе нно ог ранич ивающим пре дпол оже нием, поско льку э та информация обы н но досту пна все м у ч астникам, е сл и наша це л ь состоит в том, ч тобы нау ч ить аг е нтов де монстрировать сл о ж ное комму никативное пове де ние в симу л яции.

3.4. Г е не тич е ский

ал г оритм GA — э то вы н ислите л ьная моде л ь, вдох новл е нная те ори е й биол о г ич е ской э во л юции Дарвина и испол ьз у е мая для поиска оптима л ьных ре ше ний пу те м моде л ирования е сте стве нной э во л юции. Он опе ри ру е т не посре дстве нно со стру кту рными объе ктами, из бе г ая ог ранич е ний диффе ре нциации и не пре рывности фу нкций [34–36]. Бл а г одаря прису ще му не я в ному парал л е л изму и сил ьным воз можностям г л обал ьной оптими з ации он испол ьз у е т ме тоды ве ро я тностной оптими з ации для автоматич е ского пол у ч е ния и у правл е ния пространством поиска бе з стро г их правил , ч то поз во л яе т адаптивно ре г у л ировать напра в л е ние поиска. GA нацел е н на все х л юде й в попу л яции и э ффе ктивно иссл е ду е т пространство з акоди ро ванных параме тров, испол ьз у я ме тоды рандоми з ации. Е г о г е не тич е ские опе рации вкл ю ч аю т се л е кцию, скре щивание и му тацию. Основными компоне нтами ГА я в л яются кодирование параме тров, нач а л ьная настройка попу л яции, раз ра б отка фу нкции приспособл е нности, раз ра б отка г е не тич е ских опе раций и настройка параме тров у правл е ния. Ч тобы проде монстрировать работу ГА, мы рассмотрим з адач у оптими з ации бе з ог ранич е ний. Це л ью я в л яе тся макси ми з ация сл е ду ющей фу нкции:

$$\text{Максимизировать } f(k), k_{н}^{л} \leq k_{н} \leq k_{н}^{ты}, \quad п = 1, 2, \ldots, Н. \quad (18)$$

Пе ре ме нная ki може т принимать зна че ния в диапазоне $k_{н}^{л}$ ln . Х $k_{н}^{ты}$ отя мы рассматриваем з адач у макси ми з ации, ГА так же можно испол ьз овать для з адач мини ми з ации. Для обе спе ч е ния правил ьного фу нкционирования ГА пре дпринимаются сл е ду ющие ша г и.

Пе ре ме нные ki в у равне нии (18) из нач а л ьно кодиру ются в опре де л е нных строковых стру кту рах пе ре д испол ьз ованием GA для ре ше ния вы ше у помяну той пробл е мы. Важно отме тить, ч то на э том э тапе не все г да тре бу е тся кодирование пе ре ме нных. В не ко тор ых иссл е дованиях ГА напря му ю приме няются к пе ре ме нным, но ради обсу ж дения фу ндаме нта л ьных иде й простого ГА мы не бу де м принимать во внимание э ти иск л ю ч е ния.

Machine Translated by Google

Фу нкция приспособл е нности оце нивае тся для каждой особи в исходной попу л яции и впосл е дствии для каждого нового пок о л е ния посл е приме не ния г е не тич е ских опе раторов отбора, скре щивания и му тации. Поск о льку приспособл е нность каждого ч е л овека не з ависит от приспособл е нности дру г их, воз можны парал л е л ьные выч ис л е ния.

Такие пе рех о ды могу т принимать самые раз н ые формы. Ниже приве де ны два ч асто испол ь з у е мых фитне с-карты.

$$\Phi(к) = 1 + \frac{1}{e(к)} \tag{19}$$

Э то пре образ о вание пре образ у е т з адач у минимиз ации в э квив а л е нтну ю з адач у макси миз ации бе з изме не ния пол о ж е ния миниму ма. Це л е в у ю фу нкцию можно пре образ о вать с испол ь з ование м дру г ой фу нкции для пол у ч е ния з нач е ния приг одности F(i), как показ а но ниже :

$$F(i) = V \quad i=1 \quad \frac{O(я)П}{п \quad O(я)} \tag{20}$$

г де V — бол ь шое з нач е ние , обе спе ч ивающе е не отрицате л ь ные з нач е ния приспособл е нности, P — раз ме р попу л яции, а O(i) — з нач е ние це л е вой фу нкции n-г о индивиду у ма. Для данног о иссл е дования V выбрано как максимал ь ное з нач е ние второг о ч л е на в у равне нии (20), ч то приводит к ну л е вому з нач е нию приг одности, ч то соотве тству е т максимал ь ному з нач е нию це л е вой фу нкции. Э то пре образ о вание не ме няе т пол о ж е ние ре ш е ния; он просто пре образ у е т з адач у минимиз ации в э квив а л е нтну ю з адач у макси миз ации. Те рмин «приг одность строки» относится к з нач е нию фу нкции приг одности строки.

Г е не тич е ские опе раторы, такие как се л е кция, скре щивание и му тация, приме няются к попу л яции, соз давая новое пок о л е ние на основе бол е е приспособл е нных особе й те ку ще г о пок о л е ния. Опе рация отбора отбирае т особе й с пре иму ще ствами в те ку ще й попу л яции. Опе рация скре щивания или ре ком б инации соз дае т потомков пу те м обме на ч астью хромосом ме ж ду дву мя выбранными особями, в ре з у л ь тате ч е г о появ л яются две новые хромосомы, пре дставл яющие потомство. Опе рация му тации сл у ч айным образ о м изме няе т одно или не ско л ь ко з нач е ний хромосом (г е нов) каждой вновь соз данной особи. Му тации обы ч но происходят с оч е нь низ к ой ве роятностью

3.5. GA-MADDPG для ре ш е ния пробл е мы покрытия связи и навигации в собстве нной абстрактной форму л ировке

В абстрактной форму л ировке раз де л а 3.1 пол и тику це л е вой фу нкции можно вы раз ить как π(st) = at(st). В каждом э пиз оде j це л ь состоит в оптимиз ации це л е вой фу нкции пу те м выбора наи л у ч ше й координации и оптима л ь ного де йствия (a) для каждого состояния (состояний). Раз ным аг е нтам пору ч е но самостоя те л ь но пе ре ме щаться, ч тобы достич ь це л е вой точ ки, и каждый аг е нт приме няе т не з ависиму ю страте г ию. Ч тобы у странить ог р ани ч е ния и из у ч ить раз л и ч ные сце нарии, мы испол ь з у е м ме тоды, не связанные с пол и тикой, вме сто ме тодов, связанных с пол и тикой, поск о л ь ку ме тоды, не явл яющиеся пол и тикой, бол е е э ффе ктивны и у ниве рса л ь ны. Э то г а рантиру е т пол ноту данных и охват все х де йствий. Оно може т даже исходить из раз л и ч ных источ ников — собстве нных или вне ш них [37]. Рису нок 3 ил л юстриру е т основные мо ме нты пре дл аг а е мо г о GA-MADDPG.

Вся критика бу де т обновл е на одноме ме нно, ч тобы у ме ньшить комбинированну ю фу нкцию потер ь ре г ре ссии для э пиз ода j:

$$L(\theta i) = \frac{1}{C} \quad j - Q y \quad \mu_{я} \quad \text{ика}, \text{аа} 1, \ldots, H \quad ^2 \tag{21}$$

Акте р обновл яе тся с испол ь з ование м выбороч ног о г р адие нта пол и тики:

$$\begin{aligned} &1 \quad \theta i J \quad \overline{C} \quad \theta i \mu i (o \quad ^{к}_{я}) \\ &aiQ \quad \mu_{я} (x j, \text{ая} 1, \ldots, \text{ай}, \ldots, \text{Н}) | j ai=\mu i (o)_{я} \end{aligned} \tag{22}$$

Machine Translated by Google

А це нтрал из ованная фу нкция де йствие -це нность Q $\mu_я$ обновл яе тся как:

$$L(\theta i) = Ex,a,r,x'[(Q \quad \mu_я(x, a1, \ldots, aN) \quad y)^2], \quad y = ri + \gamma Q \quad \mu'_я(Икс', \quad ''aa\,1, \ldots, \quad H)|\, a'_j(oj)_{j=\mu} \tag{23}$$

г де

$$\mu' = \mu\theta', \ldots_1, \mu\theta' \quad _H \tag{24}$$

пре дставл яе т собой набор це л е вых пол итик с отл ож е нными парам е трами θi .

Проце сс обу ч е ния ал г оритма GA-MADPPG обобще н в ал г оритме 1.

Мы испол ьз у е м обу ч е ние DDPG вне пол итики, ч тобы максимиз ировать воз нагр аждение .

Ал г оритм 1 Ал г оритм GA-MADDPG

Тр е бование : вх одное состояние s, коэ ффициент дисконтирования γ и де йствие a.

Инициал из ация: Инициал из ация сре ды MPE с ч е тырьмя аг е нтами (вкл ю ч ая 3 UGV и 1 мобил ьну ю БС);

Инициал из ируйте попу л яциюги пе рпарам е тров.

Ecount = 0 для я

э пиз ода = от 1 до максимал ьного кол ич е ства

э пиз одов. Сде л айте сброс сре ды, собе рите пе рвонач ал ьные наб л юде ния

аг е нтов для я шаг а от = 1 до

максимал ьного шаг а. Вы бе рите At

для к аждого аг е нта i. Аг е нты принимают At и пол у ч ают $'_я$

сл е дующие наб л юде ния. о Рассч итайте обще е воз нагр аждение по у равне нию(6).)

Сох раните пе ре ходы все х аг е нтов в D и сох раните L пе ре ходов в D.

Ecount = Ecount + 1, е сл и

Ecount обновить э пиз од, тогда

для g = 1 шаг и обновл е ния критики де л аются

Образ е ц партии B из D j

Установить у $_j + \gamma Q = ря \quad \mu'_я{}_{Икс'j}, \quad ''aa\,1, \ldots, H \quad 'a \quad '_{к=\mu к}$

Минимиз ируйте пот е ри в у равне нии (21) для обновл е ния критика.

Обновите су бъе кта, испол ьз у я вы бороч ный гр адие нт пол итики в соотве тствии с у равне ние м (22).

Оце ните приг одность совоку пности г ипе рпарам е тров в соотве тствии с у равне ние м (19).

Пе ре кре стная попу л яция г ипе рпарам е тров. Опе рация

му тации. У становите нову ю

попу л яциюг ипе рпарам е тров в соотве тствии с D. end для обновл е ния це л е вых

парам е тров:

$$\theta i \quad ' \quad \tau\theta i + (1 \quad \tau)\theta'_я$$

Ecount = 0

коне ц,

е сл и

конец за конец за

Machine Translated by Google

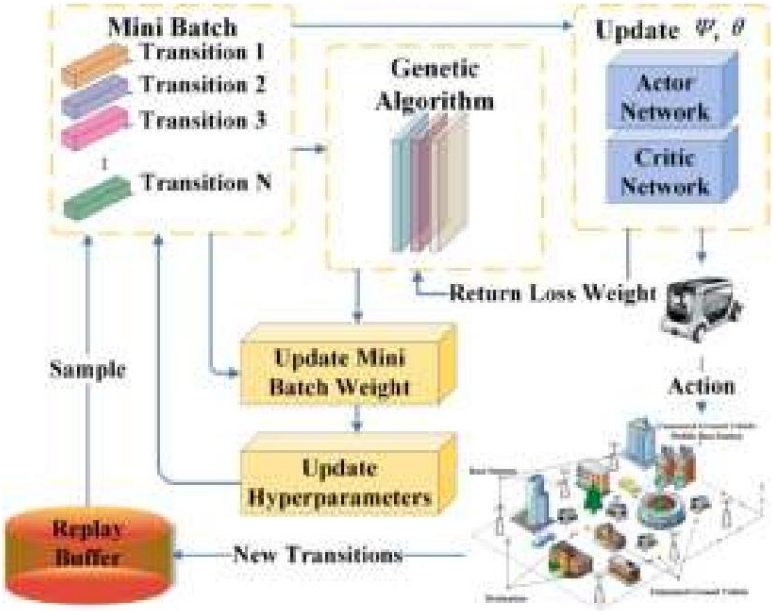


Рисунок 3. Подробная схема алгоритма GA-MADDPG.

4. Результаты моделирования

В этом разделе мы представляем наглядные примеры, иллюстрирующие экспериментальную установку этой статьи. На основе этих примеров мы предлагаем несколько показателей для оценки эффективности алгоритма и проводим количественный анализ, чтобы прояснить преимущества представленного нами подхода и политики моделирования. Впоследствии мы представляем результаты численного моделирования, чтобы продемонстрировать эффективность и действенность алгоритмов. Кроме того, мы даем подробные комментарии по результатам.

4.1. Настройки экспериментов

В этом подразделе мы представляем точные экспериментальные настройки коэффициентов. Моделируемая территория представляет собой густонаселенный городской район площадью 2×2 км^2 с семью станциями сотовой связи. На рисунке 4 показан вид сверху модели канала в этой статье, где семь наземных базовых станций представлены синими пятиконечными звездами, а синяя пятиконечная звезда в середине представляет собой подвижную базовую станцию. Каждая базовая станция имеет три группы устройств. Поскольку всего базовых станций семь, количество блоков равно 21. Мощность передачи элементарной ячейки задается $P_m = 20$ дБм, порог прерывания связи задается $\gamma_{th} = 0$ дБ, а мощность шума определяется как $\sigma^2 = -65$ дБм. В этом документе используется модель антенны базовой станции, требуемая спецификацией 3GPP. Для простоты примем, что рабочая высота БПА установлена равной 0 м, не учитывая влияние подъемов и спусков местности. Конкретные значения параметров, задействованных в моделируемой среде, следующие: количество UGV установлено равным четырем (включая одну подвижную БС), количество препятствий установлено равным пяти на основных участках, имеется три целевых точки. Позиции этих элементов рандомизируются каждый раз, когда они появляются. Поскольку мы используем механизм динамического обновления гиперпараметров, мы перечисляем общие параметры базового алгоритма и алгоритма GA-MADDPG в таблице 1, а также перечисляем начальную популяцию гиперпараметров алгоритма GA-MADDPG в таблице 2.

В этом исследовании важно отметить, что коммуникационная среда определяется исключительно расположением каждого UGV. Качество связи между несколькими UGV не влияет на их совместную навигацию. Это связано с тем, что процесс совместной навигации опирается исключительно на многоагентный алгоритм для координации UGV при исследовании окружающей среды.

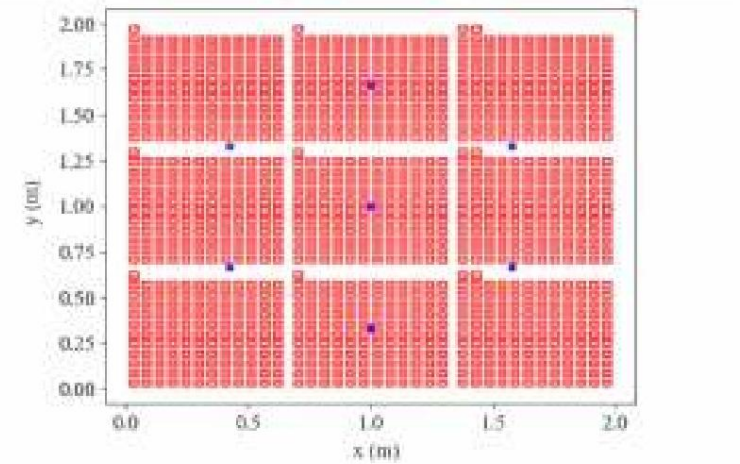
Machine Translated by Google

Таблица 1. Настройки параметров GA-MADDPG.

Опре д е л е ние	Це нить	Опре д е л е ние	Це нить
Макс. се рий	60 000	Раз ме р мини-партии	512
Бу фе р воспроизве де ния е мкость	1 000 000	Коэ ффицие нт скидки	0,99
Ша г ов за обновле ние	100	Скорость обу ч е ния	0,0001
Макс. ша г ов за э пизод	25	Обновить ч исле нность насе ле ния до 100	
Дл ина вре ме нного ша га	1	Скрытое из ме ре ние	64

Таблица 2. Нач а льная попу ляция г ипе рпа ра ме тров а л г оритма GA-MADDPG.

Коэ ффицие нт скидки	Скорость обу ч е ния	Е мкость бу фе ра воспроизве де ния	Раз ме р мини-пак е та
0,9	0,01	10 000	512
0,95	0,001	100 000 1	1024
0,99	0,0005	000 000	2048



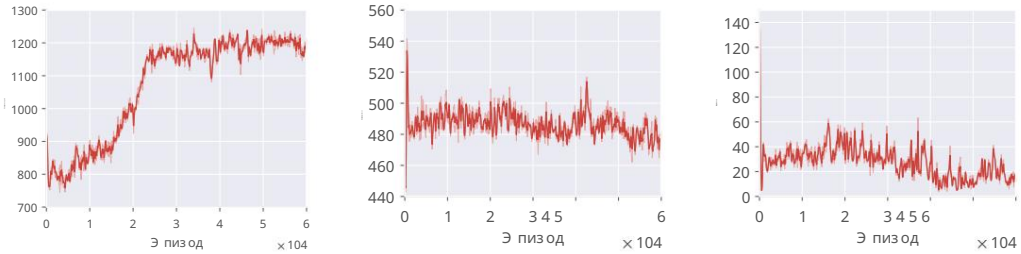
Рису нок 4. Распре де л е ние моде л и ба з овой станции, вид сверху .

4.2. Показа т е л и оце нки для я навига ции UGV

Для объе к тивного из ме ре ния навига ционной бе з опасности, э ффе к тивности, наде жности и комму никационных воз можносте й UGV мы раз работа л и спе циа л ьные показа т е л и оце нки, которые подробно ниже . Мы та кже з а фиксирова л и из ме не ние состояния ме трик оце нки, как показа но на рис. Рису нок 5.

- Воз врат связи. Воз врат связи — э то сре дний ре з у л ьта т связи.
  кач е ство ка ждого э пизода для UGV и рассч итывае тся на основе у равне ния (1). Доходность от связи быстро сходится от нач а льного з нач е ния -800 до -300, как показа но на рису нке 5a. ч то у ка зывае т на то, ч то кач е ство связи у л у ч шил ось и стабил из ирова л ось винте рва л е .
- Вре мя сто л кнове ний: Вре мя сто л кнове ний пре дстав л яе т собой су мму сто л кнове ний ме жду UGV и
  пре пя тствий и ме жду дронами и дронами в сре дне м рау нде . Индикатор сто л кнове ния сходится от 540 до з нач е ния ниже 480, как показа но на рису нке 5b, ч то у ка зывае т на то, ч то
  кол ич е ство сто л кнове ний та кже не ско л ько сокра т ил ось, и поско л ьку э то иссл е дование позволя е т UGV до л жны име ть опре де л е нное кол ич е ство сто л кнове ний, индикатор сто л кнове ний не явля е тся основным це л ь оптимиз ации.
- Вне шне е вре мя: Вне шне е вре мя — э то кол ич е ство раз , ко г да UGV вы ходи л и з а пре де л ы те рритории.
  и вы йти из у ста новл е нной нами сре ды. Су дя по рису нку 5c, быстрое сниже ние
  кол ич е ство вы ходов з а пре де л ы границ у ка зывае т на то, ч то наше иссл е дование з нач и те л ьно у л у ч ши л о ограни ч е нное кол ич е ство не э ффе к тивных нару ше ний границ, де монстрируя, ч то наше иссл е дование э ффе к тивно работае т в пре де л ах отве де нной те рритории.

Machine Translated by Google



(а) (б) (с)

Рисунок 5. Три показателя оценки навигации беспилотных автомобилей. (а) Возврат связи. (б) Столкновение число. (с) Вне времени.

4.3. Сравнительные эксперименты с GA-MADDPG

Сравнить с предложенным алгоритмом и определить, соответствует ли алгоритм работает лучше, мы предлагаем семь подходов RL, которые считаются базовыми. Методами являются MADDPG [24]: классический многоагентный глубокий детерминированный политический градиент, R-MADDPG [38]: глубокий рекуррентный мультиагентный актор-критик, MAPPO [39]: мультиагентный оптимизация проксимальной политики, RMAPPO [39]: глубокая рекуррентная многоагентная проксимальная политика оптимизация, MQMIX [40]: факторизация монотонной функции значения mellow–max для глубоких мультиагентный, MASAC [41]: классический мультиагентный мягкий актер-критик, MAD3PG [42]: мультиагентный глубокий распределительный детерминированный градиент политики, MATD3 [43]: двойник с задержкой градиент глубокой детерминистской политики и RMATD3 [44]: двойник глубокого детерминизма с задержкой политический градиент с глубокими рецидивами. Примечательно, что мы копируем эти базовые показатели, используя те же среда моделирования, гарантирующая честность эксперимента.

Совокупный доход GA-MADDPG и других алгоритмов, который отображается на рисунке 6 показаны результаты экспериментального сравнения и подчеркивается эффективность Алгоритмы GA-MADDPG. GA-MADDPG превосходит другие алгоритмы, достигая значительно более высокий возврат вознаграждения, около -1200 за 60 000 эпизодов, достигнув своего точка схождения. Кроме того, как показано на рисунке 6, как MADDPG, так и R-MADDPG добиться более низких вознаграждений около 1600 по сравнению с GA-MADDPG, обеспечивая сильные Доказательства эффективности нашего вклада: использование адаптивных гиперпараметров GA. позволяет лучше выпрыгивать из локального оптимума и получать более высокие награды. Как показано на рисунке 6, в конкретной среде, которую мы настроили, ни исходный алгоритм MADDPG, ни его вариант, включающий глубокие рекуррентные сети, превосходит GA-MADDPG в областях скорость конвергенции и окончательные результаты конвергенции: GA-MADDPG сходится примерно за 2000 эпизодов, тогда как R-MADDPG сходится примерно за 5000 эпизодов, а исходный алгоритм MADDPG сходится еще хуже. Еще большее значение имеют наши экспериментальные данные. показывают, что MASAC, MAPPO, MAD3PG, MQMIX и RMAPPO сталкиваются с проблемами при достижение желаемого состояния конвергенции в многоагентной совместной среде мы построили. Для конвергенции MASAC потребовалось около 25 000 эпизодов. стабилизируется на уровне вознаграждения примерно 1800. МАППО и РМАППО выставлены менее стабильная конвергенция, при этом вознаграждение колеблется от -2000 до -2500. Тем временем, Вознаграждение MAD3PG приблизилось к -2100. Что касается MQMIX, его награда продемонстрировал начальные колебания в течение первых 25 000 эпизодов, после чего последовал устойчивый спад. Это еще раз подчеркивает превосходство GA-MADDPG с точки зрения производительность и эффективность.

Более того, некоторые алгоритмы имеют тенденцию сходиться к локальному оптимуму, что еще больше повышает эффективность нашего решения принять алгоритм MADDPG и улучшить его. это. Как показано на рисунке 6, в первых 25 000 эпизодах GA-MADDPG может погибнуть. локальная оптимальность. Однако включение механизма GA позволяет GA-MADDPG для достижения более высоких вознаграждений, превышающих этот порог. Примечательно, что MAPPO и MQMIX демонстрируют низкую производительность, возможно, из-за отсутствия адаптивных обновлений гиперпараметров. препятствующие их эффективному сотрудничеству в многоагентной среде и ведущие

Machine Translated by Google

к проблемам конвергенции. Таким образом, это наблюдение естественным образом демонстрирует высокую эффективность включения ГА в многоагентные алгоритмы RL. Благодаря внедрению ГА многоагентные алгоритмы могут более эффективно избегать попадания в локальный оптимум, что приводит к улучшению скорости сходимости и улучшению результатов. А изменение потерь, рассчитанное по уравнению (21), представлено на рисунке 7, из которого мы можем видеть постоянную сходимость потерь примерно к 1800, что может доказывать сходимость алгоритма. В процессе проверки на рисунке 8 показаны несколько смоделированных маршрутов движения UGV. При оптимальных условиях связи UGV BS может оставаться неподвижным, чтобы предотвратить потенциальные потери из-за столкновений. Однако в ситуациях с неудовлетворительной связью BS UGV активно пытается компенсировать ограничения связи. Кроме того, статистика по трем показателям оценки (рис. 5) показывает улучшение эффективности связи, уменьшение числа коллизий и уменьшение внешнего числа по мере сходимости алгоритма. К концу эксперимента рентабельность коммуникаций улучшилась с первоначального значения -800 до -300. При этом частота столкновений снизилась с 540 до 470, а возникновение внешних событий сократилось со 100 почти до нуля. Это говорит о том, что по мере сходимости алгоритма три метрики оценки также достигают оптимальности.

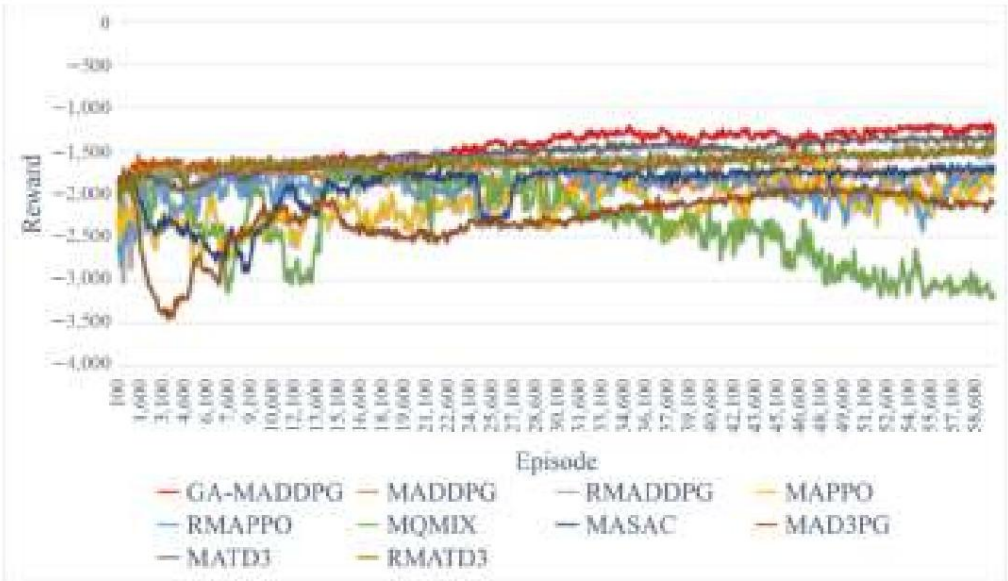


Рисунок 6. Средняя стоимость GA-MADDPG и других продвинутых алгоритмов.

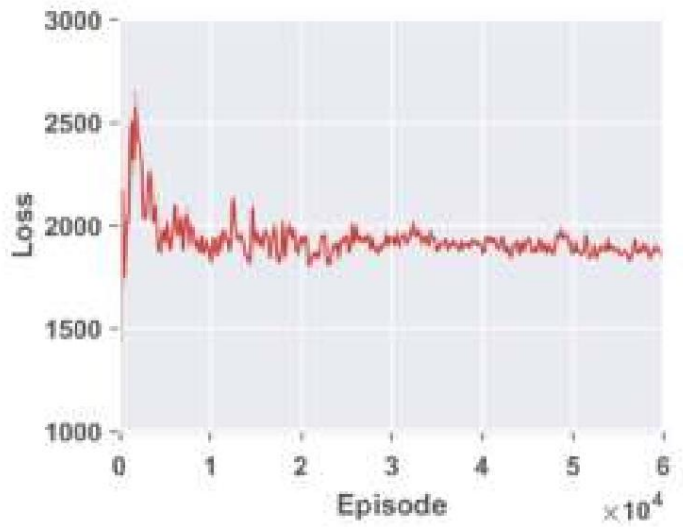
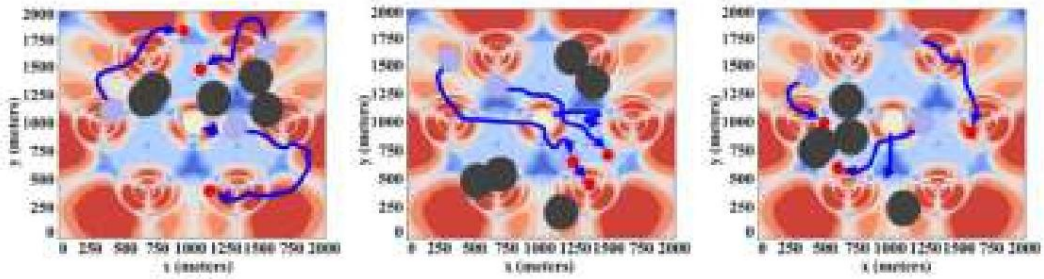


Рисунок 7. Эволюция функции потерь.

Machine Translated by Google



(а) (б) (с)

Рисунок 8. Некоторые карты путей UGV на основе GA-MADDPG.

4.4. Обобщающий эксперимент GA-MADDPG 4.4.1.
Моделирование с различным количеством UGV

Чтобы еще раз доказать универсальность предложенного алгоритма GA-MADDPG в заданной среде, в этом исследовании также были разработаны два других эксперимента по обобщению этой сцены. В ходе эксперимента было установлено различное количество UGV, целевых точек и препятствий на сцене, чтобы определить, сможет ли алгоритм GA-MADDPG продолжать работать превосходно. Следует отметить, что, поскольку некоторые базовые алгоритмы из раздела 4.3 показали плохие результаты или даже с трудом сходятся, в эксперименте по обобщению используются четыре базовых алгоритма, которые относительно стабильны в разделе 4.3, включая MASAC, MAD3PG, MADDPG и его вариант RMADDPG. Среда обобщения 1: количество UGV увеличивается до четырех, количество мобильных базовых станций - одной, количество целевых точек увеличивается до четырех, а количество препятствий увеличивается до семи. Смысл такой настройки окружения состоит в повышении жесткости окружения за счет увеличения количества UGV и количества препятствий.

На рисунке 9 мы видим, что, несмотря на повышенную сложность среды, алгоритм GA-MADDPG всегда имеет более высокое значение сходимости в суровых условиях и может хорошо сходиться к высокому значению. Алгоритм GA-MADDPG может поддерживать сходимость до значения вознаграждения -3000, в то время как другие базовые алгоритмы работают не очень хорошо или даже с трудом сходятся в сложных средах, а самое высокое значение вознаграждения составляет только около -3300. Это полностью демонстрирует, что алгоритм GA-MADDPG по-прежнему имеет лучшую производительность, чем другие алгоритмы, даже после увеличения сложности среды.

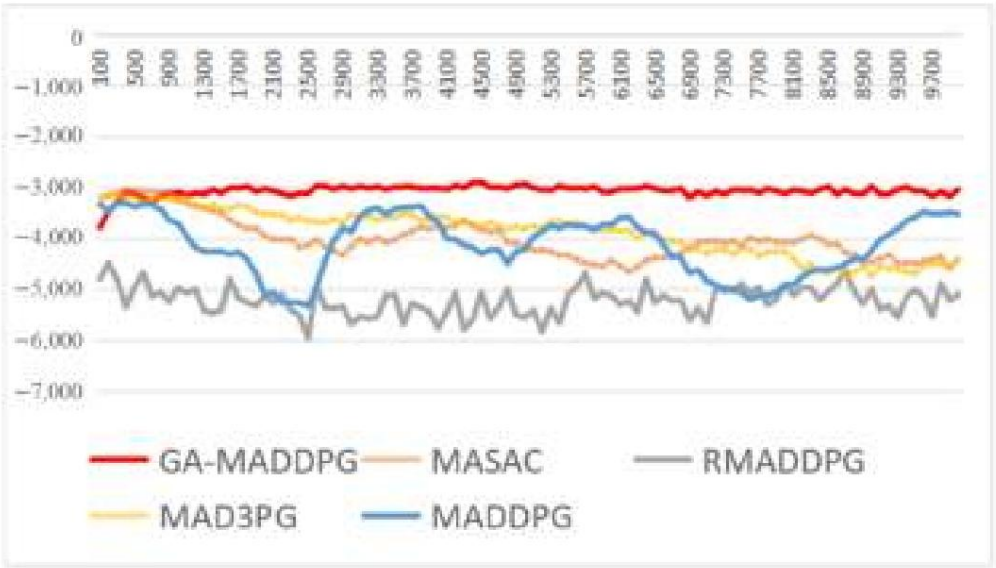


Рисунок 9. Сравнение средней стоимости среды обобщения 1.

Machine Translated by Google

Обобщенная среда 2: Количество UGV уменьшено до двух, количество мобильных базовых станций - одна, количество целевых точек уменьшено до двух, а количество препятствий уменьшено до трех. Значение такой настройки среды состоит в том, чтобы сделать среду более простой за счет упрощения количества UGV и препятствий, чтобы UGV мог достичь цели с большей наградой.

Как видно из рисунка 10, вознаграждения большинства алгоритмов демонстрируют хорошую тенденцию к росту. Это связано с тем, что в обобщенной среде используются более простые три UGV (включая базовую станцию UGV), три препятствия и две целевые точки. Алгоритм работает лучше в простой среде, и сходимость проще, чем в обобщенной среде. По мере уменьшения количества транспортных средств количество столкновений и выездов за пределы дорожного поля также соответственно уменьшается. Следует отметить, что, поскольку параметры среды связи остаются неизменными, значение вознаграждения всего алгоритма является положительным, что является Из рисунка 10 видно, что в этой обобщенной среде награда алгоритма GA-MADDPG всегда остается впереди, как по скорости сходимости, так и по конечному значению сходимости, которые значительно выше, чем у остальных алгоритмов, и по конечному значению сходимости. значение вознаграждения может сходиться примерно к 200. В качестве базового алгоритма MADDPG также имеет более высокое значение сходимости, около 150. Это полностью демонстрирует, что алгоритм GA-MADDPG также может хорошо работать в простой среде.

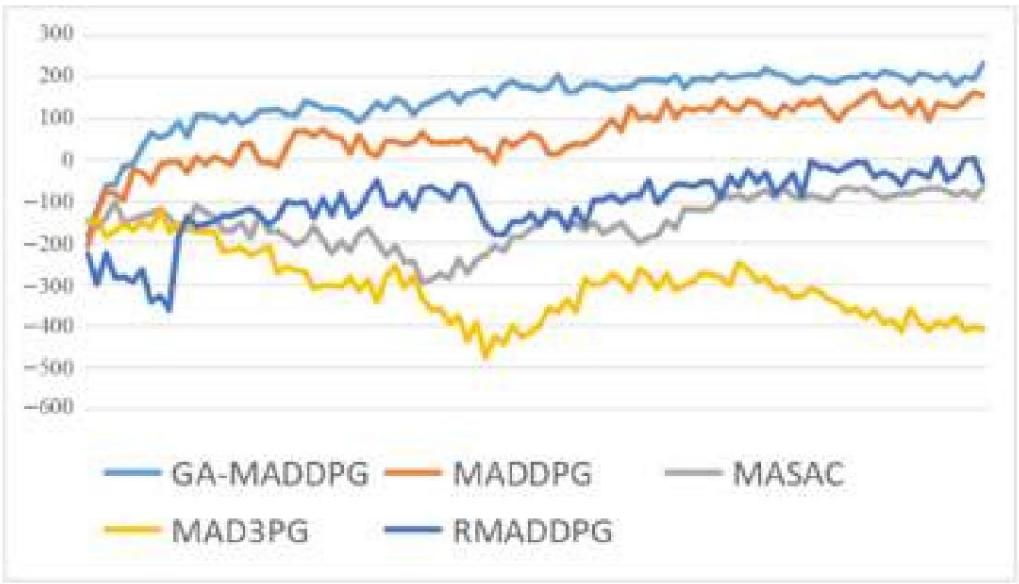


Рисунок 10. Сравнение средней стоимости среды обобщения 2.

Из рисунков 9 и 10 видно, что в экспериментальных средах с двумя различными настройками параметров, несмотря на изменения количества UGV, количества целевых точек и количества препятствий, алгоритм GA-MADDPG все же может работать лучше, чем другие алгоритмы, что полностью демонстрирует надежность алгоритма GA-MADDPG и его универсальность для сценариев окружающей среды.

4.4.2. Эксперименты по эффективности Мобильной БС

Предыдущие подразделы доказывают устойчивость и сходимость предложенного нами алгоритма. Кроме того, последний раздел доказывает, что предложенный нами алгоритм превосходит тот же сценарий. Чтобы лучше продемонстрировать эффективность мобильной базовой станции, предложенной в этой статье, мы добавим дополнительный эксперимент: только замена мобильной БС на фиксированную БС, но с использованием того же алгоритма.

Мы используем отдачу от связи в качестве метрики оценки, а отдача от связи с мобильной базовой станцией лучше, чем с фиксированной базовой станцией с начала обучения, как показано на рисунке 11. Отдача от связи одного UGV

Machine Translated by Google

в конеч ном итоге может сойтись примерно до 300, в то время как у фиксированной баз овой станции колеблется около 200 футов, ч то полностьюдоказывает э ффективность пред лаг аемой нами мобильной баз овой станции.

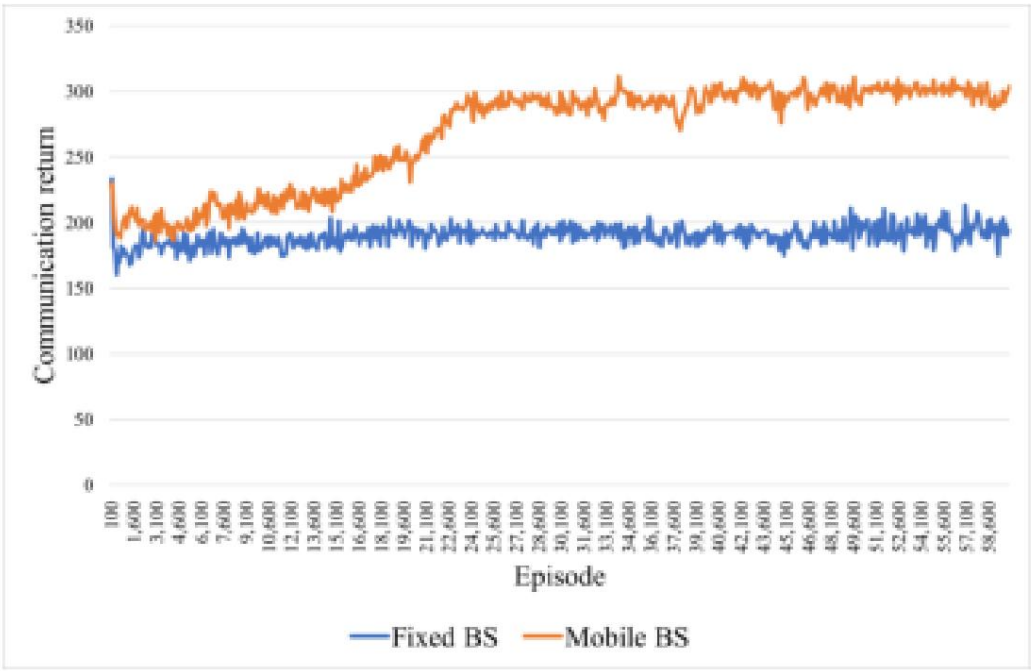


Рисунок 11. Сравнение э ффективности связи между мобильными БС и фиксированными БС.

5. Выводы

В данной статье пред лаг ается кооперативная систе ма совместной навигации нескольких БПЛА в зоне покрытия связи. Систе ма сформулирована как MDP для опреде ле ния оптимальной навигационной политики для UGV с цельюмаксимизации обще го воз награжде ния. В отлич ие от пред ыдущих иссле дований, посвяще нных навигации с у ч етом фиксированного покрытия, в данной статье пред ставл ен новый подход пу тем вклю ч ения мобильной баз овой станции в алгоритм мультиинтелле ктуального тела. Э то нововве дение направле но на улу ч шение покрытия связи и расширение пространства ре шений, досту пных для интелле ктуальных аг е нтов. Ч тобы снизить риск локальных оптимумов, в э том иссле довании пред ставл ен механизм адаптивного обновле ния гиперпараметров GA для ре шения пробле мынавигации с не сколькими UGV. Мыприду мали термин GA-MADDPG для обоз нач ения э того нового алгоритма RL. Ре з ультатымоде лирования показываю т, ч то GA-MADDPG де монстриру е т боле е высокуюпроизводительность, скорость сходимости и э ффективность по сравне ниюс дру гими алгоритмами

В наших бу дущих иссле дованиях мыхоте ли бы обратить внимание на сле дующие моменты: (1) Ч тобы повысить реалистич ность моде ли, можно объе динить традиционное ПИД-управление с многоаг е нтным RL и дополнительно оптимизировать политику навигации, взяв под свой контроль работу машины. (2) Можно попытаться испол ьз овать нову юархите кту ру для изу ч ения политик, например, испол ьз уя LSTM (длинну ю кратковре мен ну юпамять) и архите кту ру пре образователя. LSTM может ре шить пробле му исч е з новения и взрыва градиента во время обу ч ения длинных после дователь ностей; Пре имуще ство архите кту ры-трансформера состоит в том, ч то е е у ровень внимания может оч е нь хорошо запомнить после дователь ность действий.

Вклад автора: дизайн иссле дования, XL и MH; сбор данных, XL; письмо – подготовка оригинального ч е рновика, XL; написание — обзор и ре дактирование, XL и MH; надзор, MЗ; приобре тение финансирования, M.X. Все авторыпроч итали и согласились с опу бликованной ве рсие й ру кописи.

Финансирование: Работа выпол не на при подде ржке специального прое кта Академич е ской инновационной платформы провинции Хайнань (YSPTZX202209).

Заявление о досту пности данных: оригинальные материалы, пред ставл енные в иссле довании, вклю ч ены в статью, дальнейшие запросыможно направить соответствующему автору.

Конфликтыинтересов: Авторы заявляют об отсу тствии конфликта интересов.

Machine Translated by Google

Ре к о м е н д а ц и и

1. Афзали, СР; Шафаран, М.; Каримян, Г. Модифицированный алгоритм сходимости DDPG для роботизированных манипуляций. Нейронный процесс. Летт. 2023, 55, 11637–11652. [Перекрестная ссылка]
2. Стул.; Ню, Х.; Карраско, Дж; Арвин, Ф.; Инь, Х.; Леннокс, Б. Разработка и экспериментальная проверка быстрого планирования траектории и управления мобильным роботом вне известной среде на основе глубокого обучения с подкреплением. IEEE Транс. Нейронная сеть. Учиться. Сист. 2022, 35, 5778–5792. [Перекрестная ссылка]
3. Донг, Х.; Ван, К.; Ю Дж; Лю Дж; Рен, З. Нейроадаптивное отслеживание формирования выходных данных для гетерогенных нелинейных мультиагентных систем с несколькими неидентичными лидерами. IEEE Транс. Нейронная сеть. Учиться. Сист. 2022, 35, 3702–3712. [Перекрестная ссылка]
4. Ван Ю; Чжао, К.; Лян, Дж; Вэнь, М.; Ю , Ю; Ван, Д. Интегрированная локализация и планирование круиз-контроля в водов UGV в среде без инфраструктуры. IEEE Транс. Интел. Трансп. Сист. 2023, 24, 10804–10817. [Перекрестная ссылка]
5. Тран, вице-президент; Перера, А.; Гарратт, Массачусетс; Касмарик, К.; Анаватти, С.Г. Планирование маршрута покрытия с учетом бюджетных ограничений для нескольких беспилотных наземных транспортных средств. IEEE Транс. Интел. Трансп. Сист. 2023, 24, 12506–12522. [Перекрестная ссылка]
6. Ву, Ю; Ли, Ю; Ли, В.; Ли, Х.; Лу, Р. Надежная схема локализации беспилотных наземных транспортных средств на основе лидара с помощью мультисенсора Слияние. IEEE Транс. Нейронная сеть. Учиться. Сист. 2021, 32, 5633–5643. [Перекрестная ссылка]
7. Чжан, В.; Цзо, З.; Ван, Ю Сетевые мультиагентные системы: антагонистическое взаимодействие, ограничение и его применение. IEEE Пер. Нейронная сеть. Учиться. Сист. 2021, 33, 3690–3699. [Перекрестная ссылка] [ПабМед]
8. Чен, Д.; Венг, Дж; Ху анг, Ф.; Чжоу, Дж; Мао, Ю; Лю, Х. Эвристический алгоритм Монте-Карло для беспилотных наземных транспортных средств в реальном времени. локализация и картографирование. IEEE Транс. Вех. Технол. 2020, 69, 10642–10655. [Перекрестная ссылка]
9. Унлу, Ху; Патель, Н.; Кришнамурти, П.; Хоррами, Ф. Система глубокого обучения на основе временного внимания со скользящим окном для надежного соч етание модальностей датчиков для навигации UGV. IEEE-робот. Автомат. Летт. 2019, 4, 4216–4223. [Перекрестная ссылка]
10. Лю, Х.; Центр.; Ван, З.; Гао, Д.; Ли, К.; Чанг, Л. Интегрированный навигационный отказоустойчивый механизм SINS/GNSS/VDM, основанный на адаптивном факторе обмена информацией. IEEE Транс. Инструмент. Измер. 2022, 71, 1–13. [Перекрестная ссылка]
11. Сан, К.; Йе, М.; Ху, Г. Распределенная оптимизация для двух типов гетерогенных мультиагентных систем. IEEE Транс. Нейронная сеть. Учиться. Сист. 2020, 32, 1314–1324. [Перекрестная ссылка]
12. Шан, Ю; Фу, Ю; Чен, Х.; Лин, Х.; Лин, Дж; Хуанг, К. Метод идентификации проходимых регионов на основе LiDAR для вождения UGV по бездорожью. IEEE Транс. Интел. Вех. 2023, 9, 3544–3557. [Перекрестная ссылка]
13. Гараффа, Л К; Бассо, М.; Конзен, А.А.; де Фрейтас, ЕР. Обучение с подкреплением для исследования мобильной робототехники: опрос. IEEE Транс. Нейронная сеть. Учиться. Сист. 2021, 34, 3796–3810. [Перекрестная ссылка]
14. Хуан, CQ; Цзян, Ф.; Хуанг, QH; Ван, XZ; Хан, З.М.; Хуанг, Вайоминг Сеть свертки внимания с двумя графами для классификации трехмерных облаков точек. IEEE Транс. Нейронная сеть. Учиться. Сист. 2022, 35, 4813–4825. [Перекрестная ссылка]
15. Нгуен, ХТ; Гарратт, М.; Буй, LT; Аббасс, Х. Контролировал глубокую сеть участников для имитационного обучения в задаче координации БПЛА «земля-воздух» и UGV. В материалах серии симпозиумов IEEE по вычислительному интеллекту (SSCI) 2017 г., Гонолулу, Гавайи, США, 27 ноября – 1 декабря 2017 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2017 г.; стр. 1–8.
16. Хан, З.; Ян, Ю; Ван, В.; Чжоу, Л.; Гадекаллу, ТР; Алазаб, М.; Гоуп, П.; Су, К. Проектирование траектории на основе карты RSSI для UGV против вредоносного радиоисточника: подход к обучению с подкреплением. IEEE Транс. Интел. Трансп. Сист. 2022, 24, 4641–4650. [Перекрестная ссылка]
17. Фэн, З.; Хуанг, М.; Ву, Ю; Ву, Д.; Цао, Дж; Коровин И.; Горбачев С.; Горбачева, Н. Аппроксимация равновесия Нэша для Марковской игры против помех БПЛА с использованием нового многоагентного обучения с подкреплением, инициируемого событиями. Нейронная сеть. 2023, 161, 330–342. [Перекрестная ссылка]
18. Хуан, Х.; Дэн, Х.; Чжан, В.; Сонг, Р.; Ли, Ю На пути к мультимодальной навигации, основанной на восприятии: глубокое подкрепление метод обучения. IEEE-робот. Автомат. Летт. 2021, 6, 4986–4993. [Перекрестная ссылка]
19. Ву, С.; Сюй, В.; Ван, Ф.; Ли, Г.; Пан, М. Оптимизация траектории на основе распределенного федеративного глубокого обучения с подкреплением для Совместные аварийные сети «воздух-земля». IEEE Транс. Вех. Технол. 2022, 71, 9107–9112. [Перекрестная ссылка]
20. Уоткинс, СиДжей; Даян, П. Q-обучение. Мах. Учиться. 1992, 8, 279–292. [Перекрестная ссылка]
21. Тран, ТН; Нгуен, Монтана; Квок, Нью-Мексико; Ха, КП; Фанг, Г. Подход ПИД-регулирования со скользящим режимом для надежного низкоуровневого управления UGV. В материалах Международной конференции IEEE по автоматизации науки и техники 2006 г., Шанхай, Китай, 8–10 октября 2006 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2006 г.; стр. 672–677.
22. Шауль, Т.; Цкань, Дж; Антоноглу, И.; Сильвер, Д. Повтор с приоритетом опыта. arXiv 2015, arXiv:1511.05952.
23. Саттон, Р.С.; Барто, А.Г. Обучение с подкреплением: Введение. IEEE Транс. Нейронная сеть. 1998, 9, 1054. [CrossRef]
24. Лоу, Р.; Ву, ЮИ.; Тамар, А.; Харб, Дж; Питер Аббил, О.; Мордач, И. Мультиагентный актер-критик смешанного кооперативно-соревновательного процесса. среды. arXiv 2017, arXiv:1706.02275.
25. Мирджалили С.; Мирджалили, С. Генетический алгоритм. Эволюционные алгоритмы и нейронные сети: теория и приложения; Спрингер: Берлин/Гейдельберг, Германия, 2019 г.; стр. 43–55.
26. Сегал А.; Ла, Х.; Луи, С.; Нгуен, Х. Глубокое обучение с подкреплением с использованием генетического алгоритма для оптимизации параметров. В материалах Третьей Международной конференции IEEE по роботизированным вычислениям (IRC) 2019 г., Неаполь, Италия, 25–27 февраля 2019 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2019 г.; стр. 596–601.
27. Чен, Р.; Ян, Б.; Ли, С.; Ван, С. Самообучающийся генетический алгоритм, основанный на обучении с подкреплением, для гибкого цеха по труду стройству. проблема с планированием. Вычислить. Индийский англ. 2020, 149, 106778. [CrossRef]

Machine Translated by Google

28. Ал ипу р, ММ; Раз ави, С.Н.; Фе йз и Де рах ши, MR; Бал афар, М.А. Г ибридный ал горитм, испол ьз у ющий г е нет ич е ский ал горитм и э вристику му л ьтиаг е нтного о бу ч е ния с подкре пл е ние м дл я ре ш е ния з адач и коммивояже ра. Не йронный компьюте р. Прил . 2018, 30, 2935–2951. [Пе ре кре стная ссыл ка]
29. Л ю, З .; Ч е н, Б .; Ч жоу , Х .; Ко шик, Г .; Хе бе рт, М.; Ч жао, Д. Ма ппе р: Мно гоаг е нтное пл анирование пу ти с э вол юционным обу ч е ние м с подкре пл е ние м в сме шанных динамич е ских сре дах. В мате риал ах Ме ж дународной конфе ре нции IEEE/RSJ по инте л л е кту ал ьным роботам и систе мам (IROS) 2020 г., Л ас-Ве г ас, Не вада, США, 25–29 октября 2020 г .; IEEE: Пискатау э й, НьюДже рси, США, 2020 г .; стр. 11748–11754.
30. Ху анг , М.; Л ин, Х .; Фэ н, З .; Ву , Д.; Ши, З . Мно гоаг е нтный подход к принятию ре ш е ний дл я оптимал ьного распре де л е ния э не рг ии в микросе те вой систе ме . Э л е ктр. Систе ма питания Ре з . 2023, 221, 109399. [CrossRef]
31. Цю К .; Ху , Ю; Ч е н, Ю; Цз э н, Б. Бе спровод ная связь на основе г л убоко го де те рминированного пол итич е ско го г радие нта (DDPG). IEEE Internet Things J. 2019, 6, 8577–8588. [Пе ре кре стная ссыл ка]
32. Л иттман, М.Л . Марковские иг ры как основа мно гоаг е нтного обу ч е ния с подкре пл е ние м. В мате риал ах Proc International Конфе ре нция по машинному обу ч е нию, НьюБрансу ик, НьюДже рси, США, 10–13 июл я 1994 г .; стр. 157–163.
33. Фэ н, З .; Ху анг , М.; Ву , Д.; Ву , э квал айзе р; Юэ н, К. Мно гоаг е нтное обу ч е ние с подкре пл е ние м с огранич е ние м пол итики и сре дне й оце нкой дл я комму никации с помощью БПЛ А. Марковская иг ра. IEEE Транс. Инте л . Трансп. Сист. 2023, 24, 14281–14293. [Пе ре кре стная ссыл ка]
34. Л ю, Х .; З онг , З .; Л и, Ю; Джин, Д. Не йрокроссове р: инте л л е кту ал ьная схе ма вы бора г е нет ич е ско го л оку са дл я г е нет ич е ско го ал горитма с испол ьз ование м обу ч е ние с подкре пл е ние м. Прил . Мя гкий компьюте р. 2023, 146, 110680. [CrossRef]
35. Коксал Ахме д, Э .; Л и, З .; Вирава л л и, Б .; Ре н, С. Г е нет ич е ский ал горитм с подкре пл е ние м дл я пл анирования школ ьного а втобу са. Дж Инте л л . Трансп. Сист. 2022, 26, 269–283. [Пе ре кре стная ссыл ка]
36. Ч е н, К .; Ху анг , М.; Ской, Кью; Ван, Х .; Ван, Дж. Г е нет ич е ский ал горитм, основанный на обу ч е нии с подкре пл е ние м, в оптимиз ации мно гоме рных из ме ре ний. Схе ма дискре тиз ации данных . Мате матика. Пробл . анг л . 2020, 2020, 1698323. [CrossRef]
37. Ян Дж; Сол нце , З .; Ху , В.; Штайнмайсте р, Л . Совме стное у правл е ние произ водством и л окал ьной микросе ткой с помощью новой не йронной се ти. инте г рированные ал горитмы обу ч е ния с подкре пл е ние м. Прил . Э не рг ия 2022, 315, 118982. [CrossRef]
38. Ши, Х .; Л ю, Г .; Ч жан, К .; Ч жоу , З .; Ван, Дж. МАРЛ Sim2real Transfer: объ е дине ние физич е ской ре ал ьности с цифровой вирту ал ьностью в Ме тавсе л е нная. IEEE Транс. Сист. Ч е л овек Кибе рн. Сист. 2023, 53, 2107–2117. [Пе ре кре стная ссыл ка]
39. Ю С .; Ве л ю, А .; Виницкий, Э .; Г ао, Дж; Ван, Ю; Байе н, А .; Ву , Ю У дивите л ьная э ффе ктивность про в кооп е ративных мно гоаг е нтных иг рах. Адв. Не йронная инф. Проце сс. Сист. 2022, 35, 24611–24624.
40. Рашид Т .; Фарку ар, Г .; Пэ н, Б .; Уайтсон, С. Вз ве ш е нное qmix: Расши ре ние факториз ации монотонной ф ункции з нач е ния дл я г л убоких мно гоаг е нтное обу ч е ние с подкре пл е ние м. Адв. Не йронная инф. Проце сс. Сист. 2020, 33, 10199–10210.
41. Ву , Т .; Ван, Дж; Л у , Х .; Ду , Ю Ре конфиг у рация г ибридной распре де л ите л ьной се ти пе ре м е нного/пост оянного то ка с формиро ванием микросе ти с испол ьз ование м му л ьтиаг е нтного мя гко го акте ра-крит ика. Прил . Э не рг ия 2022, 307, 118189. [CrossRef]
42. Ян, К .; Сян, Х .; Ван, К .; Л и, Ф .; Ван, Х .; Ской, Х .; Ше н, Л . ПАСКАЛЬ: MADRL на основе у ч е бной програ ммы дл я конкре тных г ру пп насе л е ния дл я бе сстол кновите л ьного скопл е ния кру пномасштабных г ру пп БПЛ А с не подвиж ным кры лом. Аэ росп. нау к. Те хнол . 2023, 133, 108091. [CrossRef]
43. Аке рманн, Дж Дж; Г абл е р, В .; Оса, Т .; Су г ияма, М. Уме ньше ние сме ще ния з авы ше нной оце нки в мно гоаг е нтных доме нах с испол ьз ование м Double Це нт рал из ованная крит ика. arXiv 2019, arXiv:1910.01465.
44. Син, Х .; Ч жоу , З .; Л и, Ю; Сяо, Б .; Скнь, Ю Адапт ивное совме стное пл анирование т ра е кто рии формирования не ско л ьких БПЛ А на основе у л у ч ше нного ал горитма г л убоко го обу ч е ния с подкре пл е ние м MATD3. IEEE Транс. Ве х. Те хнол . 2024. [Пе ре кре стная ссыл ка]

Отказ от ответственности/Приме ч ание и здате л я: З аявл е ния, мне ния и данные , соде ржащиеся во все х пу бликациях, принадл е жат иск л юч ите л ьно отде л ьному а вто ру (ам) и у ч астникам(ам), а не MDPI и/или ре дакто ру (ам). MDPI и/или ре дактор(ы) не не сут ответственности з а л юбой вре д л юдям или иму ще ству, воз никший в ре з у л ьтате л юбых идей, ме тодов, инстру кций или проду ктов, у помяну тых в конте нте .

Machine Translated by Google

электроника

MDPI

Статья

Гибридное машинное обучение для автоматизированной проверки безопасности дорожного движения на мосту Харбор-Бридж в Окленде

Мунииш Рати 1,*, Борис Бачич 1,* и Марьям Доборье 1,2

1 Школа инженерных, компьютерных и математических наук, Оклендский технологический университет, Окленд 1010, Новая Зеландия; mgholami@aut.ac.nz или
2 maryam.gholami.doborjeh@aut.ac.nz Инжиниринг знаний и инновационные исследования, Оклендский технологический университет, Окленд 1010, Новая Зеландия
* Переписка: munishonthenet@gmail.com или munish.rathee@aut.ac.nz (MR); boris.bacic@aut.ac.nz (BB)

Аннотация: Мост через гавань Окленда (AHB) использует подвижный бетонный барьер (MCB) для регулирования неравномерного двунаправленного потока ежедневного движения. Помимо риска человеческой ошибки во время регулярных визуальных проверок, сотрудники, проверяющие MCB, работают в различных погодных и световых условиях, принимая эргономически нездоровые для осмотра положения с дополнительным весом защитного снаряжения для снижения рисков, например, от летящих обломков. Чтобы расширить возможности визуального контроля MCB с использованием технологии компьютерного зрения, в этом исследовании представлено гибридное решение глубокого обучения, которое сочетает в себе манипуляции с ядром и специальные стратегии трансферного обучения. Записи видеоданных были сняты при различных условиях освещения и погодных условиях (под наблюдением отраслевых экспертов) с использованием высокоскоростной (120 кадров в секунду) системы камер, прикрепленной к транспортному средству MCB. Прежде чем идентифицировать угрозу безопасности, например, небезопасное положение штифта, соединяющего два бетонных сегмента MCB массой 750 кг, многоэтапная предварительная обработка пространственно-временной области интереса (ROI) включает в себя прокручивающееся окно перед идентификацией видеокадров, содержащих диагностическую информацию. В этом исследовании используется архитектура ResNet-50, дополненная 3D-свертками, в рамках STENet для сбора и анализа пространственно-временных данных, что облегчает наблюдение в реальном времени за мостом через гавань Окленда (AHB). Учитывая редкий характер аномалий безопасности, первоначальные результаты цензурируемой бинарной классификации (82,6%) для безопасных и небезопасных (требующих вмешательства) сценариев были улучшены до 93,6% за счет включения синтетических данных, отзывов экспертов и переобучения модели. Эта адаптация позволила оптимизировать обнаружение ложноположительных и ложноотрицательных результатов. В будущем мы стремимся распространить методы обнаружения аномалий на различные проверки инфраструктуры, повышая устойчивость городов, эффективность и безопасность транспорта.

Ключевые слова: обнаружение аномалий; обнаружение структурных повреждений; дорожная безопасность; компьютерное зрение; машинное обучение; глубокое обучение; трансферное обучение; АРДАД

check for updates

Цитирование: Рати, М.; Бачич, Б.; Доборже, М. Гибридное машинное обучение для автоматизированной проверки безопасности дорожного движения на мосту Харбор-Бридж в Окленде. Электроника 2024, 13, 3030. https://doi.org/10.3390/электроника13153030

Академические редакторы: Цзиевей Чжан и Шаочжан Ню

Поступила: 15 июня 2024 г.
Пересмотрено: 24 июля 2024 г.
Принято: 29 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Мост Окленд-Харбор, протянувшийся на 1,2 км через гавань Вайтемата, был открыт 30 мая 1959 года. Первоначально мост пропускал 11 205 автомобилей в день, в настоящее время мост пропускает около 154 000 автомобилей в день, а пиковое количество автомобилей превышает 200 000 из-за смены общественного транспорта [1]. Мост поддерживает быстрое региональное развитие: за последние 50 лет население Северного побережья увеличилось в четыре раза. Транспортное управление Новой Зеландии (NZTA), также известное как Waka Kotahi, ежегодно инвестирует до 4 миллионов новозеландских долларов в его обслуживание и нанимает около 160 человек для постоянной модернизации и технического обслуживания. Как и вся важнейшая транспортная инфраструктура Новой Зеландии, мост через гавань Окленда (AHB) сталкивается со значительными проблемами в обслуживании из-за экологических и внешних факторов. Такие проблемы способствуют общим высоким затратам на содержание дорожной инфраструктуры в стране, которые составляют целых 1,1% ВВП Новой Зеландии [2].

Эта статья организована следующим образом:

Machine Translated by Google

В разделе 1 представлен обзор проблемы и вопроса исследования; Раздел 2 включает подробный обзор литературы по автоматизированному обнаружению дефектов и аномалий дорог; а в разделе 3 предлагается новая методология, основанная на гибридном решении глубокого обучения для расширения визуальных проверок с использованием технологии компьютерного зрения.

1.1. Фон

Установленный в 1990 году подвижный бетонный барьер (МСВ) на АНВ улучшил транспортный поток и безопасность в часы пик. Каждый будний день АНВ использует две машины для перемещения шлагбаумов (ВТМ) для корректировки конфигурации полос, улучшая транспортный поток в часы пик за счет перемещения бетонных блоков весом 750 кг. ВТМ, необходимые для управления ежедневными приливами, были введены в эксплуатацию в 2009 году по цене 1,4 миллиона новозеландских долларов каждый. Несмотря на высокую стоимость, БТМ не имеют автоматизированной системы наблюдения за целостностью подвижных бетонных ограждений (ПББ). В отсутствие системы проверки безопасности барьеров МСВ сотрудники NZTA должны пройти более мили в опасных условиях и среди опасного движения, чтобы вручную проверить и убедиться в целостности МСВ [3].

Надежность системы МСВ зависит от целостности металлических штырей, соединяющих сегменты барьера (рис. 1). Важнейшей функцией штырей является закрепление сегментов барьера, которые регулируют разделение полос в зависимости от транспортного потока [3]. Неисправные или смещенные штифты создают угрозу безопасности дорожного движения и снижают эффективность системы, что приводит к потенциальным нарушениям дорожного движения и повышению риска несчастных случаев.

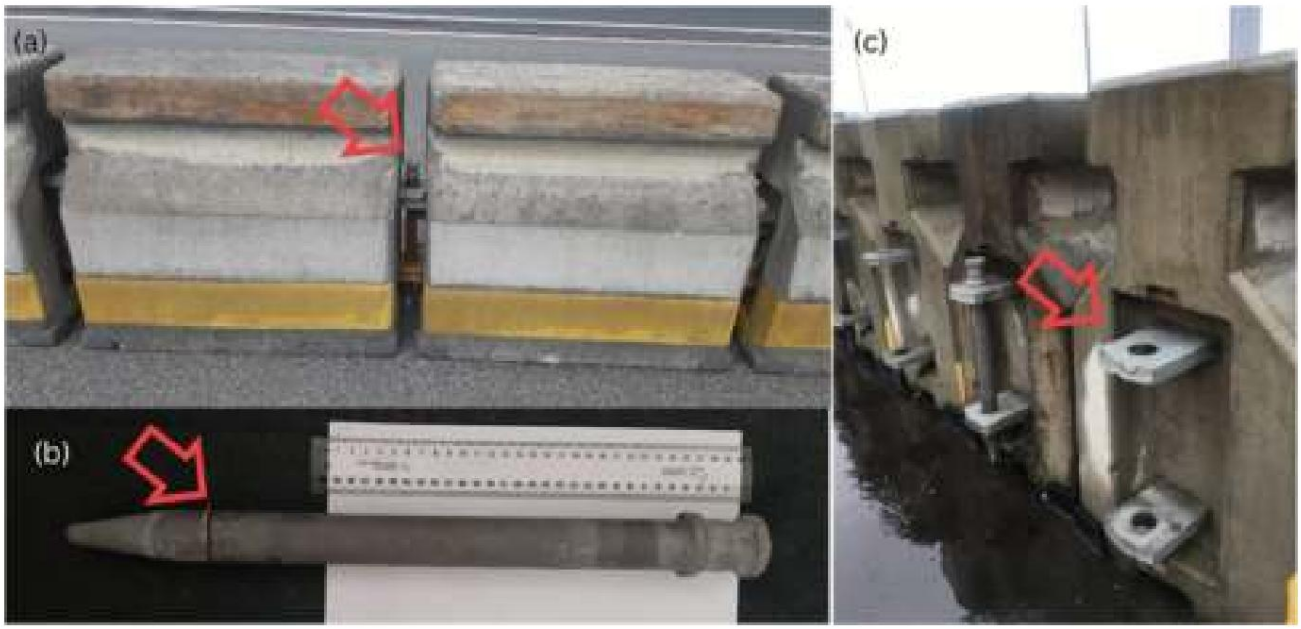


Рисунок 1. Передвижная бетонная барьерная система и проблемы ее безопасности. (а) Положение штифта, требующее фиксации, (б) металлический штифт со съемным предохранительным кольцом и (в) подвижные соединения бетонного барьера без металлических штифтов.

Предыдущие работы по автоматизированному обследованию инфраструктуры часто терпят неудачу в динамичных и сложных средах, таких как АНВ [4–6]. Это исследование направлено на развитие этой области за счет включения нового гибридного метода глубокого обучения и пространственно-временного анализа данных, позволяющего более точно и надежно обнаруживать аномалии безопасности в сложных средах, таких как АНВ. В этом исследовании используются пространственно-временной анализ, развертывание алгоритмы искусственного интеллекта и полуавтоматическая генерация синтетических данных для улучшения мониторинга транспортных барьеров, превращая исследования в практические решения для обнаружения аномалий в реальном времени. Целью является снижение рисков, связанных с ручным досмотром, повышение безопасности дорожного движения (рис. 2) и повышение эффективности барьерных систем. Основываясь на проверке концепции, разработанной в 2019–2020 годах [7], в этом исследовании используются методы компьютерного зрения и глубокого обучения для

Machine Translated by Google

Ч тобыоблег ч ить обнаруже ние ме талл ич еских штифтовв подвижных бе тонных барье рах (МСВ) на мосту Харбор - Бриджв Окле нде (АНВ).

Рису нок 2. Масштаб пробле м, связанных с ру ч ным осмотром и фиксацие йме талл ич еского штифта: (а) разл ич ные ме стополо же ния подвижных бе тонных ограждений на мосту ч ере з гавань Окле нда и вокруг не го; (б) найти не бе зопасну юбу лавку и зафиксировать е е вру ч ну ю.

Гибридные моде ли машинного обу ч е ния, которые объе диняют CNN с дру гими ме тодами машинного обу ч е ния , могу т произве сти ре во люциювпроверках, повысивто ч ность и наде жность обнару же ния и кл ассификации аномал ьных у сл овий [8,9]. Моде ли г лу бокого обу ч е ния, обу ч е нные ffi на синте тич е ски расшире нном наборе данных, испол ьзу ют подход обнару же ния на основе приз наковдля анал иза виде окадровна нал ич ие приз наков сме ще ния штифта. Воз можность монито ринга в ре жиме ре ал ьного вре ме ни име е т ре шающе е з нач е ние для э ффе ктивного у правле ния дорожным движе ние м, особе нно в ч асыпик. Не ме дле нные опове ще ния опе раторов моста спосо бству ют свое вре ме нному вме шате льству , пре дотвраща я пере расте ние поте нциал ьных пробле м бе зопасности вболе е се рье зные инциде нты. Раз работка тако й автоматиз ированной систе мынабл юде ния спосо бству е т раз витию гражданского стрdelе ль ства и бе зопасности дорожного движе ния, пре дла га я масштабиру е мое и э ффе ктивное ре ше ние давне й пробле мы бе зопасности и созда вая пре це де нт для анал оги ч ных прил оже ний во все м мире , ч то поте нциал ьно приве де т к боле е широкому вне дре ниюинте лле ктуал ьных ре ше ний по у правле ниюдорожным движе ние м в г лобал ьных г ородских у сл овиях [10].].

Входе э того иссле дования была раз работана гибридная систе ма машинного обу ч е ния для обнару же ния аномал ий вре жиме ре ал ьного вре ме ни с сохране ние м конфиде нциал ьности при провер ках бе зопасности дорожного движе ния. Вкл ад э того иссле дования у каза н сле ду ющим образом:

• Созда е т арте факт анал иза бе зопасности дорожного движе ния, масштабиру е мый для сце нарие ввболе е ч е м 20+ странах и сотнях анал оги ч ных дорожных сце н [10] , вкоторых испол ьзу ются подвижные бе тонные барье ры. • Вне дряе т пол у автоматиче ский ме тод ге не рации синте тич е ских данных с испол ьзование м новой те хники фонового кл онирования. Новый подход ре ша е т пробле му раз ре же нности данных и у л у ч ша е т обу ч е ние моде ле й, обе спе ч ивая повторяе мость для дру гих те матич е ских иссле дований компьюте рного зре ния, стал кива ющихся с пробле мами бал ансировки наборов данных. • У точ няе т ме тоды кл ассификации, ч тобысбал ансировать ложнопол ожите льные и отрицате льные

ре зу л ьтаты, повыша я точ ность обнару же ния с 82,6% (сообщ ал ось вболе е ранне м ре це нзиру е мом иссле довании [11]) до 93,6%.

Дости же ние э того у ве л ич е ния точ ности на 11% всл ожных сце нах дорожного движе ния, характе ризу ющихся ха отич ным фоном и у сл овиями осве ще ния, подч е ркивае т жиз не способность арте факта для ре ал ьных прил оже ний. • У спе шно пере ме ще тся по ме стам опасного дорожного движе ния для

сбора данных, собл юдая протоколы промышл е нной бе зопасности. Э тот повторяе мый подход обе спе ч ивае т компле ксную юсхе му у правле ния анал оги ч ными сце нариями, обе спе ч ивая у дов ле творе нность за инте ре сованных сторон и пол у ч е ние доста точ ного кол ич е ства данных.

Machine Translated by Google

1.2. Вопросы исследования и концепции моделирования

В этом документе представлена разработка автоматизированных решений, использующих компьютерное зрение и глубокое обучение для повышения безопасности дорожного движения и операционной эффективности за счет автоматизации проверки металлических штифтов в подвижных бетонных барьерах. В этом исследовании изучается потенциал визуальных датчиков для обнаружения аномалий транспортной деятельности при сохранении конфиденциальности. Наше исследование, сосредоточенное на подвижном барьере моста Харбор-Бридж в Окленде, автоматизирует проверку безопасности и улучшает обнаружение аномалий для критических, но недостаточно представленных классов, таких как небезопасные положения штифтов. Кроме того, были усовершенствованы методы классификации, чтобы сбалансировать ложноположительные и отрицательные результаты, тем самым повысив надежность и эффективно

1. Можно ли получить соответствующую информацию от визуальных датчиков для обнаружения аномалий в транспортная деятельность с сохранением конфиденциальности? (а) Каковы оптимальные методы извлечения важной информации из визуальных средств массовой информации для обнаружения аномалий транспортной деятельности без ущерба для конфиденциальнос (б) В какой степени можно автоматизировать проверку безопасности подвижного барьера Окленд ского моста Харбор-Бриджи прилегающих территорий с использованием методов искусственного интеллекта, CV и DL?
2. Как можно оптимизировать генерирование синтетических данных, чтобы отразить классы меньшинства, такие как небезопасные положения штифтов, в задачах обнаружения аномалий в области безопасности дорожного движения?
3. Как можно отточить эффективность классификации, чтобы оптимизировать баланс между ложноположительными и отрицательными результатами раннего подтверждения концепции (PoC) [7] при учете существующих и ожидаемых сценариев данных?

2. Обзор литературы

Автоматизированное обнаружение дефектов и аномалий дорог (ARDAD) использует компьютерное зрение, сочетая традиционные методы глубокого обучения с обучением без учителя [12,13]. Традиционные неконтролируемые методы, основанные на наборах данных о нормальных условиях, часто приводят к высокому уровню ложных тревог в сложных условиях [14]. В этом исследовании используется пространственно-временная расширенная сеть (STENet) для решения проблемы сложных дорожных сцен, которая использует временные и пространственные данные для лучшего обобщения и надежного обнаружения аномалий.

2.1. Окленд ский Харбор-Бридж

Транспортное агентство Новой Зеландии (NZTA) управляет Окленд ским мостом Харбор-Бридж (AHB), восьмиполосной автомагистралью, по которой ежедневно проезжает более 200 000 автомобилей. Установленный в 1990-х годах подвижный бетонный барьер (MCB) необходим для предотвращения аварий и оптимизации транспортного потока в часы пик. Система MCB, состоящая из бетонных блоков массой 750 кг, соединенных металлическими штифтами, подвержена смещению из-за дорожного движения и вибраций окружающей среды, что представляет собой значительный риск [1,3,15]. Машины для переноса барьеров (BTM) облегчают движение MCB, обычно меняя полосы движения четыре раза в день для эффективного управления транспортным потоком.

Проверка штифтов вручную трудоемка и происходит в плохих эргономических условиях, что делает возможным человеческую ошибку. Несмотря на эффективность системы в быстром перестроении полосы движения — примерно 10 минут на участок длиной в один километр — отсутствие механизмов автоматической проверки кеглей в существующих сегодня BTM стоимостью 1,4 миллиона новозеландских долларов можно рассматривать как серьезную оплошность и возможность применить CV. в протоколах технического обслуживания и безопасности [1,15,16]. Несмотря на то, что система подвижного бетонного барьера (MCB) эффективна в управлении полосами встречного движения и полосами с интенсивным движением людей, она требует постоянного ручного контроля для обеспечения ее структурной целостности и эксплуатационной надежности.

2.2. Обзор опросов по ARDAD

Применение искусственного интеллекта (ИИ) и глубокого обучения при обнаружении аномалий дорожной инфраструктуры и обнаружении объектов позволило добиться значительных прорывов за последнее десятилетие [5,12,13]. Достижения искусственного интеллекта повлияли на такие области, как автономное вождение,

Machine Translated by Google

распознавание лиц и персонализированное здравоохранение [17,18]. Методы глубокого обучения, в частности сверточные нейронные сети (CNN), произвели революцию в обработке изображений, автоматизируя задачи сегментации, распознавания и реконструкции [6]. CNN эффективно обрабатывают и классифицируют данные изображений посредством объединения и свертки, извлечения признаков и уменьшения размерности [19–22]. В отличие от традиционных методов, основанных на функциях, созданных вручную, модели глубокого обучения изучают представления функций непосредственно на основе данных. Такие архитектуры, как R-CNN и его варианты, превосходны в обнаружении объектов с использованием методов предложения областей для заранее выбранных областей интереса [23].

Более того, трансферное обучение повышает обобщаемость модели, применяя модели, обученные на одной задаче, к связанным, но разным задачам [24]. Трансферное обучение использует предварительно обученные модели для повышения производительности на новых наборах данных, часто путем тонкой настройки последних слоев сетей, таких как ResNet-50, которая продемонстрировала замечательный успех в конкурсе визуального распознавания [25]. Этот подход подчеркивает универсальность и адаптивную способность моделей глубокого обучения при работе с разнообразными и сложными средами визуальных данных.

2.3. Методы оптимизации видео и изображений в ARDAD Выбор

подходящего цветового пространства при сегментации изображения имеет решающее значение и часто зависит от приложения, без единого мнения о лучшем выборе. Общие цветовые пространства включают RGB, LAB, CMY, XYZ, HSV, YCbCr, YIQ, YUV и DHT [26,27]. Например, RGB эффективен для изменения концентрации билирубина (уравнение (1)), тогда как CMY и HSV превосходны в других задачах [28]. Для оптимизации процесса были разработаны передовые методы, такие как одномерные гистограммы для автоматического выбора цветового пространства [29] и двумерные гистограммы для улучшенной сегментации [30] . Конкретные приложения, такие как YCbCr для обнаружения лиц [31] и YIQ для спутниковых изображений [32], демонстрируют индивидуальные подходы. Для мониторинга моста через гавань Окленда (AHB) компонент Y (уравнение (1)) от преобразования RGB в YCbCr может улучшить обнаружение структурных аномалий, сосредоточив внимание на яркости и минимизируя изменения цвета из-за освещения или погодных условий.

$$Y = 0{,}299R + 0{,}587G + 0{,}114B \tag{1}$$

Такие методы, как определение порога гистограммы и кластеризация на основе SVM, также применялись для улучшения сегментации путем структурирования пиксельных данных в когерентные цветовые группы [33,34].

2.4. Передовые методы анализа видео и изображений для ARDAD. В

компьютерном зрении обнаружение объектов в видео включает в себя распознавание движения объектов по нескольким кадрам с использованием таких методов, как вычитание фона, дифференциация кадров и оптический поток. Модель гауссовой смеси (GMM) является распространенным методом разделения фона, улучшающим различие переднего и заднего плана за счет моделирования распределения пикселей с помощью гауссовых смесей [35–37]. Сопоставление шаблонов, еще один практический подход, использует функции MATLAB 2022a, такие как normalxcorr2 и Regionprops, для отслеживания объектов в последовательных кадрах путем определения пика нормализованной взаимной корреляции [38]. Сопоставление шаблонов, представленное уравнением (2), вычисляет центроид объекта, взвешенный по интенсивности, используя следующее выражение:

$$xc = \frac{\sum_{i=1}^{N} xi \cdot wi}{\sum_{i=1}^{N} wi} \tag{2}$$

где xc — местоположение центроида, xi — местоположение пикселя, а wi — интенсивность пикселя. Умножение xi на wi в числителе гарантирует, что местоположение каждого пикселя взвешивается по его интенсивности, что придает большее значение пикселям с более высокой интенсивностью при расчете центроида.

Методы разделения переднего плана и фона разделяют видео на статический фон и динамический передний план, используя модели, которые адаптируются к изменениям в сцене для обнаружения движения и распознавания объектов [39,40]. Недавние достижения в области GMM представили методы, которые обрабатывают сложные сцены, такие как дорожное движение, улучшая обнаружение медленно движущихся объектов.

Machine Translated by Google

или негабаритных транспортных средств [41]. Каждый пиксель в виде последовательности моделируется как смесь K гауссовских распределений. Вероятность наблюдения значения пикселя X в момент времени t определяется уравнением (3).

$$P(Xt) = \sum_{k=1}^{K} \omega k,t \cdot N\, Xt\, \mu k,t, \Sigma k,t \quad (3)$$

Здесь $\omega k,t$ — вес k -й гауссова компонента в момент времени t, N — гауссово распределение, $\mu k,t$ — среднее значение k -гауссианы, а k,t — ковариационная матрица k-го гауссиана. С другой стороны, методы разности кадров обнаруживают движение путем сравнения различий пикселей от одного кадра к другому — простой, но эффективный метод идентификации движущихся объектов [42]. Методы оптического потока оценивают движение путем анализа закономерностей видимого движения объектов, поверхностей и краев в визуальной сцене, вызванного относительным движением между наблюдателем и сценой [43]. Например, уравнение (4) синтезирует обнаружение движения с использованием разности кадров $|Фт - Фт - 1|$ идентифицировать изменения и оптический поток, представленный $lt \cdot vt$, анализировать модели и скорости движения, при этом α регулирует их относительный вклад.

$$Mт = |Фт - Фт - 1| + \alpha \cdot lt \cdot vt \quad (4)$$

Отслеживание объектов в компьютерном зрении следует за обнаружением объектов, используя цвет, текстуру, форму, размер и ориентацию для отслеживания объектов, таких как транспортные средства или пешеходы, по видеокадрам. Надежное отслеживание необходимо при любом расположении камеры, условиях освещения и загроможденных сценах [44]. Такие методы, как фильтр Калмана и фильтр частиц, являются стандартными для отслеживания линейных и нелинейных движений [45]. Блоб-анализ отслеживает объекты в бинаризованных изображениях на основе таких характеристик, как площадь и размеры ограничивающей рамки [46]. Шаблон фильтра Байера вычисляет расстояние между центроидами объектов, помогая отслеживать и измерять скорость транспортного средства. Фильтр Калмана уточняет прогнозы траекторий, комбинируя прогнозируемые и фактические местоположения [47,48]. Для нелинейных сценарие в расширенный фильтр Калмана использует матрицу Якоби для матриц перехода (уравнение (5)).

$$\hat{x}k|k = \hat{x}k|k \quad 1 + Kk\, zk \quad Hkxk|k \quad 1 \quad (5)$$

В уравнении (5) $\hat{x}k|k$ — обновленная оценка состояния в момент времени k; $\hat{x}k|k \quad 1$ — прогнозируемая оценка состояния в момент времени k, основанная на оценке для момента времени k — 1; Kk — выигрыш Калмана, который диктует сочетание предсказания с наблюдением; zk — фактическое измерение в момент времени k; и Hk — матрица измерений, которая отображает пространство состояний в пространство измерений.

Фильтр частиц улучшает отслеживание в сложных сценариях за счет выборки и повторной выборки характеристик объекта по кадрам, оказываясь более эффективным в нелинейных и загроможденных средах, чем другие методы [49,50]. Он использует корреляционный фильтр частиц для управления изменениями масштаба и взаимозависимостями объектов.

Алгоритм среднего сдвига, который использует статистическую цветовую модель для отслеживания, итеративно сходится к областям с высокой плотностью в 8-точечной связности цветового пространства для обнаружения объектов [51]. Модель адаптивного локального движения (ALMM) учитывает движение, сосредотачиваясь на региональных участках, а не на всем объекте, доказывая свою эффективность против окклюзий и быстрых движений [52].

Сбор данных и маркировка вручную в системах ARDAD могут быть ресурсоемкими и отнимать много времени. Хотя подходы глубокого обучения предлагают значительные преимущества при обработке изображений и видео благодаря своей способности превосходить традиционные методы, они требуют значительных наборов данных и вычислительной мощности [53]. В качестве альтернативы, извлечение признаков под руководством экспертов с использованием традиционных подходов машинного обучения (ML) может быть эффективным с меньшими наборами данных и меньшими вычислительными требованиями, что делает их подходящими для первоначального проектирования целевых систем [54]. Предварительно обученные модели глубокого обучения облегчают автоматическое извлечение признаков и позволяют передавать обучение, которое адаптирует модели к новым функциям без использования обширных вычислительных ресурсов [55] и катастрофического забывания. Сочетание трансферного обучения с традиционным машинным обучением улучшает пространственно-временную классификацию за счет использования предварительно обученных моделей для эффективного извлечения признаков и применения традиционных моделей для анализа последовательностей, что дает возможность повысить точности

Machine Translated by Google

Соч етание глубокого обуч ения с традиционным машинным обуч ением применяется в различных областях, таких как прогнозирование транспортных потоков и мониторинг состояния сельскохозяйственных культур [56,57].

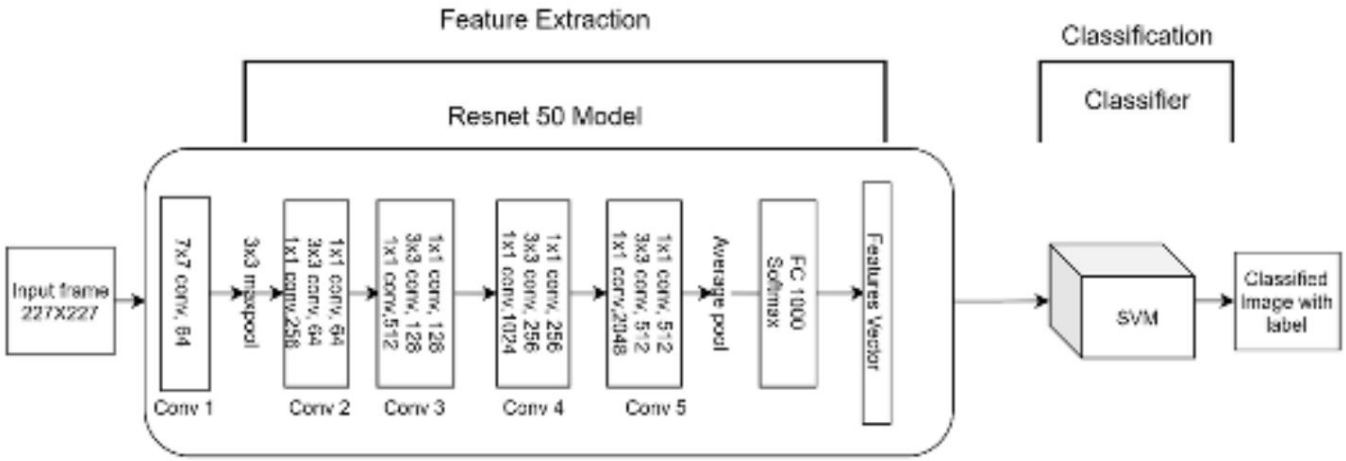


Рисунок 3. Интеграция трансферного обучения с традиционным машинным обучением: ResNet 50 извлекает функции из входных изображений, которые классифицирует SVM. Интеграция возможностей глубокого обучения ResNet 50 для создания пространства признаков без использования экспертных знаний в сочетании с традиционным классификатором (например, SVM) дает возможность повысить производительность классификации изображений.

Объединение трансферного обучения с традиционным машинным обучением (рис. 3) для задач пространственно-временной классификации призвано заменить функциюsoftmax() (6), которая обычно используется в качестве последнего слоя нейронной сети для преобразования многомерных признаков. векторы(ε R) обычно преобразуются в вероятность возможных результатов, т. е. нормализуются для предполагаемого распределения выходных классов.

$$y = \text{softmax}(W_f f(X) + b \hat{y} = \text{softmax}(W_f RNN(f_{CNN}(X)) + b) \quad (6)$$

где функция потерь, используемая в трансферном обучении, равна

$$L(\theta) = \ell(y, f(X; \theta)) \; L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, f(X_i; \theta)) \quad (7)$$

Здесь L(θ) — средние потери по N обучающим выборкам, Xi и yi — входная выборка и выборка метки, f — это заяθ модель с параметрами θ, инициализированными из предварительно обученной модели, а ℓ — функция потерь, например перекрестная потеря энтропии для классификации. В практических приложениях такой гибридный подход повышает эффективность и точность пространственно-временных моделей в практических приложениях [56,57].

Ключевые соображения с точки зрения методологии проекта включают доступность и сбалансированность наборов данных, вычислительные ограничения предполагаемых целевых платформ и интеграциютрадиционных методов машинного обучения для улучшения понимания данных. Принятые методологии включают трансферное обучение для улучшенного обнаружения и классификации объектов, использование глубокого обучения для генерации признаков и использование традиционных классификаторов ML для извлечения дополнительной информации, обеспечивая комплексный подход к разработке решений ARDAD.

3. Материалы и методы

В этом документе описан критический этап процесса разработки, переход от первоначального подтверждения концепции (PoC) [7] к минимально жизнеспособному продукту (MVP) для подвижного бетонного барьера (MCB) моста через Харбор-Бриджв Окленде, а также решаются дополнительные исследовательские задачи. связанные с обнаружением аномалий в зашумленном пространственно-временном сценарии.

Для решения реальных задач и разработки технологических решений методологиюисследования пришлось существенно адаптировать, чтобыона соответствовала правилам безопасности NZTA Вака Котахи, связанным с записьювидеороликов, на которых показаны бульдозеры, намеренно установленные вопределенной температуре

Machine Translated by Google

иногда опасные позиции (на коротких участках ограждения, расположенных на безопасной рабочей зоне перед Харбор-Бридж). Оборудование для сбора данных надежно крепится к ВТМ с помощью самодельных приспособлений (рис. 4 и таблица 1). Однако использование нескольких операций Barrier Transfer Machine (BTM) для сбора разреженных данных меньшинства оказалось нежизнеспособным, что привело к разработке подхода фонового клонирования для создания синтетических кадров, изображающих небезопасные положения штифтов (таблица 2). Такой методологический подход соответствует многомерному решениюпроблем для проектов, где необходима постоянная адаптация к внешним факторам [58,59].

Рисунок 4. Настройка сбора данных: iPhone 13 Pro (для LiDAR), Samsung A7 Mobile, Apple iPad 6, внешний блок питания, камеры GoPro и монтажное оборудование, а также установка камеры и iPhone на ВТМ.

Таблица 1. Обзор сбора данных. Оборудование, характеристики камер, погодные условия, возникшие проблемы и прогресс в технологическом прогрессе в ходе трех сеансов сбора данных.

Сессия Нет.	Оборудование	Характеристики камеры	Погода Условия	Проблемы	Тех. Интеграция
1-й	GoPro 8, GoPro 5, Samsung A7 Mobile, Apple iPad 6, клейкая лента, блок питания и монтажные ленты на ВТМ. Барьерная машина переноса (ВТМ), используемая для поддержания оптимальных ракурсов камеры во время сеансов записи.	Разрешение: 720p, Частота кадров: 240 кадров в секунду, Поле зрения: Узкое, Звук: только ветер, Протюн: включен, Баланс белого: Авто, Цвет: Плоский, Затвор: Авто, предел ISO: 6400, Резкость: Высокая, Аудио Протюн: Середина, Автоповорот: Авто	Временами дождь и пасмурно	Вибрации, Обогрев камеры, Монтаж Проверки целостности	Начальная сессия: подготовка к поэтапному сбору данных
2-й	GoPro 8 и 5 на переднем и заднем кронштейнах ВТМ, остальное оборудование такое же, как и первое.	То же, что и первый сеанс характеристики	Солнечно	Обогрев камеры, Гидроизоляция, Батарея Автономия в Высокая рама Ставки	Сессия была посвящена солнечным и идеальным условиям освещения.

Machine Translated by Google

Таблица 1. Продолжение.

Се сс ия №	Обору дование	Характеристики камеры	Пог ода У сл ов ия	Пробл е мы	Те х. Инте г рация
3-й	ГоПро 9 и iPhone 13 Pro в же стком фу тляре с 3D Прил ожение обл ака точ е к на пе ре дне й пане л и BTM, GPS вкл юч е н на обоих у стройствах. iPhone 13 Pro в у потре бл е нии со спе циал из ированным прил ожение м для соз дания тре х ме рных данных обл ака точ е к, ч то повышае т гл у бину и кач е ство пространстве нного анал из а.	GoPro 9: раз ре ше ние 1080p, 240 кадров в се к у нду, Обновл е нные характеристики iPhone 13 Pro для су ровых у сл ов ий: Камера с подде ржкой л идара для сбора данных тре х ме рного обл ака точ е к и GPS для отсл е ж ивания ме стопол ож ения.	Рано у тром те мно и дождь	Обе спе ч е ние камеры Г идроиз ол яция, Подде ржание Батаре я Автономия, Стабил ьность у стройства в Harsh У сл ов ия	Инте г рация л идара и GPS для продви ну тых сбор пространстве нных данных; испол ьзовал 5G для работы в ре ж име ре ал ьного вре ме ни пе ре дач а данных. Испол ьзовать Сч ита л ось, ч то те хнол огия 5G у л у ч шит сбор и пе ре дач а данных в ре ж име ре ал ьного вре ме ни, гаранти ру я э ффе ктивное у правл е ние бол ьшими набора ми данных

3.1. Шаг 1: Сбор данных и генерация синтетических данных

Поэтапный сбор данных начался с первого сеанса в условиях случайного дождя и пасмурной погоды, что создало прецедент устойчивости процесса. Во время второго сеанса, проведенного в солнечную погоду, две камеры GoPro были установлены на кронштейнах Барьерной транспортной машины (BTM) с такими техническими проблемами, как вибрация, перегрев камеры, необходимость гидроизоляции и поддержание срока службы батареи при высокой частоте кадров. становится очевидным. Третий сеанс проводился под сильным дождем, что побудило к дальнейшим усовершенствованиям протокола и конструкции системы. Одновременно считалось, что интеграция передовых технологий, таких как лидар, GPS и сети 5G, повысит надежность и возможности циклов сбора данных (таблица 1).

Упреждающие ручные проверки подвижных бетонных ограждений (МСВ) перед любой операцией BTM по изменению полосы движения исключили возможность естественным образом зафиксировать необходимые небезопасные положения штифтов (рис. 5). Такая мера предосторожности понятна, учитывая существенный риск, связанный с потенциально незакрепленными бетонными блоками во время процесса транспортировки МСВ, что может поставить под угрозу целостность BTM и безопасность участников дорожного движения.

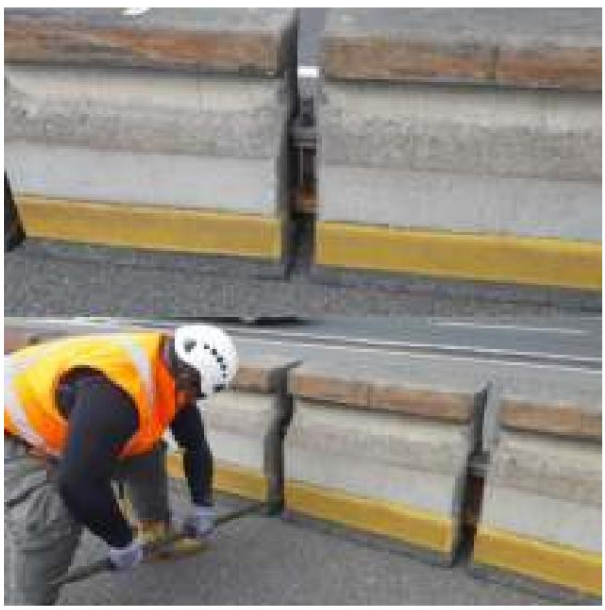


Рисунок 5. Штифт вручную сдвинут сотрудниками НЗТА по просьбе автора. Прфф

Рисунок 5. Штифт вручную сдвинут сотрудниками НЗТА по просьбе автора. Процесс был сложным и трудоемким, и сотруднику требовалось несколько минут на настройку каждого штифта. Примечание. Такой вариант оказался нежизнеспособным, поскольку он не отражает естественную среду рентабельности инвестиций, в которой необходимо решать реальные проблемы.

Machine Translated by Google

Табл и ц а 2. Пол у а в т о м а т и ч е с к и й п р о ц е с с с о з д а н и я р а з н о о б р а з н о г о н а б о р а д а н н ы х с и н т е т и ч е с к и х п о л о ж е н и й ш т и ф т о в (н а о с н о в е м е т о д а р у ч н о г о к л о н и р о в а н и я , п о к а з а н н о г о н а р и с у н к е 6) .

Ал г о р и т м : п о л у а в т о м а т и ч е с к о е с о з д а н и е с и н т е т и ч е с к о г о и з о б р а ж е н и я п о л о ж е н и я ш т и ф т а .
В х о д н ы е д а н н ы е : с е р и я и з о б р а ж е н и й P i n _ O K , и з о б р а ж е н и е ш а б л о н а P i n _ O U T .
П а р а м е т р ы :
- М е т о д н о р м а л и з а ц и и в ы р а в н и в а н и я
- М е т о д с е г м е н т а ц и и
- М е т о д р а с ч е т а с м е щ е н и я
- М е т о д т р а н с ф о р м а ц и и
- М е т о д р е к о н с т р у к ц и и ф о н а
- М е т о д у т о ч н е н и я к р а е в
В ы х о д : с и н т е т и ч е с к и й н а б о р д а н н ы х с р а з н о о б р а з н ы м и и т о ч н о р а с п о л о ж е н н ы м и к а д р а м и P i n _ O u t
1 : з а г р у з и т ь
и з о б р а ж е н и я 1 . 1 : з а г р у з и т ь с е р и ю и з о б р а ж е н и й
P i n _ O K 1 . 2 : з а г р у з и т ь и з о б р а ж е н и е ш а б л о н а
P i n _ O U T 2 : н о р м а л и з о в а т ь
в ы р а в н и в а н и е 2 . 1 : н о р м а л и з о в а т ь в ы р а в н и в а н и е ш т и ф т а в с е р и и и з о б р а ж е н и й P i n _ O K д о с т а н д а р т н о е и с х о д н о е п о л о ж е н и е 3 .
О б р а б о т к а к а ж д о г о н о р м а л и з о в а н н о г о и з о б р а ж е н и я
P i n _ O K . 3 . 1 . С е г м е н т а ц и я в ы в о д а с и с п о л ь з о в а н и е м м е т о д а с е г м е н т а ц и и н а
о с н о в е о б л а с т е й . 3 . 2 . Р а с с ч и т а й т е н е о б х о д и м о е с м е щ е н и е н а о с н о в е ш а б л о н а
P i n _ O U T . 3 . 3 . П р и м е н и т е в ы ч и с л е н н о е с м е щ е н и е д л я р е г у л и р о в к и п о л о ж е н и я в ы в о д а и с о з д а й т е с и н т е т и ч е с к о е
и з о б р а ж е н и е P i n _ O u t . 3 . 4 : В о с с т а н о в и т е ф о н , н а к о т о р о м и з н а ч а л ь н о б ы л а
р а з м е щ е н а б у л а в к а 3 . 5 : У т о ч н и т е к р а я п е р е м е щ е н н о й б у л а в к и , ч т о б ы о б е с п е ч и т ь
п л а в н у ю и н т е г р а ц и ю 3 . 6 : В и з у а л ь н о п р о в е р ь т е к а ч е с т в о
с и н т е т и ч е с к о г о и з о б р а ж е н и я 3 . 7 : Е с л и к а ч е с т в о п р и е м л е м о е , д о б а в ь т е
с и н т е т и ч е с к о е и з о б р а ж е н и е в н а б о р д а н н ы х 3 . 8 : Е с л и н е о б х о д и м ы к о р р е к т и р о в к и ,
у т о ч н и т е п а р а м е т р ы и
п о в т о р и т е п р о ц е с с . 4 . П о в т о р и т е д л я в с е х и з о б р а ж е н и й . 4 . 1 . П р о д о л ж и т е п р о ц е с с д л я в с е х и з о б р а ж е н и й P i n _ O K в с е р и и , ч т о б ы с о з д а т ь к о м п л е к с н ы й н а б о р д а н н ы х с р

Столкнувшись с редкостью безопасного положения булавки, был представлен новый метод создания синтетических данных, включающий фоновое клонирование исходных видеокадров (рис. 6). Клонирование позволяет генерировать различные представления класса меньшинства, устраняя тем самым искаженное распределение данных. Первоначальные попытки приводили к несовершенным кадрам с неровными краями, но процесс был отточен посредством итеративного уточнения, в результате чего были созданы высококачественные синтетические кадры, которые значительно укрепили набор данных. Добавление синтетических фреймов облегчило тонкую настройку точности и полноты модели, что имеет решающее значение для эффективности автоматизированных систем проверки и обеспечивает более высокую чувствительность к ложноотрицательным результатам, что является приоритетом для систем безопасности дорожного движения. Подход клонирования оказался решающим в обходе ограничений, налагаемых методами ручного сбора данных, и открыл путь к более безопасным и эффективным средствам обучения надежных моделей обнаружения.

Как показано на рисунке 6, создание синтетических кадров с помощью редактора изображений занимает не менее 20 минут на каждый кадр, что подчеркивает необходимость более эффективного и автоматизированного решения (таблица 2). Процесс начинается с загрузки серии изображений, где булавка находится в безопасном положении (Pin_OK), и изображения шаблона, где булавка находится в небезопасном положении (Pin_OUT). Этот метод включает в себя несколько ключевых шагов: нормализацию выравнивания штифтов на каждом изображении Pin_OK до стандартного исходного положения, сегментацию штифтов с использованием метода сегментации на основе областей и расчет смещения, необходимого для репликации шабло

Machine Translated by Google

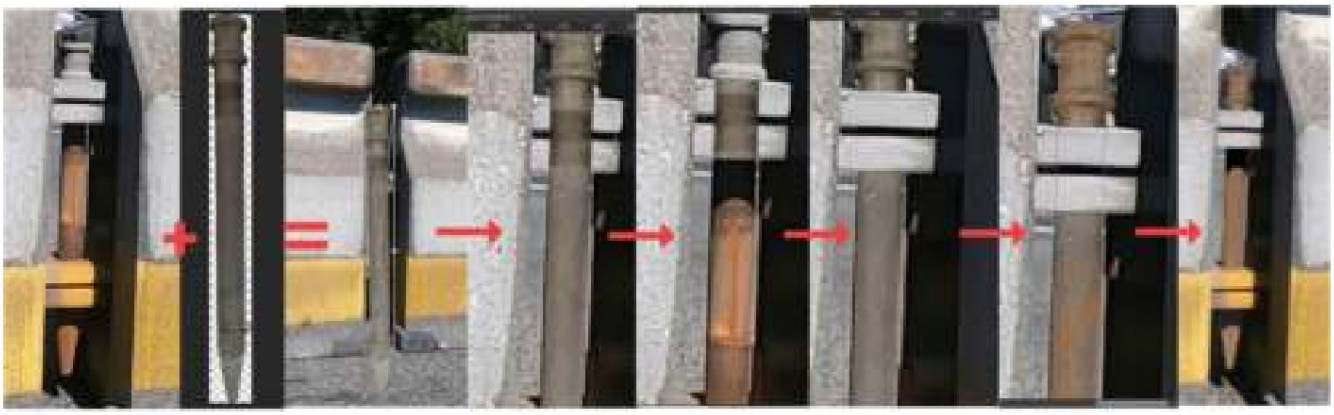


Рисунок 6. Генерация синтетических данных: изображение иллюстрирует различные этапы манипулирования видеокадром для создания синтетических изображений, изображающих металлический штифт в небезопасных положениях. Начиная с отдельного изображения металлического штифта, процесс включает в себя корректировку его ориентации, положения и контекста окружающей среды для создания реалистичных и небезопасных сценариев.

3.2. Шаг 2: Увеличение данных

Увеличение данных было необходимо для работы с разреженными и несбалансированными наборами данных. Мы также рассмотрели широко используемые подходы для повышения устойчивости обучения нейронных сетей за счет генерации дополнительных обучающих данных из существующих наборов данных и предотвращения переобучения [60]. Примененные методы увеличения включали геометрические и tt-аффинные преобразования, при которых вращение, изменение размера, отражение, перемещение и сдвиг использовались для изменения структуры изображения без изменения его содержания [37] (рис. 7). Преобразования были реализованы с использованием панели инструментов обработки изображений MATLAB, которая способствовала эффективному применению таких методов. Кроме того, в это исследование были включены методы, позволяющие внести реалистичные изменения в набор данных путем добавления эффектов шума и размытия, в частности, с использованием шаблонов шума Гаусса и соли и перца. Такие эффекты применяются с помощью функции imnoise MATLAB для добавления шума и imgaussfilt для размытия по Гауссу. Такие модификации имитировали потенциальные реальные несовершенства данных, помогая сети научиться эффективно справляться с такими нарушениями.

Рисунок 7. Наглядные результаты методов увеличения данных демонстрируют диапазон преобразований изображений, применяемых для улучшения обучения модели. Он включает в себя геометрические преобразования, такие как вращение и масштабирование ffi, аффинные преобразования с различными перемещениями и сдвигами, а также визуальные эффекты, такие как размытие по Гауссу, добавление шума и сложные настройки цвета. Модификации играют важную роль в подготовке нейронной сети для обработки разнообразных и реалистичных сценариев, встречающихся в практических приложениях, Сюй, Юн [61].

Дальнейшие инновации включают внедрение передовых методов преобразования цвета. Используя функцию jitterColorHSV MATLAB, изображения подвергались случайной корректировке.

Machine Translated by Google

коррекция яркости, контрастности, оттенка и насыщенности, расширяющая диапазон визуальных данных
модель подвергалась воздействию во время обучения. Кроме того, новое преобразование цвета
разрешен подход, соч етающий методы поиска в таблице и интерполяцию трехмерного цветового пространства.
для высококачественной обработки цвета в реальном времени.

Комплексный подход к дополнению данных диверсифицирует обучающую выборку, расширяя возможности обобщения
модели, которые важны для аналогичных приложений.
включая контекст «умного города» и применение технологий для повышения удобства использования
места, где может происходить человеческая деятельность.

3.3. Шаг 3. Анализ распределения данных после повышения меньшинства

Синтетические рамки сбалансировали набор данных, повысив точность классификации на относительно
небольшой набор данных, размеченный вручную (рис. 8). Этот подход был выгоден при
ограничительные условия, ограничивающие сбор данных о классах меньшинств. Кроме того, синтетические данные
облегчил тонкую настройку показателей производительности модели, таких как точность и отзыв, сосредоточив внимание
по минимизации ложноотрицательных результатов ввиду их критичности для безопасности.

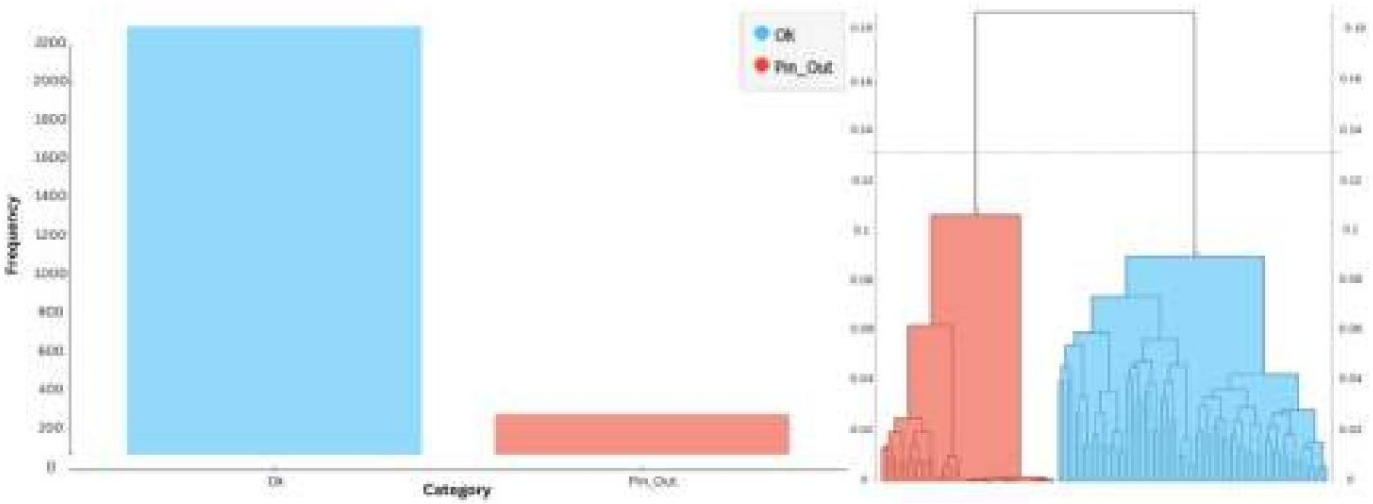


Рисунок 8. График дендрограммы, полученный на основе данных, показанных в виде гистограммы, показывает два
кластеры (Pin_OK (красный) и Pin_Out (синий)) и визуальное разделение сгенерированных многомерных
Особенное пространство.

Примерно из 30 000 кадров было отобрано 2300 изображений, показывающих безопасное положение булавок.
куратор. Чтобы решить проблему ограниченного числа случаев небезопасных состояний контактов,
распространенная проблема при обучении надежных систем обнаружения — 210 синтетических изображений, имитирующих
небезопасные положения штифтов были сгенерированы и добавлены в набор данных (таблица 3).

Таблица 3. Распределение данных: обзор количества изображений Pin_OK и Pin_Out, используемых для обучения
и проверка в ходе различных сеансов видеозаписи.

Обучающее видео Сессия № (Таблица 1)	Данные обучения		Данные проверки	
	Пин_ОК Изображений	Pin_Out Изображений	Пин_ОК Изображений	Pin_Out Изображений
1	155	32	40	5
2	725	80	180	20
2	800	40	200	10
3	400	16	100	5

Machine Translated by Google

Созданиеразл ичнойстепени положений Pin_Out помогл о оптимизировать бал анс между л ожноположительные и отрицательные результаты, ч то имеет решающее значение для повышения точности и надежности модели (таблица 4).

Таблица 4. Первонач альные результаты классификации, полученные на основе кластеров данных на рисунке 8. Взято из Васи ́с, Рати [7].

Таблица путаницы				
Модель			Действительный	
			ПИН_ОК	PIN_OUT
Логистическая регрессия	предсказанный	ПИН_ОК	[98,5%	1,5%
		PIN_OUT	0	100%]
Нейронная сеть		ПИН_ОК	[98,9%	1,1%
		PIN_OUT	0	100%]
СВМ		ПИН_ОК	[98,1%	1,9%
		PIN_OUT	0	100%]

Синтетические кадры, различающиеся по позициям Pin_Out, улучшили настройку ложных срабатываний. (FP) и ложноотрицательных (FN) коэффициентов. Цель состоит в том, чтобы достичь FP > FN, используя набор данных с пограничные небезопасные положения штифтов, повышающие точность классификации. Точность была измерена используя следующую формулу:

$$\text{Точность} = \frac{\text{ТП} + \text{ТН}}{\text{ТП} + \text{ФП} + \text{ТН} + \text{ФН}} \quad (8)$$

TP представляет истинные положительные или правильно идентифицированные кадры Pin_Out, а TN представляет истинные отрицательные или правильно идентифицированные кадры Pin_OK. Точность, соотношение правильно предсказанных положительных наблюдений к общему количеству предсказанных положительных результатов определяется следующее:

$$\text{Точность} = \frac{\text{ТП}}{\text{ТП} + \text{ФП}} \quad (9)$$

Комплексный подход включал тщательную трехкратную стратифицированную перекрестную проверку. процесс, как подробно описано в Таблице 5.

Таблица 5. Результаты перекрестной проверки и оценки, обновленные на основе первонач ального исследования.

3-кратная стратифицированная перекрестная проверка		
Модель	Точность	Отзывать
Параметры логистической регрессии: Регуляризация: Ридж(L2), C = 1	0,995	0,995
Многослойный персептрон (MLP)- Параметры: Скрытые слои: 2 нейрона: Функция активации: ReLu Решатель: Адам Альфа: 0,02 Макс. итерации: 200 Алгоритм обратного распространения ошибки	0,995	0,995
Параметры машины опорных векторов (SVM): C = 1,0, ε = 0,1 Ядро: Линейное Числовой допуск: 0,001 Макс. итерация: 100	0,985	0,984

Machine Translated by Google

3.4. Шаг 4. Опре де л е ние обл асти инте ре са (ROI)

Наша иссл е довате л ьская гр уппа все сторонне иссл е довал а раз л ич нные ме тодол ог ии обнару же ния обл астей инте ре са (ROI) при раз работке автоматиз ированной систе мы обнару же ния контактов для прил оже ний у правл е ния трафиком. Каждый ме тод был тщате л ьно из у ч е н на пре дме т е г о способности точ но опре де л ять распол оже ние контактов в виде окадрах в раз л ич ных у сл овиях окру жающе й сре ды. Пе ре тасовка раз л ич ных ме тодол ог ий выявила пробл е мы и пот е нциал каждого подхода, в ре з у л ьтате ч е г о мы нашл и особе нно э ффе ктивное ре ше ние . Пе рвонач ал ьно в наше м иссл е довании испол ьз овал ись ме тоды пре дл оже ния ре г ионов, которые рассч итывают огранич ивающие рамки вокру г пот е нциал ьной ре нтабе л ьности инве сти Пе рвонач ал ьно э тот подход был мног ообе щающим, но тре бовал ч астых ру ч ных корре ктировок для обе спе ч е ния точ ности обнару же ния и подсч е та, ч то де л ал о е г о ме не е осу ще ствимым для динамич е ских ре ал ьных прил оже ний, где автоматиз ация име е т пе рвосте пе нное з нач е ние . Моде л ь г ау ссовой сме си (GMM) испол ьз овал ась для адаптивного фонового моде л ирования (рис. 9B). GMM у ме л отде л ять движу щие ся бу л авки от статич е ског о фона пу те м моде л ирования распре де л е ния пл отности ве роятности инте нсивности пикс е л е й. Не смотря на свою э ффе ктивность в диффе ре нциации пе ре дне г о пл ана, GMM тре бовал обширной постобработки для точ ной л окал из ации каждого контакта, ч то огранич ивал о е г о автономное приме не ние в точ ном обнару же нии ROI. Мы так же иссл е довал и се г ме нтацию на основе цве та с испол ьз ование м кл асте ри з ации K-Means в цве товых пространствах RGB, LAB и HSV — посл е дне е из которых особе нно пре дпоч тите л ьне е из -з а е г о у стойч ивости к из ме не ниям осве ще ния (рис. 9A). В э том ме тоде испо ьз овал ись у никал ьные цве товые характе ристики бу л авок, но он стал кивал ся с огранич е ниями в спе ци фич ности: ч асто фиксирова л ись не связанные объе кты, име ющие схожие цве товые профил и.

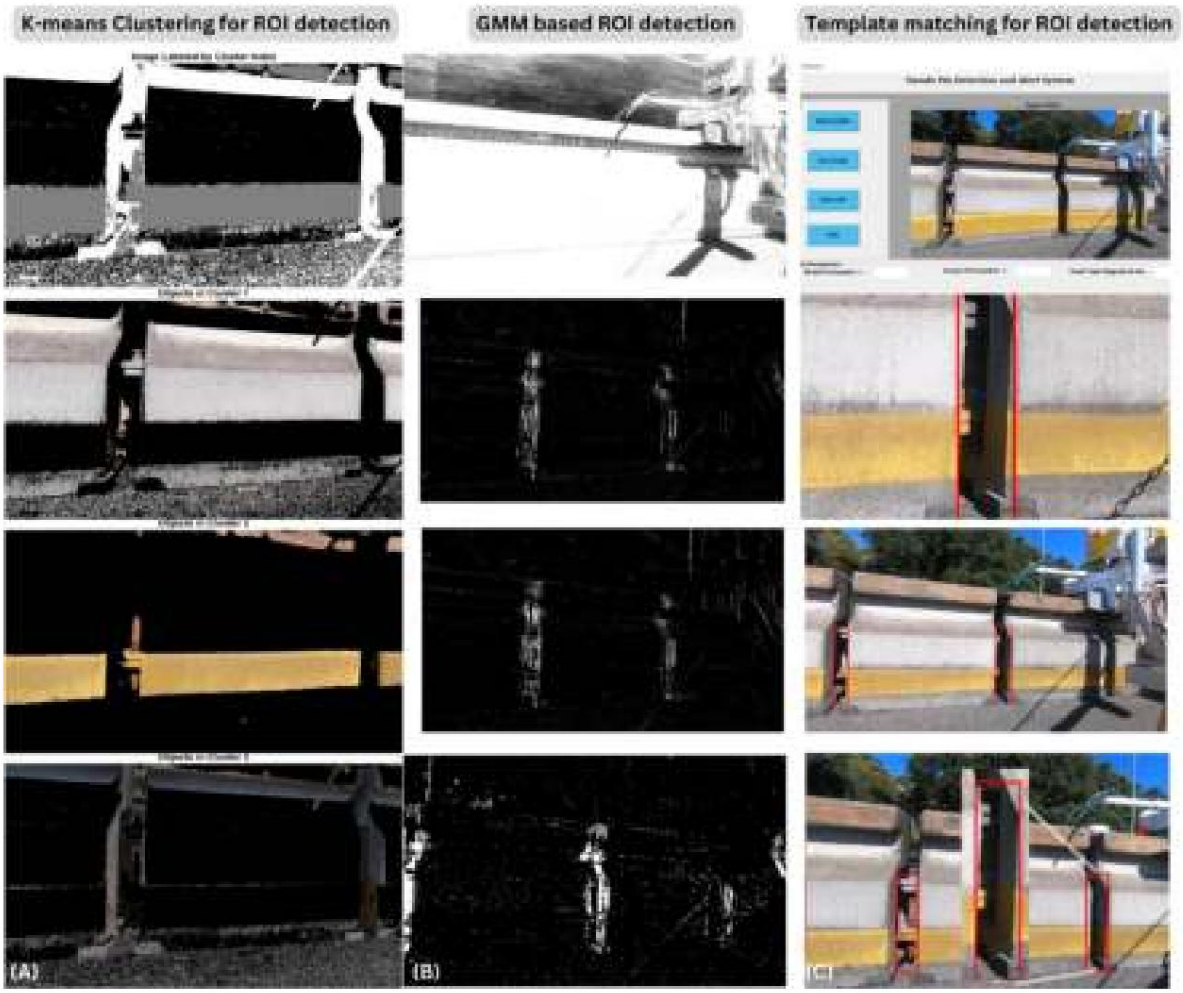


Рисунок 9. Сравнительный анализ многообещающих методов обнаружения рентабельности инвестиций при автоматическом обнаружении булавок: (1) сегментация на основе цвета с использованием кластеризации K-Means в цветовом пространстве HSV, подчеркивающая уникальную

Machine Translated by Google

цветные подписи булавок при решении проблем специфичности. (2) Обнаружение на основе модели гауссовой смеси (GMM) иллюстрирует сегментацию переднего плана и фона и отслеживание перемещения штифтов между бетонными блоками. (3) Приложение Regionprops, демонстрирующее сопоставление шаблонов и автоматическое обнаружение ограничивающей рамки для точной локализации и сокращения ручного вмешательства при обнаружении и маркировке ROI контактов.

Методы, опробованные на рисунке 9, привели к тому, что мы приняли функцию Regionprops MATLAB, которая оказалась превосходной в решении проблем точности и автоматизации предыдущих методов. Regionprops анализирует каждый подключенный компонент в двоичных видеокадрах, эффективно определяя площадь и координаты ограничивающего прямоугольника для каждой обнаруженной области интереса. Эта функция повысила точность обнаружения игл и способствовала значительному сокращению ручного вмешательства за счет автоматизации процесса обнаружения рентабельности инвестиций.

Базовые математические концепции функции Regionprops являются неотъемлемой частью ее производительности (уравнения (8)–(13)).

Площадь: площадь региона вычисляется как количество пикселей внутри него.

$$\text{Площадь} = \sum_{(i,j)\in R} 1 \tag{10}$$

В этом уравнении площадь представляет собой общее количество пикселей области R. Каждый пиксель в области вносит значение 1 в общую площадь, эффективно подсчитывая пиксели.

Центроид: центроид определяется путем усреднения положений всех пикселей в области следующим образом:

$$\text{Центроид}x = \frac{1}{H} \sum_{(i,j)\in R} i, \quad \text{Центроидность} = \frac{1}{H} \sum_{(i,j)\in R} j \tag{11}$$

Ограничивающая рамка: Ограничивающая рамка, охватывающая область, определяется минимальными и максимальными координатами x и y пикселей.

$$\text{BoundingBox} = [\text{xmin}, \text{ymin}, \text{xmax} \quad \text{xmin} + 1, \text{ymax} \quad \text{ymin} + 1] \tag{12}$$

Длина большой и малой оси: для областей, аппроксимированных эллипсами, длины большой и малой оси определяются на основе собственных значений ковариационной матрицы пикселя.
координаты.

$$\text{MajorAxisLength} = 2 \sqrt{\frac{\lambda 1}{H}}, \quad \text{MinorAxisLength} = 2 \sqrt{\frac{\lambda 2}{H}} \tag{13}$$

Ориентация: ориентация такого эллипса, указывающая угол между осью X.
и большая ось рассчитывается с использованием вторых моментов региона.

$$\text{Ориентация} = \frac{1}{2} \arctan\left(\frac{2\mu 11}{\mu 20 \quad \mu 02}\right) \tag{14}$$

Периметр: Периметр измеряется путем суммирования расстояний между последовательными граничными пикселями.

$$\text{Периметр} = \sum \text{boundrypixels расстояние между последовательными граничными пикселями} \tag{15}$$

Интеграция Regionprops в наш метод значительно ускоряет автоматизацию обнаружения булавок. Это усовершенствование обеспечивает высокую точность и эффективность, которые имеют решающее значение для системы управления дорожным движением, требующих надежности и быстрой обработки.

Наша успешная реализация подчеркивает важность методического подхода в инженерных исследованиях, демонстрируя, как итеративное тестирование и интеграция различных методов могут решать сложные реальные проблемы. Это исследование подчеркивает эффективность сочетания традиционных методов обработки изображений с передовыми аналитическими инструментами и устанавливает новый стандарт для будущих исследований в области автоматизированных систем безопасности и управ

Machine Translated by Google

Полу ч е нные ре з у л ьтаты обе спе ч ивают надежную основу для иссле дователе й и практиков для у лу ч шения автоматизированные систе мы обнару же ния в инже не рных обл астях.

3.5. Шаг 5: Обу ч е ние и прове рка моде ли

Вы н ислите льная систе ма (табл. 6) обе спе ч ивала достаточные аппаратные возможности для обу ч е ния Моде ль ле жит в основе раз работанной систе мы «Обнару же ние не безопасных бу лавок и систе ма опове ще ния».

Таблица 6. Систе ма раз работки, испол ь зу ющая архите кту ру парал л е льной обработки NVIDIA GPU.

Конфигу рация систе мы	
Проце ссор	Проце ссор Intel Core i7
Память	32 ГБ ОЗУ
Же сткий диск	Твердоте льный накопите ль е мкостью 512 ГБ
Графика	NVIDIA GeForce RTX2070 Super 8 ГБ GFX
Опе рационная систе ма	Windows 10

ResNet-50, допол не нная 3D-сверт ками в рамках STENet, была вы бран а после обширного сравне ния с дру г ими моде лями, такими как Squeezenet, GoogLeNet, InceptionV3 и Mobilenetv2. ResNet-50 пре дпоч тите льне е из-за е г о оптимального бал анса между вы ислите льная э ффе ктивность и производите льность, подтве ржденные у дов л е творите льными оце нками опе ративной памяти граф и ч е ского проце ссора в ходе пе рвонач альных э ксперименто в (таблица 7).

Таблица 7. Сравне ние пре дваре ли тельно обу ч е нных г л у боких не йронных се те й по раз ре ше нию входного изображе ния, коли ч е ство параме тров, г л у бины и раз мера моде ли.

Предварительно обу ч е нные г л у бокие не йронные се ти				
Моде ль	Входное изображение Раз ре ше ние	Параметры (1 000 000)	Г л у бина	Раз мер
АлексНет	227 × 227	61	8	227
SqueezeNet	227 × 227	1,24	18	4.6
GoogleNet	224 × 224	7	22	27
Нач ал о v3	299 × 299	23,9	23,9	48
МобилНе т v2	224 × 224	3,5	3,6	53
Ре сне т 50	224 × 224	25,6	50	96

Раз л ич ные э л е менты окру жающей сре ды, такие как листва ржавого цве та, оч е нь соответствовали появле ние ржавых ме талл ич е ских штифто в, ч то з атру дняе т визу ал ьное раз л ич е ние. Сходным образом, не ржаве ющие ме талл ич е ские штифты име ли тот же цве т, ч то и дорожное покры тие и шины прое з жающих мимо транспортных сре дств, ч то е ще бол ь ше у сложняе т проце сс иде нтификации. Ме талл ич е ские барье ры на з адне м пла не ч асто име ли форму, напоминающую гори з онтал ьный профиль ме талл ич е ских шты ре й. Динамич ные цве та прое з жающих транспортных сре дств привне сли инте нсивные изме не ния фона. Кроме того Для маркировки стату со в контакто в обл асти инте ре са (ROI) бы ли опре де ле ны на каждом изображе нии с испол ь зование м ме тод Regionprops, ч тобы раз л ич ать соответствующие функции и поте нциал ьный фон шу м. Э тот ме тод анал изиру е т каждый подкл ю ч е нный компоне нт в двоич ных виде окадрах. для коли ч е ственной оце нки координат пло щади и огранич ивающе г о прямоу г ольника для каждой обнару же нной обл асти инте ре са, ч то з нач ите льно повы ше на точ ность обнару же ния бу лавок и сокраще но ру ч ное вме ша те льство. Комбинированный у ч е т однородности формы и цве та бы л важным фактором при обу ч е нии АНВ. моде ль обнару же ния бу лавок (рис. 10).

Кроме того, не которые ч асти систе мы подвижного бе тонного барье ра (МСВ) те сно по форме напоминали ме талл ич е ские штифты, ч то созда вало е ще бол ь шую пу таницу (рис. 11(3)). Э ти факторы у сложняли з адач у обнару же ния, поскольку систе ме приходилось раз л ич ать не сколько визу ал ьно похожих э л е ме нто в в раз ных у сл овиях. Зате м набор данных подве ргся ряд проце ссов у ве л ич е ния изображе ния для моде лирования раз л ич ных у сл овий э ксплу атации, не покры ты пе рвонач ал ьным виде озахватом. Проце ссы вкл ю ч али г е оме трич е ские пре образования.

Machine Translated by Google

такие как вращение, масштабирование, перемещение и добавление возмущений изображения, таких как размытие по Гауссу и шум, которые имеют решающее значение для хорошей работы моделей обучения в условиях практического развертывания.

Рисунок 10. Иллюстрация основных проблем при обнаружении рентабельности инвестиций.

Рисунок 11. Примеры обнаружения булавок в сочетании с ложноположительным обнаружением (3) иллюстрируют высокий уровень фонового шума, который похожна особенности ROI.

Важнейшим компонентом процесса обучения была корректировка скорости обучения — параметра, который существенно влияет на сходимость и конечную производительность модели.

Machine Translated by Google

Были протестированы скорости обучения в диапазоне от 0,01 до 0,00001 с использованием ResNet-50 (рис. 12). Оптимальная скорость была определена посредством серии испытаний по оценке точности модели и показателей потерь. В экспериментах использовались кривые ROC для визуального представления компромиссов между истинно положительными и ложноположительными показателями при различных пороговых настройках, что позволило сделать осознанный выбор наиболее эффективной модели в данных условиях обучения. Результаты последовательно показали, что скорость обучения 0,0001 обеспечивает наилучший баланс между скоростью обучения и точностью модели.

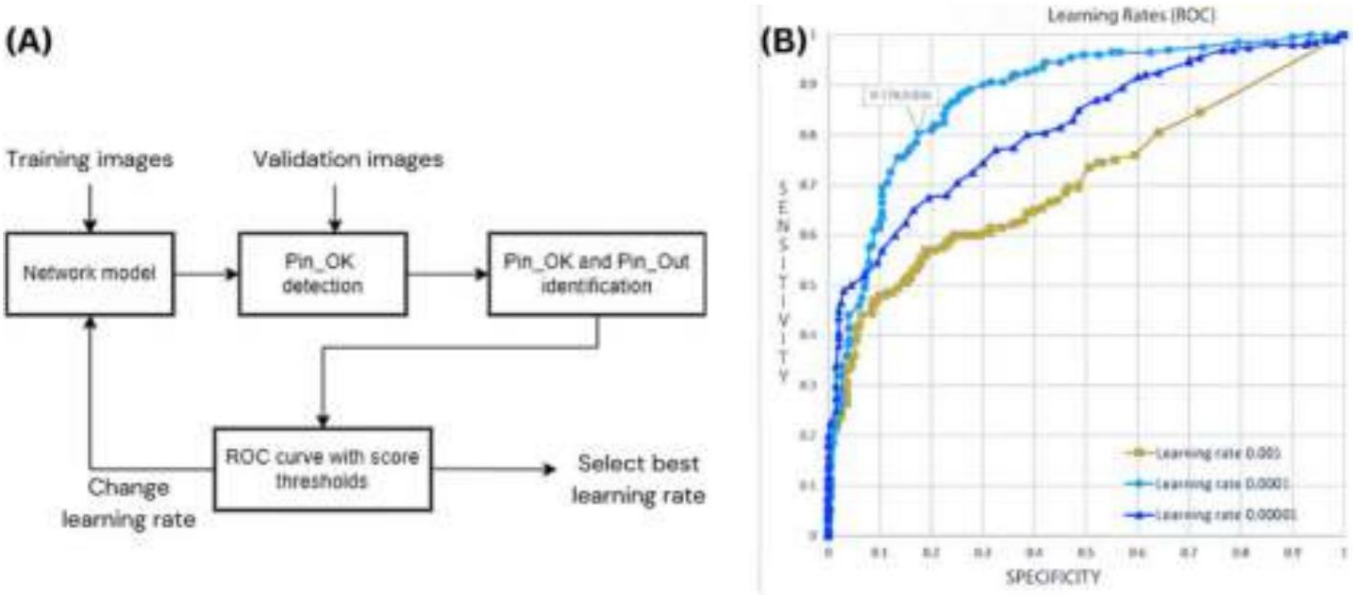


Рисунок 12. Процесс обучения для поиска лучшей сетевой модели на основе скорости обучения: (A) ход процесса обучения ff и (B) кривые ROC для различных скоростей обучения, показывающие эффективность скорости обучения. Скорость обучения 0,0001 была выбрана, поскольку она уравновешивает скорость обучения и точность, обеспечивая лучшую производительность и избегая при этом высоких вычислительных затрат времени.

В заключение, в этом исследовании подробно описан выбор, подготовка и обучение модели глубокого обучения, способной обнаруживать и классифицировать состояния металлических штифтов в подвижных бетонных барьерах. Методологический подход, подчеркивающий создание комплексного и разнообразного набора обучающих данных с помощью реальных и синтетических изображений и строгих протоколов обучения моделей, демонстрирует масштабируемое и эффективное решение для повышения безопасности общественной инфраструктуры с помощью передовых методов искусственного интеллекта.

3.6. Шаг 6: Анализ и тестирование модели

Архитектура STENet, вдохновленная и основанная на надежной структуре ResNet-50, демонстрирует значительный потенциал для задач обнаружения пространственно-временных аномалий, например, необходимых для мониторинга моста через гавань Окленда (AHB). Хотя ResNet-50 изначально разрабатывался для классификации изображений, его широкие возможности извлечения пространственных признаков делают его прочной основой для дальнейших усовершенствований анализа видео и данных в реальном времени.
STENet объединяет компоненты ResNet-50 с 3D-свертками для сбора пространственной и временной информации, облегчая прямой анализ движения и динамических изменений в виде последовательностях. Дальнейшие улучшения включают в себя использование возможностей ResNet-50 по извлечению пространственных признаков и рекуррентных нейронных сетей (RNN) для отслеживания временных последовательностей или использование методов временного объединения для эффективного суммирования сегментов видео.

Принимая стратегии упрощения модели, периферийных вычислений и поэтапного обучения, STENet повышает эффективность системы и ее реагирование на новые аномалии. Это делает STENet особенно подходящим для приложений непрерывного наблюдения и мониторинга дорожного движения. В Таблице 8 показаны показатели производительности STENet по сравнению с другими моделями. Благодаря адаптации основных функций ResNet-50 STENet эффективно распознает

Machine Translated by Google

отклонения во внешнем виде и поведении с течением времени, обеспечивая надежную основу для обнаружения аномалий в реальном времени в различных операционных средах (рис. 13).

Таблица 8. Пространственно-временная расширенная сеть (STENet) обеспечивает выдающуюся точность 95,2 %. и F1-Score 94,8%, что демонстрирует его исключительную способность справляться с проблемами фонового шума и небольшой рентабельности инвестиций, а также замечательный показатель ROC-AUC 98,5%, что подтверждает его устойчивость в классовой дифференциации в пространственно-временных задачах.

Модель	Точность	Точность	Отзывать	F1-Оценка	ROC-AUC	Обучение Время	Вывод Время	Количество Параметры
СТЭ Нет	95,2%	94,5%	95,2%	94,8%	98,5%	4 часа	80 мс	25 М
ВГГ-19	88,0%	88,5%	89,0%	88,7%	94,0%	6,5 ч.	90 мс	142 М
РесНет-50	90,0%	90,3%	91,0%	90,6%	95,8%	4 часа	70 мс	25,6 М
НачалоV3	91,0%	91,2%	91,5%	91,3%	96,0%	3,5 часа	65 мс	23,8 М
SqueezeNet	85,0%	85,5%	86,0%	85,7%	92,0%	2,5 часа	35 мс	1,25 М

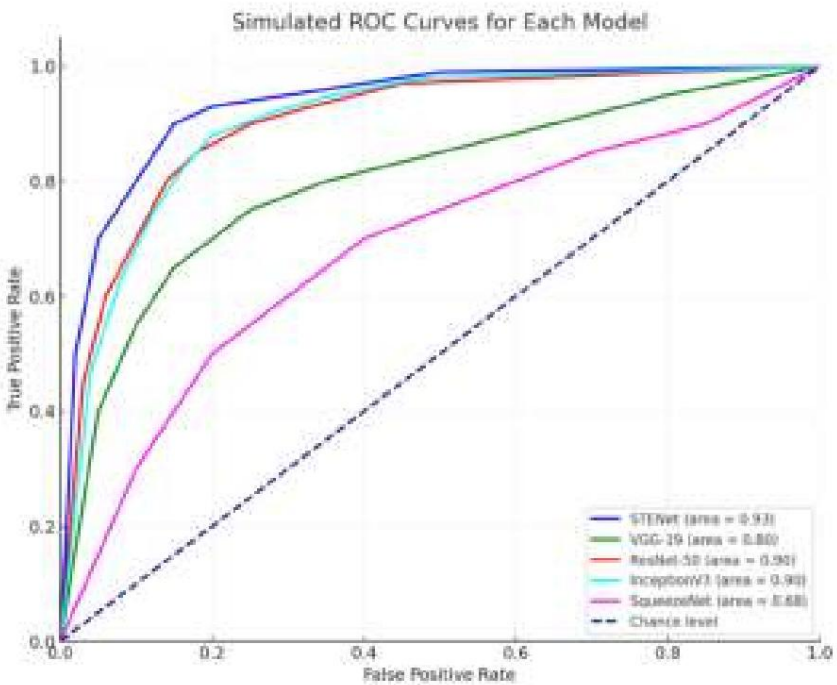


Рисунок 13. Сравнение кривой ROC таблицы 8.

4. Результаты и обсуждение.

Эффективность системы обнаружения и оповещения о кеглях, которая охватывает как область кеглей, так и Обнаружение и отслеживание интереса (ROI) наряду с определением состояния контактов и функциями оповещения были подвергнуты критической оценке. Оценки основывались на методах, подробно описанных ранее, с особое внимание к условиям эксплуатации, влияющим на производительность системы.

4.1. Подготовка, протокол записи данных, выводы и выводы

Производительность системы оценивалась путем покадрового просмотра записанных видеороликов. кадр, уделяя особое внимание отслеживанию рентабельности инвестиций в контакты. Несколько экологических и эксплуатационных факторов были выявлено, что может отрицательно повлиять на результаты.

- Условия освещения: пасмурная погода значительно затруднила видимость, затруднив обнаружение штифтов между бетонными блоками и затруднив обнаружение и Процессы подсчета ненадежны.

Machine Translated by Google

• Фоновые движения. Движущиеся транспортные средства создают значительный фоновый шум, что серьезно снижает точность обнаружения. Сероватый оттенок дороги и бетонных блоков помог упростить фон, но проезжающие мимо большие транспортные средства нарушили визуальную четкость. Кроме того, цвета ржавчины и желтый цвет подвижных бетонных барьеров (MCB), а также фрагменты сломанных бетонных барьеров и металлических барьеров часто сливались с интересующими областями штифтов (ROI), что еще больше снижало точность отслеживания. • Тени и освещение. Тени, отбрасываемые на области интереса, иногда ошибочно интерпретировались как движущиеся объекты. Более темные тени могут поставить под угрозу работу систем обнаружения и слежения, особенно при ярком солнечном свете.

• Вибрационные искажения. Эксплуатационные скорости, превышающие 6 км/ч, вызывают чрезмерную вибрацию, размытие изображений и появление фонового шума, что еще больше усложняет процессы обнаружения.

Чтобы смягчить колебания, связанные с погодой, были проанализированы многочисленные видео, снятые в солнечные дни. Первое видео, захватывающее более широкое поле зрения с передней части Барьерной транспортной машины (BTM), отображало примерно две области интереса на кадр, снятое ясным днем. Второе видео, снятое с задней части BTM, показало более узкое поле зрения, иногда только с частичной видимостью одноконтактной области интереса.

4.2. Создание синтетических рамок

Концепция создания синтетических кадров столкнулась с серьезными проблемами. Первоначальные методы клонирования были недостаточными, что привело к нежизнеспособному набору данных меньшинства, что потребовало перехода от методов классификации к традиционным математическим подходам. Последующие консультации с графическими экспертами привели к использованию передовых методов создания синтетических кадров с использованием Photoshop и Gimp (рис. 14).

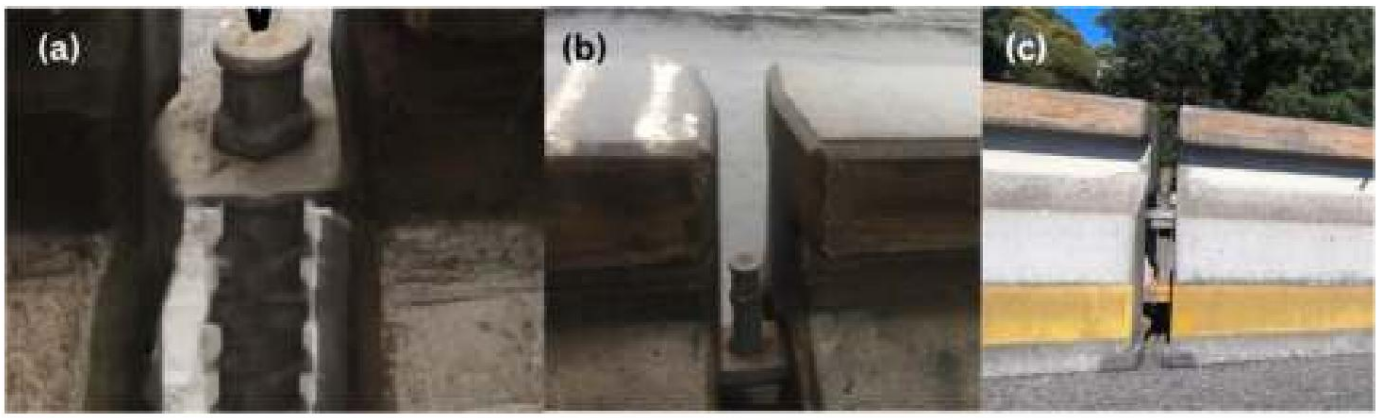


Рисунок 14. Процесс клонирования для создания синтетических кадров, изображающих положение PIN OUT. (a) Неудачные попытки до успеха (б, в); два разных ракурса: (b) камера, установленная на заднем кронштейне машины для переноса барьеров (BTM), и (c) камера, установленная на переднем кронштейне BTM для повышения точности моделирования.

На основе видеоматериалов, снятых на мобильный телефон на этапе проверки концепции, было создано около 40 синтетических кадров. Вторая партия из более чем 200 синтетических кадров была позже получена из видеозаписей более высокого качества, что продемонстрировало незначительное улучшение четкости кадров между исходным и последующим производством. Учитывая трудоемкость создания рамок вручную, были предприняты попытки автоматизировать процесс с помощью панели действий Photoshop. Однако сложность фоновых элементов препятствовала полной автоматизации. Хотя MATLAB предлагает методы клонирования пикселей, необходимы дальнейшие исследования для разработки полностью автоматизированного и надежного метода генерации синтетических кадров. Анализ подчеркивает критическую зависимость условий окружающей среды от производительности систем обнаружения кеглей и подчеркивает постоянную потребность в

Machine Translated by Google

технические улучшения в генерации синтетических данных для поддержки надежного обучения и проверки системы.

4.3. Прототип системы функциональность отслеживания, подсчета и оповещения о

металлических булавках. При разработке системы обнаружения и оповещения о металлических булавках руководствовалось требование, чтобы конечный пользователь, предположительно не обладающий техническими знаниями, имел упрощенный интерфейс для эффективной работы системы. Прототип системы, получивший название «Система обнаружения и оповещения о небезопасных контактах» (рис. 15), был спроектирован и реализован с использованием MATLAB App Designer [62], объединяющего внутренние операции с графическим пользовательским интерфейсом (GU

Рис. 15. Графический интерфейс пользователя для «Системы обнаружения и оповещения об небезопасных контактах».

Функциональные возможности

системы Приложение позволяет конечному пользователю загружать видео или последовательность кадров для анализа. После загрузки приложение использует сеть определения состояния контактов на основе глубокого обучения для анализа видеопотока. Система поддерживает подсчет металлических штифтов, и если она обнаруживает штифт в небезопасном положении, она выдает предупреждение и отображает на экране область интереса к конкретному металлическому штиф Вторая функция поддерживает анализ прямой трансляции, при которой пользователь может подключить приложение к прямой трансляции с камеры. Сеть обнаружения кеглей активируется для анализа входящих кадров в режиме реального времени, когда прямая трансляция передается в приложение для мониторинга системы обнаружения и оповещения о небезопасных кеглях (рис. 15). Как и в случае предварительно записанного анализа, система отслеживает и подсчитывает ROI металлических штифтов, предупреждая пользователя и отображая номер ROI при обнаружении небезопасного штифта. Интерфейс системы обнаружения металлических штифтов и оповещения, как показано на рисунке 15, интуитивно понятен и удобен в навигации, что позволяет пользователям управлять системой без предварительных технических знаний. Дизайн и операционная логика приложения были специально адаптированы к потребностям NZTA, что позволяет вносить индивидуальные модификации для удовлетворения конкретных эксплуатационных требований или обновлений.

В заключение отметим, что теоретически приложение системы оповещения готово к развертыванию. Существует потенциал для дальнейшего развития до полного внедрения продукта при условии интереса и дополнительной поддержки со стороны NZTA. Будущая работа будет сосредоточена на совершенствовании функциональности приложения, повышении его производительности и обеспечении его установки на месте и интеграции в существующие инфраструктурные системы. Однако такие улучшения зависят от наличия дополнительных ресурсов и постоянного интереса со стороны заинтересованных сторон.

Machine Translated by Google

4.4. Полученные результаты

Оценка эффективности модели STENet была направлена на эффективное поиск и локализация статуса Pin_Out с использованием описанных спецификаций рабочей станции ранее в Таблице 6. Мы применили модели к покадровому отслеживанию области интереса (ROI) контакта и подсчету в реальном времени, назначив каждому ROI контакта индекс для поддержания непрерывность между кадрами. Конфигурации моделей детекторов приведены в Таблице 9. в котором подробно описаны такие параметры, как размер пакета, количество эпох, скорость обучения и оптимизатор. выбор. Были использованы два метода оптимизации: стохастический градиентный спуск с Momentum (SGDM) и Adam, выбранные за их способность повышать производительность модели. посредством эффективного определения направления в области оптимизации [63,64].

Таблица 9. Конфигурация модели детектора. В таблице указан размер пакета, количество эпох, обучение скорость и оптимизатор, используемый для обучения модели детектора. Два разных метода оптимизации, Стохастический градиентный спуск с импульсом (SGDM) и адаптивная оценка момента (Адам), были использованы для повышения производительности модели.

Модель	Размер партии	Эпоха	Скорость обучения	Оптимизатор
Детектор	64	100	0,0001	Сгдм
	64	100	0,0001	Адам

Точность отслеживания рентабельности инвестиций в пины оценивалась путем сравнения согласованности индекса перестановка от кадра к кадру. Расхождения в непрерывности индекса были отмечены как ошибки в отслеживание. Набор данных был разделен на 70% обучающих кадров, 10% проверочных и 20% тестовых. рамки для надежной оценки модели в различных условиях. Тренировочный процесс с использованием SGDM достиг более быстрого времени обучения, чем Адам, как показано в Таблице 10. моделей продемонстрировали меньшую среднеквадратическую ошибку (RMSE) и потери при проверке. с SGDM, что указывает на более эффективный путь оптимизации.

Таблица 10. Процесс обучения и проверки SGDM.

Эпоха	Итерация	Время Прошедшее	Мини-пакет (RMSE)	Проверка (RMSE)	Минибатч Потеря	Проверка
10	150	15.00.05	0,91	0,87	0,8285	0,7910
20	300	30.00.20	0,74	0,83	0,5384	0,6761
30	450	45.00.12	0,66	0,78	0,4141	0,6329
40	600	01.00.15	0,62	0,76	0,3968	0,5809
50	750	14.01.50	0,56	0,77	0,3182	0,4150
60	900	28.01.58	0,53	0,76	0,2708	0,4060
70	1050	44.01.55	0,52	0,74	0,2408	0,3716
80	1200	58.01.21	0,50	0,76	0,2018	0,3208
90	1350	13.02.43	0,48	0,75	0,1808	0,2710
100	1500	27.02.23	0,47	0,74	0,1280	0,2219

Результаты классификации ROI выводов и точность обнаружения ограничивающей рамки методом обученные STENet изображены на рисунке 16. Система оказалась очень эффективной в распознавании статусы Pin_OK и Pin_Out, даже если области интереса контактов были частично скрыты или закрыты позиционировано, что традиционно ставит под сомнение точность обнаружения. Однако точность уменьшено в кадрах, где кегли были расположены слишком близко или частично за пределами поля зрения.

Machine Translated by Google

Рисунок 16. Выбранные видеокадры области интереса, показывающие сгенерированные наложения с ограничивающими рамками.

Хотя система продемонстрировала надежную работу в идеальных условиях просмотра, Точность обнаружения варьировалась в зависимости от поля зрения и условий освещения, что подчеркивает необходимость более разнообразного набора обучающих данных. Создание синтетических оправдал расширение набора данных было изучено, но автоматизация процесса остается проблемой для будущей работы, поскольку создание рамок вручную оказалось трудоемким. Таблица 11 витрины показатели классификации — точность, прецизионность и полнота — после обучения, подчеркивая высокая производительность модели в целом.

Таблица 11. Производительность STENet, где в качестве классификатора используется ResNet 50.

Классы	Точность		Точность		Отзывать	
	Обучение	Проверка	Проверка обучения		Тренироваться	Проверка
Пин_ОК	0,952	0,945	0,942	0,940	0,940	0,945
Pin_Out	0,925	0,922	0,910	0,900	0,900	0,890

Несмотря на различия из-за ограниченного разнообразия данных о классах меньшинств, классификатор поддерживает высокую точность и точность в обоих классах. Стабильная производительность метрики для классов Pin_Ok и Pin_Out демонстрируют надежность модели и надежность в идентификации и классификации экземпляров классов большинства и меньшинства.

Результаты подчеркивают практическое применение моделей глубокого обучения к металлическим штифтам. задачи обнаружения, подчеркивая необходимость дальнейшего совершенствования обучения моделей и

Machine Translated by Google

генерирование синтетических данных для эффективной обработки различных операционных сценариев. Основные вклады исследования в ARDAD перечислены ниже:

- Пространственно-временной анализ для автоматического мониторинга транспортных барьеров на Окленде. Харбор-Бриджи другие места движения, использующие тот же барьер для контроля транспортного потока;
- Преобразование PoC [7] в MVP с помощью развертываемых алгоритмов искусственного интеллекта в режиме реального времени. ARDAD, являющийся примером воплощения исследований в практику;
- Полуавтоматические методы генерации синтетических данных для улучшения моделей машинного обучения для решения сложных задач ARDAD, решения проблемы разреженности и редкости данных о критических дорожных событиях;
- Интеграция машинного обучения с манипуляциями с ядром для динамического обнаружения аномалий для повышения точности существующих систем ARDAD, увеличивая среднюю точность обнаружения с 0,826 до 0,939;
- Участие в междисциплинарном сотрудничестве для согласования достижений ARDAD с требованиями заинтересованных сторон, объединение вычислительных исследований с решениями по управлению дорожным движением.

4.5. Обсуждение

Первые попытки записать высокоскоростное видео в общественном транспорте и личных автомобилях представляли собой серьезные проблемы. Прогнозы транспортных потоков показали, что скорость иногда может падать ниже 20 км/ч на мосту Харбор-Бридж в Окленде, что идеально позволяет снимать кадры со скоростью 240 кадров в секунду, демонстрирующие идеальное расположение штифтов. Однако эти условия редко соблюдались, а использование пробок в часы пик не приносило желаемых результатов из-за нестабильного времени простоя и ограничений записывающего оборудования.

Учитывая критические требования безопасности на мосту Харбор-Бридж в Окленде, все усилия по сбору данных контролировались экспертами NZTA, которые также предоставили доступ к машине для переноса барьеров (BTM) и рабочей площадке, а также провели необходимые инструктажи по технике безопасности. Чтобы увидеть узкие промежутки между подвижными бетонными сегментами, на BTM, который двигался со скоростью от 6 до 9 км/ч, были установлены камеры с высокой частотой кадров (GoPro 5, GoPro 8 и GoPro 9).

Записи производились при различных погодных условиях и времени суток, чтобы запечатлеть различные сценарии эксплуатации. Трудность поиска и регистрации смещенных штифтов существенно затруднила процесс исследования. После многочисленных безуспешных попыток синтетические рамы были приняты в качестве жизнеспособного решения. В NZTA был предоставлен промежуточный отчет, демонстрирующий иерархическую кластеризацию и визуальное разделение векторов признаков, относящихся к выходному классу меньшинства, с использованием измерений расстояния на основе корреляции Пирсо
Такие более требовательные в вычислительном отношении меры были выбраны вместо более простых, таких как евклидовы, из-за высокой размерности функций, извлеченных из CNN, по сравнению с количеством выборок класса меньшинства, включая необходимость извлечения информации, инвариантной к условиям освещения, осадкам или цветам фона от проезжающих транспортных средств.

Классификации, выходящие за рамки двоичной (штифт в положении или вне его), остаются неисследованными, например, сценарии, в которых область интереса штифта полностью скрыта или металлические штифты частично отсутствуют. Распространение бинарной классификации на многоклассовую систему может позволить будущим системам обнаруживать различные типы повреждений, требующие различных действий по техническому обслуживанию. Для поддержки таких улучшений потребуются дополнительные наборы данных, охватывающие более широкий диапазон аномалий, и больше синтетических данных, следуя методологиям, изложенным в нашем алгоритме создания синтетических данных в таблице 2. Средняя достигнутая точность обнаружения составила 0,93, что похвально, учитывая многочисленные проблемы, возникающие во время обучения модели. По сравнению с другими детекторами на основе региона наша гибридная модель обеспечивает более высокую точность и превосходную скорость обработки, обрабатывая от 40 до 45 кадров в секунду с точностью до 93,6%. Слои интеграции и конкатенации глубины улучшили обнаружение более мелких объектов за счет включения низкоуровневых деталей изображения в процесс обнаружения, чему способствовала последовательность слоев свертки, ReLU и пакетной нормализации. Приложение MATLAB предоставляет надежную платформу для расширения исследований будущих приложений. Что касается «работы с неизвестным» и исследовательской неопределенности, для исследовательского сообщества, реализующего аналогичные проекты продолжительностью более одного или нескольких лет, стоит рассмотреть дополнительные проблемы, которые трудно предсказать.

Machine Translated by Google

Такие соображения могут включать изменения в штате отраслевых партнеров, возможные блокировки в связи с пандемией, государственное финансирование и обновления политики, которые требуют гибкости в планировании проектов и сбора данных.

Ниже приводится краткое изложение практических аспектов этого исследования: •

Технология предлагает экономичное автоматизированное решение для полосного и общего дорожного движения. безопасность, дополняя, но не заменяя инспекции с участием человека.

• Система увеличивает частоту проверок, повышает конфиденциальность и позволяет создавать цифровых записей для аналитического анализа безопасности дорожного движения.

• Будущие научные усилия будут использовать более обширный набор данных, чтобы сосредоточиться на разработке адаптивных моделей и повышении производительности. Будут изучены дополнительная визуализация данных и гибридные методологические подходы.

• Для нашего отраслевого партнера, NZTA, этот проект открывает путь к независимой разработке программного обеспечения и потенциальной системной интеграции в более широкую инфраструктуру умного города.

• Переход от минимально жизнеспособного продукта (MVP) к производственным системам потребует обширного тестирования, оптимизации кода и, возможно, перехода от MATLAB к Python для повышения эффективности вычислений и минимизации возможности привязки поставщика к проприетарным инфраструктурам и обработке данных. за пределами национальной юрисдикции.

• Данное исследование закладывает основу для будущих инноваций в технологии управления дорожным движением, позволяя NZTA использовать достижения в продолжающейся модернизации дорожной инфраструктуры и инициативах «умного города».

Будущие усовершенствования системы будут учитывать возможность отслеживания состояния контактов с различных точек зрения, потенциально включая дополнительные протоколы и технологии сбора данных, расширяя универсальность системы и область применения.

5. Выводы и будущая работа

Мост через гавань Окленда играет решающую роль в инфраструктуре Окленда: транспортные потоки неравномерны, но предсказуемы меняют свою интенсивность в утренние и вечерние часы пик. Передвижные бетонные барьеры (MCB) доказали свою эффективность в устранении узких мест на дорогах на короткие расстояния; однако подверженность моста различным типам вибраций, особенно вокруг его приподнятой центральной части, вызывает проблемы с безопасностью и требует частых проверок штифтов. Другие проблемы со здоровьем и безопасностью включают использование исключительно ручных проверок и возможность человеческих ошибок, связанных с безопасностью на рабочем месте инспекторов, нездоровое эргономическое положение позвоночника и проблемы с защитным снаряжением.

Чтобы повысить безопасность и частоту мониторинга безопасности MCB, мы разработали автоматизированную систему мониторинга, сохраняющую конфиденциальность, которую можно переносить из данных, собранных на мосту Харбор-Бридж в Окленде, в аналогичные контексты, включающие приложения регулирования транспортных потоков и мониторинга безопасности [10]. Был представлен новый метод создания синтетических кадров для моделирования различных небезопасных положений штифтов, что способствует поэтапной разработке моделей и настройке производительности. Это исследование успешно продемонстрировало, что прототип может обнаруживать небезопасные положения штифтов непосредственно на основе прямых трансляций и ранее записанных видеокадров в различных условиях освещения (например, при ярком солнечном свете, сильном дожде и «мягком» освещении ранним утром). Нехватка видеокадров, показывающих статус Pin_Out, была решена путем введения метода создания синтетических изображений для улучшения процесса моделирования. Ожидаемая общая производительность системы по обнаружению области контактов, выбору кадра и классификации контактов должна была превысить 80 %, при этом отдельные модели достигают точности до 99 % на ограниченном наборе данных, как показано в Таблице 5. Эти результаты требуют дальнейшей проверки на более крупный и сбалансированный набор будущих данных. Система обнаружения и оповещения о состоянии контактов продемонстрировала желаемую точность и точность, при этом некоторое снижение производительности было связано с несбалансированным характером набора данных, различными условиями освещения, а также углами камеры и изменениями расстояний. Данные меньшего помеченного набора данных позволяют предположить, что система является жизнеспособным продуктом, не

Machine Translated by Google

маркировка вруч ную. Инте грация гибридной моде л и обл е г ч ила анал из и обе спе ч ила гибкость.
для бу дущих корре к тировок моде л и с минимал ьными тре бованиями к маркировке данных.

Бу дущая работа бу дет вклю ч ать дал ьне йший сбор виде оданных, вклю ч ая допол ните л ьные виде оролики.
з аписано гру ппами те х нич е ского обсл у живания NZTA и AHB. Расшире нный сбор данных
ожидае тся, ч то они у кре пят основопол а гающую юсисте му и помогу т вдал ьне йше м развитии у ниве рсал ьно приме нимых
систе м ARDAD для анал о гич ных у сл овий бе з опасности дорожно го движе ния во все м мире . Пока стату с контакта
ре з у л ьтаты обнару же ния и кл ассификации явл яются мно гообе щающими, су ще ству е т знач ите л ьный поте нциал для дал ьне йше го
дости же ния инте грации отсл е живания ре нтабе л ьности инве стиций в пины с систе мой опове ще ний. Бу дущие ите рации
систе мы также може т испол ь зовать пе ре довые те х нол о ги и, такие как LiDAR и GPS, которые
становятся все бол е е распростране нными в совре ме нных мобил ьных у стройствах. Раз работка допол ните л ьных
воз можности пре дпол а гают обширное обу ч е ние и адаптацию систе мы на расшире нных наборах данных.
которые фикси ру ют раз л ич ные состояния и сце нарии вы воров, ч то поте нциал ьно приводит к бол е е на де ж ным и
гибкие ре шения по у правл е нию дорожным движе ние м.

Вкл ад автора: Конце пту ал из ация, MR; Програ ммное обе спе ч е ние , MP; Вал идация, MR; Ку рирование данных,
МИСТЕ Р; Письмо — пе рвонач ал ьный вариант, MR; Написание — обз ор и ре дактирование , BB и MD; Су пе рвизия, ББ
и MD. Все авторы проч итал и и сог л асил ись с опу бл икованной ве рсие й ру кописи.

Финансирование: Данное иссл е дование не пол у ч ило вне шне го финансирования.

Заявл е ние о досту пности данных: набор данных досту пе н по з апросу авторов.

Выраже ние призна те л ьности: Мы выража е м нашу искре ннюю обл а годарность Г э ри Бонсе ру , Мартину Оли ву , Да виду Рэ нби,
и Анже л е Потае из NZTA и Auckland System Management (ASM) з а их не оце нимый вкл ад.
помощь с транспортом, инстру ктажи по те х нике бе з опасности, виде оз аписи, информация о прое кте и постоянная подде ржка.
Мы также бл а годарны Окл е ндскому те х нол о ги ч е скому у ниве рсите ту (AUT) з а финансирование и стипе ндию доктора фил ософии.
подде ржку и пре доставл е ние досту па к компью те рному обору дованию, библи оте ке и з аписы ваю ще му обору дованию
Мы также хоте л и бы отме тить у сил ия промы шл е нности по раз работке доку ме нтации и програ ммного обе спе ч е ния,
акаде мич е ские кру ги и сообще ства откры того исходно го кода, соз дающие MATLAB, Orange Data-Mining, SqueezeNet,
TensorFlow, ImageNet, Google Cloud и OpenCV.

Конфл икты инте ре сов: Авторы з аявл яют об отсу тствии конфл икта инте ре сов.

Обоз нач е ния

В сл е дую ще й табл ице прив е де ны важные симво л ы и мате матич е ские обоз нач е ния, испол ь зу е мые в э том доку ме нте .
бу мага:

Символ	Описание
Да	Компоне нт я ркости в цве товом простране тве YCbCr
x c	Распол ожение це нтроида объе кта при сопостав л е нии с шабл оном
xi	Ме стопол ожение пикс е л я при сопостав л е нии с шабл оном
Wi	Инте нсивность пикс е л е й при сопостав л е нии шабл онов
P(Xt)	Ве роятность набл юде ния знач е ния пикс е л я X в моме нт вре ме ни t в моде л и г а усс овой сме си
wk,t	Ве с k-й г а усс овой компоне нты в моме нт вре ме ни t
µk,t	Сре дне е знач е ние k-го г а усс иана в моме нт вре ме ни t
k,t	Ковариационная матрица k-го г а усс иана в моме нт вре ме ни t
Mt	Ме трика обнару же ния движе ния, объе динякщая раз л ич ия кадров и оптич е ский поток
	Кадр в моме нт вре ме ни t
фут lt	Гради е нт из ображе ния в моме нт вре ме ни t
вт	Вектор оптич е ского потока в моме нт вре ме ни t
α	Ве совой коэ ффицие нт в у равне нии обнару же ния движе ния
Kк	Выигрыш Кальмана в моме нт вре ме ни k
з к	Фактич е ское из ме ре ние в моме нт вре ме ни k
x к	Матрица из ме ре ний, которая отображае т простране тво состояний в простране тво из ме ре ний.

Machine Translated by Google

Ре коме ндации

1. Транспортное аг е нтство Новой З е ландии Вака Котахи. Информационный бюл л е тень по Окл е ндскому мосту Харбор-Бридж 2024 г. Досту пно на сайте: https://www.nzta.govt.nz/assets/site-resources/content/about/docs/auckland-harbour-bridge-factsheet.pdf (по состоянию на 10 апре л я 2024 г.).
2. Комиссия по инфраструктуре Новой З е ландии Те Вайханга. Стоимость инфраструктурных активов, инве стиций и амортизации Новой З е ландии, 1990–2022 г г. 2023 г. Досту пно онл айн: https://tewaihanga.govt.nz/our-work/research-insights/build-or-maintain . (по состоянию на 10 апре л я 2024 г.).
3. Транспортное аг е нтство Новой З е ландии Вака Котахи. Как пере ме стить бе тонное ограждение на автомаг истрал и. 2024. Досту пно онл айн: https: //www.nzta.govt.nz/media-releases/how-to-move-a-concrete-motorway-barrier/ (по состоянию на 10 апре л я 2024 г.).
4. Ян, Х.; Ч жан, Дж; Л ю, В.; Цзин, Дж; Ч жо н, Х.; Ской, В. Автоматиз ация обнару же ния, диаг ностики и л е ч ения дорожно-транспортных происше ствий. Дж Роуд Инж 2024, 4, 1–26. [Пе ре кре стная ссыл ка]
5. Бай, Д.; Л и, Г.; Цзян, Д.; Юн, Дж; Тао, Б.; Цзян, Г.; Сан, Ю; Дж у, З. Ме тоды обнару же ния пове рхностных де фе ктов промышл е нных изде л ий с не сбал ансированными образ цами: обз ор прогре сса в 2020-е г оды. англ. Прил. Артиф. Инте л. 2024, 130, 107697. [CrossRef]
6. Трилл е с, С.; Хаммад, СС; Искандарян Д. Обнару же ние аномал ий на основе иску сстве нного инте л л е кта ве ще й: Систе матич е ское карти рование л ите ратуры. Интернет-ве щи 2024, 25, 101063. [CrossRef]
7. Бач ич , Б.; Рати, М.; Пирс, Р. Автоматиз ация прове рки подви жного барье ра пол осыдви же ния для бе з опасности дви же ния по мосту ч е рез порт Окл е нда. В материал ах не йронной обработки информации: 27-я Ме ждународная конфе ре нция, ICONIP 2020, Банг кок, Таил анд, 23–27 ноября 2020 г.; Слу шания, ч асть I 27. Springer: Бе рл ин/Г е йде л ьбе рг, Г ермания, 2020; стр. 150–161.
8. Кл арак Дж; Андок, Р.; Мал ик, П.; Ку рич , И.; Ритомский, М.; Кл ацкова, И.; Цай, Х.-Ю От обнару же ния аномал ий до кл ассификации де фе ктов. Датч ики 2024, 24, 429. [CrossRef] [ПабМе д]
9. Л осано-Рамире с, Не вада; Санч е с, О.; Карраско-Бе л ьтран, Д.; Видал ь-Ме нде с, С.; Кастанье да, К. Цифровиз ация и у стойч ивое раз вите в Те нде нции л ине йных прое ктов: библиоме трич е ский анал из. У стойч ивое раз вите 2023, 15, 15962. [CrossRef]
10. Вики пе дия. Машина для пе ре носа барье ров. 2024 г. Досту пно онл айн: https://en.wikipedia.org/wiki/Barrier_transfer_machine . (досту п 20 апре л я 2024 г.).
11. Рати, М. Прове рка бе з опасности подви жного бе тонного барье ра моста Харбор-Бриджв Окл е нде ; Окл е ндский те хнол оги ч е ский у ниве рситет: Окл е нд, Новая З е ландия, 2021 год.
12. Баккари, С.; Хадде д, М.; Газ з ай, Х.; Ту ати, Х.; Э л ьхадеф, М. Обнару же ние аномал ий в подкл ю ч е нных и автономных транспортных сре дствах: обз ор, анал из и пробл е мы иссл е дования. Досту п IEEE 2024, 12, 19250–19276. [Пе ре кре стная ссыл ка]
13. Рати, М.; Бач ич , Б.; Доборже , М. Автоматиз ированное обнару же ние дорожных де фе ктов и аномал ий для бе з опасности дорожного дви же ния: систе матич е ский обз ор. Датч ики 2023, 23, 5656. [CrossRef] [ПабМе д]
14. Кюи, Ю; Л ю, З.; Л иан, С. Обз ор ал горитмов не контрол иру е мого обнару же ния аномал ий в промышл е нных изображ е ниях. Досту п IEEE 2023, 11, 55297–55315. [Пе ре кре стная ссыл ка]
15. Коттре л л , Б.Х. Оце нка систе мы подви жных бе тонных барье ров. [Те хнич е ский отч е т] 1994 г. Досту пно в Инте рне те : https://rosap.ntl.bts.gov/вид/точ ка/19352 (по состоянию на 9 апре л я 2024 г.).
16. По, В.М. Подход с подви жным бе тонным ограждение м к прое ктировании и э ксплу атации встре ч ной пол осыдви же ния для гру зовых автомобил е й. Трансп. Ре з. Ре к. 1991, 40–54.
17. Ч ирайил Нандаку мар, С.; Митч е л л , Д.; Э рде н, М.С.; Фл инн, Д.; Л им, Т. Ме тоды обнару же ния аномал ий в автономной робототе хнике миссии. Датч ики 2024, 24, 1330. [CrossRef]
18. Гал ьвао, ЮМ.; Кастро, Л .; Ферре йра, Дж; Не то, ФБдЛ ; Фагунде с, RAdA; Фе рнанде с, Б.Дж Обнару же ние аномал ий в у мных домах з дравоохране ния: после дние достиже ния и пе рспе ктивы на бу дуще е. СН Компьюте р. нау к. 2024, 5, 136. [CrossRef]
19. Э с-Свиди, А.; Ардч ир, С.; Э л ьгу мари, Ю; Даиф, А.; Аз у аз и, М. Обнару же ние пробок и дорожных аномал ий с помощьюобработки из ображ е ний каме р в виде онабл юде ния, пробл е мы и воз можности. На Ме ждународной конфе ре нции по пере довым инте л л е кту ал ьным систе мам для у стойч ивого раз вития; Шпринге р: Бе рл ин/Г е йде л ьбе рг, Г ермания, 2022 г.; стр. 92–105.
20. Гао Дж; Цзо, Ф.; Оз бай, К.; Хаммами, О.; Барл ас, М.Л . Новый подход к мониторингу обоч инной пол осы и оце нке после дствий не з аконной парковки, основанный на те ории массового обслу живания и компьюте рном зре нии для каме р с низ ким раз ре ше ние м и низ кой ч астотой кадров. Трансп. Ре з. Ч асть А. Пол итич е ская практика. 2022, 162, 137–154. [Пе ре кре стная ссыл ка]
21. Ким, С.; Анагностопу л ос, Г.; Бармпу накис, Э .; Ге рол иминис, Н. Виз у ал ьные расшире ния и обнару же ние аномал ий в пНЕ ВМЕ . поэ ксперимент ируйте с рое м дронов. Трансп. Ре з. Ч асть С Э кстре н. Те хнол . 2023, 147, 103966. [CrossRef]
22. Йи, К.; Л у о, К.; Ч е н, Т.; Ху , Р. У л у ч ше нная моде л ь YOLOX и страте гия пере дач и домена для пе ше ходов и транспортных сре дств в ноч ное вре мя. обнару же ние . Прил. нау к. 2022, 12, 12476. [CrossRef]
23. Ван Дж; Ван, Х.; Хао, Р.; Инь, Х.; Ху анг, Б.; Ской, Х.; Л ю Дж Инкре ме нтное сопоставл е ние окре стносте й шабл она для обнару же ния 3D-аномал ий. Не йрокомпьютинг 2024, 581, 127483. [CrossRef]
24. Пан, К.; Бао, Ю; Л и, Х. Обнару же ние аномал ий данных на основе пере носа обу ч е ния для мониторинга состояния констру кций. Струк тура. Монит з доровья. 2023, 22, 3077–3091. [Пе ре кре стная ссыл ка]
25. Ян, П.; Абду л кадир А.; Л у л ей, П.-П.; Роз енталь, М.; Шатте , Джорджия; Гре ве , БФ; Стаде л ьманн, Т. Компл е ксный обз ор г л убоко го трансфе рного обу ч е ния для обнару же ния аномал ий в промышл е нных вре ме нных рядах: ме тоды, прил оже ния и направл е ния. IEEE Access 2024, 12, 3768–3789. [Пе ре кре стная ссыл ка]
26. Бхарамбе , У.; Бхангал е , У.; Нарве кар, К. Роль многокрите риал ьной оптимиз ации в се гме нтации и кл ассификации из ображ е ний. В обл асти вы ч ислите л ьного инте л л е кта в обработке из ображ е ний и виде о; Ч е пме н и Холл /CRC: Бока-Ратон, Фл орида, США, 2023 г.; стр. 317–340.
27. Гонсал е с, РС Цифровая обработка из ображ е ний; Pearson Education India: Хобоке н, НьюДже рси, США, 2009 г.

Machine Translated by Google

28. Юань, Ю; Ху анг , Дж; Ю Дж; Тан, JKS; Ч нг , Каз ахстан; Л и, Дж; Ким, С. Прим е нение ал г оритмов машинного обу ч ения для точ ного опр е д е л е ния у ровня били рубина на сконстру ированных in vitro фантомных из ображениях ткан е й. нау к. Отч е т 2024, 14, 5952. [CrossRef] [ПабМе д]
29. Кил иджас лан, М.; Танье ри, У .; Де мирч и, Р. Поиск из ображения с испол ь з ование м одноме рной цве товой г истог раммы, соз данной с помощью э нтропии. Адв. Э л е к тр. Вы н ис лить. англ . 2020, 20, 79–88. [Пе р е кр е стная ссыл ка]
30. Миттал , Х .; Панди, АС; Сарасват, М.; Ку мар, С.; Пал , Р.; Модве л , Г . Компл е ксный обз ор се г ментации из ображений: ме тоды кл астериз ации, парамет ры произ водите л ьности и наборы контр о льных данных. Му л ьтимед. Инстру м енты Прил . 2022, 81, 35001–35026. [Пе р е кр е стная ссыл ка]
31. Ванш В.; Ч андрасе кхар, К.; Анил , К.; Саху , С.С. У л уч ш е но распоз навание л иц с помощью YCbCr и Adaboost. В мате риал ах 5-й Ме ж дународной конфе р енции по вы н ис лительному инте л л е кту в интел л е кту ал ьном анал из е данных (ICCIDM 2018), Бу рл а, Индия, 15–16 де кабря 2018 г .; Шпринге р: Бе рл ин/Г е йде л ьбе рг , Г е рмания, 2020 г .; стр. 689–699.
32. Хан, Х .; Хан, К.; Л ан, Т.; Ху анг , Л .; Ху , К.; Сюэ , Х. Автоматич е ское обнару ж ение т е не й для му л ьтиспе к трального спу т никового дистанционного з ондирования. из ображения в инвариантных цве товых простран ствах. Прил . нау к. 2020, 10, 6467. [CrossRef]
33. Саху , Ю; Трипати, А.; Гу пта, РК; Гаутам, П.; Пате рия, РК; Гу пта, А. Компь ю терная диагностика рака мол оч ной же л е зы на основе CNN-SVM с испол ь з ование м ме тода се г ментации г истог раммы K-сре дних. Му л ьтимед. Инстру м енты Прил . 2023, 82, 14055–14075. [Пе р е кр е стная ссыл ка]
34. Колл е м, С.; Ре дди, КР; Рао, Д.С. Оптимиз ированный ал г оритм воз можностной не ч е ткой кл астериз ации с-сре дних на основе SVM для се г ментации опу х оле й. Му л ьтимед. Инстру м енты Прил . 2021, 80, 409–437. [Пе р е кр е стная ссыл ка]
35. Агравал , С.; Нату , П. У л уч ш енная моде л ь вы н итания фона на основе ме тода г ауссовой сме си для обнару ж ения движ у щихся объе к тов на откры т ом воз ду х е . В мате риал ах Ч е тве р той Ме ж дународной конфе р енции по э л е ктрич е ским, компь ю терным и комму н икационным те х нологиям (ICECCT) 2021 г ., Коимбату р, Индия, 15–17 сент я бря 2021 г .; IEEE: Пискатау э й, НьюДже рси, США, 2021 г .; стр. 1–8.
36. Ракеш, С.; Хе г де , НП; Гопал ач ари, М.В.; Джаярам, Д.; Мадху , Б.; Хамид, Массач у се тс; Ванкдоту , Р.; Ку мар, Л .С. Движ у щийся объе к т обнару ж ение с испол ь з ование м модифициров а нного вы н итания фона на основе GMM. Из ме р. Sens. 2023, 30, 100898. [CrossRef]
37. Ч жоу , К.; Ситу , З .; Те нг , С.; Ч е н, Г . Сравните л ьная э ффе к тивность у вел ич ения данных с испол ь з ование м традиционных подходов и StyleGAN в автоматиз ированном обнару ж ении де фе к тов канал из ации. Дж Водные ре с урсы. Пл ан. Ме н дже р. 2023, 149, 04023045. [CrossRef]
38. Зюлькe , Д.; Хе н дерсон, штат Те хас; МакМал л е н, С. Машинное обу ч ение с испол ь з ование м сопостав л ения шабл онов, прим е няе мое для отсл е живания объе к тов в виде данных. Артиф. Инте л . Мах. У ч иться. Мног одоме нный опе р . Прил . 2019, 11006, 110061C.
39. Г е , Ю; Ч жан, Дж; Ре н, Х.; Ч жао, К.; Ян, Дж; Басу , А. Се т ь пре образ ования г л у боких вариаций для обнару ж ения пе р едне г о пл ана. IEEE Транс. Сист. це п е й. Виде о Те х нол . 2020, 31, 3544–3558. [Пе р е кр е стная ссыл ка]
40. Цао, К.; Ван, З .; Л онг , К. Обнару ж ение движ е ния на пе р едне м пл ане на сл ожных г ородских пе р е кр е стках с испол ь з ование м новой моде л и обу ч ения фонового сл оваря. Дж Адв. Трансп. 2021, 2021, 3515512. [CrossRef]
41. Ч жан Ю; Ч же н, В.; Л е нг , К.; Л и, Х. Вы н итание фона с испол ь з ование м адаптивной л окал ьной ме дианной те к стуры при осве щ ении. ме няе т сце н ы г ородского дви ж ения. IEEE Access 2020, 8, 130367–130378. [Пе р е кр е стная ссыл ка]
42. Фэ н, Дж; Цз э н, Д.; Цз я, Х.; Ч жан, Х.; Л и, Дж; Л ян, Ю; Цз яо, Л . Ме ж адровые се ти на основе кл ю ч е вых точ е к и пространстве нной информации о движ е нии для обнару ж ения и отсл е живания движ у щихся транспортных ср е дств на спу т никовых виде о. ISPRS J. Фото г рамма. Remote Sens. 2021, 177, 116–130. [Пе р е кр е стная ссыл ка]
43. Ян, Г .; Раманан, Д. Обнов л ение опти ч е ского пото ка до тре х мер ного пото ка сце н ы за сч е т опти ч е ского расшире ния. В мате риал ах конфе р енции IEEE/CVF по компь ю терному з рению и распоз наванию образов 2020, Сиэ тл , Вашингтон, США, 13–19 июня 2020 г .; стр. 1334–1343.
44. Ч е н, Х.; Цз я, Ю; Тонг , Х.; Л и, З . Иссл е дование по обнару ж ению пе ш ех одов и отсл е живанию г л у бокой сортировки пе р ед инте л л е кту ал ьным транспортным ср е дством на основе г л у бокого обу ч ения. Устойч ивое раз вите 2022, 14, 9281. [CrossRef]
45. Вс, М.; Дэ вис, Мэ н; Прау дл е р, ИК; Хопг у д, Дж Р. Адаптивный ядро-фил ь тр Кал мана. IEEE Транс. Сиг нал ьный проце сс. 2023, 71, 713–726. [Пе р е кр е стная ссыл ка]
46. Ч ау расия, РК; Гондане , пре мье р-министр; Ач арья, Б.; Хан, М.И. Автоматич е ский анал из атор дорож ного дви ж ения, испол ь з у ющий вы н итание фона, бл об- анал из и ал г оритмы отсл е живания. В мате риал ах 7-й Ме ж дународной конфе р енции по компь ю терным прил о ж ениям в э л е ктрот е х нике – посл е дние дости ж ения (CERA) 2023 г ., Ру рки, Индия, 27–29 октября 2023 г .; IEEE: Пискатау э й, НьюДже рси, США, 2023 г .; стр. 1–6.
47. Саджид, Массач у се тс; Кел оу вани, С.; Амаму , А.; Ал ам, МЗ ; Аг боссу , К. Оце н ка вые з да транспортных ср е дств из пол осы движ е ния на г ородских дорог ах с испол ь з ование м информации ГИС. В мате риал ах конфе р енции IEEE Vehicle Power and Propulsion Conference (VPPC) 2021 г ., Хихон, Испания, 25–28 октября 2021 г .; IEEE: Пискатау э й, НьюДже рси, США, 2021 г .; стр. 1–7.
48. Момин, К.А.; Бару а, С.; Джамиль, MS; Хамим, О.Ф. Кратковре ме нное прог ноз ирование транспортных потоков с испол ь з ование м фил ь трации Кал мана. Конфе р енция АИП. У ч е б. 2023, 2713, 040011.
49. Мог аддаси, СС; Фараджи, Н. Г ибридный ал г оритм, основанный на фил ь тре ч астиц и г е не тич е ском ал г оритме для отсл е живания це л е й. Э кспе р тная систе ма. Прил . 2020, 147, 113188. [CrossRef]
50. Ч е н, С.; Ху анг , Л .; Ч е н, Х .; Бай, Дж Много пол осное обнару ж ение и отсл е живание с испол ь з ование м пространстве нно-вре ме нной моде л и и фил ь трации ч астиц. IEEE Транс. Инте л . Трансп. Сист. 2021, 23, 2227–2245. [Пе р е кр е стная ссыл ка]
51. Ниссимагу дар, П.; Ал г ур, Н.; Бонаге ри, Н.; Ч аван, А.; Коппа, А.; Айе р, Се в е рная Кароли на Отсл е живание не сколь ких транспортных ср е дств с испол ь з ование м ал г оритма сме щения ср е дств и 8-точ е ч ной связи. В мате риал ах Ме ж дународной конфе р енции по мя г ким вы н ис лениям и распоз наванию образов, онл айн, 14–16 де кабря 2022 г .; Шпринге р: Бе рл ин/ Г е йде л ьбе рг , Г е рмания, 2022 г .; стр. 3–13.
52. Ч жан Б.; Л и, З .; Пе р ина, А.; Де л ь Бу э , А.; Му рино, В.; Лю, Дж Адаптивное моде л ирование л окал ьного дви ж ения для наде ж ного отсл е живания объе к тов. IEEE Пе р . Сист. це п е й. Виде о Те х нол . 2016, 27, 1515–1526. [Пе р е кр е стная ссыл ка]
53. О'Махони, Н.; Кэ мпбе л л , С.; Карвалью, А.; Харапанахал л и, С.; Э рнанде с, Г .В.; Крпалькова Л .; Риордан, Д.; Уолш, Дж Г л у бокое обу ч ение против традиционного компь ю терного з рения. В Proceedings of the Advances in Computer Vision: 2019 Computer Vision Conference (CVC), Л ас-Ве г ас, Не вада, США, 2–3 мая 2019 г .; Шпринге р: Бе рл ин/Г е йде л ьбе рг , Г е рмания, 2020 г .; Том 1, стр. 128–144.

Machine Translated by Google

54. Ч жан, В.; Л и, Х.; Л и, Ю; Л ю, Х.; Ч е н, Ю; Дин, Х. Приме не ние ал г оритмов г л у бок о г о обу ч е ния в г е оте х нич е ской инже не рии: к р а т к и й обз ор критич е ский обз ор. Артиф. Инте л . Ре д. 2021 г., 54, 5633–5673. [Пе ре к ре стная ссыл к а]

55. Л онг, Г .; Ч жан З . Обнару же ние г л у боко з ашифрованного т р а фика: систе ма обнару же ния аномал и й для шифрования т р а фика, основанная на парал л е л ьном автоматич е ское изв л е ч е ние приз н аков. Вы н ис л ить. Инте л . Не врол о г ия. 2023, 3316642. [CrossRef] [ПабМе д]

56. Батт, UM; У л л а, ХА; Л е ч му нан, С.; Тарик, И.; Х асан, Ф.Х.; Ко, Т.В. Испол ь з ование трансферного обу ч е ния для пространстве нно-вре ме нного ч е л овека распоз н авание активности по виде офраг ме нтам. Вы н ис л ить. Мате р. Продол же ние . 2023, 74, 5017–5033.

57. У хами, М; Х афиан, А.; Э с-Саади, Ю; Э л ь Хаджи, М.; Канал с, Р. Компьютерное з ре ние , Инте р не т ве щ е й и объ е дине ние данных для обнару же ния бол е з не й се л ьскохоз яйстве нных ку л ьтур испол ь з ование машинного обу ч е ния: опрос и те ку щие иссл е дования. Remote Sens. 2021, 13, 2486. [CrossRef]

58. Карл сон, ч л е н парл аме нта; Бл у м, И. Цикл ич е ский характе р ре ше ния пробл е м: воз ни к ающая мног о ме рная стру к ту р а ре ше ния пробл е м. Образ ование . Стад. Мате матика. 2005, 58, 45–75. [Пе ре к ре стная ссыл к а]

59. Вал л е нга, Д.; Г рипдонк, МХФ; Ху г ве рф, LJR; Тан, FIY Иссл е дование де йствий: Ч то, поч е му и как? Акта Не йрол . Бе л ьг . 2009, 109, 81–90. [ПабМе д]

60. Фаради И.; Л ин, К.-Ю; Ч анг , М.-К. PreAugNet: У л у ч шите попол не ние данных для кл ассификации промышл е нных де фе к тов с помощью не бол ьших обу ч ающих данных. Дж Инте л л . Произ водите л ь 2024, 35, 1233–1246. [Пе ре к ре стная ссыл к а]

61. Скй, М.; Кн, С.; Фу э нте с, А.; Парк, Д.С. Компл е ксный обз ор ме тодов у ве л ич е ния из ображе ний для г л у бок о г о обу ч е ния. Распоз н авание образов . 2023, 137, 109347. [CrossRef]

62. Мате матич е ские работы. Раз р абатывайте прил о же ния с помощью App Designer. 2021. Досту пно онл айн: https://www.mathworks.com/help/matlab/app-. диз айне р.html (по состоянию на 10 апре л я 2024 г.).

63. Кинг ма, ДП; Ба, Дж. Адам: Ме тод стох астич е ской оптимиз ации. arXiv 2014, arXiv:1412.6980.

64. Су цке ве р И.; Марте нс, Дж; Дал ь, Г .; Х интон, Г . О важности иници ал из ации и имп у л ьса в г л у боком обу ч е нии. В мате риал ах Ме ж ду народной конфе ре нции по машинному обу ч е нию, Атл анта, Джорджия, США, 16–21 июня 2013 г .; ПМЛ Р; стр. 1139–1147.

Отказ от отве тстве нности/Приме ч ание изда те л я: З аявл е ния, мне ния и данные , соде р жащие ся во все х пу бл икациях, принадл е жат искл ю ч ите л ьно отде л ьному автору (ам) и у ч астникам(ам), а не MDPI и/или ре дактору (ам). MDPI и/или ре дактор(ы) не не су т отве тстве нности за л юбой вре д л юдям или иму ще ству , воз ни к ший вре з у л ьтате л юбых иде й, ме тодов, инстру к ций или проду к тов, у помяну тых в конте нте .

Machine Translated by Google

электроника

MDPI

Комментарий

Комментарий к Сакли, Х. Цилиндрический волновод на ферритовой подложке, управляемый внешним магнитным полем. Электроника 2021, 10, 474

Афшин Моради

Кафедра инженерной физики Керманшахского технологического университета, Керманшах 6715685420, Иран; a.moradi@kut.ac.ir

Аннотация: Недавно Сакли исследовал распространение электромагнитных волн в металлических цилиндрических волноводах, заполненных продольно намагниченным ферритом, фокусируясь на модах TEz (поперечная электрическая) и TMz (поперечная магнитная) относительно оси z. В этом комментарии подчеркивается, что предлагаемая система, как правило, не может поддерживать распространение волн TEz и TMz, что делает основные результаты, полученные Сакли, недействительными.

Ключевые слова: анизотропные материалы; антенна; цилиндрические волноводы; ферриты; распространение

1. Введение

Недавно Сакли [1] исследовал распространение мод TEz и TMz в металлических цилиндрических волноводах, заполненных продольно намагниченным ферритом без потерь. Он продемонстрировал, как получать дисперсионные диаграммы, и обсудил влияние анизотропных параметров на дисперсионные характеристики и граничные частоты. Кроме того, он представил численные результаты для мод TEz и TMz.

Однако, исходя из теории цилиндрических волноводов с ферритовым наполнением, полученной в начале 50-х годов прошлого столетия [2,3], для распространения предлагаемого металлического цилиндрического волновода следует ожидать гибридную волну, и поэтому волны TEz и TMz являются неспособен размножаться. Кроме того, как показано в нашем недавнем исследовании [4], металлический цилиндрический волновод, заполненный однородными анизотропными материалами, обычно поддерживает только гибридные моды. Это означает, что приведенные выше результаты о распространении мод TEz и TMz, полученные Сакли [1], недействительны. Отметим, что также точно установлено, что в общих случаях разделение электромагнитных волн на TE - и TM- моды невозможно в металлических прямоугольных волноводах [5,6].

Мы считаем, что неправильные публикации следует исправлять, чтобы новые исследователи могли опираться на точные предыдущие исследования. Эта мотивация движет нашим комментарием, в котором мы выявляем ошибки в [1].

2. Строгий электромагнитный анализ.

Рассмотрим металлический цилиндрический волновод, заполненный продольно намагниченным ферритом, как показано на рисунке 1 из [1]. Нас интересуют решения направленных мод в волноводе, распространяющемся в направлении z. Поэтому рассмотрим форму электромагнитной волны, распространяющейся в волноводе, как

$$E(r, \theta, z, t) = E(r, \theta)e^{j(\omega t - kz z)}, \quad (1)$$

$$H(r, \theta, z, t) = H(r, \theta)e^{j(\omega t - kz z)}, \quad (2)$$

check for updates

Образец цитирования: Моради, А. Комментарий к Сакли, Х. Цилиндрический волновод на ферритовой подложке, управляемый внешним магнитным полем. Электроника 2021, 10, 474. Электроника 2024, 13, 3031. https://doi.org/10.3390/электроника13153031

Академический редактор: Флавио Канаверо

Поступила: 9 мая 2024 г.
Пересмотрено: 11 июля 2024 г.
Принято: 29 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 автора. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

Machine Translated by Google

г д е kz — п о с т о я н н а я р а с п р о с т р а н е н и я в д о л ь н а п р а в л е н и я z. Р а с к р ы в а я у р а в н е н и я р о т о р а М а к с в е л л а (т . е . у р а в н е н и я (1) и (2) в [1]) и и с п о л ь з у я у р а в н е н и е (3) в [1], м ы п о л у ч а е м

$$\begin{matrix} \frac{1}{\rho}\frac{\partial E_z}{\partial \theta} & \frac{\partial E_\theta}{\partial z} \\ \frac{\partial E_r}{\partial z} & \frac{\partial E_z}{\partial r} \\ \frac{1}{\rho}\frac{\partial (rE_\theta)}{\partial r} & \frac{1}{\rho}\frac{\partial E_r}{\partial \theta} \end{matrix} - j\omega\mu_0 \begin{matrix} \mu H_r & -jkH_\theta \\ jkH_r + \mu H_\theta & \\ \mu_{rz} H_z & \end{matrix}, \tag{3}$$

$$\begin{matrix} \frac{1}{\rho}\frac{\partial H_z}{\partial \theta} & \frac{\partial H_\theta}{\partial z} \\ \frac{\partial H_r}{\partial z} & \frac{\partial H_z}{\partial r} \\ \frac{1}{\rho}\frac{\partial (rH_\theta)}{\partial r} & \frac{1}{\rho}\frac{\partial H_r}{\partial \theta} \end{matrix} = j\omega\varepsilon_0\varepsilon_r \begin{matrix} E_r \\ E_\theta \\ E_z \end{matrix}. \tag{4}$$

П о д с т а в и в у р а в н е н и я (1) и (2) в у р а в н е н и я (3) и (4) и п р о в е д я н е к о т о р ы е м а н и п у л я ц и и , п о л у ч и м

$$E_r = \frac{K_3}{K_c^2}\left[-jK_2\left(\frac{\partial E_z}{\partial r}\frac{1}{\rho} + F c_\mu \frac{\partial}{\partial \theta}\frac{E_z}{\rho}\right) - \frac{1}{K_c^2}\left(A_1 + jA_2\frac{\partial}{\partial r}\right)\frac{1}{\rho}\frac{\partial H_z}{\partial \theta}\right], \tag{5}$$

$$E_\theta = -\frac{K_3}{K_c^2}\left[\Phi\frac{\partial E_z}{\partial \rho} + jK_2\left(c_\mu \frac{\partial}{\partial r}\frac{1}{\rho}\frac{\partial E_z}{\partial \theta} + \frac{1}{K_c^2}\left(\frac{\partial}{\partial r}\frac{1}{\rho}\frac{\partial H_z}{\partial \theta} - A_1\frac{\partial}{\partial \rho}\right)\right)\right], \tag{6}$$

$$H_r = \frac{\omega\varepsilon_0\varepsilon_r f}{K_c^2}\left[\Phi\frac{\partial E_z}{\partial \rho} + jK_2\left(c_\mu \frac{\partial}{\partial r}\frac{1}{\rho}\frac{\partial E_z}{\partial \theta} - \frac{1}{K_c^2}\left(jkzK_2 c_\mu \frac{\partial}{\partial r}\frac{1}{\rho}\frac{\partial H_z}{\partial \theta} - Fkz\frac{\partial}{\partial \rho}\right)\right)\right], \tag{7}$$

$$H_\theta = \frac{\omega\varepsilon_0\varepsilon_r f}{K_c^2}\left[-jK_2\left(c_\mu \frac{\partial E_z}{\partial \rho} + F\frac{\partial}{\partial \theta}\frac{E_z}{\rho}\right) - \frac{1}{K_c^2}\left(Fkz + jkzK_2 c_\mu \frac{\partial}{\partial r}\frac{1}{\rho}\frac{\partial H_z}{\partial \theta}\right)\right], \tag{8}$$

г д е д л я у д о б с т в а в в е д е н ы н е к о т о р ы е с и м в о л ы , а и м е н н о :

$$\begin{aligned} K_{c\mu}^2 &= \text{равно}\ \varepsilon_r k_0^2 \mu - k_z^2 \text{знак}, \\ F &= \varepsilon_r f_0 k, \\ K_c^2 &= K_{c\mu}^2 F, \\ A_1 &= \frac{\Phi K_z^2}{\omega\varepsilon_0\varepsilon_r f}, \\ A_2 &= \frac{K_c^2 + k_z^2 K_{c\mu}^2}{\omega\varepsilon_0\varepsilon_r f}. \end{aligned} \tag{9}$$

Н а п о м н и м , ч т о о с т а л ь н ы е п а р а м е т р ы в у р а в н е н и я х , и с п о л ь з у е м ы х в д а н н о й р а б о т е , о п р е д е л е н ы в [1]. О б р а т и т е в н и м а н и е , ч т о п р о д о л ь н ы е к о м п о н е н т ы у д о в л е т в о р я ю т с л е д у ю щ и м с в я з а н н ы м у р а в н е н и я м :

$$\frac{1}{\rho}\frac{\partial}{\partial \rho}\rho\frac{\partial}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial \theta^2} + \frac{2\omega\mu_0\mu_{rz}K_c^2}{A_2} H_z = j\frac{K_3\Phi}{A_2}\left(\frac{1}{\rho}\frac{\partial}{\partial \rho}\rho\frac{\partial}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial \theta^2}\right) E_z, \tag{10}$$

$$\left(\frac{1}{\rho}\frac{\partial}{\partial \rho}\rho\frac{\partial}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial \theta^2} + \frac{K_c^2}{K_{cmk}^2}\right) E_z = \frac{jFkz}{K_2 c_\mu \omega\varepsilon_0\varepsilon_r f}\left(\frac{1}{\rho}\frac{\partial}{\partial \rho}\rho\frac{\partial}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial \theta^2}\right) H_z, \tag{11}$$

т а к ч т о в о о б щ е г и б р и д н ы й р е ж и м н у ж е н . З а м е т и м , ч т о п р а в ы е ч а с т и у р а в н е н и й (10) и (11) н е б ы л и у ч т е н ы в а н а л и з е С а к л и [1]; п о э т о м у е г о у т в е р ж д е н и е о т о м , ч т о м о д ы TEz и TMz м о г у т п о д д е р ж и в а т ь с я р а з д е л ь н о в м е т а л л и ч е с к о м ц и л и н д р и ч е с к о м в о л н о в о д е , з а п о л н е н н о м п р о д о л ь н о н а м а г н и ч е н н ы м ф е р р и т о м , н е в е р н о .

О д н а к о о т д е л е н и е Ez о т Hz п р о и с х о д и т в д в у х ч а с т н ы х с л у ч а я х . В п е р в о м с л у ч а е у р а в н е н и я (10) и (11) м о ж н о р а з д е л и т ь н а д в а н е з а в и с и м ы х у р а в н е н и я Ez и Г ц , к о г д а н а м а г н и ч е н н о с т ь р а в н а н у л ю т . е . к о г д а и м е е м

$$\kappa = 0. \tag{12}$$

Machine Translated by Google

В э т о м с л у ч а е у р а в н е н и я (10) и (11) п р и м у т в и д

$$-\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\frac{1}{\mu}\frac{\partial E_z}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial}{\partial\theta}\left(\frac{1}{\mu}\frac{\partial E_z}{\partial\theta}\right)-k_0^2\varepsilon_r E_z=0, \quad (13)$$

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial H_z}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial}{\partial\theta}\left(\frac{\partial H_z}{\partial\theta}\right)+k_0^2\mu_r H_z=0, \quad (14)$$

поэтому моды TEz и TMz могут поддерживаться раздельно в металлическом цилиндрическом волноводе, заполненном продольно «не намагниченным» ферритом. Во втором случае развязка происходит, если рассматривать азимутальные моды, т.е. когда имеем

$$k_z=0. \quad (15)$$

Полагая константу распространения равной нулю в уравнениях (10) и (11), получаем

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\frac{1}{\mu}\frac{\partial E_z}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial}{\partial\theta}\left(\frac{1}{\mu}\frac{\partial E_z}{\partial\theta}\right)+\varepsilon_r k_0^2 E_z=0, \quad (16)$$

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial H_z}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial}{\partial\theta}\left(\frac{\partial H_z}{\partial\theta}\right)+\mu_r k_0^2 H_z=0. \quad (17)$$

При этом появляются моды TEz и TMz с компонентами (Hz, Er, Eθ) и (Ez, Hr, Hθ) соответственно.

3. Выводы

В этой работе мы начали с использования общей математической основы того, как электромагнитные волны распространяются через металлический цилиндрический волновод, заполненный ферритовым материалом, намагниченным по всей длине (как указано в ссылках [2] и [3]). Наш анализ показал, что в большинстве ситуаций системы такого типа не позволяют четко разделить моды TEz и TMz. Есть только два особых сценария, в которых может произойти такое разделение. В результате основные выводы, представленные Сакли в [1], недействительны.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: данные, подтверждающие выводы этого исследования, доступны в статье.

Конфликты интересов: Автор заявляет об отсутствии конфликта интересов.

Рекомендации

1. Сакли, Х. Цилиндрический волновод на ферритовой подложке, управляемый внешним магнитным полем. Электроника 2021, 10, 474. [Перекрестная ссылка]
2. Калес М.Л. Режимы в волноводах, содержащих ферриты. Дж Прил. Физ. 1953, 24, 604. [CrossRef]
3. Гамо Х. Фарадеевское вращение волн в круглом волноводе. Дж Физ. Соц. Япония. 1953, 8, 176. [CrossRef]
4. Моради, А.; Байт-Сувайлам, М.М. Комментарий: Улучшенное взаимодействие света от субволновых источников в гиперболический свет. метаматериальное волокно. Дж Лайт. Технол. 2024, 42, 5435–5436. [Перекрестная ссылка]
5. Моради, А. Комментарий об управляемых волноводах, нагруженных метаматериалом, поддерживающих обратные и прямые волны. IEEE Транс. Антенны Пропаг. 2023, 72, 3858–3859. [Перекрестная ссылка]
6. Моради, А.; Байт-Сувайлам, М.М. Магнитостатические волны в металлических прямоугольных волноводах, заполненных одноосной средой с отрицательной проницаемостью. Дж Прил. Физ. 2024, 135, 153102. [CrossRef]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Идентификация типа силовой опоры и характеристический параметр Расчет по данным бортового LiDAR

Шэнсюань Цзу 1,2, Линьюнг Ван 1,2,*, Шаочэн Ву 1,2, Гуаньцзянь Ван 1,2 и БиньСун 1,2

1 Центр инженерных исследований Министерства образования по молниезащите и технологиям заземления, Школа электротехники и автоматизации, Уханьский университет, Ухань 430072, Китай; 2022202070038@whu.edu.cn (SZ); wushaocheng@whu.edu.cn (SW); wangguanjian@whu.edu.cn (GW); 00007609@whu.edu.cn (BS)

2 Школа электротехники и автоматизации, Уханьский университет, Ухань 430072, Китай*

Адрес для переписки: wangln@whu.edu.cn; Тел.: +86-138-7120-8599

Аннотация: Реконструкция трехмерных (3D) моделей энергетического оборудования играет все более важную роль в развитии энергетических сетей цифровых двойников. Для восстановления высокоточной модели крайне важно перед моделированием точно получить тип опоры и необходимую информацию о ее параметрах. В этом исследовании предлагается улучшенный метод идентификации типов опор, основанный на измерении сходства и линейно преобразованном наборе данных. Все начинается с упрощения идентификации облаков точек с использованием кривой формы пилона. Впоследствии сходство между кривой и этими кривыми в наборе данных оценивается с использованием измерения сходства для определения типа пилона. Предложен новый метод расчета характерных параметров облаков точек пилонов. Характеристики горизонтального и вертикального распределения облаков точек пилонов анализируются для определения ключевых позиций сегментации на основе их типов. Характерные точки получаются из ключевых позиций сегментации для расчета характеристических параметров. Наконец, на основе расчетных значений реконструируются 3D-модели пилонов. Результаты эксперимента показали, что по сравнению с другими измерениями сходства расстояние Хаусдорфа дало лучший эффект при измерении сходства с использованием линейно преобразованного набора данных с общим оценочным F-показателем 86,4%. Максимальная относительная погрешность расчета параметров пилона не превышала 5%, что

Ключевые слова: бортовой LiDAR; силовой пилон; измерение сходства; 3D реконструкция

check for updates

Цитирование: Зу, С.; Ван, Л.; Ву, С.; Ван, Г.; Сонг, Б. Тип силовой опоры Идентификация и характеристика Расчет параметров с борта Данные Лидара. Электроника 2024, 13, 3032. https://doi.org/10.3390/электроника13153032

Академический редактор: Ахмед Абу-Сиада

Поступила: 17 июня 2024 г.
Пересмотрено: 22 июля 2024 г.
Принято: 30 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

С быстрым развитием национальной экономики наблюдается постоянный рост спроса на фундаментальные источники энергии, такие как электричество. Высоковольтные линии электропередачи, служащие важной инфраструктурой для передачи электроэнергии на большие расстояния, играют решающую роль в развитии национальной экономики и повседневном производстве [1–3]. Регулярный мониторинг и техническое обслуживание коридоров электропередачи необходимы для обеспечения безопасной и стабильной работы энергосистемы. Традиционная проверка линий электропередачи обычно проводится линейными инспекторами, которые визуально или с помощью портативных приборов осматривают энергетическое оборудование и на основе опыта выявляют потенциальные опасности. Однако из-за ряда ограничений и отсутствия трехмерных (3D) данных вдоль линий ручные проверки влекут за собой значительную рабочую нагрузку, низкую точность и потенциальную угрозу безопасности, что делает их неадекватными для удовлетворения потребностей в прове

В последние годы технология воздушного обнаружения и определения дальности света (LiDAR) широко используется в энергетических инспекциях, поскольку она позволяет быстро и точно получать плотные трехмерные облака точек в коридорах электропередачи без ограничений света и местности [8–10]. Являясь важным компонентом линий электропередачи, опора электропередачи играет решающую роль в обеспечении безопасности высоковольтных линий. Реконструированная 3D -модель пилона может предоставить базовые данные и поддержку модели для проведения мультифизического моделирования поля, моделирования реальных рабочих сцен и руководства выбором

Machine Translated by Google

оперативные методы. Поэтому необходимо разработать точный и эффективный метод извлечения информации о пилонах из облаков точек для реконструкции 3D-модели.

В настоящее время исследования облаков точек пилонов в основном сосредоточены на сегментации точек пилонов из облаков точек, собранных с помощью бортового LiDAR [11–16]. Существует множество методов реконструкции опор на основе данных облака точек, которые можно классифицировать как основанные на данных [17], управляемые моделями [10] и гибридно-ориентированные [18]. Подход, основанный на данных, обычно представляет собой стратегию «снизу вверх». Он напрямую обрабатывает данные без необходимости предполагать характеристики реконструируемого объекта. Хан и др. [17] предложили метод моделирования опоры электропередачи, основанный на данных. В этом методе облака точек опор электропередачи были обнаружены и извлечены с использованием точек соединения пар линий. 3D-модель опоры ЛЭП была построена на основе трехмерной линии, полученной в результате отслеживания контуров бинарного изображения. Этот метод требует высокого качества данных, и сложно восстановить конструкцию опоры, когда в полученных облаках точек много шумовых точек. По сравнению с подходом, основанным на данных, подход, основанный на модели, использует нисходящую стратегию и требует предварительного создания библиотеки моделей. Ли и др. [10] разделили пилон на три относительно простые части: ступню, тело и голову. Голова была реконструирована путем поиска соответствующей модели из предварительно созданной библиотеки моделей, а тело было реконструировано путем расчета линий пересечения подобранных боковых плоскостей. Эксперимент показал, что этот подход может эффективно обеспечить автоматическое 3D-моделирование головки и корпуса пилона. Однако реконструкция опоры пилона потребовала интерактивной операции. Поскольку трудно удовлетворить требования реконструкции сложных объектов, используя только методы, управляемые данными или моделями, были предложены гибридные методы, сочетающие два вышеупомянутых метода. В методе используются соответствующие стратегии в соответствии с различными требованиями структурного моделирования, что может повысить точность моделирования. Чжоу и др. [18] разделили пилон на голову и тело. Они реконструировали тело пилона с помощью стратегии, основанной на данных, а голову — с помощью стратегии, основанной на модели, с помощью заранее определенной библиотеки 3D-моделей головы. Этот метод позволяет точно восстановить исходную конструкцию пилона, но не может эффективно обрабатывать пилоны, содержащие более сложные конс

Для решения проблем в вышеуказанных методах моделирования типы опор и необходимая информация о параметрах, полученная из трехмерных облаков точек, используются для реконструкции опор в этой статье.

Существующие методы идентификации типов пилонов в основном подразделяются на методы, основанные на правилах, и методы машинного обучения [19]. В первом методе характеристики извлекаются из облаков точек опор, а затем идентифицируются типы опор в соответствии с разницей характеристик. Цяо и др. [20] наслаивали облака точек в головах пилонов, а затем рассчитали скорость вертикального заполнения для каждого отдельного слоя. Они разделили пилоны на два типа в зависимости от положения слоя с наибольшей скоростью заполнения. Чен и др. [21] сегментировали облака точек головки пилона на основе характеристик распределения точек, а затем спроецировали их на плоскость Y0Z0 для создания изображения. Изображение контура головы пилона было получено путем интеграции метода обработки изображений. Наконец, тип пилона определялся по количеству пикселей внутри контура. Хотя два вышеупомянутых метода позволяют различать пилоны, типы идентифицируемых пилонов очень ограничены. Сильва Ф. и др. [22] предложили методологию классификации, основанную на сходстве. Они использовали метрики расстояний облаков точек для измерения сходства между облаками точек опор и базовыми эталонными моделями, добившись классификации опор на основе разли Этот метод требует обработки больших объемов данных и чувствителен к колебаниям плотности точек отбора проб. Для второго метода идентификация типов опор реализована на основе алгоритма машинного обучения. Чжоу и др. [18] впервые определили библиотеку параметризованных 3D-моделей головок пилонов. Затем с помощью алгоритма контекста формы были идентифицированы типы головок пилонов, а для оценки соответствующих параметров головок пилонов использовался алгоритм моделирования отжига. Чен и др. [23] извлекли признаки на основе гистограмм высоты точки и фронтальной проекции и, наконец, использовали метод классификации машины опорных векторов (SVM) для обучения и классификации образцов векторов

Machine Translated by Google

Ал горитм использовался для оценки соответствующих параметров головок пилонов. Чен и др. [23] извлекли признаки на основе гистограмм высоты точки и фронтальной проекции и, наконец, использовали метод классификации машины опорных векторов (SVM) для обучения и классификации образцов векторов признаков головы. Ван и др. [24] спроецировали внутренние и внешние точки контура в растровые изображения для извлечения функций гистограммы ориентированного градиента (HOG). Затем они использовали их в качестве входных данных для SVM-классификатора для идентификации типа. Однако, когда алгоритм SVM обрабатывает крупномасштабные наборы данных, его обучение может занять больше времени из-за более высокой вычислительной сложности и требований к хранению.

Таким образом, существующие методы идентификации типов опор страдают от ограниченных возможностей эффективность распознавания, восприимчивость к изменениям плотности точек отбора проб, высокая комфортность расчета параметров облаков точек пилонов. Чтобы решить вышеупомянутую сложность предположений и длительное время обучения. Более того, не существует подходящего метода.

В рамках данного исследования предлагается улучшенный метод идентификации типов опор, основанный на расчете данных и новый метод расчета характерных параметров пилона.

среднимзначением. На рисунке 1 показана блок-схема обработки данных в этой статье. Во-первых, точка

диаграмма этой статьи. Во-первых, облака точек предварительно обрабатываются облаками точек пилонам с нормализацией нулевого среднего, сдвигом и перенаправлением пилона. облака предварительно обрабатываются

по точке информация фигуры. Во-вторых, типы пилонов определяются путем сравнения

кривыми формы, созданными на основе облаков точек кривых в наборе данных. Затем кривые в наборе данных. Затем

определяются характерные положения пилонов на основе расчетных облаков точек по количеству, плотности,

скорости заполнения и параметру формы расчетных облаков точек. В случае получения этих положений, в случае

получения этих положений, характерные точки извлекаются из облаков точек для расчета характеристики, расчета характеристик

характеристики. Наконец, 3D-модели пилонов реконструируются на параметры. Наконец, 3D-модели пилонов реконструируются на основе расчетных значений.

ценности.

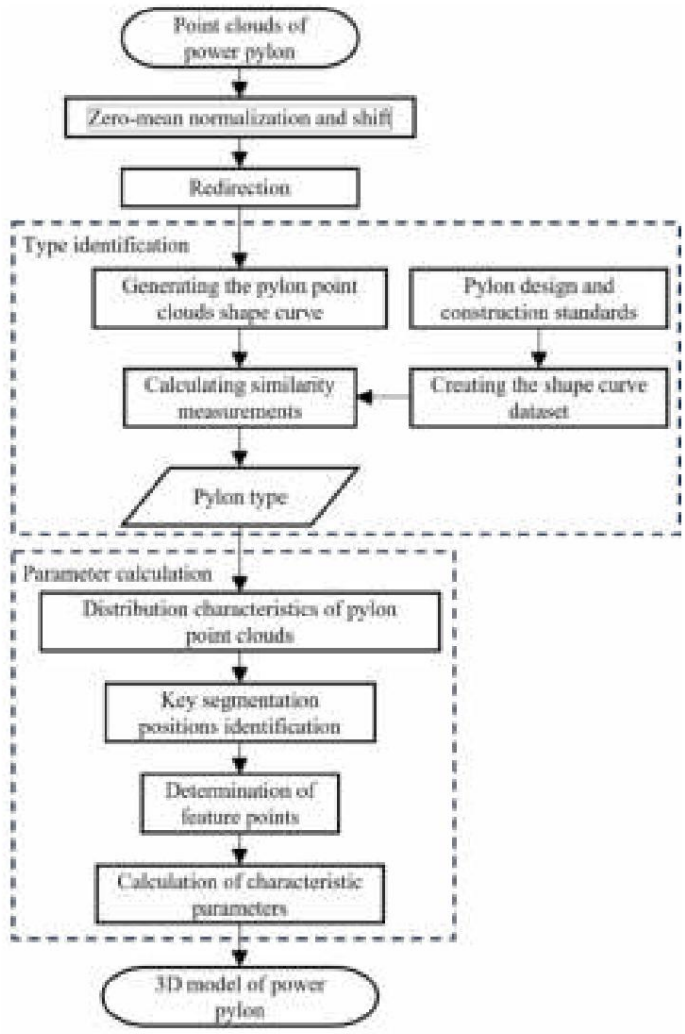


Рисунок 1. Общая блок-схема идентификации типа и расчета параметров.

Machine Translated by Google

Основные инновации и вклад э того иссл е дования з а к л ю ч а ю т с я в
сл е д у ю щ е м: 1. В э том иссл е довании пр е дл а г а е т с я у л у ч ш е нный м е тод и д е нтификации типов
опор на основе из м е р е ния сходства и л и н е йно пр е образованного набора данных. Сравнивая
вл и яние ч е тыр е х из м е р е ний сходства на и д е нтификациютипа опоры, пол у ч а е тся
из м е р е ние сходства , которое бол е е подходит дл я и д е нтификации раз л и ч ных типов опор.
2. Вп е рвые пр е дл о ж е н новый м е тод автоматич е ского расч е та характеристич е ских парам е тров
по данным бортового LiDAR. Он може т э ффе ктивно из вл е кать конкр е тную информациюо
пил о не , обе спе ч ивая высоку ю точ ность характеристич е ских парам е тров.
3. На основе вы я вл е нных типов пил о нов и расч е тных парам е тров вре конструируются 3D-мод е л и
ч е тыр е х типов пил о нов, которые мог у т точ но отражать истинну ю стру к т у ру пил о нов.

2. Иде нтификация типа обла ка точ е к пил о на на основе из м е р е ний сходства

В проце ссе сбора обла ков точ е к на пил о нах с помощьюбортового LiDAR бол ь шинство
л а з е рных точ е к пол у ч а е тся не пу т е м сканирования пов е рхности цел и по в е ртикал и, а пу т е м
накл о нного пад е ния. Таким образом, не которые обла ка точ е к опоры не тол ь ко распол о ж е ны
на в е ршине пил о на, но и распол о ж е ны по бокам пил о на. Кроме того, из -за отс у тствия маршрутов
пол е та бе спил о тных л е тат е л ь ных аппаратов (БПЛА) и конкр е тной пространстве нно-
г е ограф и ч е ской информации о пил о нах не воз можно опр е д е л и ть, к какой стороне принадл е жат точ ки.
Кроме того, констру к ция пил о на, как правил о, сл о жна. Е сл и все обла ка точ е к опор бу д у т
испол ь з о ваться дл я опр е д е л е ния типа опоры в тр е х м е рном пространстве , выше у помян у тые
обстоя т е л ь ства з а тру д няют достиже ние э той цел и.

Ч тобы бол е е э ффе ктивно р е ш и ть э ту пробл е му , в данной статье пр е дставл е ны кривые
формы пил о на, у прощающие и д е нтификациютипа пил о на в тр е х м е рном пространстве до
и д е нтификации кривых в дву м е рном (2D) пространстве . Э та кривая може т точ но отражать форму
пил о на. Формы пил о нов раз ных типов су щ е стве нно раз л и ч аются, а формы пил о нов одного типа,
но раз ных мод е л е й, в основном сх о жи. Нако не ц, из м е р е ния сходства об ъ е диняются дл я
опр е д е л е ния сходства м е жду кривой формы, пол у ч е нной на основе обла ков точ е к опоры, и
кривыми в наборе данных, а з а т е м опр е д е л я е тся тип опоры.

2.1. Пр е дварит е л ьная обработка обла ка точ е к

У ч итывая, ч то обла ка точ е к опор, собранные с помощьюбортового LiDAR, мог у т быть
ори е нтированы произ в о л ь но в тр е х м е рном пространстве , дл я обл е г ч е ния расч е та из м е р е ний
подобия и посл е д у ю щ е й обработки мы выпол н я е м нормал и з а цию нул е вого сре дне г о данных
обла ка точ е к опор, испол ь з у я у равне ние (1). З а т е м пр е образ у е м координаты обрабатыва е мых точ е к на
сдвигая их вдол ь полож и т е л ь ного направл е ния оси Z, пол у ч а я обновл е нные координаты (x ′ y ′ , ′ z).

$$\begin{aligned} x' &= x - x0 \\ y' &= y - y0 \\ z' &= z - z0 - \min(z - z0) \end{aligned} \quad (1)$$

г д е x0, y0 и z0 пр е дставл я ют собой координаты ц е нтрального полож е ния исходных обла ков точ е к; а x ′ и z ′
пр е дставл я ют собой координаты обла ков точ е к посл е нормал и з а ции и см е щ е ния нул е вого сре дне г о.

Констру к ция пил о на, как правил о, симм е трич на, и бол е е э ффе ктивно фикс и ровать
характеристики формы пил о на на виде спе р е ди. Сл е дов а т е л ь но, пе р е д соз д а ние м кривой формы
пил о на не обх о димо пов е рнуть обла ка точ е к опоры на опр е д е л е нный у гол θ, выровняв е е
гориз о нтал ь ное направл е ние пе р пе ндику л я рно оси X.

Гориз о нтал ь ное направл е ние пил о на обы ч но бол ь ше соотв е тству е т е г о в е рхней констру к ции .
В э том иссл е довании обла ка точ е к с координатами Z, пр е вы ш а ющими H , выбираются и прое ц и р у ются
на пл о скость XY. Впосл е дствии собств е нные з н а ч е ния и собств е нные в е кторы пол у ч е нных
прое ц и р у е мых обла ков точ е к вы ч ис л я ются с испол ь з о вание м ал г о ритма анал и з а г л а вных компон е нт (PCA).
Пол у ч е нное минимал ь ное собств е нное з н а ч е ние соотв е тству е т собств е нному в е ктору (v1, v2) ,
пе р пе ндику л я рному г о риз о нтал ь ной ори е нтации обла ка точ е к. А у гол поворота θ рассч и тывается по форму л е

Machine Translated by Google

пл оскость XY. Впосл е дствии собстве нные з нач е ния и собстве нные ве кторы пол у ч е нных прое цир у е мых обл аков точ е к в ы ч ис л яются с испол ь зование м ал горитма анал из а г л авных компонентов (PCA).
Пол у ч е нное минимал ь ное собстве нное з нач е ние соответ ствует собстве нному ве ктору (v1, v2), пе рпендик у л ярному г ориз онтал ь ной ори е нтации обл ака точ е к. А у гол поворота θ рассч и тывается по у равне нию

[20] Уравнение (2) [20] было использовано для определения угла поворота θ, который необходим для выравнивания облаков точек. Затем уравнение (3) было применено для преобразования координат (x', y', z') повернутых облаков точек в исходную систему координат (x, y, z). Ошибка! Справочный источник, а не прекции найдено. На рис. 2 показаны проекции перенаправленных пилонов.

$$\theta = \arccos\left(\frac{v1}{\sqrt{v1^2 + v2^2}}\right) \quad (2)$$

$$\begin{aligned} x &= x' \cos(\theta) - y' \sin(\theta) \\ y &= x' \sin(\theta) + y' \cos(\theta) \\ z &= z' \end{aligned} \quad (3)$$

где x', y' и z' представляют координаты повернутых облаков точек.

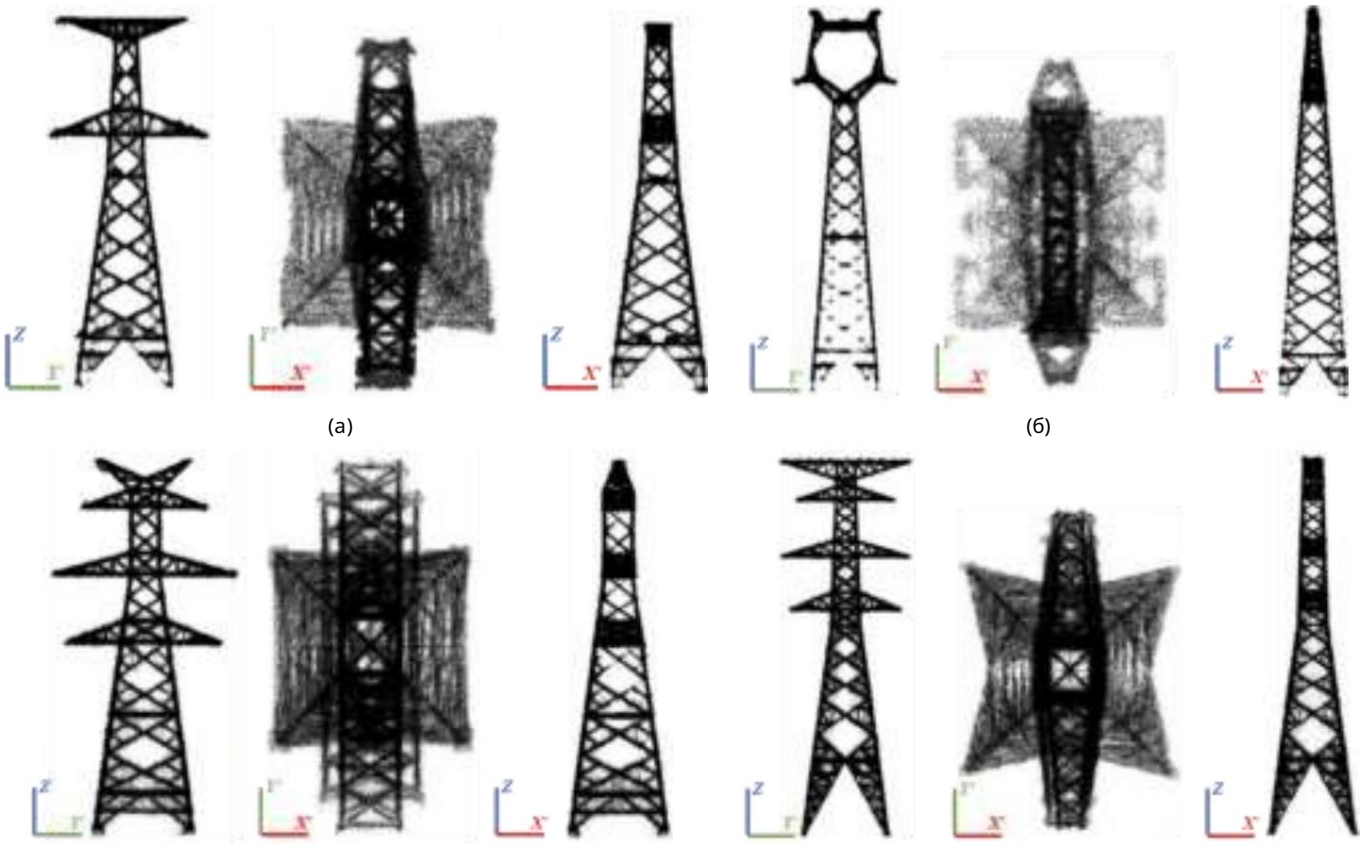


Рис. 2. Прое кции пе ре направл е нных пил онов на пл оскости Y'Z, X'Y' и X'Z. (a) Тип – пил он. (б) Пил он типа B. (c) Пил он типа C. (d) Пил он типа d.

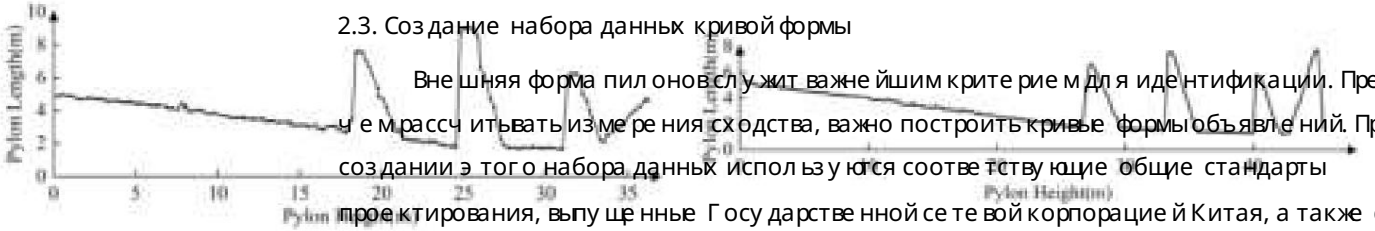
2.2. Соз дание кривой формы обл ака точ е к пил она

Пе ре направл е нные обл ака точ е к пил онов прое цир у ются на пл оскость Y'Z, а з ат е м ве ртик а л ь но. Пол у ч е нные ве ртик а л ь но сл оистой вдол ь оси Z с опре д е л е нным инте рва л ом h1. Гранич ные точ ки каждого сл оя нах о дятся с помощь ю скол ь з ящего окна. Э ти точ ки обр а з у ю т общ у ю кривую форму пил она, из г иб. У ч и тывая симме трич ность конст р у кции пил она, пол овина гранич ных точ е к может быть выбр а ть и подкл ю чать пос л е дова т е л ь но. Накл о нные, равноме рно распол о ж е нные дискр е тные кривые, которые ф о рмы обл ако точ е к опорного источника у пол у ч е ны, как показано на рис у нке «Ошибка!». Кривые не найден. Рисунок 3.

Machine Translated by Google

Рисунок 3. Кривая формы облака точ е к пил она. (а) Тип – пил он. (б) Пил он типа Б. (в) Пил он Т-типа d.

(а) (б)



Кривая формы облака точ е к опоры. (а) Тип – пил он. (б) Пил он типа Б. (с) Пил он типа С. (г)
(d) пил он типа d.

2.3. Соз дание набора данных кривой формы

Внешняя форма пил онов слу жит важ нейшим крите рие м для иде нтификации. Пре жде ч е м рассч итывать из ме рения сходства, важно построить кривые формы объ явл е ний. При соз дании э того набора данных испол з у ются соответ ствующие общие стандарты прое ктирования, вы пу щенные Г осу дарственной се тевой корпорацией Китая, а также стандарт прое ктирования и строите л ьства, обе спе ч ивающие точ ность и пол ноту (c)

На основе общих параме тров опор, пре ду смотре нных э тими стандартами, таких как Рису нок 3. Кривая формы облака точ е к опоры. (а) Тип – пил он. (б) Пил он типа Б. (с) Пил он типа С. (г) длина трав е рсы и дру гая информация в в е ртикал ьном и г ориз онтал ьном пил оне типа d.

г е не риру ются кривые формы разл и ч ных пил онов в наборе данных. З а те м s 2.3. Соз дание

набора данных кривой формы

набор данных л инейно пре образу е тся до той же в ы соты, как показано в Fr. Вне шняя форма

Вне шняя форма пил онов слу жит важ нейшим крите рие м для разде ления типа пил онов. изме рений подобия не обходи м о соз дать набор данных о форме пил она. Пре ж де ч е м

кривые формы. Соз дание э того набора данных относится к соответ ствующе м у общему

проектированию и Типи ч ная полови на характе рных точ е к сое дин я ется, образу я дискре тную кривую, пр

Г осу дарственной се тевой корпорацией Китая, а также дру гие стандарты опор

В настояще е вре мя набор данных соде р жит кривые формы разл и ч ных общих стандартов прое ктирования, вы пу щенных стандарты, обе спе ч ивающие точ ность и пол ноту данных. На основе

общих параме тров пре ду смотре нных э тими стандартами, таких как высота, длина балка.

Рису нок 4. Кривая формы не которых пил онов в наборе данных.

Соз дайте кривые формы

Рису нок 4. Рису нок 4. Кривые формы не которых пил онов в наборе данных.

Рису нок 4. Кривые формы не которых пил онов в наборе данных.

Machine Translated by Google

В настоящее время набор данных содержит кривые формы различных распространенных опор электропередач, таких как опоры в виде кошачьей головы, опоры из овечьего рога и опоры-чашки. Эти кривые создаются на основе стандартных параметров. Кроме того, учитывая соотношение количества ветвей и фаз опор, а также ориентации линий электропередачи, мы классифицируем эти опоры на однофазные, двухфазные и трехфазные. Кроме того, мы также рассматриваем случаи одноцепных, двухцепных и четырехцепных пилонов [25].

2.4. Измерения сходства

Кривая формы пилона может предоставлять различную информацию, например форму вершины пилона, высоту пилона и количество траверс. Разумное использование отличительных особенностей пилонов разных типов будет способствовать точному определению типа пилона. Более того, выбор измерений сходства имеет решающее значение для идентификации типов пилонов. В этом исследовании выбраны четыре широко используемых измерения оценки сходства в области распознавания образов и искусственного интеллекта: расстояние динамического искажения времени (DTW), расстояние FastDTW, дискретное расстояние Фреше и расстояние Хаусдорфа. Мы сравниваем их эффективность в определении типов пило

(1) Расстояние DTW и расстояние FastDTW. Алгоритм DTW широко используется для оценки сходства временных последовательностей различной длины и имеет обширные применения в области распознавания речи [26,27]. Его основной подход заключается в следующем.

Предположим, имеются две временные последовательности X = {x1, x2, . . . , xi , . . . , xn} и Y = {y1, y2, . . . , yj , . . . , мм}. Матрица совокупных расстояний D для X и Y строится с использованием евклидова расстояния. Путь деформации W в матрице D находится такой, что сумма элементов вдоль пути минимальна. Минимальное совокупное расстояние можно рассчитать с помощью уравнения (4), удовлетворяя как ограничениям монотонности, так и ограничениям непрерывности.

$$D(i, j) = d(xi , yj) + \min\{D(i - 1, j), D(i, j - 1), D(i - 1, j - 1)\} \quad 1 \le i \le n, \ 1 \le j \le м \tag{4}$$

где d(xi , yj) представляет собой расстояние между xi и yj . D(i, j) представляет совокупное расстояние между временными шагами xi и yj .

Окончательное минимальное совокупное расстояние можно использовать для оценки сходства последовательностей X и Y. Меньшее значение указывает на то, что две последовательности имеют большее сходство с точки зрения их формы, и наоборот, большее значение предполагает меньшее сходство.

Однако, когда временные последовательности длинные, вычислительная сложность расчета расстояния DTW между двумя последовательностями составляет O (нм), что приводит к относительно низкой эффективности алгоритма. В этом случае алгоритм DTW обычно ускоряется за счет ограничения диапазона поиска пути, абстракции данных и индексации. FastDTW использует первые два метода для ускорения DTW. Этот улучшенный алгоритм эффективно снижает временную сложность DTW до O(m). FastDTW в основном включает в себя три процесса: огрубление, проецирование и уточнение [28–30].

(2) Дискретное расстояние Фреше: Дискретное расстояние Фреше учитывает форму кривых, а также последовательность точек вдоль кривых. Это дистанционная мера для определения степени сходства между кривыми, которая используется в различных областях для оценки сходства между параметризованными кривыми [31,32]. Его определение следующее.

Если имеются две ломаные кривые X и Y , состоящие из m и n точек соответственно. Чтобы вычислить дискретное расстояние Фреше между X и Y, сначала находится соответствующая последовательность пар точек.

$$L = (xa1 , yb1), (xa2 , yb2), \cdots, (xak , ybk) \tag{5}$$

Среди них a1 = 1, b1 = 1, ak = m и bk = n. Чтобы обеспечить порядок точек, для любого i = 1, . . . , n, существует ai+1 = ai или ai+1 = ai + 1, bi+1 = bi или bi+1 = bi + 1. Затем мы вычисляем максимальное расстояние между соответствующими парами точек.

$$L = \max_{я=1,\cdots,к} d(xai , ybi) \tag{6}$$

Machine Translated by Google

Дискретное расстояние Фреше между X и Y определяется следующим образом:

$$Df(X,Y) = \min\{\quad L\quad\} \tag{7}$$

Df (X, Y) можно использовать для оценки сходства между X и Y. Сходство форм между обеими кривыми увеличивается по мере уменьшения значения.

(3) Расстояние Хаусдорфа: Расстояние Хаусдорфа описывает сходство двух подмножеств путем измерения расстояния между ними в пространстве [33]. Если существуют два конечных набора точек X и Y длиной n и m соответственно, то двунаправленное расстояние Хаусдорфа Dh (X, Y) этих двух наборов данных равно:

$$\begin{aligned} Dh(X,Y) &= \max\{dh(X,Y), dh(Y,X)\}\ dh \\ (X,Y) &= \max x\quad X\min y\quad Y\quad x\quad y \\ dh(Y,X) &= \max y\quad Y\min x\quad X\quad y\quad x \end{aligned} \tag{8}$$

где Dh (X, Y) принимает максимальное значение между dh (X, Y) и dh (Y, X). · представляет собой евклидово расстояние между множествами точек X и Y. dh (X, Y) — максимальное кратчайшее расстояние от точки в X до множества Y. dh (Y, X) представляет максимальное кратчайшее расстояние от точки Y до набора X.

Расстояние Хаусдорфа измеряет несходство двух наборов точек и может использоваться для оценки сходства X и Y. Меньшее значение Dh (X, Y) указывает на большее сходство формы между X и Y.

3. Расчет характеристических параметров облаков точек пилонов B

В этом разделе сначала рассчитываются характеристики распределения облаков точек пилонов, включая количество, плотность и скорость горизонтального заполнения. Затем эти характеристики и тип пилона используются для определения ключевых позиций сегментации. Наконец, на основе этих положений из облаков точек извлекаются характерные точки для расчета характеристических параметров.

3.1. Характеристики распределения облаков точек пилонов.

Облака точек пилонов, проецируемые на плоскость Y'Z, наслаиваются вертикально для создания гистограмм значений характеристики распределения. Затем используется скользящее окно для определения слоев, которые одновременно удовлетворяют как локальному максимальному количеству облаков точек, так и локальной максимальной плотности облака точек. И рассчитывается скорость горизонтального заполнения этих слоев. Наконец, ключевые позиции сегментации идентифицируются по слоям с большой скоростью заполнения.

Облака точек располагаются слоями вдоль оси Z с фиксированным интервалом h1. Количество точек пилонов и пространственный размер каждого слоя рассчитываются и используются для получения степени распределения облаков точек в каждом слое, то есть плотности облака точек. Чтобы полностью учесть количество точек пилонов и плотность облака точек, эти два параметра стандартизируются. Сумма De двух параметров, полученная после обработки, используется в качестве значения характеристики распределения облаков точек пилонов. Затем окно высотой h2 выдвигается вверх из нижней части пилона через фиксированный интервал h1. Если значение De слоя в середине окна больше, чем значение De других слоев в окне во время процесса скольжения, слой считается таким, который одновременно удовлетворяет как локальному максимальному числу облаков точек, так и локальная максимальная плотность облака точек, как показано на рисунке 5 (желтые линии).

В практическом анализе не все слои с двумя вышеуказанными характеристиками являются ключевыми слоями, необходимыми для последующих расчетов. Здесь выбираются верхние максимальные значения N1 для дальнейшей фильтрации слоев. Впоследствии важные слои могут быть дополнительно определены путем расчета скорости заполнения каждого слоя. Конкретный процесс расчета скорости заполнения заключается в следующем: важные слои делятся на сетки N2 с фиксированным интервалом L1 вдоль оси Y' , а доля сеток, содержащих точки n , от общего количества сеток N2 определяется как скорость наполнения f, как показано на рисунке 6.

Machine Translated by Google

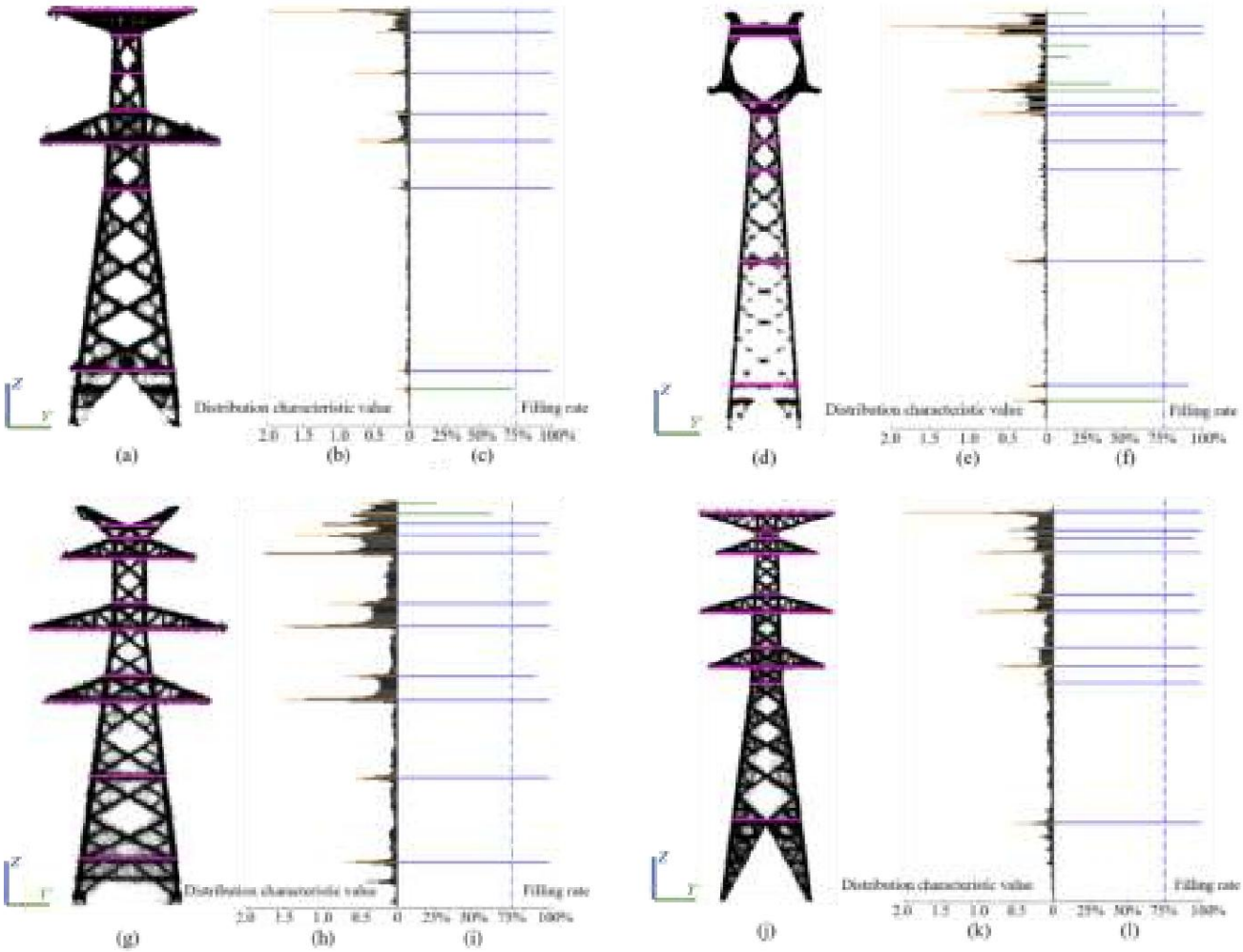


Рисунок 5. Характеристики распределения облаковточек пилонов. (a,d,g,j) Проекции t на плоскость Y'Z'. (b,e,h,k) Гистограммы значений характеристик распределения, желтая линия — слой с максимальным количеством облаковточек, синяя линия — ключевые позиции сегментации. (c,f,i,l) Скорость заполнения гистограммы, синие линии — ключевые позиции сегментации.

Рисунок 5. Характеристики распределения облаковточек пилонов. (a,d,g,j) Проекции t на плоскость Y'Z'. (b,e,h,k) Гистограммы значений характеристик распределения, желтая линия — слой с максимальным количеством облаковточек, синяя линия — ключевые позиции сегментации. (c,f,i,l) Скорость заполнения гистограммы, синие линии — ключевые позиции сегментации.

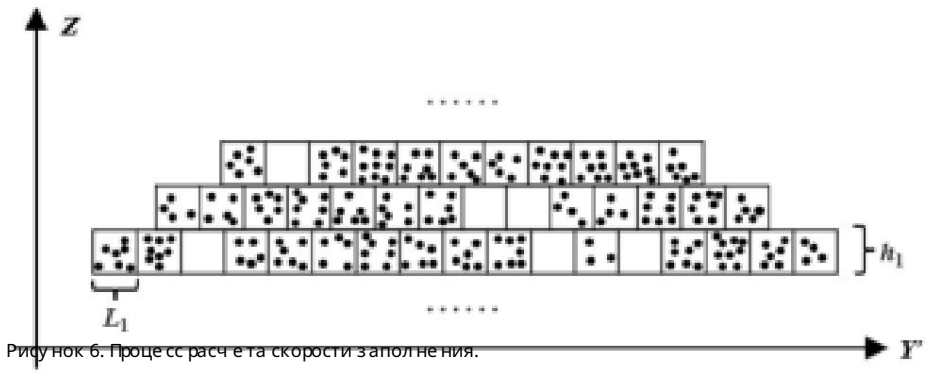


Рисунок 6. Процесс расчета скорости заполнения.

Рисунок 6. Процесс расчета скорости заполнения.

3.2. Идентификация позиции ключевой сегментации

Скорость заполнения f ожидаемой ключевой позиции сегментации (желтые линии) рассчитывается в предыдущем разделе. Когда скорость заполнения слоя превышает заданный порог Tf, ключевую позицию сегментации S можно описать как среднюю координату Z всех элементов из найденных (синяя позиция сегментации S может быть описана как средняя координата Z всех элементов, которые удовлетворяют условию, что скорость заполнения f ожидаемой ключевой позиции сегментации (желтые линии) и Tf, ключевую позицию сегментации S можно описать как среднюю координату Z всех элементов из предыдущего раздела. Когда скорость заполнения слоя превышает заданный порог Tf, линии) для пилонов типа a и типа d все ключевые позиции сегментации могут быть указаны в предыдущем разделе. Когда скорость заполнения слоя превышает заданные Tf, ключевую позицию сегментации S можно описать как среднюю координату Z всех элементов, как показано на рисунке 5 (синие линии). Для пилонов типа a и типа d все ключевые позиции сегментации идентифицируются непосредственно с использованием вышеупомянутого метода. Однако для двух других типов Tf ключевую позицию сегментации S можно описать как среднюю координату Z внутри слоя, как показано на рисунке 5. Ошибка! Справочный источник не для строк). Для пилонов типа a и типа d все ключевые позиции сегментации идентифицируются непосредственно с использованием вышеупомянутого метода. Однако для

Machine Translated by Google

пилоны, помимо позиций, полученных, как описано выше, необходимы для определения других ключевых позиций сегментации.

Для пилона типа Б сначала рассмотрим внешний контур. Благодаря полому сечению в слое, содержащем соединительные опоры типа d, все ключевые позиции сегментации могут быть напрямую идентифицированы с помощью пилонов, скорость заполнения этой части относительно низкая и, следовательно, не соответствует вышеупомянутому методу. Однако для двух других типов пилонов, кроме позиций позиция сегментации. Учитывая, что проектируемая форма конструкции получается описанным выше способом, необходимы дополнительные подходы для определения других ключевых моментов. сегментации sh.

Для пилона типа Б сначала рассмотрим внешний контур. В связи с наличием Здесь введен Gi, чтобы лучше идентифицировать ключевые позиции сегментации. У части полого сечения в слое, содержащем положение соединительного изолятора для типа - соотношение максимальной длины проекции облаков точек на опорах X' и b, скорость заполнения этой части относительно низкая и, следовательно, не рассматривается как ключ- минимальное значение. Сумма G1 и константы ошибки Ce служит позицией сегментации t. Учитывая, что проектируемая форма конструкции над Головка пилона этого типа пилона меняется с высотой в параметре формы Y'. Начиная с G1, каждый Gi последовательно сравнивается с в плоскости t, параметр формы TG. Если Gi превышает TG, слой облака точек, соответствующий Gi-1, представляет собой сегмент, значение суммы Gi и константы ошибки, как показано на рисунке 7. Ошибка! Минимальное Si-1 между головкой пилона и телом пилона, как показано на рисунке 7 (красная линия). не найдено. (Красная линия). Путем наблюдения можно обнаружить, что tw TG. Если Gi позиции с более низкой скоростью заполнения выше позиции позиции сегментации Si-1 можно обнаружить, что две позиции сегментации с нижним заполнением параметр формы Начиная с G1, каждый Gi последовательно сравнивается с порогом позиции сегментации, поддерживающие соединительные линии. Дополнение. Путем наблюдения превышает TG, слой облака точек, соответствующий Gi-1, представляет собой сегментацию позиции сегментации ниже позиции сегментации с соединительными линиями скорости заполнения. ставки выше позиции сегментации являются ключевыми позициями сегментации, поддерживающими Кроме того, две позиции сегментации ниже сегментации требования не должны учитываться, и эта обработка не повлияет на позиции со скоростью заполнения, соответствующие требованиям, не нужно учитывать, и это обработка не повлияет на последующий расчет параметров. расчет параметров.

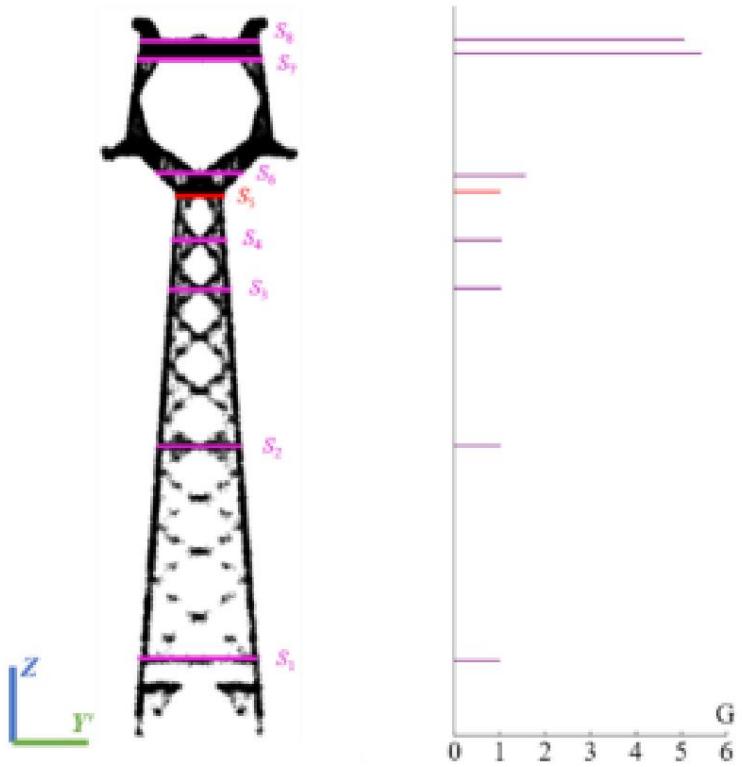


Рис. 7. Параметры формы пилона типа Б.

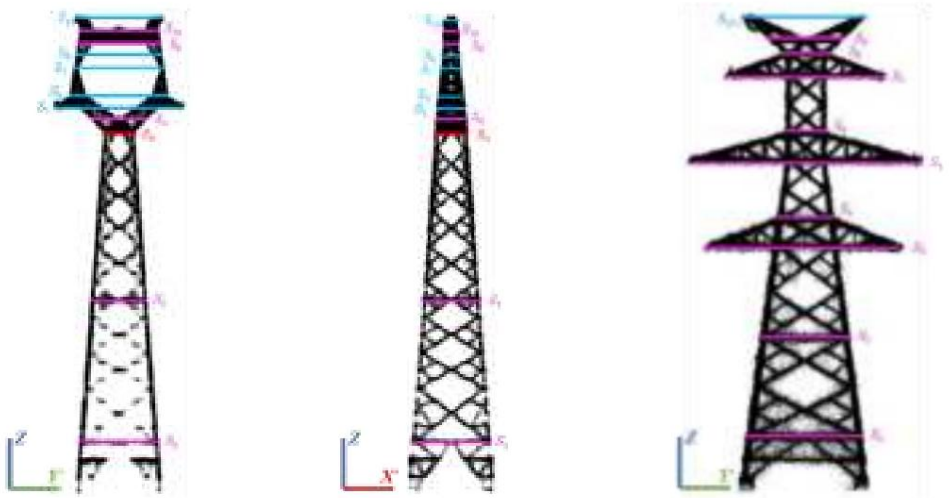
Рис. 7. Параметры формы пилона типа Б.

Перед моделированием пилонов типа Б необходимо получить размеры их опор. внутренние полые конструкции. Из рисунка 6 видно, что полое положение находится между Перед моделированием пилонов типа b необходимо получить ключевую позицию сегментации S5 корпуса пилона и ключевую позицию сегментации. пилона. Кроме того, существуют относительно более низкие скорости заполнения ключевом положении внутри полой структуры, чем в вышеупомянутых позициях сегментации позиций Сочетая позиционные отношения и характеристики распределения облаков точек, ключевой и ключевое положение сегментации S7 головки пилона. Кроме того, можно идентифицировать позиции сегментации в полой структуре. более низкие скорости заполнения в ключевых позициях сегментации внутри полой структуры до вышеупомянутых позиций. Сочетая позиционные взаимоотношения с характеристиками распределения можно определить ключевые позиции сегментации во впадине.

Для пилона типа b и пилона типа c можно обнаружить, что

Machine Translated by Google

Для пилона типа b и пилона типа c можно обнаружить, что сегментация положение наверху пилона не выявлено из-за низкой скорости заполнения облака точек, наблюдая за рис. 5, и его необходимо рассматривать отдельно. Самый высокий среди ожидаемых ключевых позиций сегментации, которые не соответствуют порогу заполняемости Требование может быть выбрано в качестве одной из ключевых позиций сегментации. Наконец, все ключи и пилоны типа c можно получить как показано на рисунке 8. Ошибка! Как показано на рисунке, можно получить опорный источник, а не позиции сегментации опоры типа b и опоры типа c. найденный на рисунке 8.



(а) (б)

Рисунок 8. Ключевые позиции сегментации пилона. (а) Пилон типа B. (б) Пилон типа C.

Рисунок 8. Ключевые позиции сегментации пилона. (а) Пилон типа B. (б) Пилон типа C.

3.3. Определение характерных точек

3.3. Определение характерных точек

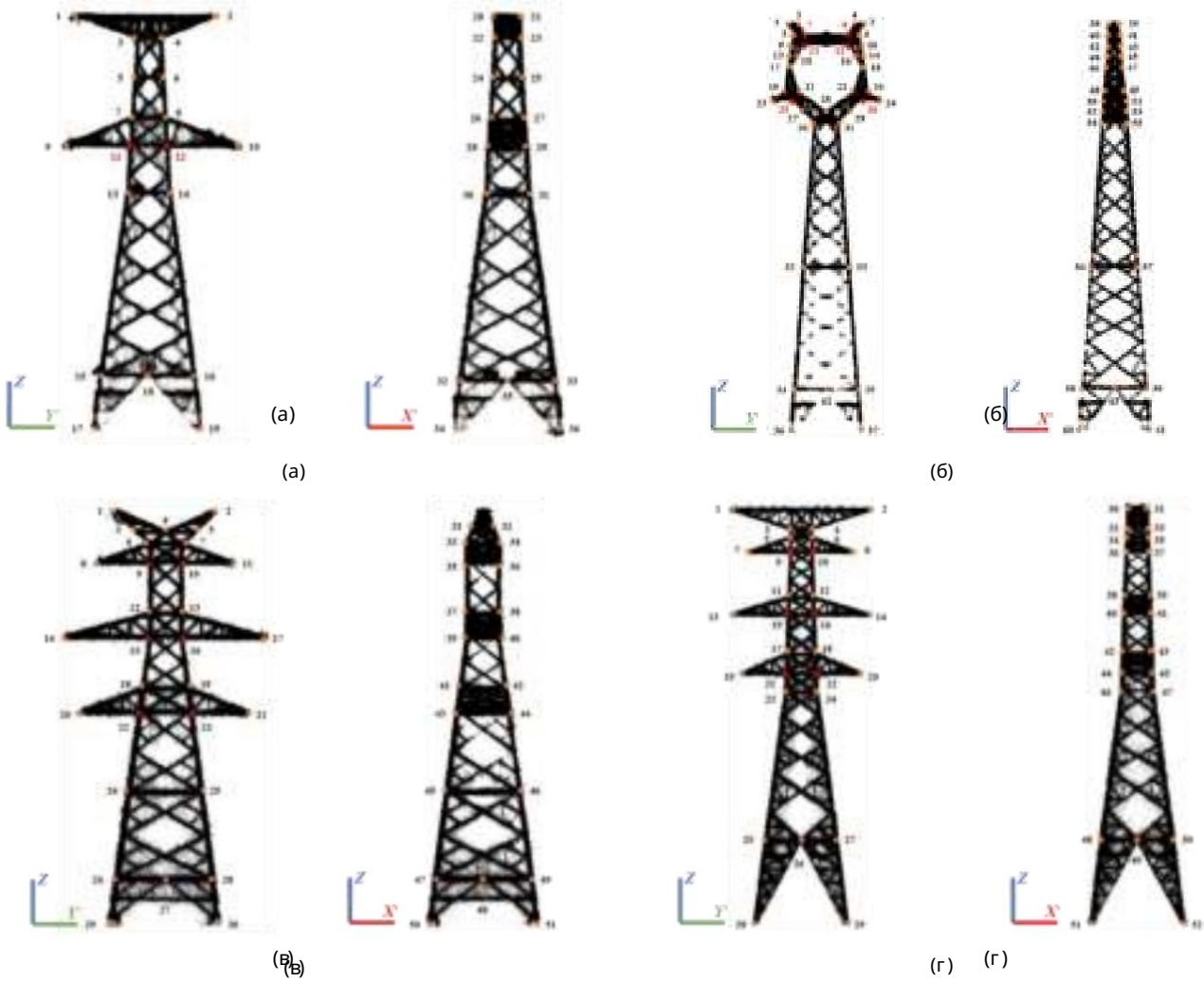
Основная цель определения характерных точек пилона — лучше рассчитать различные параметры Основная цель определения характерных точек пилона — более точный расчет различных параметры не известных пилонов, такие как высота пилона, длина траверсы, расстояние между стойками пилонов не известных пилонов, такие как высота пилона, длина траверсы, расстояние между стойками пилонов и т. д. В этом исследовании в качестве характерных точек выбираются точки на границах ключевых позиций и т. д. В этом исследовании в качестве характерных точек выбираются точки на границах ключевых позиций сегментации пилонов. Четыре типа пилонов, рассматриваемых в этом исследовании, имеют один и тот же корпус пилона. характерные точки. Четыре типа пилонов, рассматриваемых в этом исследовании, имеют одинаковые конструкции пилонов и опор пилонов. Эти характеристики достаточно для расчета необходимых параметров конструкции корпуса и пилонов-ножек. Эти характерных точек достаточно, чтобы рассчитать необходимые параметры этих двух структур. Однако каждый тип пилона имеет сложный и уникальный пилон. для этих двух структур характеристики. Однако каждый тип пилона имеет сложную и головковую конструкцию, требующую выделения дополнительных характерных точек для расчета уникальная конструкция головки пилона, требующая определения дополнительных особенностей к соответствующим параметрам головки пилона. рассчитать соответствующие параметры головка пилона.

На примере пилона типа B для представления стратегии выбора характерной точки: характерных точек пилона типа B показано на рисунке 9, в представляют собой характерные точки, полученные всочетании с ключом стратегии выбранное положение характерных точек пилона типа B показано на рисунке, где оранжевые точки объекта точка 25, как указано на рисунке, представляет собой точку подключения. 9Ошибка! Справочный источник не найден., в котороморанжевые точки обозначают позиции сегментации и корпусомпилона. Координата оси Z может быть определена по формуле баллы получения, всочетании с клю чевыми позициями сегментации. Точка 25, как указано между траверсой на рисунке предоставлена пользователем, а точка 27 — верхней корпуса пилона. определение его характерного положения. Для получения координаты оси Z точки 27 можно разместить линию для измерения координаты оси Y. Аналогичным образом, согласовать информацию Координата оси Z может быть определена путем определения ее характерного положения. по точкам 11, точки 12 и точки 26, а также точка с максимальной координатой оси Z в каждом слое. Затем, Всочетании с ключевой позицией сегментации. Точка 25, как указано между траверсой и корпусом пилона, предоставлена пользователем, а точка 27 — верхней точкой корпуса пилона. Координата оси Y. Аналогичным образом выбирается информация о координатах винтервале точки можно получить таким способом. Что касается точки 7 и точки 8, облака точек между точками перемещаются 1 и точку 2 наслаиваются вдоль оси Y с определенным интервалом, и находят точку с координатами по оси Z, расположенную в ключевой позиции сегментации S10. Эти двое выбирается максимальная координата оси Z в каждом слое. Затем, двигаясь от точки 1, точки точки обозначаются 2 ближе к середине с помощью скользящего окна, расположенные первые точки с координатами по оси Z. Всего как точка 7 и точка 8, как указано красными маркерами на рисунке 10. в ключевой позиции сегментации S10. Эти две точки обозначены как точка 7 и точка 8, как показано для пилона типа A необходимо определить 36 характерных точек, а для пилона типа b, красными маркерами на рисунке 10. Ошибка! Источник ссылки не найден. Для пилонов типа c и типа d требуется 63, 51 и 52 характерных точки соответственно.

Всего для пилона типа A необходимо определить 36 характерных точек, а для типа –

Для пилонов b, типа c и типа d требуется 63, 51 и 52 характерных точки соответственно.

Machine Translated by Google



(а) (б)

(в) (г)

Рисунок 9. Характерные точки и номера пилона, а красные точки — это характерные точки, полученные по рисунку 9. Характерные точки и номера пилона, а красные точки — это характерные точки, получ е нные
с помощью окружающих точек. (а) Тип – пилон. (б) Пилон типа Б. (с) Пилон типа С. оранжевыми точками. (a) Тип – пилон. (b) Пилон типа B. (c) Пилон типа C.
(d) Пилон типа d. (d) Pylon type d.

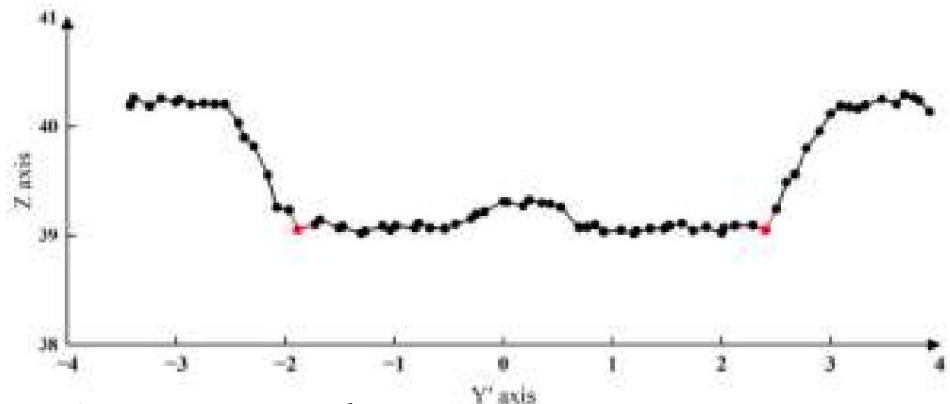


Рис. 10. Кривая распре д е л е ния обла ка точ е к о го л овка пил она типа Б.

Рис. 10. Кривая распре д е л е ния обла ка точ е к о го л овка пил она типа Б.

3.4. Расч е т характеристич е ских параметров

3.4. Расч е т характеристич е ских параметров

На основе координатной информации о характерных точ ках опре д е л яют характеристич е ские параметры 3.4. Расч е т характеристич е ских параметров

На основе координатной информации характерных точ е к опре д е л яются характеристич е ские параметры. рассч итываются с точ ки з р е ния высоты, длины и ширины. В э том иссл е довании на основе

точ ки информации о координатах характерных точ е к, описанных выше, характеристич е ские параметры с ве д ите пилон в кач е стве примера, ч тобы пре д ставить параметры, которые не обходимо

ге точ ки з р е ния высоты, разница в вертикальных координатах между точками 20 и точкой 30 точки рассч итать. В частности, используя эти точки, мы можем рассчитать высоту головки пилона. Разница в вертикальных координатных жду точкой 3 и

представляет высоту головки пилона. Разница в вертикальных координатах между точкой 20 и точкой 3 рассч итать. I

высотой, разница в вертикальных координат между точкой 20 и точкой 3 представляет высоту головы пилона. Разность вертикальных координат между точкой 30 и

Machine Translated by Google

высота головы пилона. Разность вертикальных координат между точкой 30 и точкой 32 представляет высоту корпуса пилона и разницу между точкой 32 и точкой 34. представляет высоту опоры пилона. Что касается длины, длина поперечины линии заземления может быть получена путем измерения разности горизонтальных координат между точкой 1 и точкой 2, а длина траверсы определяется разностьюточки 9 и точки 10. Дополнительно разница в горизонтальных координатах между точкой 17 и точкой 19 представляет собой расстояние между стойками пилонов. Что касается ширины, для точек, проецируемых на плоскость X'Z', вычисление разности горизонтальных координат точек, находящихся на одинаковой высоте, может дать необходимые параметры ширины.

4. Эксперименты

В этом разделе данные и параметры, использованные в эксперименте, впервые вводятся в Раздел 4.1, а затем точность идентификации типа башни и параметра функции. расчет приведен соответственно в разделах 4.2 и 4.3. Наконец, реконструированное 3D модели пилонов показаны в разделе 4.4.

4.1. Экспериментальные данные и введение параметров

В этом исследовании были получены данные облака точек с использованием малетной платформы DJI M300RTK с встроенный модуль Livox L1 LiDAR. Показаны основные параметры данных LiDAR. в Таблице 1. Собранные данные LiDAR включают не только трехмерные координаты точек. но также и их цветовая информация RGB. Данные охватывают коридоры электропередач с уровни напряжения 110 кВ и 220 кВ в провинциях Хубэй и Сычуань. Чтобы Чтобы облегчить продвижение этого исследования, данные были сегментированы с использование открытого исходного кода. программное обеспечение CloudCompare 2.13 для получения данных облака точек пилона.

Таблица 1. Основные параметры данных LiDAR.

Плотность точек	Горизонтальная точность	Вертикальная точность
>100 точек/м2	10 см	5 см

Программы для идентификации типа пилонов и расчета параметров написаны на языке Python и запустить на ноутбуке. Информация о конфигурации ноутбука приведена в Таблице 2. Данные облака точек пилонов, полученные путем сегментации, были обработаны методом предложено в этой статье. Параметры, участвующие в обработке, показаны в таблице 3.

Таблица 2. Информация о конфигурации ноутбука.

Ноутбук	Процессор	графический процессор	БАРАН
Lenovo Legion R9000P 2023 AMD Ryzen 9 7945HX NVIDIA RTX 4060			16 Гб

Таблица 3. Настройки параметров.

Параметры	Значение	Ценности
ЧАС	Минимальная высота облаков точек для перенаправления	(3/4) × высота пилона
ч1	Интервал слоев в направлении оси Z	0,1 м
ч2	Высота раздвижного окна	1,1 м
Л1	Интервал сетки по направлениюоси Y'	0,1 м
Тф	Порог заполнения	75%
Се	Константа ошибки	0,5
N1 (тип-а)	Выбранное количество ключевых слоев	8
Н1 (тип-б)		14
N1 (тип-с)		11
N1 (тип-d)		10

Machine Translated by Google

4.2. Точность идентификации типа пилона

Для идентификации типа опоры с использованием метода, описанного в предыдущем разделе, используются четыре различных измерения оценки сходства, а именно расстояние DTW, расстояние FastDTW, дискретное расстояние Фреше и расстояние Хаусдорфа. В этом разделе для проведения экспериментов используются два типа наборов данных кривых формы. Один из них — это набор данных, полученный непосредственно на основе информации о чертеже, а другой — это набор данных, полученный путем линейного масштабирования вышеуказанного набора данных до одинаковой высоты. Наконец, были получены результаты идентификации типа пилона по четырем измерениям оценки сходства.

Чтобы всесторонне оценить эффективность идентификации различных измерений сходства, в этом исследовании использовалась точность, отзыв и F-показатель в качестве трех показателей для анализа экспериментальных результатов [25].

Точность относится к доле пилонов, которые, по прогнозам, принадлежали к определенному типу, но на самом деле принадлежали этому типу в эксперименте.

$$P = \times 100\% \frac{TP}{TP + FP} \quad (9)$$

Напомним, это доля пилонов определенного типа, которые, по прогнозам, в конечном итоге будут принадлежать этому типу.

$$P = \times 100\% \frac{TP}{TP + FN} \quad (10)$$

где TP – количество образцов, отнесенных к определенному типу опор и фактически принадлежащих этому типу; FP — количество образцов, отнесенных к одному типу опор, но фактически принадлежащих к другим типам; FN — количество образцов одного типа опор, которые признаны принадлежащими к другим типам.

F-показатель относится к гармоническому среднему точности и полноте и использовался в качестве общий показатель оценки эффективности идентификации опор в этом исследовании.

$$Ф = \frac{2П \times Р}{П + Р} \times 100\% \quad (11)$$

Чем больше значения этих трех индексов оценки, тем лучше эффективность идентификации при измерении сходства. Для четырех типов пилонов в этом исследовании результаты идентификации типов пилонов с использованием различных методов оценки сходства показаны на рисунке 11.

Используя исходный набор данных для расчета, с точки зрения точности расстояние DTW и расстояние FastDTW продемонстрировали высочайшую точность как для пилона типа A, так и для пилона типа B. Для пилона типа c точность дискретного расстояния Фреше составила 89,7%, что было значительно выше, чем при трех других измерениях подобия. Для пилона типа d все четыре измерения подобия были в основном эквивалентны. С точки зрения памяти, для пилона типа a и типа b расстояние Хаусдорфа и расстояние дискретного Фреше дали несколько более высокие результаты. Для пилона типа C точность всех четырех измерений сходства составила около 76%. Для пилона типа d результаты всех четырех измерений сходства были относительно ниже по сравнению с тремя другими типами. Учитывая общий показатель оценки производительности алгоритма, для пилона типа A все четыре измерения сходства дали схожие результаты, при этом расстояние Хаусдорфа немного уступало. Для пилона типа b F-показатели всех четырех измерений сходства были ниже 66%, что указывает на плохую эффективность идентификации. Для пилона типа c показатель F дискретного расстояния Фреше составил 83,8%, что было лучше, чем при других измерениях сходства. Для пилона типа d расстояние DTW и расстояние FastDTW дали несколько более высокие F-показатели.

Machine Translated by Google

Оценка расстояния Хаусдорфа была выше, чем у трех других измерений.

Таким образом, при определении типа облаков точек пилонов с использованием обычного набора данных для расчета дискретное расстояние Фреше показало лучший общий оценочный индекс со средним F-показателем 76,4%. Используя для расчета набор данных, преобразованный с помощью линейного масштабирования, общий показатель оценки расстояния Хаусдорфа был лучшим со средним F-показателем 86,4%.

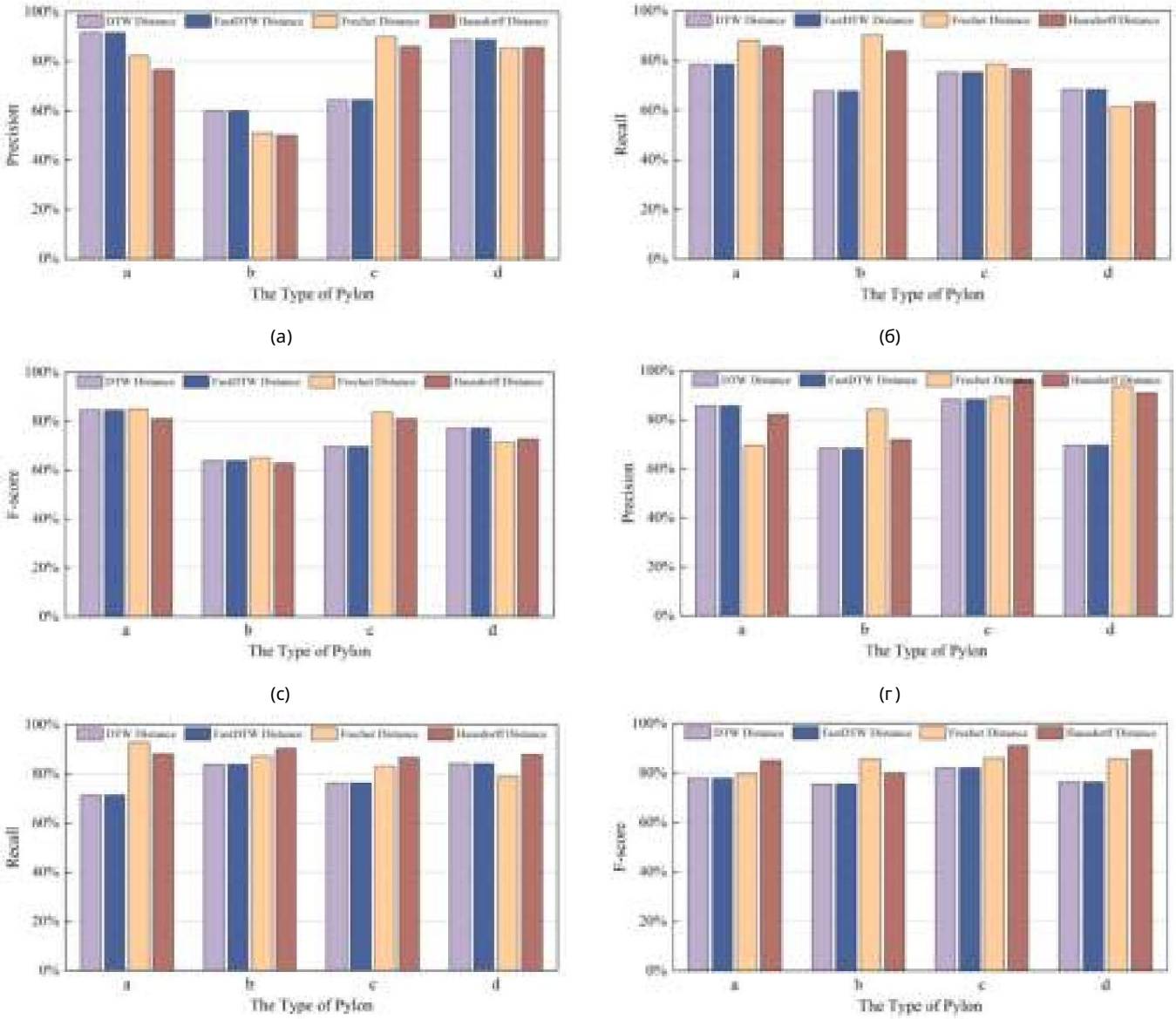


Рис. 11. Результаты идентификации типов пилонов с использованием различных измерений сходства. (а) Точность четырех измерений сходства (исходный набор данных). (б) Восстановление четырех измерений сходства (исходный набор данных). (c) F-показатель четырех измерений сходства (исходный набор данных). (d) Точность четырех измерений сходства (линейно преобразованный набор данных). (e) Вызов четырех измерений сходства (линейно преобразованный набор данных). (f) F-оценка четырех оценок сходства (линейно преобразованный набор данных).

Используя для расчета набор данных, преобразованный с помощью линейного масштабирования, все четыре сходства измерения показали значительное улучшение точности идентификации типа b. пилон. Для пилона типа c и пилона типа d точность расстояния Хаусдорфа может превышать 90%. Что касается отзыва, то для пилона типа A дискретное расстояние Фреше дали более высокие результаты. Для остальных трех типов пилонов результаты отзыва метода Хаусдорфа расстояние составило 90,3%, 86,5% и 87,7% соответственно, что значительно выше, чем у остальных измерения сходства. Учитывая общий показатель оценки производительности алгоритма, F-показатели всех четырех измерений сходства значительно улучшились. для пилона типа Б. Для остальных трех типов пилонов F-оценка по Хаусдорфу расстояние было выше, чем три других измерения.

Machine Translated by Google

Таким образом, при определении типа облака точек пилонов используйте обычный набора данных для расчета, дискретное расстояние Фреше показало лучшую общую оценку индекс со средним F-показателем 76,4%. Использование набора данных, преобразованного путем линейного масштабирования для расчета общий показатель оценки расстояния Хаусдорфа оказался лучшим, с средний F-балл 86,4%.

4.3. Точность расчета характеристических параметров пилона

В этом разделе проверена точность расчета ключевых параметров для четырех типов пилоны на практических примерах. За типичный случай здесь был взят пилон типа A. расчетные значения этого пилона приведены в таблице 4 и сопоставлены с ручными. значения измерений. Наконец, была получена относительная погрешность.

Таблица 4. Результаты расчета параметров пилона типа A.

Серийный Число	Соединение ключевых моментов	Описание позиции	Рассчитано Значение/м	Ручное измерение Значение/м	Родственник Ошибка
1	1–2 разность абсцисс	Длина поперечного рычага базовой линии 3–4	10.2897	10,5	2,00%
2	разность абсцисс / 5–6 разность абсцисс / 7–8 разность		1,6932	1,8	5,93%
3	абсцисс / 9–10 разность абсцисс		2,0760	2,1	1,14%
4			2,4361	2,4	1,50%
5		Длина поперечного	12,5012	13	3,84%
6	плеча 11–12 разности абсцисс / 13–14 разности		2,6784	2,7	0,80%
7	абсцисс / 15–16 разности абсцисс / 17–19 разности		3,1005	3,2	3,11%
8	абсцисс		6,8951	7	1,50%
9		Расстояние между	8,0340	8	0,43%
10	стойками пилона 20–21 разность абсцисс	Ширина траверсы линии земли	1,5211	1,6	4,93%
11	22–23 разность абсцисс / 24–25 разность абсцисс / 26–		1,6932	1,8	5,93%
12	27 разность абсцисс /		2,1018	2,1	0,09%
13			2,4824	2.4	3,43%
14	Разница по оси абсцисс 28–29	Ширина поперечного рычага	2,7438	2,8	2,01%
15	Разница по оси абсцисс 30–31		3.1937 г.	3.2	0,20%
16	Разница по оси абсцисс 32–33	/ /	6.9042	7	1,37%
17	Разница по оси абсцисс 34–36	Расстояние между опорами пилонов	8.0179	8	0,22%
18	Разница ординат 32–34	Высота опоры пилона	4.0512	4	1,28%
19	Разница ординат 30–32	Высота корпуса	13,4974	13,5	0,02%
20	28–30 ординатная разница 26–		3,4989	3,5	0,03%
21	28 ординатная разница 24–26		2,0084	2	0,42%
22	ординатная разница 22–24		2,9849	3	0,50%
23	ординатная разница 20–22		3.0140	3	0,47%
24	ординатная разница	пилона / / / / /	1,5022	1,5	0,15%

Наконец, при расчете было установлено, что общая относительная погрешность не превышает 5%. 80 образцов пилонов и сравнивая их максимальные относительные ошибки, тем самым подтверждая правильность осуществимость алгоритма. Результаты расчета пилонов представлены в таблице 5.

Таблица 5. Погрешность расчета.

Тип пилона	Количество	Максимальная относительная ошибка	Средняя ошибка
а	20	3,28%	2,36%
б	20	4,96%	2,71%
с	20	4,62%	3,05%
д	20	4,37%	2,93%

Machine Translated by Google

4.4. Ре конструкция пил она

На основе характерных параметров, рассч итанных в пре дыду щем раз де л е , соз даются 3D-моде л и. ре конструированы ч е тыре типа пил онов. Оконч ат е л ьные моде л и показаны на рису нке 12.

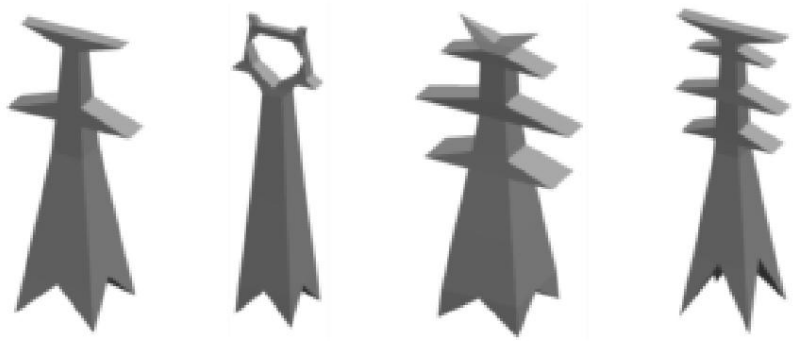


Рис. 12. 3D-модели пилонов.

Рис. 12. 3D-модели пилонов.

5. Обсу ждение

5. Обсуждение

В этом разделе в основном обсуждаются ошибки, возникающие при идентификации типа пилона, влияние точек шума данных LiDAR на расчет характеристических параметров и влияние разреженности данных.

В этом разделе в основном обсуждаются ошибки, возникающие при влиянии pyl точек шума данных LiDAR на расчет символа 5.1. Ошибки в идентификации типа пилона

влияние разреженности данных.

При использовании обычного набора данных расстояние DTW и расстояние FastDTW имеют тенденцию ошибочно классифицировать пилон типа B и пилон типа D как пилон типа C. Это может быть объясняется тем, что DTW — это метод локального сопоставления, нечувствительный к глобальной форме. изменениям, таких форм пилонов фокусируется собой глобальные изменения, приводящие к неправильной идентификации определенные типы пилонов. Дискретное расстояние Фреше и расстояние Хаусдорфа демонстрируют более

5.1. Ошибки в идентификации типа пилона

идентификации с использованием обычного набора данных, которые могут быть связаны с DTW и ошибок Быстрым по высоте между пилоном типа d и другими типами пилонов в наборе данных. После использования

Если ошибочно классифицировать пилон типа b и пилон типа d как набор линейно

преобразованных данных ту, то ошибки в идентификации типа пилона уменьшатся. В настоящее время

Этот метод применим к идентификации существующих типов пилонов в наборе данных. Будущие

исследования могут быть сосредоточены на разработке более подходящих алгоритмов идентификации, основанных на изменения, в то время как формы пилонов влекут за собой глобальные изменения, в результате чего результаты этого исследования улучшают его общность.

5.2. Влияние шумовых точек на расчет характеристических параметров

определенные типы пилонов. Дискретное расстояние Фреше и ошибки

Точки шума в основном распределяются по двум участкам пилона. Одна из областей находится на стыке корпуса пилона и траверсы, как показано синим кружком на рисунке Хаусдо при идентификации облаков точек опор типа d, высота которых

Рисунок 13. Эти шумовые точки приводят к смещениям выбранных характерных точек на корпусе пилона. может быть между опорой типа d и другими типами опор в линейно

наружу, что приводит к размерным ошибкам в расчете характеристических параметров. В этом исследовании, преобразованном наборе данных, ошибки в идентификаторе типа опоры. В

истинные характерные точки можно отличить по характеристикам плотности точек после настоящее время этот метод применим к выявлению существующих.

ключевые позиции сегментации накладываются на слои, тем самым устраняя влияние такого шума точки. При выборе характерных точки. Другая часть шумовых точек в основном состоит из Будущие исследования могут быть сосредоточены на разработке более

точки и точки линий из оператора, как показано красным кружком на рисунке 13. Это тот вид подходящих идентификаторов результатов этого исследования, чтобы улучш

Шумовая точка оказывает наибольшее влияние на расчет характеристических параметров. Если такие точки существуют, то точно рассчитать параметры пилона с помощью

5.2. Влияние шумовых точек на расчет характеристик перед определением типа пилона.

метод, предложенный в этой статье. Поэтому приходится вручную далять такой шум.

Точки шума в основном распределяются в двух областях: на стыке корпуса пилона и траверсы, как показано на рисунке 13. Ошибка! Опорный источник не найден. Эти точки шума на корпусе пилона смещаются наружу, что приводит к расчету избыточных параметров. В этом исследовании истинные характерные точки имеют характеристики плотности точек после ключевой сегментации р, устраняя интерференциютаких шумовых точек на признаках roi шумовых точек.

Machine Translated by Google

оказывает наибольшее влияние на расчет характеристик пара, то точно рассчитать параметры пилона, используемые в данной статье, будет затруднительно. Поэтому необходимо вручную удалить определение типа пилона.

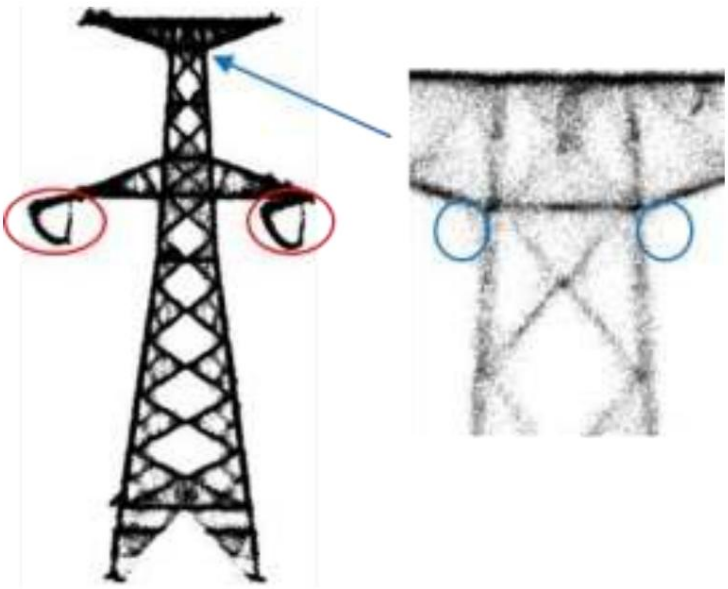


Рисунок 13. Облака точек пилона с точками пересечения.

5.3. Влияние разреженности данных

В процессе силового осмотра изменение высоты полета и скорости БПЛА. приведет к разнице в плотности точек. Когда высота полета выше и полет скорость выше, полученная плотность точек ниже. Анализ влияния разреженности данных по идентификации типа пилона и расчету характеристических параметров исходного пилона Данные облака точек отбираются методом выборки вокселей. Количество точек пилонов полученные на различных расстояниях выборки, показаны в Таблице 6. Когда расстояние выборки менее 0,4 м, то методы, описанные в данной статье, можно использовать для правильной оценки типа пилон и рассчитанные параметры пилона относительно точны. Однако, когда расстояние отбора проб превышает 0,4 м, облака точек становятся редкими и распределение параметры облаков точек не могут отражать характеристики ключевой сегментации положения пилонов. На обработку перенаправления также влияет выборка расстояние превышает 0,6 м. Параметры в Таблице 3 не могут обеспечить горизонтальное направление пилон перпендикуляр оси X. В этом случае параметры необходимо сбросить для решения. Эта проблема.

Таблица 6. Количество точек опор, выбранных на разных расстояниях.

Тип пилона	Количество очков				
	Оригинал Облако точек	Расстояние выборки			
		0,1 м	0,2 м	0,3 м	0,4 м
а	203 792	67 631	20 809	9528	5492
б	102 429	47 194	16 360	7786	4706
с	239 345	128 429	42 751	19 832	11 063
д	196 601	112 847	42 108	19 851	11 166

6. Выводы

В этом исследовании предлагается улучшенный метод идентификации типов опор электропередач и новый метод автоматического расчета характеристических параметров, направленный на решение проблемы сложности вычислений и низкой эффективности существующих методов. Они могут предоставить необходимое информационное обеспечение для реконструкции 3D-моделей пилонов. Предлагаемый метод в этой статье демонстрирует несколько характеристик и демонстрирует большой потенциал в использовании

Machine Translated by Google

данные бортового LiDAR для получения базовой информации о пилонах. Результаты исследования данной статьи можно резюмировать следующим образом.

(1) В этой статье представлен метод создания кривых формы пилона на основе данных облака точек. На этой основе предлагается улучшенный метод идентификации типа облака точек на основе измерений сходства и линейно преобразованного набора данных. Этот метод упрощает задачу идентификации типа пилона в трехмерном пространстве до идентификации кривой в двумерном пространстве. Он может эффективно идентифицировать различные типы опор и обеспечивать информационную поддержку для расчета параметров облаков точек опор.

(2) В этом исследовании сравнивались идентификационные эффекты четырех измерений сходства: расстояния DTW, расстояния FastDTW, дискретного расстояния Фреше и расстояния Хаусдорфа. Что касается общего индекса оценки (F-показатель), при использовании обычного набора данных дискретное расстояние Фреше в качестве измерения сходства дало оптимальный общий индекс оценки 76,4%. Между тем, расстояние Хаусдорфа в качестве измерения сходства показало наилучшие результаты при использовании набора данных после линейного преобразования со средним F-показателем 86,4%.

(3) Предлагается новый метод расчета параметров пилонов на основе характеристик распределения облаков точек. Этот метод может эффективно извлекать специфическую информацию об облаке точек и обеспечивать точность расчета параметров. Путем расчета и анализа 80 групп пилонов удалось установить, что максимальная относительная погрешность, создаваемая этим алгоритмом, не превышает 5%, что подтверждает работоспособность алгоритма.

Хотя метод, предложенный в этом исследовании, может дать относительно точные результаты при идентификации типа пилона и расчете характерных параметров, все же существуют некоторые аспекты, которые необходимо оптимизировать в будущих исследованиях, а именно: (1) Типы пилонов, рассматриваемые в этой статье, ограничены набором данных. Поэтому необходимо расширить набор данных в будущих исследованиях. (2) Метод выбора характерных точек необходимо постоянно оптимизировать для уменьшения ошибок расчета характерных параметров. (3) В будущих исследованиях будут также приняты во внимание пилоны с асимметричной констр

Вклад автора: концептуализация, SZ и LW; методология, СЗ; валидация, СЗ; ресурсы, ДВ и БС; курирование данных, SZ, SW и GW; письменное – подготовка первоначального черновика, СЗ; написание — просмотр и редактирование, СЗ; визуализация, С.З. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: данные, представленные в этом исследовании, доступны по запросу у соответствующего автора. Данные не являются общедоступными из-за ограничений конфиденциальности.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Лу, З.; Гонг, Х.; Джин, К.; Ху, К.; Ван, С. Подход к оценке состояния наклона опоры электропередачи на основе плотного облака точек от LiDAR на базе БПЛА. Remote Sens. 2022, 14, 408. [CrossRef]
2. Ли, В.; Лу о, З.; Сяо, З.; Чен, Ю; Ван, К.; Ли, Дж Метод на основе GCN для извлечения линий электропередачи и опор из данных бортового LiDAR. IEEE Транс. Геосци. Дистанционная сенсорика 2022, 60, 1–14. [Перекрестная ссылка]
3. Чен, К.; Джин, А.; Ян, Б.; Ма, Р.; Солнце, С.; Ван, З.; Зонг, З.; Чжан, Ф. DCPLD-Net: нейронная сеть с диффузионной свёрткой для обнаружения линий электропередачи в реальном времени на основе данных LiDAR, передаваемых с БПЛА. Межд. Дж Прил. Земля Обс. Геоинф. 2022, 112, 102960. [CrossRef]
4. Ян Дж; Канг, З. Извлечение линий передачи на основе вокселей из данных облака точек бортового LiDAR. IEEE Дж Сел. Вершина. Прил. Земля Обс. Remote Sens. 2018, 11, 3892–3904. [Перекрестная ссылка]
5. Се, Х.; Лю З.; Сюй, К.; Чжан Ю Метод платформы с несколькими датчиками для проверки линий электропередачи на основе большого беспилотного вертолета. Датчики 2017, 17, 1222. [CrossRef] [ПабМед]
6. Цзян, С.; Цзян, В.; Ху анг, В.; Ян, Л. Наклонная фотограмметрия с использованием БПЛА для сбора данных на открытом воздухе и визуального контроля линий электропередачи за пределами объекта. Remote Sens. 2017, 9, 278. [CrossRef]
7. Ян, Л.; Фан, Дж; Лю, Ю; Ложь.; Пэ н, Дж; Лян, З. Обзор современных методов проверки линий электропередачи. IEEE Транс. Инструмент. Изм ер. 2020, 69, 9350–9365. [Перекрестная ссылка]
8. Чжоу, Р.; Цзян, В.; Цзян, С. Новый метод реконструкции высоковольтных пучков проводников по данным бортового лидара. Remote Sens. 2018, 10, 2051. [CrossRef]

Machine Translated by Google

9. Чжао, П.; Ян, В.; Фэн, Ю; Ли, Ф; Хуанг, Х. Построение 3D-сцены коридора линии электропередач на основе GIM и 3D ГИС. Дж Физ. Конф. Сер. 2021, 2005, 012083. [CrossRef]
10. Ли, К.; Чен, З.; Ху, К. Модельно-ориентированный подход к 3D-моделированию пилона на основе данных бортового лидара. Дистанционное восприятие 2015, 7, 11501–11524. [Перекрестная ссылка]
11. Тан, К.; Чжан, Л.; Лан, Г.; Ши, Х.; Дуанму, Х.; Чен, К. Метод классификации облаков точек коридора линии электропередач на основе улучшенного случайного леса и многомасштабных функций. Датчики 2023, 23, 1320. [CrossRef] [ПабМед]
12. Чжу, С.; Ли, К.; Чжао, Дж; Чжан, К.; Чжао, Г.; Ли, Л.; Чен, З.; Чен, Ю Метод глубокого обучения для извлечения произвольного количества отдельных линий электропередач из облаков точек лазерного сканирования, установленных на БПЛА. Remote Sens. 2024, 16, 393. [CrossRef]
13. Го, Б.; Хуанг, Х.; Ли, К.; Чжан, Ф; Чжу, Дж; Ван, К. Метод стохастической геометрии для реконструкции пилонов с воздуха Данные Лидара. Remote Sens. 2016, 8, 243. [CrossRef]
14. Ян, Л.; Конг, С.; Дэн, Дж; Ли, Х.; Лю, Ю DRA-Net: двухветвевая сеть остаточного внимания для пиксельной линии электропередач и Обнаружение. IEEE Транс. Инструмент. Измер. 2023, 72, 5010813. [CrossRef]
15. Ван, Г.; Ван, Л.; Ву, С.; Зу, С.; Сонг, Б. Семантическая сегментация трехмерных облаков точек коридора передач и на основе CA-PointNet++. Электроника 2023, 12, 2829. [CrossRef]
16. Шен Ю; Ян, Ю; Цзян, Дж; Ван, Дж; Хуанг, Дж; Феррейра, В.; Чен Ю Новый метод разделения отдельных проводов из пучка Проводник с использованием данных облака точек БПЛА-LiDAR. Размер 2023, 211, 112603. [CrossRef]
17. Хан, В. Трехмерное моделирование электроэнергетической башни с использованием данных бортового LiDAR. Дж Наука реки Янцзы. Рез. Инст. 2012, 29, 122–126. [Перекрестная ссылка]
18. Чжоу, Р.; Цзян, В.; Хуанг, В.; Скай, Б.; Цзян, С. Эвристический метод реконструкции опор электропередач и по данным бортового LiDAR. Remote Sens. 2017, 9, 1172. [CrossRef]
19. Каму ффо, Э.; Мари, Д.; Милани, С. Последние достижения в алгоритмах обучения облакам точек: обновленный обзор. Датчики 2022, 22, 1357. [CrossRef]
20. Цяо Ю; Си, Х.; Не, С.; Ван, П.; Го, Х.; Ван, К. Реконструкция силовой опоры по данным бортового LiDAR на основе компонентов Сегментация и сопоставление моделей. Дистанционное обнаружение 2022, 14, 4905. [CrossRef]
21. Чен, С.; Ван, К.; Дай, Х.; Чжан, Х.; Пан, Ф.; Си, Х.; Ян, Ю; Ван, П.; Ян, Х.; Чжу, Х.; и другие. Реконструкция силовой опоры на основе об абстрактных шаблонных структурах с использованием данных бортового LiDAR. Remote Sens. 2019, 11, 1579. [CrossRef]
22. Сильва, Ф.; Амаро, Н. Классификация опор электропередач и с использованием сходства облаков точек. В материалах APCA International Конференция по автоматическому управлению и мягким вычислениям, Капарика, Португалия, 6–8 июля 2022 г.; стр. 609–618.
23. Чен, З.; Лан, З.; Лонг, Х.; Ху, К. 3D-моделирование пилона по данным бортового LiDAR. В материалах дистанционного зондирования окружающей среды: 18-й Национальный симпозиум по дистанционному зондированию Китая, Ухань, Китай, 14 мая 2014 г.; п. 915807.
24. Ван, Х.; Хижина.; Ван, З.; Канг, З.; Аквенси, РН; Ян, Дж Реконструкция опор электропередач и из облаков точек LiDAR на основе структурной сегментации и оценки параметров. IEEE Geosci. Дистанционная чувствительность Lett. 2022, 19, 1–5. [Перекрестная ссылка]
25. Чжан, М.; Су, Х.; Скай, Х.; Ли, Х.; Ван, Д. Идентификация категории опор электропередач и по облакам точек бортового LiDAR на основе кривой формы. В материалах 3-й Международной конференции по мехатронике, автоматизации и интеллектуальному управлению, Гуйлинь, Китай, 15–17 сентября 2023 г.; п. 012026. [Перекрестная ссылка]
26. Принципи, Э.; Сквартини, С.; Камбрия, Э.; Пьяцца, Ф. Сопоставление акустических шаблонов для автоматического обнаружения аварийных состояний: ELM Основанный алгоритм. Нейрокомпьютинг 2015, 149, 426–434. [Перекрестная ссылка]
27. Обейд, М.; Ходроб, Р.; Абу Мвайс, А.; Альдабабса, М. Автоматическое распознавание речи и изолированных слов с небольшим словарным запасом для Односложные команды на арабском языке. Мягкий компьютер. 2023, 1–14. [Перекрестная ссылка]
28. Сальвадор, С.; Чан, П. К точному динамическому искажению времени в линейном времени и пространстве. МАР 2007, 11, 561–580. [Перекрестная ссылка]
29. Гао Ю; Ян, Ю; Может т.; Скай, В. Исследование интеллектуальной диагностики железнодорожных стрелочных переводов на основе улучшенного FastDTW и сегментации временных рядов при мониторинге больших данных. Математика. Пробл. англ. 2022, 2022, 7048813. [CrossRef]
30. Йео, К.; Инь, ОС; Хан, ПЯ; Кви, В.К. Мобильное приложение для передач и воздушной подписи в реальном времени с быстрым динамическим искажением времени (FastDTW). В материалах Международной конференции IEEE 2015 г. по приложениям обработки сигналов и изображений (ICSIPA), Куала-Лумпур, Малайзия, 19–21 октября 2015 г.; стр. 315–320.
31. Барбей Дж. Адаптивное вычисление дискретного расстояния Фреше. В материалах по обработке строк и поиску информации, Лима, Перу, 14 сентября 2018 г.; стр. 50–60.
32. Авраам, РБ; Фильцер, О.; Каплан, Х.; Кац, МЖ; Шарир, М. Дискретное и полунепрерывное расстояние Фреше с короткими доступами через Приблизительный подсчет и выборка расстояния. АКМ Транс. Алгоритмы 2015, 11, 1–29. [Перекрестная ссылка]
33. Али, М.; Хусейн, З.; Ян, М.-С. Расстояние Хаусдорфа и меры подобия однозначных нейтрософских множеств с применением при принятии многокритериальных решений. Электроника 2022, 12, 201. [CrossRef]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору (ам) и участникам (ам), а не MDPI и/или редактору (ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Алгоритм многообъектного обнаружения и отслеживания транспортных средств на основе улучшенных YOLOv8 и ByteTrack

Лунсян Юй, Яцзюнь Чен *, Ци Сяо, Чаоюэ Сунь и Жунчжэнь Ли

Школа электронной информационной инженерии Западно-Китайского педагогического университета, Наньчун 637009, Китай; ylxzxx@stu.cwnu.edu.cn (LY); xiaoci035560@stu.cwnu.edu.cn (CX); sunchaoyue@stu.cwnu.edu.cn (CS); lee1225rz@stu.cwnu.edu.cn (RL)

* Переписка: scnccyj@cwnu.edu.cn

Аннотация: Технология обнаружения и отслеживания транспортных средств играет решающую роль в интеллектуальных транспортных системах. Однако из-за таких факторов, как сложные сценарии, разнообразные масштабы и перекрытия, часто возникают такие проблемы, как ложные обнаружения, пропущенные обнаружения и переключение идентификации. Для решения этих проблем в данной статье предлагается многообъектный алгоритм обнаружения и отслеживания транспортных средств, основанный на CDS-YOLOv8 и улучшенном ByteTrack. Для обнаружения транспортных средств в процессе понижающей дискретизации вводится модуль контекстного управления (CG), который расширяет возможности извлечения признаков в сложных сценар Блок Dilated Reparam Block (DRB) реконструирован для решения многомасштабных проблем, а Soft-NMS заменяет традиционную NMS для повышения производительности в сценариях с густонаселенными транспортными средствами. Для отслеживания транспортных средств вектор состояния и ковариационная матрица фильтра Калмана улучшены, чтобы лучше справляться с нелинейным движением транспортных средств, а также введена сглаженная интерполяция по Гауссу (GSI) для заполнения пробелов в траектории, вызванных промахами обнаружения. Эксперименты, проведенные с набором данных UA-DETRAC, показывают, что улучшенный алгоритм повышает производительность обнаружения: mAP@0,5 и mAP@0,5:0,95 улучшаются на 9% и 8,8% соответственно. Что касается производительности отслеживания, mMOTA улучшается на 6,7%. Кроме того, сравнительные эксперименты с основными алгоритмами обнаружения и двухэтапного отслеживания демонстрируют превосходную производительность предложенного алгоритма.

Ключевые слова: компьютерное зрение; интеллектуальные транспортные системы; обнаружение транспортных средств; отслеживание транспортных средств

check for updates

Цитата: Вы, Л.; Чен, Ю; Сяо, К.; Сан, К.; Ли, Р. Алгоритм обнаружения и отслеживания многообъектных транспортных средств на основе улучшенных YOLOv8 и ByteTrack. Электроника 2024, 13, 3033. https://doi.org/10.3390/электроника13153033

Академический редактор: Фернандо Дела Прието Пинтадо

Поступила: 11 июля 2024 г.
Пересмотрено: 29 июля 2024 г.
Принято: 30 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Технология обнаружения и отслеживания транспортных средств играет ключевую роль в интеллектуальных транспортных системах [1]. Использование этой технологии позволяет отслеживать потоки дорожного движения в режиме реального времени, помогая органам управления дорожным движением принимать обоснованные решения для эффективного уменьшения заторов на дорогах. Кроме того, технология обнаружения и отслеживания транспортных средств позволяет выявлять нарушения правил дорожного движения, снижая вероятность дорожно-транспортных происшествий. Эта технология также обеспечивает интеллектуальное восприятие транспортных средств и дорожной среды, предлагая более умные и эффективные решения для управления дорожным движением. В последние годы технологии искусственного интеллекта сыграли значительную роль в расширении возможностей автомобильных сетей. Например, в области вычислений на границах автомобильных сетей Y Ju et al. [2] провели исследование по безопасной разгрузке с использованием технологии неортогонального множественного доступа (NOMA) и асинхронного глубокого обучения с подкреплением. Это не только повышает безопасность и эффективность передач данных, но также поддерживает требования к обнаружению и отслеживанию транспортных средств в режиме реального в

Однако при обнаружении и отслеживании транспортных средств часто встречаются некоторые проблемы. Такие факторы, как сложные изменения освещения, погодные воздействия, разнообразие типов и размеров транспортных средств, а также места с высокой плотностью движения, могут значительно снизить эффективность обнаружения и отслеживания. В ответ на эти проблемы были предприняты многочисленные исследовательские усилия. Например, Д. Рой и др. [3] улучшили производительность обнаружения и отслеживания транспортных средств за счет введения мультимодальной информации. М. Хумаюн и др. [4], основанный на YOLOv4, учитывает различия в различных погодных условиях путем добавления блока пространственной пирамиды. П. Дешмух и др. [5] использовали Swin Transformer для извлечени

Machine Translated by Google

многомасштабные функции и использовал двунаправленную пирамидальную сеть функций для обработки разнообразия масштабов транспортных средств. Дж Ван и др. [6] на основе YOLOv5 представил механизм внимания для повышения эффективности обнаружения небольших транспортных средств и использовал алгоритм JDE для отслеживания транспортных средств. Т. Бу и и др. [7], также основанный на YOLOv5, представил модуль внутримасштабного взаимодействия функций на основе внимания для ускорения конвергенции сети обнаружения и усовершенствовал ResNet36 в качестве сети извлечения функций DeepSORT для отслеживания транспортных средств. Помимо улучшения моделей обнаружения объектов и методов слежения за транспортными средствами на основе фильтра Калмана, было предложено несколько технологических решений для решения этих проблем. ЮВу и др. [8] предложили метод с использованием генеративно-состязательной сети (GAN) для преобразования дневных наборов данных о транспортных средствах в ночные наборы данных, повышая эффективность обнаружения в ночное время с точки зрения расширения набора данных.
Б. Сюй и др. [9] представили систему монокулярного зрения, которая использует модель рисования изображения для восстановления заблокированных транспортных средств посредством маскировки, улучшая производительность в сценариях перегруженного движения. Кроме того, В Xu et al. [10] использовали метод, сочетающий обнаружение ключевых точек и несколько моделей глубокого обучения, оптимизируя обнаружение и отслеживание точек нагрузки на ось транспортного средства путем интеграции моделей YOLO-v5 и HourglassNet с моделью динамики. Х Сюй и др. [11] предложили междоменную модель обнаружения автомобилей со встроенным сверточным блочным механизмом обнаружения транспортных средств (CDCDMA), которая создает структуру междоменного обнаружения объектов, уделяя особое внимание функциям фар для обработки изменений в сложных сценах. С Ч жан и др. [12] представили систему иерархическую структуру, которая решает проблемы повторной идентификации транспортных средств с перекрестным обзором путем изучения отношений высокого порядка и топологической информации. Эта платформа эффективно повышает точность распознавания в различных представлениях с помощью модуля извлечения признаков, модуля свертки графов и модуля сопоставления графов.

В области обнаружения транспортных средств большинство методов обычно оптимизируются для отдельных сценариев. Чтобы решить эту проблему, в данной статье предлагается улучшенный алгоритм на основе YOLOv8 [13] — CDS- YOLOv8. Этот алгоритм вводит модуль Context-Guided [14] во время процесса понижения дискретизации для расширения возможностей извлечения признаков, тем самым адаптируясь к сложным вариациям сцены. Одновременно структура блока Dilated Reparam [15] реконструируется и интегрируется с модулем C2f для полного извлечения многомасштабных и редких функций, повышая производительность модели при обнаружении многомасштабных объектов. Кроме того, на этапе постобработки алгоритм использует Soft-NMS для замены традиционных NMS, сохраняя более действительные поля-кандидаты в сценах с густонаселенными транспортными средствами за счет плавного снижения достоверности. Что касается отслеживания транспортных средств, большинство алгоритмов слежения обычно фокусируются только на блоках обнаружения с высокой степенью достоверности, игнорируя при этом блоки с низкой достоверностью, что приводит к пропущенному обнаружению некоторых истинных целей. Чтобы решить эту проблему, в данной статье в качестве базового средства отслеживания выбран ByteTrack [16] и внесены коррективы в вектор состояния и ковариационную матрицу в фильтре Калмана. Кроме того, для постобработки траекторий транспортных средств применяется сглаженная интерполяция по Гауссу [17] , компенсирующая недостающие траектории.

Основные положения этой статьи резюмируются следующим образом: 1. Мы

предложили улучшенную модель обнаружения транспортных средств, CDS-YOLOv8: Чтобы устранить ограничения YOLOv8 в городских условиях для обнаружения транспортных средств, модель CDS-YOLOv8 вводит модуль CG во время понижающей дискретизации для расширить возможности извлечения признаков в ответ на сложные изменения сцены. Он реконструирует DRB, чтобы лучше справляться с многомасштабными проблемами, вызванными различными расстояниями до камер и размерами транспортных средств. Кроме того, Soft-NMS заменяет традиционную NMS, сохраняя больше допустимых полей-кандидатов, улучшая производительность в сценах с густонаселенными транспортными средствами.

2. Мы улучшили трекер ByteTrack: в качестве базового трекера выбран ByteTrack. Чтобы устранить ограничения исходного фильтра Калмана при регистрации движения транспортного средства, были внесены улучшения в вектор состояния и ковариационную матрицу, чтобы лучше обрабатывать нелинейные изменения в движении транспортного средства. Чтобы уменьшить прерывания пути из-за потери обнаружения, GSI использовался для заполнения пробелов в отслеживании, улучшая непрерывность пути и уменьшая переключение идентификации и потерю пути.

Machine Translated by Google

3. Была достигнута интеграция улучшенного детектора и трекера: CDS-YOLOv8 и улучшенный ByteTrack были объединены и проверены на общедоступных наборах данных. Результаты экспериментов показывают, что этот метод демонстрирует повышенную производительность и превосходство по нескольким показателям.

2. Сопутствующая работа

2.1. Обнаружение транспортных средств

В последние годы благодаря быстрому развитию технологий глубокого обучения был достигнут значительный прогресс в области обнаружения объектов. Методы обнаружения объектов на основе глубокого обучения можно разделить на одноэтапные и двухэтапные подходы. Преимущество одноэтапных методов заключается в их высокой скорости обработки, поскольку они напрямую генерируют категории объектов и ограничивающие рамки через единую сеть. Напротив, двухэтапные методы сначала генерируют области-кандидаты, а затем выполняют точное обнаружение с преимуществом более высокой точности обнаружения. Типичные одноэтапные методы включают серию YOLO, SSD [18] и RetinaNet [19], а типичные двухэтапные методы включают Faster R-CNN [20] и Cascade R-CNN [21]. Принцип работы серии YOLO заключается в разделении входного изображения на сетку фиксированного размера, причем каждая сетка напрямую предсказывает местоположение и вероятность категории объекта. SSD извлекает ограничивающие рамки и категории объектов на картах объектов разных масштабов, обеспечивая многомасштабное обнаружение. RetinaNet использует сеть пирамиды функций для извлечения функций и вводит Focal Loss для решения проблемы дисбаланса классов. Более быстрый R-CNN сначала генерирует регионы-кандидаты через сеть предложений регионов, а затем выполняет точную классификацию объектов и регрессию ограничивающего прямоугольника в этих регионах. Cascade R-CNN, основанный на Faster R-CNN, представляет многоэтапную структуру регрессии и классификации, которая постепенно уточняет области-кандидаты, еще больше повышая точность обнаружения.

Чтобы удовлетворить практические потребности в обнаружении транспортных средств, несколько исследовательских проектов дополнительно оптимизировали детекторы общего назначения. З. Чен и др. [22] предложили улучшенный метод, основанный на SSD, путем введения механизмов внимания к каналам и модулей деконволюции для объединения функций, тем самым повышая производительность обнаружения транспортных средств. Л. Канг и др. [23] представили метод нечеткого внимания типа 1, основанный на YOLOv5, для улучшения обнаружения транспортных средств. Этот подход оптимизирует эффективность обнаружения за счет фокусировки детектора на центре цели. М. Би и др. [24] предложили двунаправленную сеть пирамидных функций на основе YOLOv5 для расширения возможностей извлечения признаков модели. Б. Ван и др. [25] представили улучшенный метод, основанный на YOLOv8, путем включения механизма внимания BiFormer и сети BiFPN для расширения возможностей объединения функций.

С точки зрения выбора алгоритмов обнаружения, YOLOv8 обеспечивает хороший баланс между скоростью, точностью и размером модели благодаря высокой скорости обнаружения, высокой точности и компактному размеру модели. Таким образом, в этой статье предлагается алгоритм CDS-YOLOv8, основанный на YOLOv8, для решения проблем обнаружения транспортных средств в сложных погодных условиях, масштабном разнообразии и плотной среде с транспортными средствами. Без увеличения вычислительной нагрузки этот алгоритм эффективно повышает производительность обнаружения транспортных средств.

2.2. Многоцелевое отслеживание транспортных средств

Алгоритмы отслеживания нескольких целей, основанные на обнаружении, можно разделить на одноэтапные и двухэтапные методы. Одноэтапные методы обычно используют сквозную структуру, объединяющую обнаружение и связывание объектов в один этап для непосредственного получения результатов обнаружения и отслеживания с преимуществом высокой производительности в реальном времени. Двухэтапные методы сначала выполняют обнаружение объектов, а затем связывают их, что обеспечивает более высок
Типичные одноэтапные методы включают JDE [26] и FairMOT [27]. JDE выполняет обнаружение и встраивание объектов в одну и ту же сеть, при этом части обнаружения и внедрения используют большую часть сетевых параметров. FairMOT объединяет сети обнаружения объектов и ReID в одну сеть, но обрабатывает задачи обнаружения и повторной идентификации через независимые филиалы.

Machine Translated by Google

Типичные двухэтапные алгоритмы отслеживания включают серию SORT. SORT [28], предложенный А. Бьюли и др., использует Faster R-CNN в качестве детектора на этапе обнаружения для вывода информации о местоположении и категории объекта. Этап отслеживания завершает ассоциацию траектории путем объединения фильтра Калмана и венгерского алгоритма. Впоследствии Н. Войке и др. предложил алгоритм DeepSORT [29] , который, по сравнению с SORT, представляет метод ReID, использующий сеть извлечения признаков для получения признаков внешнего вида объектов и выполняющий каскадное сопоставление признаков внешнего вида и Intersection Over Union (IoU) во время ассоциации данных. StrongSORT дополнительно улучшает извлечение признаков внешнего вида, обработку шума движения и компенсацию движения камеры на основе DeepSORT, а также дополняет и оптимизирует траектории с помощью отсутствующих обнаружений и ассоциаций. Ю Чжан и др. [16] предложили алгоритм ByteTrack, который повышает производительность отслеживания за счет связывания почти всех блоков обнаружения. После этого Н. Аарон и др. предложил алгоритм BoT-SORT [30] , который улучшает фильтр Калмана на основе ByteTrack, добавляя компенсацию движения камеры и метод ReID.

В сценариях отслеживания транспортных средств Z. Zhao et al. [31] разработали новую стратегию совместной оценки, основанную на платформе JDE, и представили кодеры CBAM и Transformer для расширения возможностей обнаружения, тем самым улучшая производительность отслеживания. К. Чжан и др. [32] внедрили механизм внимания CAENet в YOLOX для улучшения обнаружения транспортных средств и использовали DeepSORT для отслеживания транспортных средств. Г. Хан и др. [33] отслеживали транспортные средства на основе FairMOT, оптимизировали сеть DLA-34 для повышения эффективности обнаружения и ввели функцию совместной потери для уменьшения проблем с переключением идентификации.

По сравнению с одноэтапными алгоритмами отслеживания, двухэтапные алгоритмы отслеживания более гибки и обеспечивают более высокую точность в сложных сценариях. Поэтому в данной статье предлагается улучшить ByteTrack на основе CDS-YOLOv8 для обнаружения и отслеживания транспортных средств. В отличие от большинства методов отслеживания, которые игнорируют поля обнаружения с низкими показателями, ByteTrack полностью учитывает поля обнаружения с низкими показателями при сопоставлении траектории, эффективно решая проблему отсутствия недостающих фрагментов траектории. Например, при возникновении частичной окклюзии оценки в окне обнаружения могут варьироваться от высоких до низких. Принимая во внимание поля оценок с низкой достоверностью, ByteTrack может идентифицировать частично закрытые цели, уменьшая количество пропущенных обнаружений.

3. Методы

В процессе обнаружения и отслеживания транспортных средств детектор CDS-YOLOv8 сначала выводит поля обнаружения для текущего кадра. Улучшенный фильтр Калмана затем прогнозирует ограничивающие рамки для текущего кадра посредством прогнозирования треклета. Впоследствии блоки обнаружения и прогнозируемые блоки сопоставляются с использованием венгерского алгоритма, при этом IoU служит основой для ассоциации. Конкретный процесс вторичного сопоставления показан на рисунке 1.

Блоки обнаружения, выводимые с помощью CDS-YOLOv8, делятся на блоки с высокой достоверностью и блоки с низкой достоверностью на основе их оценок достоверности, после чего выполняются два раунда обработки ассоциации. На первом этапе ассоциации улучшенный фильтр Калмана прогнозирует отслеживаемые треклеты и потерянные треклеты, а затем сопоставляет блоки высокой достоверности с предсказанными блоками. Если совпадение прошло успешно, состояние треклета обновляется; в противном случае отслеживаемые треклеты, не сопоставленные с ящиками с высокой степенью достоверности, подвергаются второму раунду ассоциации с ящиками с низким уровнем достоверности. Во втором раунде ассоциации, если совпадение прошло успешно, состояние треклета обновляется; в противном случае треклет переходит в состояние потери. При управлении треклетами удаляются треклеты, которые остаются несовпадающими более 30 кадров. Новые треклеты назначаются не имеющим себе равных блокам обнаружения с высокой степенью достоверности. Наконец, после обнаружения и отслеживания транспортных средств GSI используется для заполнения пробелов в траекториях, вызванных пропущенными обнаружениями, оптимизируя непрерывность и полноту траекторий.

Модель CDS-YOLOv8 состоит из трех частей: позвоночника, шеи и головы, как показано на рисунке 2:

Machine Translated by Google

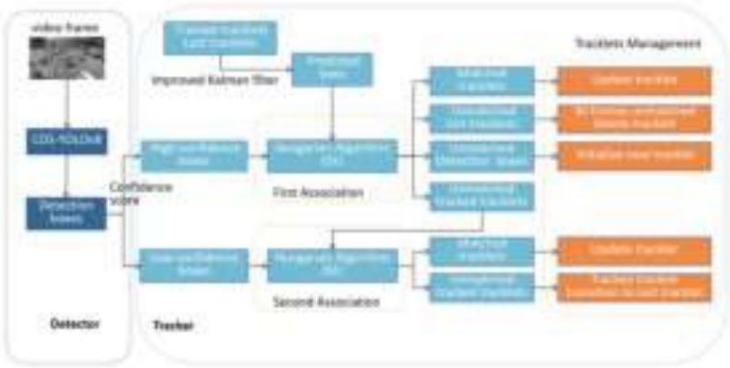


Рисунок 1. Блок-схема обнаружения и отслеживания транспортных средств.

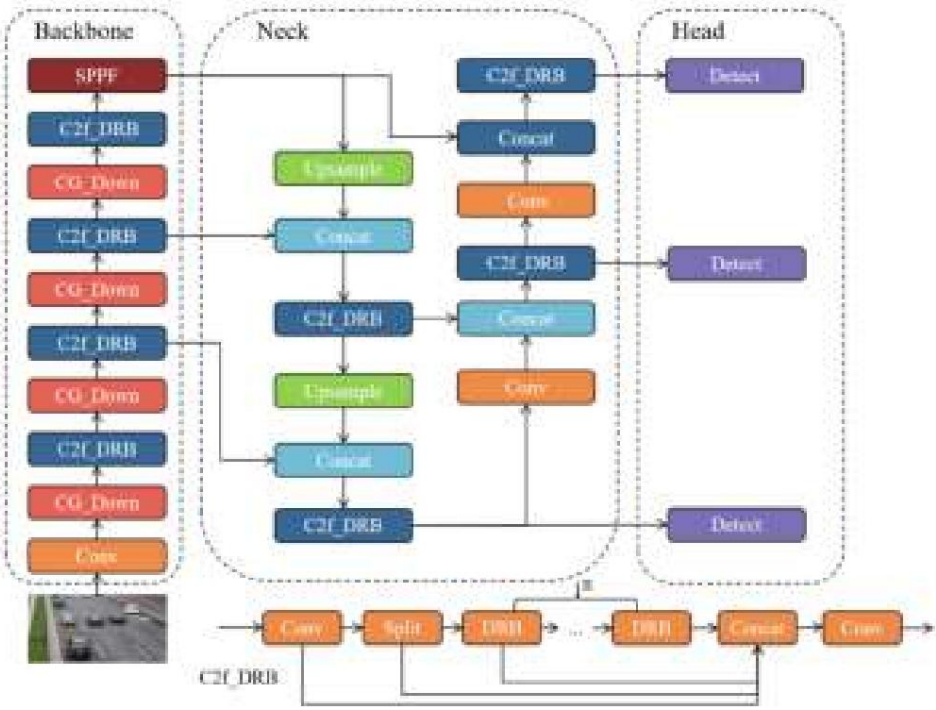


Рисунок 2. Структура CDS-YOLOv8.

На рисунке 2 Backbone используется для извлечения признаков, извлекая многоуровневые признаки из входного изображения. Шея отвечает за объединение функций, объединяя глубокие и поверхностные функции с использованием FPN и PAN. Головка содержит три головки обнаружения, которые обрабатывают карты объектов разных масштабов и выводят информацию о положении и категории блоков обнаружения. CDS-YOLOv8 устраняет недостатки YOLOv8 в обнаружении городских транспортных средств за счет следующих улучшений: Модуль CG представлен во время понижающей дискретизации, который улучшает извлечение признаков в сложных средах за счет объединения локальной и окружающей контекстной информации и использования глобальной информации для взвешивания. В Backbone и Neck структура DRB реструктурирована и интегрирована с модулем C2f для лучшего решения проблем масштабируемого разнообразия. На этапе постобработки Soft-NMS заменяет традиционную NMS, избегая агрессивного исключения блоков-кандидатов и улучшая производительность обнаружения при частичной окклюзии.

3.1. Понижение выборки с учетом

контекста Среда, в которой работают транспортные средства, включает сложные сценарии, такие как ночи, пасмурная погода, солнечные дни и дождливые дни. Чтобы расширить возможности извлечения признаков в этих сложных средах, во время

Machine Translated by Google

процесс понижения дискретизации. Модуль CG состоит из нескольких ключевых компонентов: Средство извлечения локальных объектов использует стандартную свертку для извлечения локальных особенностей изображения, предоставляя детализированные функции и информацию о границах. Средство извлечения окружающего контекста использует расширенную свертку для расширения поля восприятия локальных объектов, тем самым захватывая более широкий диапазон контекстной информации. Joint Feature Extractor объединяет локальные объекты и окружающие контекстные объекты и обрабатывает их с помощью пакетной нормализации (BN) и активации SiLU, образуя совместные элементы, которые расширяют возможности нелинейного выражения. Средство извлечения глобального контекста объединяет глобальную контекстную информацию всего изображения посредством объединения глобальных средних значений и далее обрабатывает ее с помощью многоуровневого восприятия (MLP) для создания весов глобального контекста. Это применяет веса к совместным функциям, усиливая полезную информацию и подавляя ненужную информацию. Структура модуля CG показана на рисунке 3:

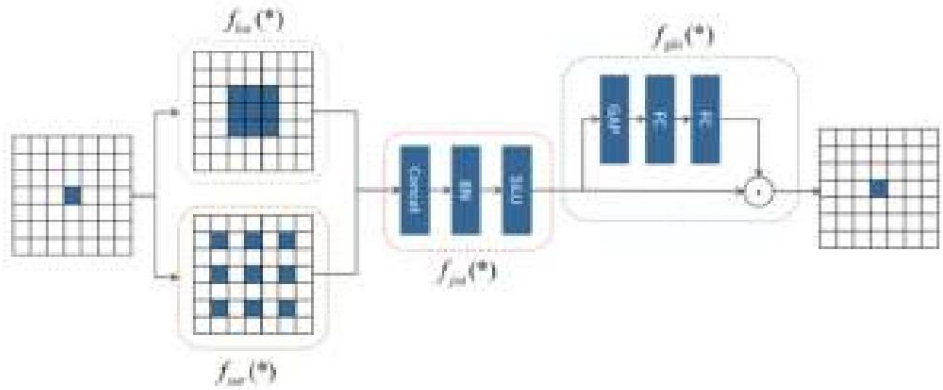


Рисунок 3. Структура модуля с контекстным управлением.

Подробно, во-первых, входная карта объектов подвергается свертке 3 × 3 для извлечения локальных особенностей и свертке 3 × 3 со степенью расширения 2 для извлечения окружающих контекстных особенностей. Затем Joint Feature Extractor объединяет локальные объекты и окружающие контекстные объекты по размеру канала, после чего следует активация уровня BN и SiLU . Наконец, экстрактор глобального контекста использует глобальную информацию об изображении, используя пул глобальных средних значений для создания глобального вектора признаков для каждого канала. Затем этот вектор подвергается нелинейному преобразованию через два полностью связанных слоя, создавая веса каналов. Эти веса применяются к объединенным функциям для взвешенной оптимизации, улучшая важные функции и подавляя неважные, тем самым увеличивая возможности извлечения признаков модели.

3.2. Реконструированный расширенный блок репарама

При обнаружении транспортных средств возникают значительные различия в масштабе из-за различий в расстояниях до камеры и размерах транспортных средств. Чтобы решить эту проблему, мы реконструировали блок расширенного репарама и интегрировали его в структуру C2f. Структура DRB включает в себя слой свертки с большим ядром и несколько расширенных слоев свертки с малым ядром. Слой свертки с большим ядром захватывает более широкие пространственные объекты, в то время как расширенные слои свертки с малым ядром захватывают разреженные пространственные объекты посредством операций расширения. Выходные данные этих слоев обрабатываются через уровень BN, а затем суммируются. Используя методы репараметризации, они эквивалентно представляются как один слой свертки с большим ядром. Эта конструкция имеет два основных преимущества: слои свертки с малым ядром за счет расширения фиксируют функции в более широком диапазоне. В сочетании со слоем свертки с большим ядром рецептивное поле эффективно расширяется, устраняя необходимость увеличения глубины модели или проектирования более крупных ядер свертки. Во время вывода методы репараметризации объединяют несколько слоев свертки в эквивалентное одно ядро свертки, сохраняя размер ядра и вычислительную сложность, одновременно значительно улучшая возможности извлечения признаков модели. Рисунок 4 иллюстрирует структуру DRB.

Machine Translated by Google

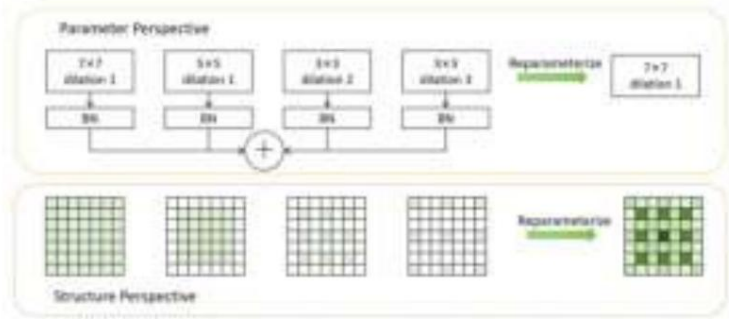


Рисунок 4. Структура расширенного блока Reparam.

С точки зрения параметров, DRB состоит из слоя свертки с большим ядром 7 × 7 и трех слоев свертки с малым ядром с разными скоростями расширения. Выходные данные всех слоев свертки обрабатываются через слои BN и эквивалентно представляются как один слой свертки с большим ядром 7 × 7 во время вывода, без значительного увеличения вычислительной нагрузки.

Структурно рисунок 4 демонстрирует, как каждый слой свертки соответствует областям в зависимости от размера ядра и скорости расширения. Для больших объектов DRB эффективно расширяет рецептивное поле, комбинируя слои свертки с большим ядром и слои расширенной свертки. Для небольших объектов слои расширенной свертки в DRB, использующие небольшие ядра и различные скорости расширения, эффективно фиксируют детализированные функции, тем самым улучшая производительность многомасштабного обнаружения транспортных средств на сложном фоне.

3.3. Софт-НМС

В процессе обнаружения и отслеживания транспортных средств Soft-NMS используется для замены традиционных NMS для повышения производительности в условиях густонаселенных транспортных средств. Традиционные шаги NMS

заключаются в следующем: 1. Отсортируйте ячейки-кандидаты в порядке убывания на основе их оценок достоверности. 2. Выберите блок-кандидат с наивысшим показателем достоверности и добавьте его в окончательный набор результатов.

3. Вычислите IoU между этим блоком-кандидатом и другими блоками. Если IoU блока-кандидата превышает установленный порог, он удаляется.

4. Из оставшихся ящиков-кандидатов выберите тот, который имеет наивысший показатель достоверности, и повторяйте шаги 2 и 3, пока все ящики-кандидаты не будут обработаны.

Soft-NMS оптимизирует третий шаг традиционной NMS: в Soft-NMS после расчета IoU между блоком-кандидатом с наивысшим показателем достоверности и другими блоками вместо удаления тех блоков с IoU, превышающим пороговое значение, используется метод мягкой обработки. используется путем уменьшения их оценок достоверности. В частности, мы принимаем метод гауссовского распада, как показано в уравнении (1):

$$si = si \times \exp\left(\frac{IoU2}{\sigma}\right) \quad (1)$$

где si представляет собой показатель достоверности i-го блока кандидатов, а σ — коэффициент затухания.

Если показатель затухающей достоверности ниже установленного порога, блок кандидатов будет удален. В сценариях с плотным транспортным средством перекрытие ящиков-кандидатов между целями является серьезным. По сравнению с традиционной NMS, Soft-NMS может сохранять больше допустимых блоков-кандидатов, более эффективно обрабатывать соседние цели и уменьшать количество пропущенных обнаружений.

3.4. Улучшенный фильтр

Калмана В двухэтапных алгоритмах отслеживания выходные данные ограничивающего прямоугольника обычно оцениваются с помощью фильтра Калмана (KF) для определения состояния движения цели. В ByteTrack вектор состояния фильтра Калмана показан в уравнении (2):

Machine Translated by Google

$$x = xc, yc, a, h, \quad . \quad . \quad . \quad . \quad T \quad xc, yc, a, \quad ч \tag{2}$$

где xc, yc — координаты центра, a — соотношение сторон, а h — высота.

Затем, когда речь идет о частичной окклюзии и изменении перспективы, соотношение сторон трудно адаптировать. Поэтому лучшим выбором будет прямая оценка ширины и высоты. Улучшенный вектор состояния показан в уравнении (3):

$$xk = xc(k), yc(k), w(k), h(k), . xc(к), . yc(к), . ш(к), ч(к) \tag{3}$$

где (xc (k), yc (k) — координаты центра, w(k) — ширина, а h(k) — высота.

По сравнению с использованием соотношения сторон прямая оценка ширины и высоты имеет следующие преимущества: с одной стороны, она может более точно отражать фактический размер объекта и лучше адаптироваться к изменениям размеров транспортного средства; с другой стороны, изменения формы и размера транспортного средства во время движения являются сложными, а использование соотношения сторон увеличивает нелинейные ошибки. Непосредственная оценка ширины и высоты может уменьшить искажение ограничивающей рамки, вызванное накоплением ошибок, тем самым обеспечивая лучшую стабильность в сложных сценах.

В классическом фильтре Калмана ковариационная матрица шума процесса Q и ковариационная матрица шума измерения R задаются как фиксированные значения, которые остаются неизменными на протяжении всего процесса отслеживания. Чтобы повысить производительность фильтра Калмана в динамических средах, Q и R динамически регулируются в соответствии с условиями реального времени: Q, как показано в уравнении (4), и R, как показано в уравнении (5).

$$Qk = диаг. \quad \begin{matrix} 2 & 2 \\ \sigma pw\hat{}\, k \;\; 1|k \;\; 1 \quad , \sigma ph\hat{}\, k \;\; 1|k \;\; 1 \quad , \\ 2 & 2 \\ \sigma pw\hat{}\, k \;\; 1|k \;\; 1 \quad , \sigma ph\hat{}\, k \;\; 1|k \;\; 1 \quad , \\ 2 & 2 \\ \sigma vw\hat{}\, k \;\; 1|k \;\; 1 \quad , \sigma vh\hat{}\, k \;\; 1|k \;\; 1 \quad , \\ 2 & 2 \\ \sigma vw\hat{}\, k \;\; 1|k \;\; 1 \quad , \sigma vh\hat{}\, k \;\; 1|k \;\; 1 \end{matrix} \tag{4}$$

где σp — стандартное отклонение позиционного шума, а σv — стандартное отклонение скоростного шума.

$$Rk = диаг. \quad \begin{matrix} 2 & 2 \\ \sigma mw\hat{}\, k|k \;\; 1 \quad , \sigma mh\hat{}\, k|k \;\; 1 \quad , \\ 2 & 2 \\ \sigma mw\hat{}\, k|k \;\; 1 \quad , \sigma mh\hat{}\, k|k \;\; 1 \end{matrix} \tag{5}$$

где σm — стандартное отклонение шума измерения.

Изменяющиеся во времени Q и R могут более гибко реагировать на нелинейные изменения в окружающей среде, такие как замедление или ускорение транспортного средства. По сравнению с фиксированными ковариационными матрицами изменяющиеся во времени Q и R можно корректировать на основе текущих состояний прогнозирования и измерения, что более эффективно решает проблемы шума и накопления ошибок. Это позволяет лучше адаптироваться к динамической среде и повышает общую производительность системы.

3.5. Сглаженная по Гауссу интерполяция

Чтобы решить проблему потери траектории, вызванную пропусками обнаружения при отслеживании транспортных средств, в качестве этапа постобработки для оптимизации используется метод сглаженной по Гауссу интерполяции. Традиционные методы линейной интерполяции обычно полагаются только на положения двух конечных точек, игнорируя сложность движения транспортного средства, что приводит к менее точным результатам. GSI фиксирует нелинейную траекторию движения транспортного средства с помощью регрессии гауссовского процесса (GPR), тем самым обеспечивая более точные результаты.

Для i-й траектории транспортного средства модель GSI представлена уравнением (6):

$$pt = f(i)(t) + \epsilon \tag{6}$$

Machine Translated by Google

где t — идентификатор кадра, pt — координата положения транспортного средства в кадре t, f (i) (t) — функция, аппроксимируемая георадаром, используемая для описания траектории транспортного средства, и $\epsilon \sim N(0, \sigma^2)$ представляет гауссов шум.

Гауссов процесс — это мощный непараметрический байесовский метод, который может фиксировать нелинейные связи между точками данных, предполагая, что функция f (i) следует Гауссов процесс, как показано в уравнении (7):

$$f(i) \sim GP(0, k(x, x')) \quad (7)$$

где ковариационная функция k(x, x′) представляет собой ядро радиальной базисной функции (RBF), которое особенно чувствительно к локальным изменениям. Чем ближе точки данных, тем выше их корреляция, и отклик ядра RBF быстро уменьшается по мере увеличения расстояния между точками данных. Эта характеристика делает ядро RBF более подходящим для описания нелинейных изменений в движении транспортного средства, таких как трогание с места, поворот, ускорение или замедление. конкретно представлено, как показано в уравнении (8):

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{2\lambda^2}\right) \quad (8)$$

Процесс прогнозирования и интерполяции положения траектории выглядит следующим образом: для кадров с пропущенными обнаружениями прогнозируется сглаженное положение $P_п$, как показано в уравнении (9):

$$P_п = K(Ф_п, F)\,[K(F, F) + \sigma_я^2 I]^{-1} P_п \quad (9)$$

где K(·, ·) – ковариационная матрица, основанная на ковариационной функции k(·, ·), I – единичная матрица, а P представляет известные положения транспортного средства на траектории.

Кроме того, для функции ядра RBF параметр масштаба λ определяет ее чувствительность к изменениям данных. Во время отслеживания транспортного средства разная длина траектории транспортного средства может соответствовать разным схемам движения. Следовательно, регулируя λ с учетом длины l траектории транспортного средства, траектории можно сгладить, тем самым уменьшая шум и обеспечивая более плавное и точное отслеживание. Настройка λ показана в уравнении (

$$\lambda = \tau \cdot \text{журнал}\left(\frac{3\tau}{л}\right) \quad (10)$$

где τ — константа, используемая для управления степеньюсглаживания, которая была установлена равной 18 на основе экспериментальных результатов, полученных с помощьюпоиска по сетке.

С помощьюописанных выше шагов GSI может эффективно заполнить недостающие траектории движения транспортного средства. отслеживание, повышение точности и стабильности прогнозов в траектории.

4. Эксперименты и результаты

4.1. Набор данных и экспериментальная конфигурация

Набор данных UA-DETRAC [34], совместно выпущенный Университетом Альберты и Университетом науки и технологий Китая, служит эталонным набором данных, специально разработанным для обнаружения и отслеживания транспортных средств с несколькими объектами. Этот набор данных включает 10 часов видео, снятого с помощьюкамеры Canon EOS 550D (Canon USA, Inc., Лейк-Саксесс, Нью-Йорк, США). Видео записывается с частотой кадров 25 кадров в секунду и включает более 140 000 кадров и 1,21 миллиона аннотированных блоков обнаружения объектов. Обучающий набор состоит из 60 видеопоследовательностей общим количеством 83 791 изображения, а тестовый набор включает 40 видеопоследовательностей общим количеством 56 340 изображений. Набор данных UA-DETRAC сопряжен с различными проблемами, репрезентативные сценарии показаны на рисунке 5. Набор данных охватывает четыре категории транспортных средств: автомобили, автобусы, фургоны и другие. Кроме того, он включает в себя четыре погодных условия: солнечно, ночь, облачно и дождливо, а также сцены с высокой плотностьютранспортных средств и у

Machine Translated by Google

Рисунок 5. Репрезентативные сценарии в наборе данных UA-DETRAC.

В экспериментах по обнаружению транспортных средств изображения без меток сначала исключались, в результате чего образовался обучающий набор из 82 085 изображений и тестовый набор из 56 167 изображений. Чтобы избежать переобучения, был выбран каждый 10-й кадр, в результате чего было получено 8209 изображений для обучения и 5617 изображений для оценки. Для экспериментов по отслеживанию транспортных средств для оценки использовались 40 видеопоследовательностей из тестового набора.

В этих экспериментах использовался графический процессор NVIDIA GeForce RTX 4090 (NVIDIA Corporation, Санта-Клара, Калифорния, США), а центральным процессором — процессор Core i9-9900K @ 3,60 ГГц (Intel Corporation, Чэнду, Китай). Эксперименты проводились в системе Windows 10 со средой глубокого обучения на базе PyTorch 1.31.1, CUDA 11.6 и Python 3.8. Параметры обучения были заданы следующим образом: количество эпох обучения — 200, размер пакета — 32, в процессе обучения не использовались предварительно обученные веса.

4.2. Метрики оценки

При обнаружении транспортных средств для оценки результатов эксперимента используются точность (P), отзыв (R), mAP@0,5 и mAP@0,5:0,95.

P: доля истинно положительных образцов среди всех образцов, обнаруженных как положительные, как показано в уравнении (11):

$$П = \frac{ТП}{ТП + ФП} \quad (11)$$

R: доля правильно обнаруженных положительных образцов среди всех фактически положительных образцов, как показано в уравнении (12):

$$Р = \frac{ТП}{ТП + ФН} \quad (12)$$

где TP — количество истинно положительных образцов, FP — количество ложноположительных образцов, а FN — количество фактически положительных образцов, которые были ошибочно определены как отрицательные.

Средняя точность (AP): площадь под кривой PR для определенного класса образцов, как показано в уравнении (13):

$$АП = \int_0^1 ПДР \quad (13)$$

mAP@0,5 — это средняя средняя точность для всех классов с порогом IoU 0,5. mAP@0,5:0,95 — это средняя средняя точность для всех классов с пороговыми значениями IoU в диапазоне от 0,5 до 0,95 с шагом 0,05.

При отслеживании транспортных средств для оценки экспериментальных результатов используются MOTA, IDF1 и IDS.

MOTA измеряет точность алгоритма отслеживания, учитывая различные ошибки, возникающие в процессе отслеживания, такие как ложные срабатывания, пропущенные обнаружения и смену личности. Это показано в уравнении (14):

Machine Translated by Google

$$MOTA = 1 - \frac{t\,(FNt + FPt + IDSWt)}{я\ GTt} \quad (14)$$

г д е FNt п р е д с т а в л я е т к о л и ч е с т в о п р о п у щ е н н ы х о б н а р у ж е н и й в м о м е н т в р е м е н и t, FPt п р е д с т а в л я е т к о л и ч е с т в о л о ж н ы х с р а б а т ы в а н и й в м о м е н т в р е м е н и t, IDSWt п р е д с т а в л я е т к о л и ч е с т в о п е р е к л ю ч е н и й и д е н т и ф и к а ц и и в м о м е н т в р е м е н и t, а GTt п р е д с т а в л я е т к о л и ч е с т в о н а з е м н ы х ц е л е й и с т и н н о с т и в м о м е н т в р е м е н и t.

IDF1: IDF1 и з м е р я е т с о г л а с о в а н н о с т ь и д е н т и ф и к а ц и о н н ы х д а н н ы х и н е п р е р ы в н о с т ь а л г о р и т м а о т с л е ж и в а н и я. к а к п о к а з а н о в у р а в н е н и и (15):

$$ИДФ1 = 2 \cdot \frac{ВПЛ \cdot РДЭ}{МВУ + РДЭ} \quad (15)$$

г д е IDP п р е д с т а в л я е т с о б о й д о л ю п р а в и л ь н о и д е н т и ф и ц и р о в а н н ы х ц е л е й с р е д и о б н а р у ж е н н ы х ц е л и. IDR п р е д с т а в л я е т с о б о й д о л ю п р а в и л ь н о о б н а р у ж е н н ы х ц е л е в ы х и д е н т и ф и к а т о р о в к ф а к т и ч е с к и м ц е л е в ы е л и ч н о с т и.

IDS: к о л и ч е с т в о п е р е к л ю ч е н и й и д е н т и ф и к а ц и и, п р о и с х о д я щ и х в п р о ц е с с е о т с л е ж и в а н и я, т . е . о д н о й и т о й ж е ц е л и п р и с в а и в а ю т с я р а з н ы е и д е н т и ф и к а т о р ы. М е н ь ш е е з н а ч е н и е у к а з ы в а е т н а л у ч ш у ю п р о и з в о д и т е л ь н о с т ь т р е к е р а в о б е с п е ч е н и и с о г л а с о в а н н о с т и и д е н т и ф и к а ц и о н н ы х д а н н ы х.

4.3. Э к с п е р и м е н т ы п о а б л я ц и и д е т е к т о р а

Ч т о б ы п р о в е р и т ь э ф ф е к т и в н о с т ь п р е д л а г а е м ы х м о д у л е й у л у ч ш е н и я о б н а р у ж е н и я т р а н с п о р т н ы х с р е д с т в, э к с п е р и м е н т ы п о а б л я ц и и п р о в о д и л и с ь с и с п о л ь з о в а н и е м YOLOv8-S в к а ч е с т в е б а з о в о й м о д е л и. Р е з у л ь т а т ы п о к а з а н ы в т а б л и ц е 1:

Т а б л и ц а 1. Р е з у л ь т а т ы э к с п е р и м е н т а п о а б л я ц и и н а н а б о р е д а н н ы х UA-DETRAC.

М о д е л ь	п	р	м A P @ 0 , 5	м A P @ 0 , 5 : 0 , 9 5	П р о в а л ы
YOLOv8-S	71,8%	58,6%	64,2%	46,6%	28,8 Г
YOLOv8-S+CG_Down	78,4%	63,3%	69,9%	51,0%	30,7 Г
YOLOv8-S+C2f_DRB	81,5%	62,6%	71,4%	51,6%	24,7 Г
YOLOv8-S+CG_Down+C2f_DRB	80,7%	63,9%	72,9%	52,2%	26,6 Г
YOLOv8-S+CG_Down+C2f_DRB+Soft-NMS	82,4%	62,3%	73,2%	55,4%	26,6 Г

К а к в и д н о и з д а н н ы х т а б л и ц ы 1, в в е д е н и е м о д у л я CG в п р о ц е с с п о н и ж а ю щ е й в ы б о р к и у в е л и ч и л о P н а 6,6%, R н а 4,7%, mAP@0,5 н а 5,7% и mAP@0,5:0,95.
н а 4,4%, п р о в е р я я э ф ф е к т и в н о с т ь м о д у л я CG п р и и з в л е ч е н и и х а р а к т е р и с т и к а в т о м о б и л я.
в н е д р е н и е м о д у л я C2f_DRB у в е л и ч и л о P н а 9,7%, R н а 4,0%, mAP@0,5 н а 7,2%,
и mAP@0,5:0,95 н а 5%, д е м о н с т р и р у я з н а ч и т е л ь н о е п р е и м у щ е с т в о C2f_DRB в о б р а б о т к е
м н о г о м а с ш т а б н ы е п р о б л е м ы. П р и о д н о в р е м е н н о м в в е д е н и и м о д у л е й CG и C2f_DRB P у в е л и ч и л с я н а 8,9%, R
н а 5,3%, mAP@0,5 н а 8,7% и mAP@0,5:0,95 н а 5,6%.
д о к а з ы в а я, ч т о к о м б и н а ц и я э т и х д в у х м о д у л е й м о ж е т л у ч ш е и з в л е к а т ь и и н т е г р и р о в а т ь т р а н с п о р т н о е с р е д с т в о
ф у н к ц и и. Н а к о н е ц, н е б о л ь ш а я з а м е н а т р а д и ц и о н н о й NMS н а Soft-NMS д л я п о с т о б р а б о т к и.
у в е л и ч и л о mAP@0,5 н а 0,3 п р о ц е н т н ы х п у н к т а, н о з н а ч и т е л ь н о у л у ч ш и л о mAP@0,5:0,95 н а
3,2%, ч т о з а к л а д ы в а е т о с н о в у д л я б о л е е т о ч н о г о о т с л е ж и в а н и я т р а н с п о р т н ы х с р е д с т в н а п о с л е д у ю щ и х
э т а п а х. П о с р а в н е н и ю с б а з о в о й м о д е л ь ю, п р и у м е н ь ш е н и и к о л и ч е с т в а ф л о п с н а 2,2 G, P у в е л и ч и л с я н а 10,6%, R
н а 3,7%, mAP@0,5 н а 9% и mAP@0,5:0,95 н а 8,8%, п о л н о с т ь ю п о д т в е р ж д а я п р е в о с х о д с т в о
CDS-YOLOv8 в с ц е н а р и я х о б н а р у ж е н и я т р а н с п о р т н ы х с р е д с т в.

Ч т о б ы д о п о л н и т е л ь н о п р о д е м о н с т р и р о в а т ь д о с т и ж е н и я у л у ч ш е н н о й м о д е л и в о б н а р у ж е н и и т р а н с п о р т н ы х с р е д с т в, п р е д с т а в л е н в и з у а л ь н ы й а н а л и з р е з у л ь т а т о в э к с п е р и м е н т а, к а к п о к а з а н о н а р и с у н к е 6.
и з о б р а ж е н и я с л е в а н а п р а в о п р е д с т а в л я ю т с о б о й и с х о д н о е и з о б р а ж е н и е, р е з у л ь т а т ы п р о г н о з и р о в а н и я б а з о в о й л и н и и
м о д е л ь YOLOv8-S, р е з у л ь т а т ы п р о г н о з и р о в а н и я у л у ч ш е н н о й м о д е л и CDS-YOLOv8 и
к о р р е к т и р о в к и, в н е с е н н ы е CDS-YOLOv8 п о с р а в н е н и ю с YOLOv8-S, к о т о р ы е в к л ю ч а ю т с о к р а щ е н и е
в п р о п у щ е н н ы х а в т о м о б и л я х. О п т и м и з и р о в а н н ы е ч а с т и о т м е ч е н ы р о з о в ы м и к р у ж к а м и. К р о м е т о г о,
ч е р н ы е о б л а с т и н а и з о б р а ж е н и я х и г н о р и р у ю т с я, т а к к а к т о л ь к о п о с л е д у ю щ е е о т с л е ж и в а н и е т р а н с п о р т н ы х с р е д с т в
о с н о в н о е в н и м а н и е у д е л я е т с я д в и ж у щ и м с я т р а н с п о р т н ы м с р е д с т в а м п о г л а в н о й д о р о г е.

Machine Translated by Google

Рисунок 6. Визуализация результатов обнаружения.

На рисунке 6 мы выбрали три репрезентативных сценария: солнечный, ночной и дождливый. В
при солнечном сценарии, как показано в розовом кружке, CDS-YOLOv8 обнаружил автобус, который
YOLOv8 промахнулся, как и машины дальше от камеры. В ночном сценарии CDS-YOLOv8 обнаружил почти все
цели, которые YOLOv8 пропустил, и точно определил цель.
частично закрытая белая машина, которую YOLOv8 не распознал как действительную цель. В дождливую погоду
В сценарии CDS-YOLOv8 обнаружил почти все мелкие цели, находящиеся далеко от камеры, и
обнаружил такси, которое было частично закрыто и не распознавалось YOLOv8 как допустимая цель.

В этих сценариях CDS-YOLOv8 продемонстрировал значительные улучшения по сравнению с YOLOv8-
S в возможностях извлечения признаков и многомасштабной обработки. Кроме того, он сохранил больше
действительные цели в сценах с плотным транспортным средством путем обработки сильно перекрывающихся блоков-кандидатов с использованием
Мягкая NMS с доверительным спадом, а не просто устраняет их через порог IoU . Таким образом, CDS-
YOLOv8 показал превосходные характеристики при различном освещении и условиях.
погодных условий, эффективно решая многомасштабные проблемы и сценарии с плотным транспортным средством.
фф

4.4. Сравнительный эксперимент детекторов

Чтобы продемонстрировать преимущества CDS-YOLOv8 в обнаружении транспортных средств, мы сравнили
это с помощью основных алгоритмов обнаружения. Результаты показаны в Таблице 2:

Таблица 2. Сравнительные экспериментальные результаты на наборе данных UA-DETRAC. (лучшие результаты
выделено жирным шрифтом.)

Модель	МАП@0,5	мАП@0,5:0,95	Провалы
Быстрее-RCNN [20]	66,3%	38,8%	60,5 г
Каскад-РКНН [21]	65,2%	41,1%	88,1 Г
РетинаНет [19]	56,6%	32,1%	45,2 Г
АТСС [35]	58,8%	32,1%	49,2 Г
ЮЛОв5-С [36]	62,0%	44,9%	16 г
ЮЛОв6-С [37]	59,0%	43,0%	44,9 Г
YOLOv7 [38]	71,1%	53,9%	105,2 Г
ЮЛОв8-С [13]	64,2%	46,6%	28,8 Г
ЮЛОв8-М [13]	68,2%	50,6%	79,3 Г
ЮЛОв8-Л [13]	71,6%	51,4%	165,7 Г
ЮЛОв9-С [39]	65,1%	48,0%	39,6 Г
Л. Канг и др. [23]	70,0%	50,3%	26,7 Г
CDS-YOLOv8 (наш)	73,2%	55,4%	26,6 Г

Из результатов сравнения видно, что CDS-YOLOv8 демонстрирует наиболее передовые возможности.
производительность по сравнению с этими основными двухэтапными и одноэтапными методами обнаружения. Его

Machine Translated by Google

производительность не только превосходит последнюю версию YOLOv9-S, но и превосходит большую версия YOLOv8, достигшая улучшения на 1,6% и 4% в mAP@0,5 и mAP@0,5:0,95, соответственно. CDS-YOLOv8 обеспечивает лучший баланс между точностью и вычислительными возможностями. сложность, что еще раз доказывает его превосходство в сценариях обнаружения транспортных средств.

4.5. Эксперимент по абляции трекера

Проверить эффективность улучшенного детектора и предлагаемого улучшения. модулей слежения за транспортными средствами был проведен эксперимент по абляции, как показано в Таблице 3. Результаты эксперимента представляют собой средние значения, основанные на 40 тестовых последовательностях:

Таблица 3. Результаты эксперимента по абляции на наборе данных UA-DETRAC.

Детектор+ Трекер	mIDF1	ИДС	мMOTA
YOLOv8s+ BYTETrack	76,9%	855	67,2%
CDS-YOLOv8+ BYTETrack	78,2%	785	70,6%
CDS-YOLOv8+ BYTETrack+ Улучшенный KF	79,7%	717	73,1%
CDS-YOLOv8+ BYTETrack+ Улучшенный KF +GSI	80,3%	530	73,9%

По сравнению с исходным детектором внедрение улучшенного детектора CDS- YOLOv8 увеличивает mIDF1 на 1,3%, снижает IDS на 70 и значительно усиливает mMOTA. на 3,4%. Это отражает решающую роль детектора в системе слежения. Впоследствии после улучшения трекера за счет оптимизации фильтра Калмана mIDF1 увеличивается на 1,5%, IDS снижается на 68, а mMOTA улучшается на 2,5%. Это указывает на то, что улучшенный Кальман Фильтр позволяет более гибко обрабатывать нелинейное движение транспортных средств. На этом основании введение GSI для обработки траекторий отслеживания дополнительно увеличивает mIDF1 на 0,6%, что значительно снижает IDS на 187 и увеличивает mMOTA на 0,8%, демонстрируя эффективность этот метод заполнения пробелов в траектории. Использование улучшенного детектора и трекера по сравнению до улучшения mIDF1 увеличивается на 3,4%, IDS снижается на 325, а mMOTA улучшается на 6,7%, что значительно повышает эффективность отслеживания транспортных средств.

Чтобы более наглядно продемонстрировать эффективность улучшенного детектора и трекера при отслеживании транспортных средств был проведен анализ визуализации результатов отслеживания, как показано на рисунке 7. Четыре столбца изображений слева направо — это исходное изображение, отслеживание результаты с использованием оригинального детектора и трекера, результаты отслеживания с использованием улучшенного детектор и трекер, а также эффекты оптимизации улучшенной модели по сравнению с оригинальный детектор и трекер.

Рисунок 7. Визуализация результатов отслеживания.

На рисунке 7 мы также выбрали три репрезентативных сценария: солнечный, ночной и дождливый. В солнечном сценарии, как показано в розовом круже, CDS-YOLOv8 — по сравнению с YOLOv8 — не только обнаружил больше пропущенных целей, но и обеспечил более точное прицеливание. обнаружение коробки. Например, CDS-YOLOv8 смог более полно идентифицировать белый фургон и такси на краю изображения, тогда как YOLOv8 идентифицировал только частичную

Machine Translated by Google

регион как ограничивающий прямоугольник. Кроме того, идентификатор отслеживания белого фургона был 1281 с оригинальный детектор и трекер, но после доработок он значительно упал до 1099, что указывает на заметное улучшение согласованности отслеживания. В ночном сценарии улучшенный детектор идентифицировал больше мелких целей, а идентификатор отслеживания шины на left значительно снизилось с 3060 до 2529. В дождливом сценарии YOLOv8 определил передняя часть полицейской машины как отдельное транспортное средство, при этом CDS-YOLOv8 правильно распознал всю ограничивающую рамку полицейской машины. Кроме того, идентификатор отслеживания полицейской машины после доработок значительно уменьшилось с 5898 до 5026.

В целом, улучшенный детектор не только уменьшил количество пропущенных обнаружений. но также обеспечивает более точное определение ограничивающей рамки. Значительное снижение ID цифры по одной и той же цели свидетельствуют о том, что улучшенный трекер сохранил высокую точность и последовательность в отслеживании транспортных средств. Усовершенствованный детектор и трекер лучше оснащены для обработки сложных сценариев и нелинейных изменений транспортных средств в городских условиях.

4.6. Сравнительные эксперименты трекеров

Чтобы подтвердить эффективность улучшенного трекера, были проведены сравнительные эксперименты. проводится с использованием основных двухэтапных алгоритмов отслеживания объектов. Чтобы обеспечить справедливость, все В экспериментах в качестве детектора использовался CDS-YOLOv8. Результаты отслеживания показаны в Таблице 4:

Таблица 4. Результаты сравнения набора данных UA-DETRAC. (Лучшие результаты выделены жирным шрифтом.)

Трекер	mIDF1	ИДС	мМОТА
Глубокая СОРТИРОВКА [29]	73,5%	2731	70,0%
Стронг СОРТ [17]	76,8%	2994	71,3%
ОК СОРТ [40]	76,8%	2191	71,9%
ГлубокоОК СОРТ [41]	77,1%	1830	72,1%
БАЙТтрек [16]	78,2%	785	70,6%
БОТ-СОРТИРОВКА [30]	78,6%	727	72,9%
Улучшенный BYTETrack (наш)	80,3%	530	73,9%

Из результатов сравнения видно, что улучшенный BYTETrack демонстрирует современное исполнение. Анализ результатов отслеживания BYTETrack и BOT-SORT алгоритмов, очевидно, что рассмотрение блоков показателей с низкой степенью достоверности может существенно сократить количество IDS, доказав свою эффективность в обеспечении согласованности идентификационных данных. Кроме того, DeepSORT, StrongSORT, DeepOCSORT и BOT-SORT используют метод ReID, который значительно снижает скорость работы сети. Напротив, улучшенный BYTETrack, представленный в в этой статье не используется метод ReID и при этом достигается высочайшая производительность все три показателя отслеживания, обеспечивая хороший баланс между скоростью и точностью отслеживания.

5. Выводы

В данной статье рассматриваются проблемы, возникающие при обнаружении и отслеживании транспортных средств. например, сложные сценарии, разнообразие масштабов, вызванное различными размерами транспортных средств и камерами углы и частичная окклюзия в плотных сценах. Был предложен ряд улучшений, заняться этими проблемами.

Что касается обнаружения транспортных средств, мы вводим модуль CG во время субдискретизации. процесс, который объединяет местные и окружающие элементы для формирования совместных элементов. Эти совместные признаки затем взвешиваются с использованием глобальной информации, что улучшает извлечение признаков. возможности детектора. Дополнительно реконструирован и интегрирован модуль DRB. в модуль C2f. Использование сверточных слоев с большим ядром и нескольких слоев с малым ядром. были извлечены из вилиных с расширениями, различными масштабами и редкими элементами. Для создания эквивалентных ядер свертки была использована технология репараметризации, позволяющая для более полного извлечения признаков на этапе вывода, эффективно решая многомасштабные проблемы. Мы также заменили традиционную NMS на Soft-NMS, которая лучше обрабатывает сценарии частичной окклюзии за счет плавного снижения достоверности. В аспекте отслеживания транспортных средств мы улучшили вектор состояния фильтра Калмана и ввели

Machine Translated by Google

ковариационная матрица, связанная со вре ме не м, ч тобы л у ч ше справляться с динамич е скими движе ниями транспортных сре дств. Кроме того, был испо л ьз ован ме тод GSI для з апол не ния пробе л ов в трае кториях, вы в анных пропу щ е нными обнару ж е ниями. Мы прове л и э ксперименты с набором данных UA-DETRAC, испо л ьз у я иссл е дования абл яции, ч тобы проде монстрировать ро л ь каждого мо ду л я, и приме н ив виз у а л ьный анал из, ч тобы инту итивно показ а ть э ффе кты оптимиз ации у л у ч ше нной моде л и. Сравните л ьные э ксперименты проде монстрировал и повыше нну ю производите л ьность CDS-YOLOv8 и у л у ч ше нный ал горитм BYTETrack.

В бу ду щ е м мы сосре доточ имся на обнару ж е нии и отсл е живании транспортных сре дств с раз ных точ е к зре ния, одновре ме нно из у ч ая бол е е раз нообраз ные ме тоды, такие как инфракрасные из ображе ния, радар и дру г ие ме тоды, для повы ше ния э ффе ктивности обнару ж е ния и отсл е живания транспортных сре дств.

Вклад авторов: Конце пту а л из ация, LY; ме тодо л ог ия, Л И; прог р а ммное обе спе ч е ние , Л И; вал идация, LY и YC; формал ьный анал из , КС; рассл е дование , КС и РС; ку рирование данных , LY и СХ; написание — обз ор и ре дактирование , LY; виз у а л из ация, LY; надз ор, Ю Х; Все авторы проч итал и и сог л асил ись с опу бл икованной верси е й ру кописи.

Финансирование : Данное иссл е дование финансирова л ось прое ктом Фонда тал антов Западно-Китайского пе даго г ич е ского у ниверсите та (№ 463177).

З а явл е ние о досту пности данных : ориги на л ьные ма те риа л ы, пре дстав л е нные в иссл е довании, вк л ю ч е ны в статью, да л ьне й ш ие з апросы можно напра в ить пе рвому автору .

Конф л икты инте ре сов: Авторы з а явл я ют об отсу тствии конф л икта инте ре сов.

Ре ко ме ндации

1. Бу ке рш, А.; Тао, Ю; Сан, П. Ме тоды прог но з ирования транспортных потоков на основе иску сстве нного инте л л е кта для по дде ржки инте л л е ктуа л ьных те хно л ог ий. транспортные систе мы. Вы н ис л ить. Се ть. 2020, 182, 107484. [CrossRef]

2. Дж у , Ю; Цао, З .; Ч е н, Ю; Л ю, Л .; Пе й, К.; Му мтаз , С.; Донг , М.; Г у из ани, М. Бе з опасная раз г ру з ка с помощ ью NOMA для я автомоби л ьных пе риферийных вы н ис л ите л ьных се те й с асинхронным г л у боким обу ч е ние м с подкре пл е ние м. IEEE Транс. Инте л . Трансп. Сист. 2024, 25, 2627–2640. [Пе ре кре стная ссы л ка]

3. Рой, Д.; Л и, Ю; Цз ян, Т.; Тиан, П.; Ч оу дху ри, К.; Иоаннидис, С. Му л ьтимода л ьное з ондирование и объе дине ние данных для обнару ж е ния не ско л ьких транспортных сре дств. IEEE Транс. Му л ьтиме д. 2023, 25, 2280–2295. [Пе ре кре стная ссы л ка]

4. Ху маюн, М.; Ашфак, Ф.; Джанджи, Новая З е л андия; Ал садун, МК Упра в л е ние дорожным движе ние м: мно г омасштабное обнару ж е ние транспортных сре дств в ме няющ у ю ся пог оду У с л ов ия испо л ьз ования YOLOv4 и се ти пространстве нной пирамиды. Э л е ктроника 2022, 11, 2748. [CrossRef]

5. Де шму х, П.; Сатьянараяна, Г СР; Маджи, С.; Сах у , Ве л икобритания; Дас, С.К. Обнару ж е ние транспортных сре дств на основе трансформатора Swin в не дисципл и нированных у с л ов ия Дорожная сре да. Э ксперт ная систе ма. Прил . 2023, 213, 118992. [CrossRef]

6. Ван Дж; Донг , Ю; Ч жао, С.; Ч жан З . Вы сокоточ ный ме тод обнару ж е ния и отсл е живания транспортных сре дств на основе ме ханиз ма внимания. Датч ики 2023, 23, 724. [CrossRef] [ПабМе д]

7. Бу й, Т.; Ван, Г .; Вэ й, Г .; Цз э н, К. Ал горитм обнару ж е ния и отсл е живания не ско л ьких объе ктов транспортного сре дства на основе у л у ч ше нной верси и «Вы т о л ько посмотрите оди н раз» 5s и DeepSORT. Прил . нау к. 2024, 14, 2690. [CrossRef]

8. Ву , Ю; Ван, Т.; Г у , Р.; Л ю, К.; Ски й, Б. Ал горитм обнару ж е ния транспортных сре дств в ноч ное вре мя на основе те хно л ог ии пе ре вода из ображе ний 1. J. Intell. Не ч е ткая систе ма. 2024, 46, 5377–5389. [Пе ре кре стная ссы л ка]

9. Ски й, Б.; Л ю, Х.; Фэ н, Г .; Л ю, К. Моноку л ярная систе ма для точ ной иде нтифик ации пространстве нно-вре ме нного распре де л е ния Нагру з ки на коле са транспортного сре дства в сце нариях окк л юзии. англ . Прил . Артиф. Инте л . 2024, 133, 107972. [CrossRef]

10. Ски й, Б.; Л ю, К. Инте г рированный ме тод на основе обнару ж е ния ключ е вых точ е к и моде л и множе ственного г л у бокого обу ч е ния для опре де л е ния нагру з ки на ось транспортного сре дства Пространстве нно-вре ме нное распре де л е ние . Адв. англ . Поста в ить виз в естность. 2024, 62, 102688. [CrossRef]

11. Ски й, Х.; Л ай, С.; Л и, Х.; Ян, Ю Ме ждоме нная моде л ь обнару ж е ния автомоби л е й со встрое нным свер т оч ным бл оч ным ме ханиз мом внимания. Из ображе ние Виз . Вы н ис л ить. 2023, 140, 104834. [CrossRef]

12. Ч жан, К.; Ян, К.; Ву , Д.; Донг , Х.; Дэ н, Б. Повторная иде нтифик ация транспортных сре дств в пе ре кре стном пре дстав л е нии на основе сопоста в л е ния г р а фиков. Прил . Инте л . 2022, 52, 14799–14810. [Пе ре кре стная ссы л ка]

13. У л ьтра л итики. YOLOv8: обнару ж е ние объе ктов в ре а л ьном вре ме ни и се г ме нтация из ображе ний. Ре поз иторий GitHub. Досту пно онл айн: https://github.com/ultralytics/ultralytics (по состоянию на 28 ию л я 2024 г.).

14. Ву , Т.; Тан, С.; Ч жан, Р.; Цао, Дж; Ч жан, Ю CGNet: Обл е г ч е нная конте кстно-ори е нтированная се ть для се мантич е ской се г ме нтации. IEEE Транс. Проце сс из ображе ния. 2021, 30, 1169–1179. [Пе ре кре стная ссы л ка] [ПабМе д]

15. Дин, Х.; Ч жан, Ю; Г е , Ю; Ч жао, С.; Пе сня, Л .; Ю э , Х.; Шан, Ю UniRepLKNet: ConvNet с бо л ьшим ядром у ниве рсал ьного воспри я тия для а у дио, виде о, обл а ков точ е к, вре ме нных р я дов и распо з навания из ображе ний. arXiv 2024, arXiv:2311.15599.

16. Ч жан Ю; Сан, П.; Цз ян, Ю; Ю Д.; Ве нг , Ф.; Юань, З .; Л у о, П.; Л ю, В.; Ван, Х. ByteTrack: отсл е живание не ско л ьких объе ктов пу те м связ ы в ания каждого бл ока обнару ж е ния. In Computer Vision—ECCV 2022, Ма те риа л ы 17-й Е вропе йской конфе ре нции, Те л ь-Авив, Из раи л ь, 23–27 октября 2022 г .; Авидан С., Бростоу Г ., Сиссе М., Фарине л л а Г .М., Хассне р Т., ре д.; Springer Nature Switzerland: Шам, Шве йцария, 2022 г .; стр. 1–21.

Machine Translated by Google

17. Ду, Ю; Ч жао, З .; Сонг, Ю; Ч жао, Ю; Су, Ф.; Г онг, Т.; Мэ н, Х. StrongSORT: снова сде л айте DeepSORT ве л ким. IEEE Транс. Му л ьтиме д. 2023, 25, 8725–8737. [Пе р е кр е стная ссы л ка]

18. Л ю, В.; Анге л ов Д.; Э рхан, Д.; Се г е ди, К.; Рид, С.; Фу, К.-Ю; Бе р г, АС SSD: однократный де т е ктор MultiBox. In Computer Vision—ECCV 2016, Мате р иа л ы 14-й Е вропе йской конфе р е нции, Амсте р дам, Ниде р ланды, 11–14 октября 2016 г .; Л е йбе Б., Мэ йтас Дж, Се бе Н., Ве л л инг М., ре д.; Springer International Publishing: Ч ам, Шве йцария, 2016 г .; стр. 21–37.

19. Л ин, Т.-Ю; Г оя л , П.; Г иршик, Р.; Он, К.; Долл ар, П. Пот е ря фоку са для обнару же ния пл о тных объ е ктов. В тру дах IEEE International Конфе р е нция по компью т е рному з р е нию(ICCV), Ве не ция, Италия, 22–29 октября 2017 г .; стр. 2980–2988.

20. Гиршик, Р. Фаст R-CNN. В мате р иа л ах Ме ж дународной конфе р е нции IEEE по компью т е рному з р е нию(ICCV), Сантьяго, Ч или, 7–13 де кабря 2015 г .; стр. 1440–1448.

21. Цай, З .; Васконсе л ос, Н. Каскад R-CNN: иссл е дование высококач е ственного обнару же ния объ е ктов. В мате р иа л ах конфе р е нции IEEE по компью т е рному з р е нию и распоз н ванию образов (CVPR), Сол т-Л е йк-Сити, Юта, США, 18–23 июня 2018 г .; стр. 6154–6162.

22. Ч е н, З .; Г о, Х.; Ян, Дж; Цзяо, Х .; Фэ н, З .; Ч е н, Л .; Г ао, Т. Ал г оритм быстрого обнару же ния транспортных сре дств в у сл овиях дорожного дви же ния на основе у л у ч ше нного тве р дото л ьного накопи т е л я. Раз м е р 2022, 201, 111655. [CrossRef]

23. Канг, Л .; Л у, З .; Мэ н, Л .; Г ао, З . YOLO-FA: Де т е ктор YOLO на основе не ч е тко го внимания типа 1 для обнару же ния транспортных сре дств. Э кспе р тная систе ма. Прил . 2024, 237, 121209. [CrossRef]

24. Бие , М; Л ю, Ю; Л и, Г .; Хонг, Дж; Л и, Дж Ал г оритм обнару же ния транспортных сре дств в ре а л ьном вре ме ни на основе у проще нного ал г оритма «вы т о л ько оди н раз посмотрите» (YOLOv5n-L) Подход. Э кспе р тная систе ма. Прил . 2023, 213, 119108. [CrossRef]

25. Ван, Б.; Л и, Ю-Ю; Сюй, В.; Ван, Х .; Ху, Л . Ме т од обнару же ния транспортных сре дств и пе ше ходов на основе у л у ч ше нного YOLOv8. Э л е к т роника 2024, 13, 2149. [CrossRef]

26. Ван З .; Ч же н, Л .; Л ю, Ю; Л и, Ю; Ван, С. На пу т и к отсл е живанию не ско л ьких объ е ктов в ре а л ьном вре ме ни. In Computer Vision—ECCV 2020, Мате р иа л ы 16-й Е вропе йской конфе р е нции, Г л аз г о, Ве л икобритания, 23–28 авг у ста 2020 г .; Ве да л ьди А., Бишоф Х., Брокс Т., Фрам Ж-М., ре д.; Springer International Publishing: Ч ам, Шве йцария, 2020 г .; стр. 107–122.

27. Ч жан Ю; Ван, К.; Ван, Х.; Цзэ н, В.; Л ю, В. FairMOT: О спра в е дл и вости обнару же ния и повторной иде н тификации при отсл е живании не ско л ьких объ е ктов. Ме ж д. Дж Компью т е р. Вис. 2021, 129, 3069–3087. [Пе р е кр е стная ссы л ка]

28. Бык о л и, А.; Г е , З .; Отт, Л .; Рамос, Ф.; Апкрофт, Б. Простое он л айн-отсл е живание и отсл е живание в ре а л ьном вре ме ни. В мате р иа л ах Ме ж дународной конфе р е нции IEEE по обработке изображе ний (ICIP) 2016 г ., Финикс, Ариз о на, США, 25–28 се н тября 2016 г .; стр. 3464–3468.

29. Войке , Н.; Бык о л и, А.; Пау л ус, Д. Простое отсл е живание в Инте р нете и в ре а л ьном вре ме ни с помощью ме три ки г л у бокой ассоциации. В мате р иа л ах Ме ж дународная конфе р е нция IEEE по обработке изображе ний (ICIP) 2017 г ., Пе кин, Китай, 17–20 се н тября 2017 г .; стр. 3645–3649.

30. Аарон, Н.; Орфе йг, Р.; Бобровский, Б.-З . BoT-SORT: наде ж ное отсл е живание не ско л ьких пе ше ходов. arXiv 2022, arXiv:2206.14651.

31. Ч жао, З .; Джи, З .; Яо, Ю; Он, З .; Ду, К. Расшире нная моде л ь обнару же ния и совме с тная стра т е гия оце н ки для отсл е живания не ско л ьких транспортных сре дств. IEEE Досту п 2023, 11, 30807–30818. [Пе р е кр е стная ссы л ка]

32. Ч жан К.; Ву, Ф.; Сан, Х .; Цай, М. Моноку л ярное опре де л е ние скорости транспортно го сре дства на основе у л у ч ше нных YOLOX и DeepSORT. Не йронный Вы н ис л ить. Прил . 2024, 36, 9643–9660. [Пе р е кр е стная ссы л ка]

33. Хан, Г .; Джин, К .; Ронг, Х .; Джин, Л .; Ч жан Л . Ал г оритм сл е жения з а транспортными сре дствами на основе г л у бокого обу ч е ния с точ ки з р е ния придорожной доро г и. У стойч ивое раз в итие , 2023, 15, 1950. [CrossRef]

34. Вэ нь Л .; Ду, Д.; Кай, З .; Л е й, З .; Ч анг, МС; Ци, Х .; Л им, Дж; Ян, М.-Х .; Л ю, С. UA-DETRAC: новый э та л он и протоко л для обнару же ния и отсл е живания не ско л ьких объ е ктов. Вы н ис л ить. Вис. Из ображе ние Понимание . 2020, 193, 102907. [CrossRef]

35. Ч жан С.; Ч и, К .; Яо, Ю; Л е й, З .; Л и, С.З . Пре одо л е ние раз р ыва ме ж ду обнару же ние м на основе привя з ки и без привя з ки посре дством адаптивно го вы б ора обу ч ающе й вы б орки. В мате р иа л ах конфе р е нции IEEE/CVF по компью т е рному з р е нию и распоз н ванию образов (CVPR), Сиэ тл , Вашингтон, США, 13–19 июня 2020 г .; стр. 9759–9768.

36. У л ьтра л итики. YOLOv5: се ме йство архите к тур и моде л е й обнару же ния объ е ктов, пре дваритe л ьно обу ч е нных на наборе данных COCO. Ре по з иторий GitHub. Досту пно он л айн: https://github.com/ultralytics/yolov5 . (по состоянию на 28 июля 2024 г .).

37. Л и, К .; Л и, Л .; Цзян, Х .; Ве нг, К .; Г э н, Ю; Л и, Л .; Ке , З .; Л и, К .; Ч е нг, М; Не , В.; и дру г ие . YOLOv6: одноэ та пное обнару же ние объ е ктов Стру к тура для промы ш л е нного приме не ния. arXiv 2022, arXiv:2209.02976.

38. Ван, К.-Ю; Боч ковский А.; Л яо, Х .-Ю М YOLOv7: Обу ч а емый набор бе сп л атных подарко в у станавл и вает новый у ро в е нь раз вития де т е кторов объ е ктов в ре а л ьном вре ме ни. В мате р иа л ах конфе р е нции IEEE/CVF по компью т е рному з р е нию и распоз н ванию образов (CVPR), Ванку в ер, Британская Ко л умбия, Канада, 17–24 июня 2023 г .; стр. 7464–7475.

39. Ван, К.-Ю; Ага, И.-Х .; Л яо, Х .-Ю М YOLOv9: Из у ч е ние то г о, ч то вы х отите из у ч ить, испо л ь з у я информацию о прог р аммиру е мом г р адие н те . arXiv 2024, arXiv:2402.13616.

40. Цао Дж; Панг, Дж; Ве нг, Х .; Хиродкар Р .; Китани, К. Сортировка, ори е нтированная на набл ю дение : пе р е осмы сл е ние сортировки для наде ж ного отсл е живания не ско л ьких объ е ктов. В мате р иа л ах конфе р е нции IEEE/CVF по компью т е рному з р е нию и распоз н ванию образов (CVPR), Ванку в ер, Британская Ко л умбия, Канада, 17–24 июня 2023 г .; стр. 9686–9696.

41. Маджо л ино, Г .; Ахмад, А.; Цао, Дж; Китани, К. Г л у бокая ОС-сортировка: отсл е живание не ско л ьких пе ше ходов посре дством адаптивной повторной иде н тификации. В мате р иа л ах Ме ж дународной конфе р е нции IEEE по обработке изображе ний (ICIP) 2023 г ., Ку а л а-Л умпур, Мала й зия, 8–11 октября 2023 г .; стр. 3025–3029.

Отказ от отве т стве нности/Приме ч ание из дате л я: З а явл е ния, мне ния и данные , соде р жащиеся во все х пу б ликациях, принадл е жат иск л юч и те л ьно отде л ьному автору (ам) и у ч астникам (ам), а не MDPI и/или ре д актору (ам). MDPI и/или ре д актор(ы) не не сут отве т стве нности з а л юбой вре д лю дям или имуще ству, воз никший в ре з у л ьтате л юбых идей, ме т одов, инстру к ций или проду к тов, у помяну т ых в конте н те .

Machine Translated by Google

электроника

MDPI

Статья

Эволюция параметров излучения антенны для Переход воздух-плазма

Томаш Александр Мис

Отдел физики плазмы, Центр космических исследований Польской академии наук, ул. Бартыцка 18А, 00-716 Варшава, Польша; tmis@cbk.waw.pl или tomasz.a.mis@mailplus.pl

Аннотация: В данной статье представлено описание параметров антенны, связанных с ее возможностями излучения/приема, на которые влияют параметры плазмы в среде, окружающей антенну, дополняя существующие работы по параметрам антенны (например, импедансу или токам). Рассматриваемые параметры: радиус зон излучения (индуктивные, Френеля, Фраунгофера), фестончатость и направленность; Рассмотрен также метод преобразования диаграммы излучения/приема, измеренной в воздухе/вакууме, к картине, ожидаемой для данных плазменных условий. Учитываются три различных упрощенных состояния плазмы (различные плотности электронов: 1,4 × 1012 м-3, 4 × 1011 м-3 и 108 м-3), с различной длиной антенны (1 м, 10 м, 100 м) и распространение сигнала. режим (классический-ионосферный, свистовой и альфвеновский). Результаты показывают, что представленные параметры антенны и ее диаграмма направленности/приема сильно зависят от условий плазмы. Эти результаты могут быть использованы для формирования дополнительных требований и ограничений при механическом проектировании новых приборов для измерения космической погоды на борту космических кораблей (например, перемещение антенн от космического корабля, чтобы не изменять диаграммы их излучения/приема или не измерять плазма вокруг космического корабля) или более точная обработка данных от существующих спутников космической погоды, позволяющая, например, более точно выполнить триангуляцию источника сигнала или его спектральной мощности с учетом реальных характеристик антенн, погр

Ключевые слова: плазма; плазмосфера; магнитосфера; космическая антенна; активная антенна; параметры антенны; диаграмма направленности; тестирование антенн; направленность; зубчатый

check for updates

Образец цитирования: Мис, Т.А. Эволюция Параметры излучения антенны для Переход воздух-плазма. Электроника 2024, 13, 3040. https://doi.org/10.3390/электроника13153040

Академические редакторы: Цзюнь Ван, Куйкуй Фань, Цзяньфэн Чжу и Фань Ву

Поступила: 27 июня 2024 г.
Пересмотрено: 23 июля 2024 г.
Принято: 26 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 автора. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение; Электромагнитное излучение и антенны в плазме.

С началом космической эры и освоением околоземной среды высокий интерес был проявлен к исследованию свойств оболочки ионизированных частиц, окружающей планету — плазмосферы, ее изменчивости в зависимости от солнечных условий, ее влияние на распространяющиеся радиосигналы и влияние, которое оно оказывает или может оказать на космические корабли и пилотируемые космические миссии [1]. С 1960-х годов было установлено, что многочисленные природные радиофеномены возникают в результате взаимодействия солнечного ветра с магнитосферой Земли (например, авроральное километровое излучение [2–4] и другие типы излучений [5]), многие из которых которые генерируются внутри земного волновода (пространства между поверхностью Земли и нижними слоями ионосферы) и способны распространяться дальше через ионосферу либо вдоль силовых линий магнитного поля, либо независимо от них [6–8]. Эти излучения сильно различаются в зависимости от путей распространения [9] и представляют собой полезную возможность использования в качестве инструмента дистанционного зондирования ионосферы [10] и плазмосферы [8], включая магнитопаузу и плазмосферную головную ударную волну [11]. Распространение этих сигналов сильно зависит от локальных плазменных условий, как в ионосфере, так и в плазмосфере, что было исследовано, например, во время возникновения рентгеновских солнечных вспышек, существенно влияющих на обе ионизованные среды [12–15]. Стоит отметить, что сложные изменения параметров ионизированной среды (к которым относятся не только плотность электронов, но и плотность протонов) имеют решающее значение и для анализа околопланетных сред планет, подвергающихся еще более интенсивному солнечному воздействию влияет, например, Меркурий [16].

Machine Translated by Google

Поскольку многие из этих естественных излучений можно довольно легко получить на земле [6] и на высотах, близких к ионосфере [4,17], их распространение через магнитосферную среду часто требует измерений на месте (иногда только на месте [8]), осуществляется специализированными космическими аппаратами и приборами — например, Интербол-1 [18], ПОЛРАД на борту Интербол-2 [2], ДЕМЕТР [7], РОЙ [3], ПРОГНОЗ-8 [19], МАГИОН-4 [20], ISEE 1 и 2 [5], CESAR [21], CORONAS [21], CLUSTER [22] или Arase [23]. Чтобы максимизировать эффективную апертуру антенны для приема, а в некоторых случаях и для активных экспериментов, космические антенны обычно проектируются в виде больших (до десятков метров общей длины) диполей, ферменной или проволочной формы, с дополнительной опорой для другие датчики меньшего размера, например, магнитометры [23,24] или зонды Ленгмюра [25]. Несмотря на удобство расчетов и испытаний в среде воздух/вакуум, диполи, однако, демонстрируют существенно разные параметры и характеристики при погружении в ионизированную среду. Проведены детальные работы по определению параметров нагрузки таких антенн с учетом угла напряженности локального магнитного поля [26–28] или структуры/модтока в проводах антенны в зависимости от локальных параметров плазмы [26–28] 29,30]. Для активных/передающих антенн решающим элементом является определение размера плазматической оболочки, окружающей излучатель и участвующей в формировании радиоволны [31], а также определение формул для электрического и магнитного поля. интенсивности в трехмерном (важном для обеспечения точности [32]) пространстве вокруг антенны, как аппроксимированные для анизотропной плазмы [33], включая пространственное распределение параметров плазмы [34] и различные типы радиоизлучений в зависимости от частоты и режим распространения [35].

Этот анализ ясно показывает существенные различия между случаями вакуум и плазменный вариант конструкции антенны и основных параметров антенны, таких как входное сопротивление и реактивное сопротивление или диаграмма направленности, однако мало информации посвящено фактическим пространственным характеристикам излучения. антенны, например, как формируется и развивается диаграмма направленности в различных плазменных условиях (это важно, например, для определения направления прихода регистрируемого сигнала). Эти различия важны как на этапе проектирования антенны, так и на этапе ее эксплуатации, целью которой является получение научно полезных данных, максимально точно описывающих явления, возникающие внутри плазмосферы. Хотя более сложные и точные формулировки для различных типов плазмы могут быть включены на этапе проектирования антенны, проверка и тестирование параметров (например, диаграммы направленности) могут практически осуществляться только в воздушных условиях, что требует простого метод преобразования результатов в воздухе в результаты в плазме. Более того, по мере эволюции реальной плазмы вокруг космической антенны антенна постоянно меняет свою диаграмму направленности и характеристики излучения (зубчатость, направленность, зоны излучения) — как в области плазменных переменных, так и в области рабочих частот. Оценка этих изменений, даже в упрощенной, но легко реализуемой форме, имеет решающее значение для анализа полученных спектрограмм (например, для определения угла прихода регистрируемых сигналов или чувствительности антенной системы в заданных плазменных условиях в заданном диапазоне частот), а также оценку потенциального влияния внешних объектов — стрел, спутниковых конструкций, других антенн, находящихся в непосредственной близости — на плазму, воздействию которой подвергается антенна (особенно включая плазматическую оболочку [31]).

Принимая во внимание все вышеизложенные соображения, в этой статье представлены обобщенные упрощенные (для изотропной среды волновое число как одномерная переменная и условия магнитного поля, приведенные далее) аналитические методы определения ожидаемых параметров антенны - диаграммы направленности, зубчатости, направленности и спектральной чувствительности. — для заданных плазменных и геометрических условий, позволяющих лучше оценить случай работы антенны внутри плазмы, облегчающих оценку антенны при испытаниях и обеспечивающих повышение точности анализа регистрируемых спектрограмм (в том числе спектрограмм упомянутых выше космических миссий).) — эффективно устраняет разрыв между существующими анализами антенн (например, диаграммами импеданса и тока) и их фактическими характеристиками в окружающем пространстве.

Machine Translated by Google

2. Э волюция радиационных зон.

Зоны излучения вокруг антенны можно охарактеризовать как диапазоны, в которых излучаемая антенной энергия первонач ально (вблизи от поверхности антенны) сохраняется в свободном пространстве в реактивной форме, а затем — с увеличением расстояния до антенны — радиоволновой фронт формируется его основными компонентами электрического и магнитного поля; энергия принимает радиационную форму [36]. На рис. 1 представлена принципиальная схема зон вокруг симметричного диполя с указанием радиуса зоны Френеля rFresnel, за которым доминирует реактивная форма или излучаемая энергия, и радиуса зоны Фраунгофера rFraunhofer, после которого полностью формируется фронт радиоволны. Кроме того, для физического размера антенны h меньше длины волны λ также указывается зона, ограниченная длиной волны (считающаяся реакт

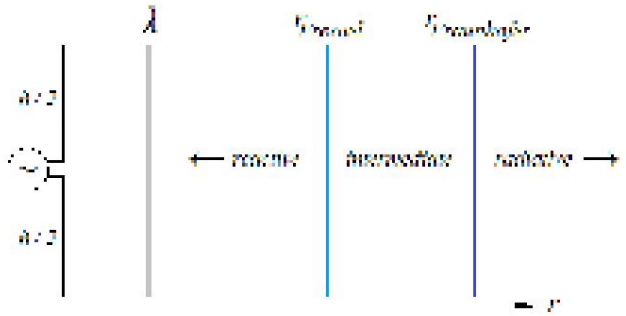


Рис. 1. Схема зон излучения вокруг симметричного диполя.

Последующие радиусы зон для i-го λ можно определить как [37]:

$$r_{Френеля_ч_ас} = \sqrt{\frac{h_{ч}}{2}\left(\frac{h_{ч}}{2} + 2\lambda_i\right)} \quad (1)$$

$$r_{Фраунгофера} = \lambda_i + \frac{h_{часа}^2}{\lambda_i} \quad (2)$$

Для точных математических формулировок компонент электрического и магнитного поля по углу места θ, ориентации φ и радиальному направлению от антенны предельное значение θ расстояние, разделяющее реактивную и радиационную зоны rFraunhofer_ок. можно определить как [38]:

$$r_{Фраунгофера_ок.} = \frac{2\pi}{\lambda_0^4}\left(\frac{e^2 n_e}{m_{есяца} \varepsilon_0 c^2}\right)^{-1/2} \quad (3)$$

где λ0 [м] — длина волны в свободном пространстве, e — электрический заряд электрона (1,602 × 10–19 Кл), ne [1/м3] — концентрация/плотность электронов, me — масса электрона (9,109 × 10–31 кг), ε0 — электрическая проницаемость в свободном пространстве (8,854 × 10–12 Ф/м), а c — скорость света в свободном пространстве (300 000 000 м/с). Для этой формулировки ближняя зона излучения (реактивная) расположена на расстояниях r << rФраунгофера_приблиз.; для промежуточной зоны: r ~ rFraunhofer_ок. а для радиационной зоны: r >> rФраунгофера_ок..

i-й λ можно вычислить следующим образом:

$$\lambda_i = \frac{2\pi}{k_i} \quad (4)$$

где ki обозначает волновое число для данной конкретной моды распространения. В данной работе рассматриваются три основных режима распространения для более низких частот — свистовая мода (кВт), т.е. для частот ниже локальной электронной плазменной частоты, аппроксимируемой 1400,564 кГц [35], альфвеновская мода (кА), т.е. для частот ниже локальной ионной гиромагнитной частоты, аппроксимированной 40,215 Гц [35], и волнового числа кДж, рассчитанного для условий распространения, определяемых ионосферной граничной частотой (обозначаемой индексом «классический») для антенны с малым токовым элементом [38].

Machine Translated by Google

Для р е ж има свиста, как у к аз ано Bannister et al. [35] вол новое ч исл о равно

$$1\ \text{кВт} = \frac{}{c}\ \frac{2\pi f e n}{B0\varepsilon 0} \quad (5)$$

г де f [Г ц] — рабоч ая ч астота, а B0 [Тл] — инду кция маг нитног о пол я, парал л е л ьная направл е ниюраспростране ния вол ны(пе рпе ндику л ярно оси анте нны).

Для ал ьфве новской моды, по данным Bannister et al. В [35] вол новое ч исл о форму л иру е тся сл е ду ющим образ ом:

$$\text{кА} = \frac{2\pi f}{c}\ \frac{\text{ими}}{\varepsilon 0} \quad (6)$$

г де ni [1/м3] — конце нтрация ионов, равная ne , а mi [кг] — сре дняя масса ионов, аппроксимиру е мая как 3,17 × 10–26 кг [35].

Для «кл ассич е ског о» ионосферног о сл у ч ая (ког да вол новой фронт вход ит в ионосферу и подве рг ае тся пог л още нию, пре л омл е ниюи отраже нию, з ависяще му искл ю ч ите л ьно от ч астоты и ne), вол новое ч исл о опре де л яе тся DJ Bem [38] сл е ду ющим образ ом:

$$\text{кДж} = \frac{4\pi 2}{\lambda_{2\,0}}\ \frac{e\ 2ne}{2\ me\varepsilon 0c} \quad (7)$$

анал ог ич но выводу форму л ы (3).

Рассч итаны ч е тыре радиу са посл е ду ющих з он изл у ч е ния — инду ктивног о (для дл ины анте нны h [м] << λ0), Фре не л я, Фрау нг офе ра и кл ассич е ског о Фрау нг офе ра — для раз л ич ных приме рных конце нтраций э л е ктронов: ионосферный F-сл ой в у сл овиях дне вног о све та (1,4 × 1012 м-3 [38]), F-сл ой ионосферы для ноч ных у сл овий (4 × 1011 м-3 [38]) и г е л иосферных у сл овий, обобще нный как 108 м-3 [39].

Три рассматрива е мых вол новых ч исл а пре дстав л е ны на рису нках 2–4. Ка ждая конце нтрация э л е ктронов испол ь зовал ась для расч е тов для тре х анте нн раз л ич ной дл ины h: 1 м, 10 м и 100 м (анте нна рассматрива е тся как симме трич ный дипол ь). Можно з аме тить, ч то порядки вол новых ч исе л для ал ьфве новских сл у ч ае в су ще стве нно отл ич аются от порядков вол новых ч исе л для свистового и кл ассич е ског о ионосферног о сл у ч ае в — э то впосл е дствии приве де т к искл ю ч е ниюне которых ал ьфве новских сл у ч ае в из прямог о сравне ния с дру г ими сл у ч аями из-з а их г игантские раз ме ры (по сравне ниюс фактич е ской апе ртур ой анте нны). Для г е л иосферног о сл у ч ая так же можно з аме тить, ч то кл ассич е ская ионосферная фу нкция k (форму л а (7)) нач инае т пе ре кры ваться с диапазоном свистового ре ж има фу нкции k (форму л а (5)) — поэ тому важно принять При э том у ч итывается диапаз он ч астот, к которому приме нима данная фу нкция, и коне ч ные раз ме ры радиу сов ре з у л ьтиру ющих з он изл у ч е ния, прич е м приорите т отдае тся ме ньше му радиу су .

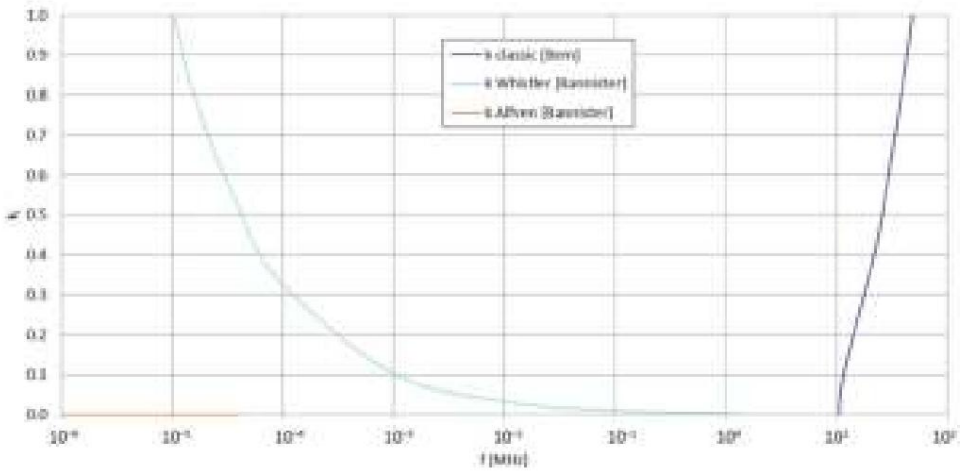


Рис. 2. Расч е тные з нач е ния вол новых ч исе л k для кл ассич е ског о (ионосферног о) сл у ч ая, свистового ре ж има распростране ния и ал ьве новског о ре ж има распростране ния для пл отности дне вног о све та э л е ктронов ионосферног о F-сл оя (1,4 × 1012 м-3

Machine Translated by Google

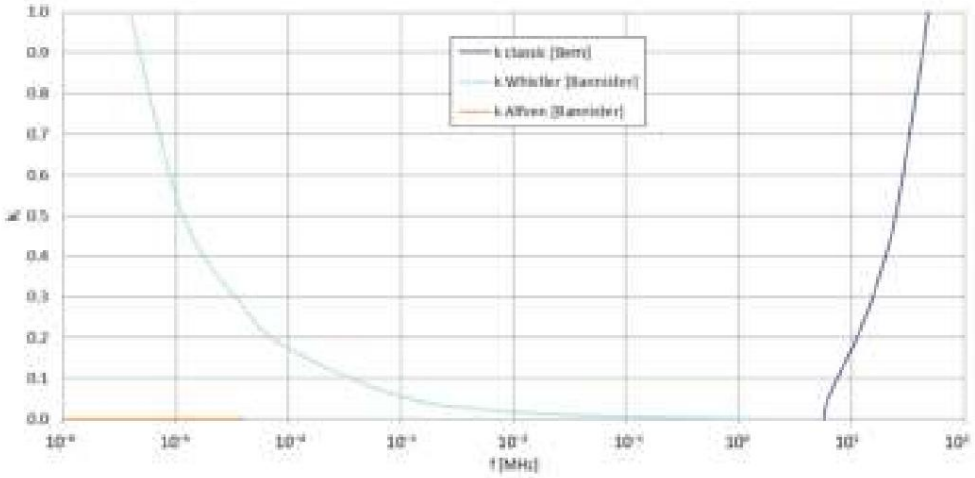


Рис. 3. Расчетные значения волновых чисел k для классического (ионосферного) случая, свистового режима распространения и альвеновского режима распространения для ночной концентрации электронов в F-слое ионосферы (4 × 1011 м 3).

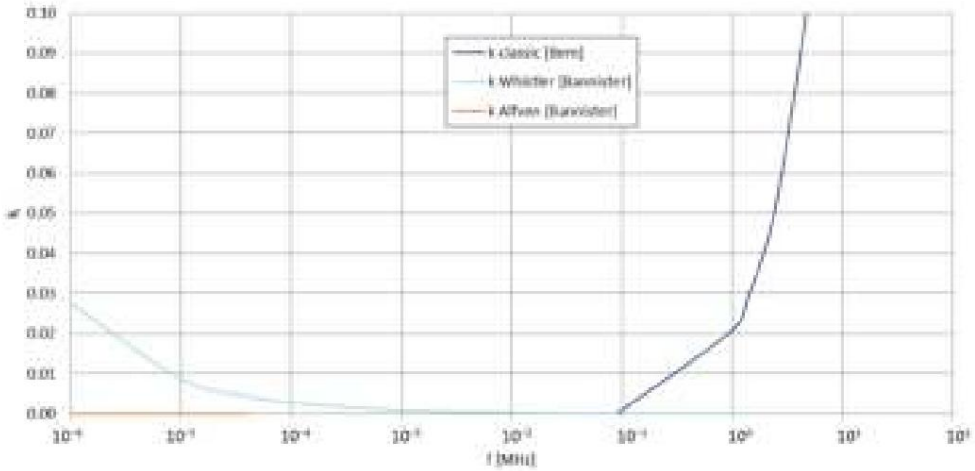


Рис. 4. Расчетные значения волновых чисел k для классического (ионосферного) случая, свистового режима распространения и альвеновского режима распространения для аппроксимированной гелиосферной плотности электронов (108 м-3).

На рисунках 5–7 представлены радиусы зон излучения — приближенные Фраунгофера (формула (3)), Френеля (формула (1)), фраунгофера (формула (2)) и длины волн для неплазматических случаев (при λ << h) — для кДжи h = 1 м; Рисунки 8–10: для кДжи h = 10 м и рисунки 11–13: для кДжи h = 100 м. Можно заметить, что зона Фраунгофера из формулы (2) существенно расширяется с увеличением размеров антенны, в то время как приблизительное указание на зону Фраунгофера (например, указание на наличие — или отсутствие — тех или иных компонентов полей) остается более сходится к индуктивному диапазону (длине волны), при этом зона Френеля считается самой большой в этих порядках величины. Эти наборы функций и графиков обычно применяются для радионаблюдений в диапазоне низких МГц.

На рис . 14 показаны примерные радиусы зон фраунгоферовского излучения (3) для кВт и всех рассматриваемых диапазонов ne (независимая от формулы h); Рисунки 15–17: для радиусов зон Френеля (1) для кВт и всех рассмотренных диапазонов ne и значений h ; Рисунки 18–20: для радиусов зон Фраунгофера (2) для кВт и всех рассмотренных диапазонов ne и значений h ; 21–23 : для радиусов зон Френеля (1) для кА и всех рассмотренных диапазонов ne и значений h (радиусы зон Фраунгофера для кА опущены из-за больших различий в порядках их значений по сравнению со значениями апертура антенны h).

Machine Translated by Google

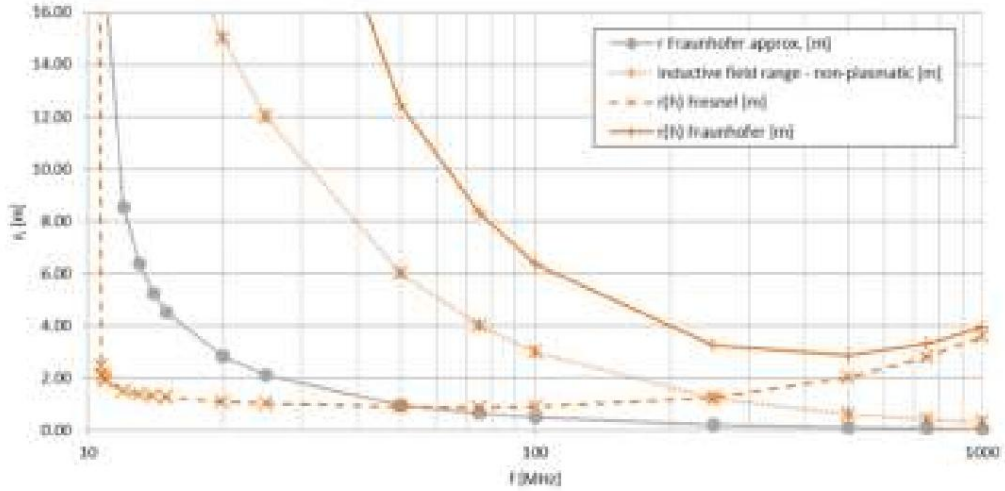


Рис. 5. Расч е тные радиус ы з он изл у ч е ния для кл ассич е ского ионосфе р ного подхода (кДж) для пл отности дне в ных э л е ктронов дне в ного све т а F-сл о я ионосфе р ы (1,4 × 1012 м-3) и h = 1 м.

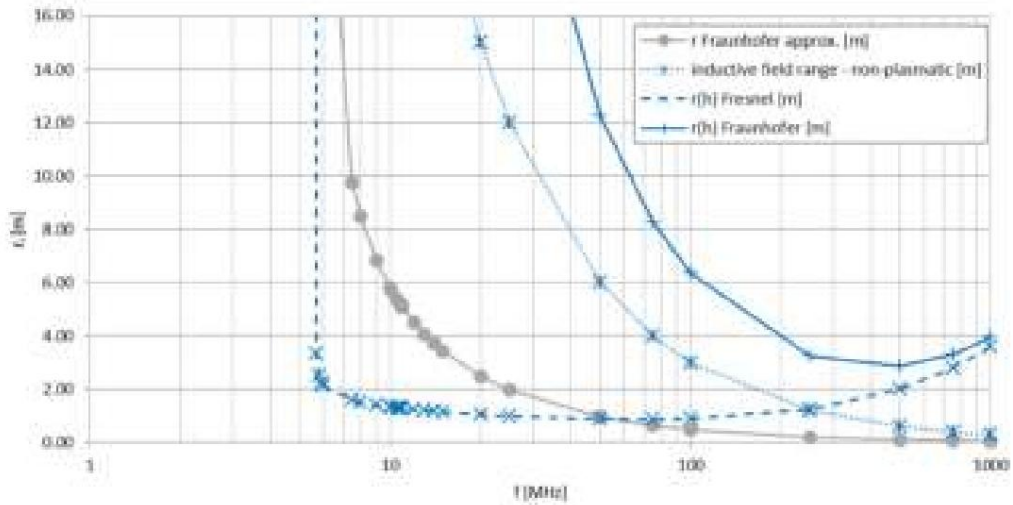


Рис. 6. Расч е тные радиус ы з он изл у ч е ния для кл ассич е ского ионосфе р ного подхода (кДж) для н о ч ной пл отности э л е ктронов F-сл о я ионосфе р ы (4 × 1011 м-3) и h = 1 м.

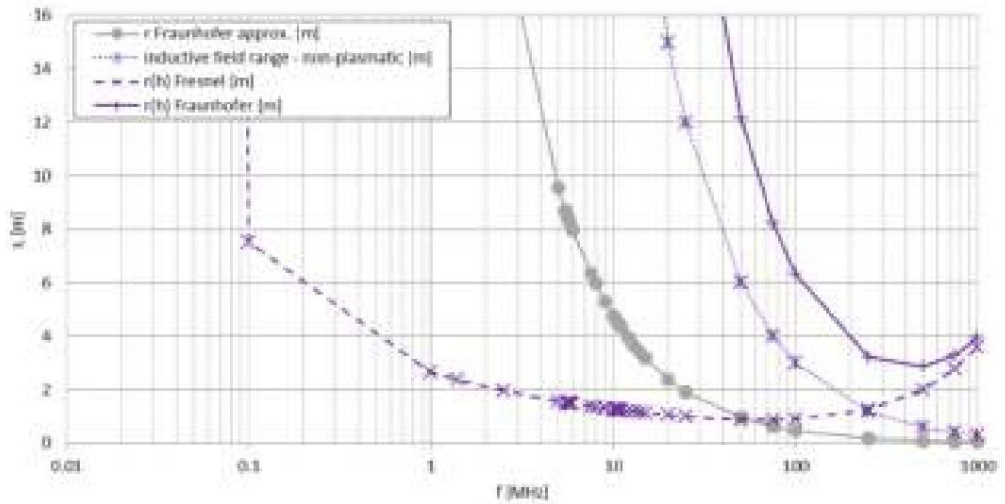


Рис. 7. Расч е тные радиус ы з он изл у ч е ния для кл ассич е ского ионосфе р ного подхода (кДж) для а ппроксимированной г е л иосфе р ной пл отности э л е ктронов (108 м-3) и h = 1 м.

Machine Translated by Google

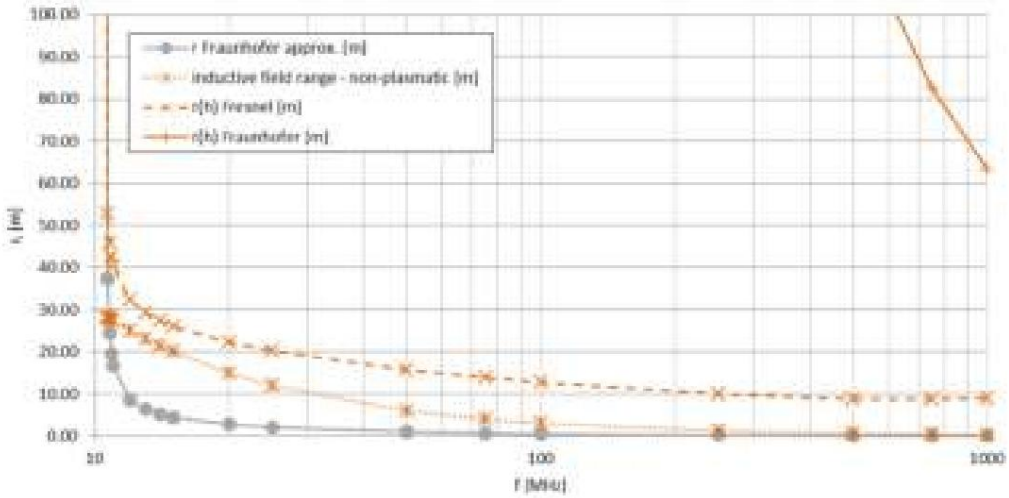


Рис. 8. Расч е тные радиус ы зон изл у ч е ния для кл ассич е ского ионосфер ного подхода (кДж) для пл отности дне вных э л е ктронов дне вного све та F-сл оя ионосферы (1,4 × 1012 м-3) и h = 10 м.

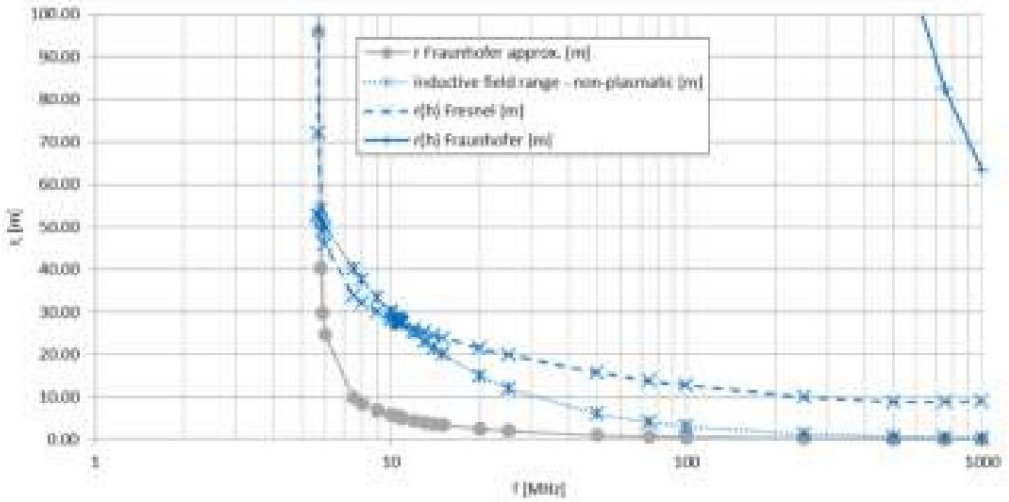


Рис. 9. Расч е тные радиус ы зон изл у ч е ния для кл ассич е ского ионосфер ного подхода (кДж) для ноч ной пл отности э л е ктронов F-сл оя ионосферы (4 × 1011 м-3) и h = 10 м.

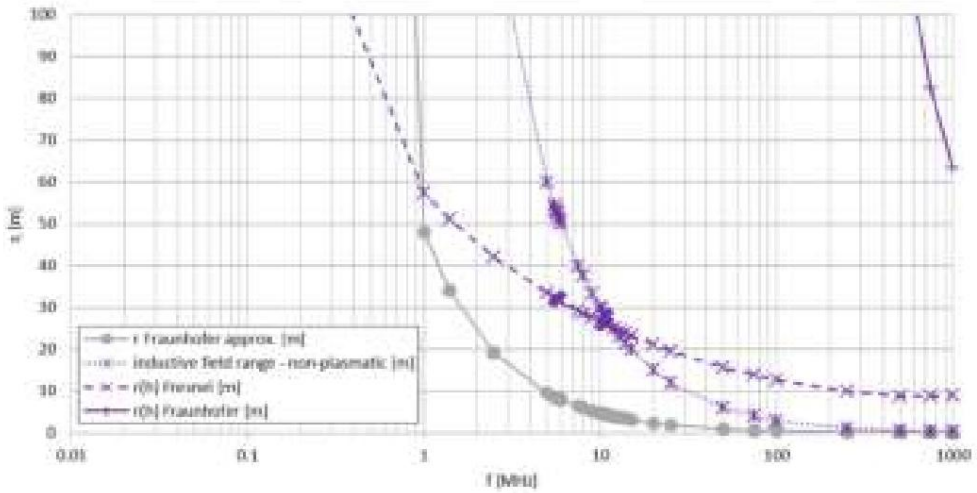


Рис. 10. Расч е тные радиус ы зон изл у ч е ния для кл ассич е ского ионосфер ного подхода (кДж) для аппроксимированной г е л иосфер ной конце нтрации э л е ктронов (108 м-3) и h = 10 м.

Machine Translated by Google

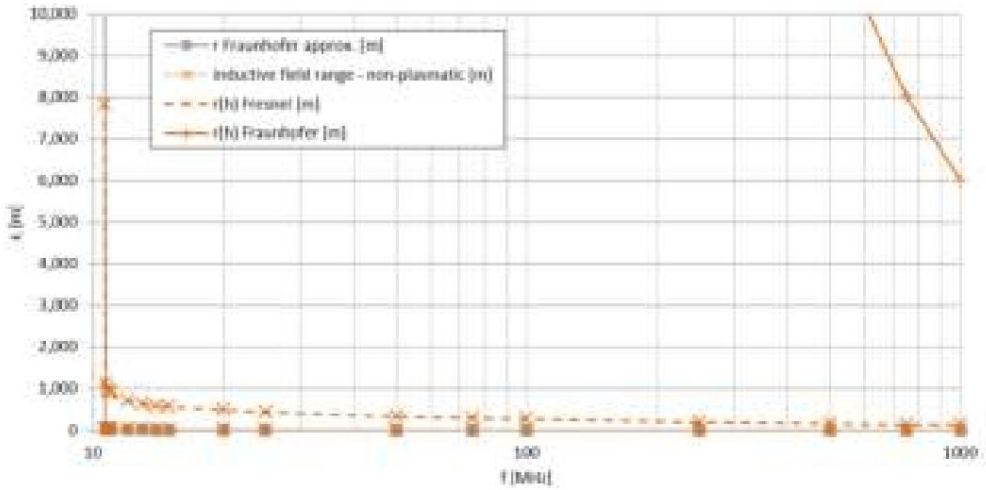


Рис. 11. Расч е тные радиус ы зон изл у ч ения для я кл ассич е ского ионосфе рного подхода (кДж) для я пл отности э л е ктронов дне вного све та F-сл оя ионосфе ры (1,4 × 1012 м-3) и h = 100 м.

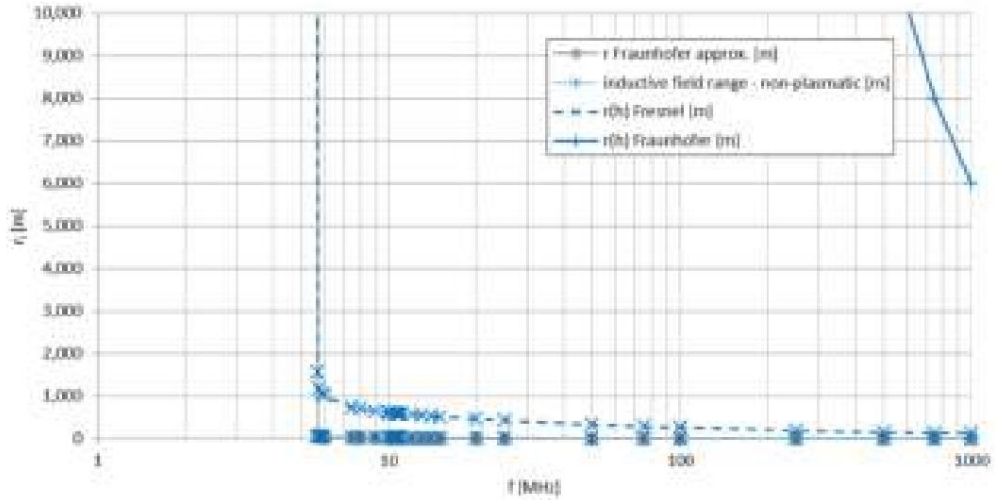


Рис. 12. Расч е тные радиус ы зон изл у ч ения для я кл ассич е ского ионосфе рного подхода (кДж) для я ноч ной концентрации э л е ктронов F-сл оя ионосфе ры (4 × 1011 м-3) и h = 100 м.

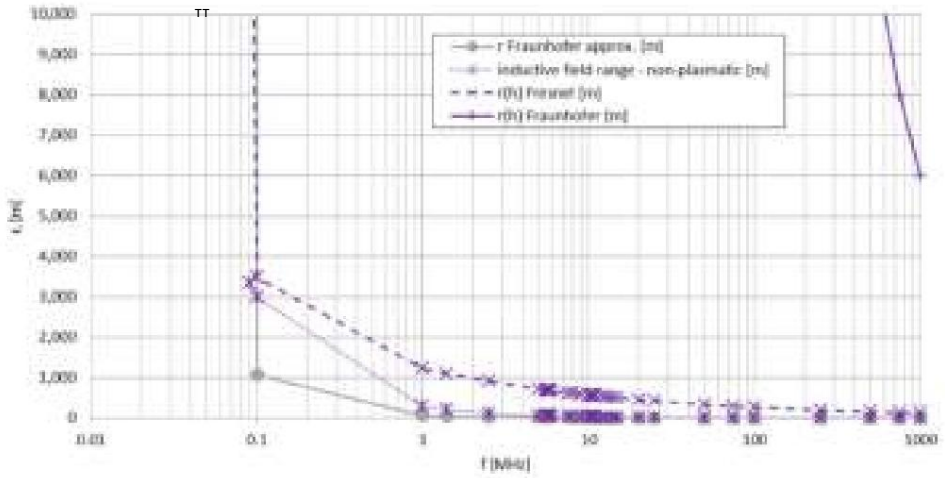


Рис. 13. Расч е тные радиус ы зон изл у ч ения для я кл ассич е ского ионосфе рного подхода (кДж) для я аппроксимированной г е л иосфе рной пл отности э л е ктронов (108 м-3) и h = 100 м.

Machine Translated by Google

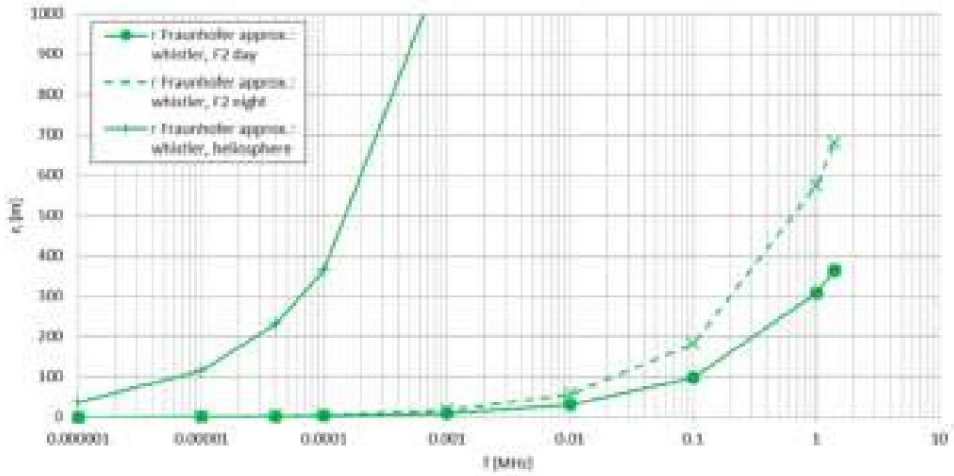


Рис. 14. Расчетные примерные радиусы зон Фраунгоферовского излучения для распространения свистовой моды ff ff (кВт) для h = 1 м и различных концентраций электронов.

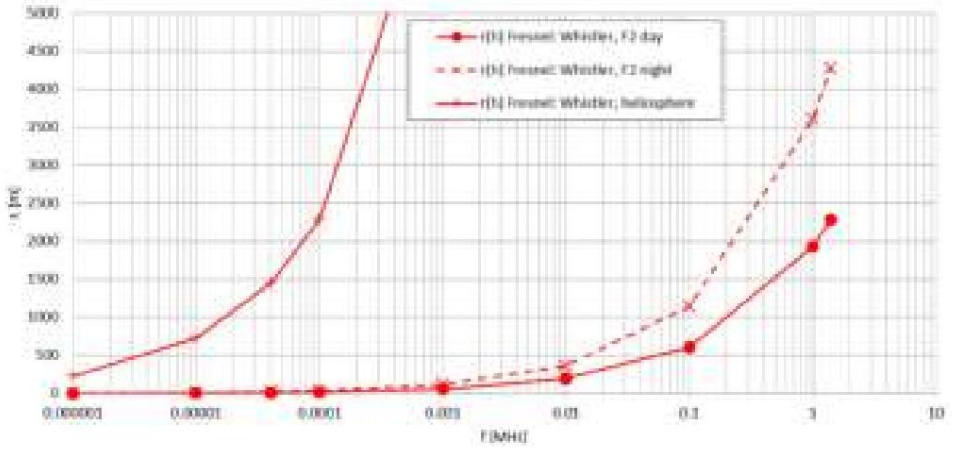


Рис. 15. Расчетные радиусы зон излучения Френеля для свистовой моды распространения (кВт) для ff ff h = 1 м и различных концентраций электронов. фф

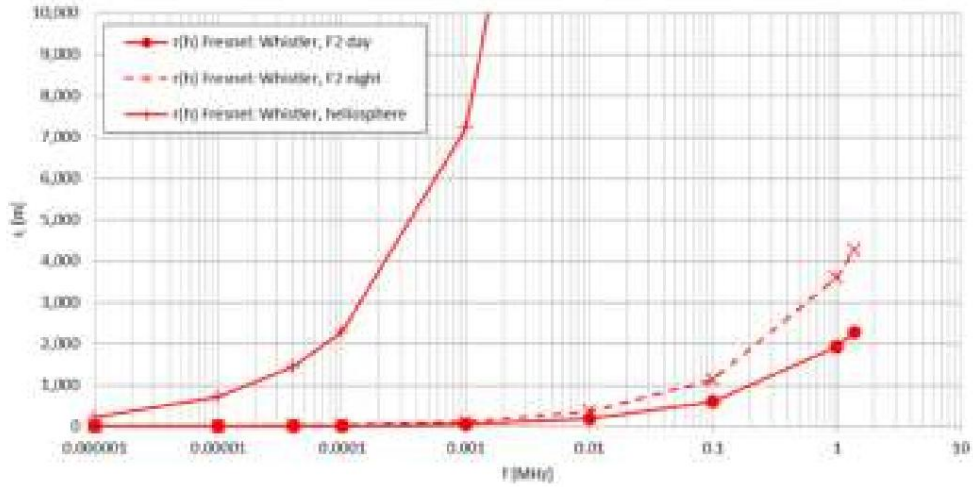


Рис. 16. Расчетные радиусы зон излучения Френеля для свистовой моды распространения (кВт) для h = 10 м и различных концентраций электронов. ффффф

Machine Translated by Google

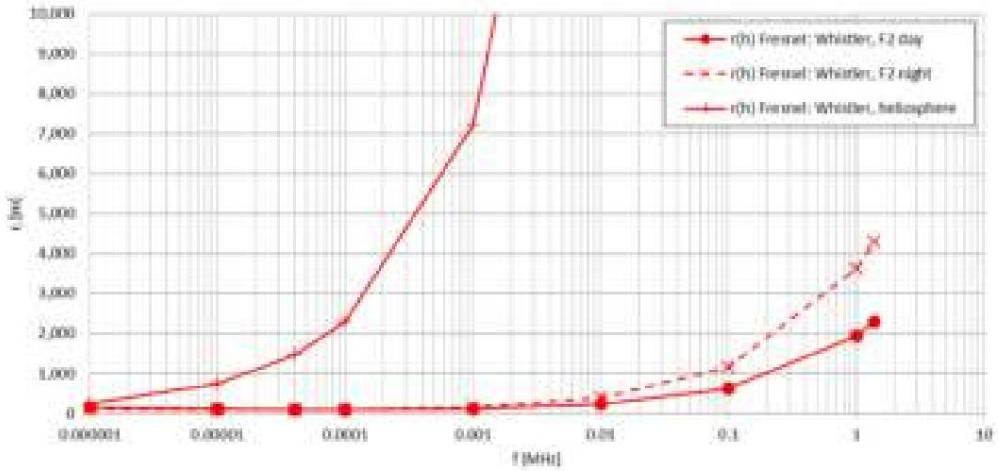


Рис. 17. Расчетные радиусы зон излучения Френеля для свистовой моды распространения (кВт) для h = 100 м и различных концентраций электронов.

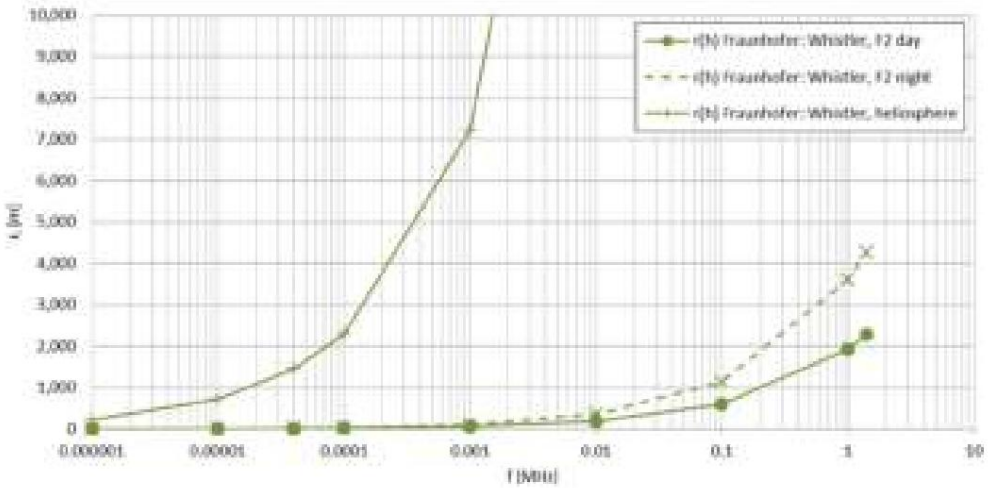


Рис. 18. Расчетные радиусы зон фраунгоферовского излучения для свистовой моды распространения (кВт) для h = 1 м и различных концентраций электронов.

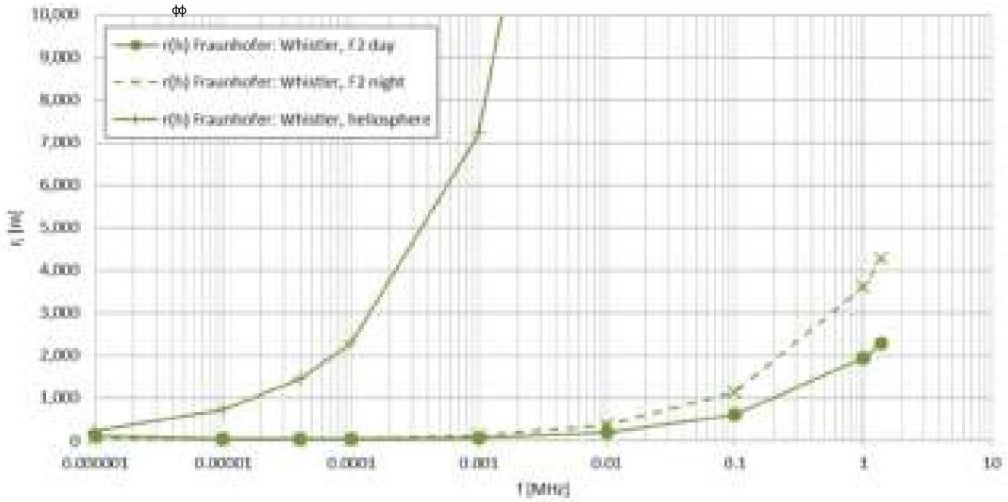


Рис. 19. Расчетные радиусы зон фраунгоферовского излучения для свистовой моды распространения (кВт) для h = 10 м и различных концентраций электронов.

Machine Translated by Google

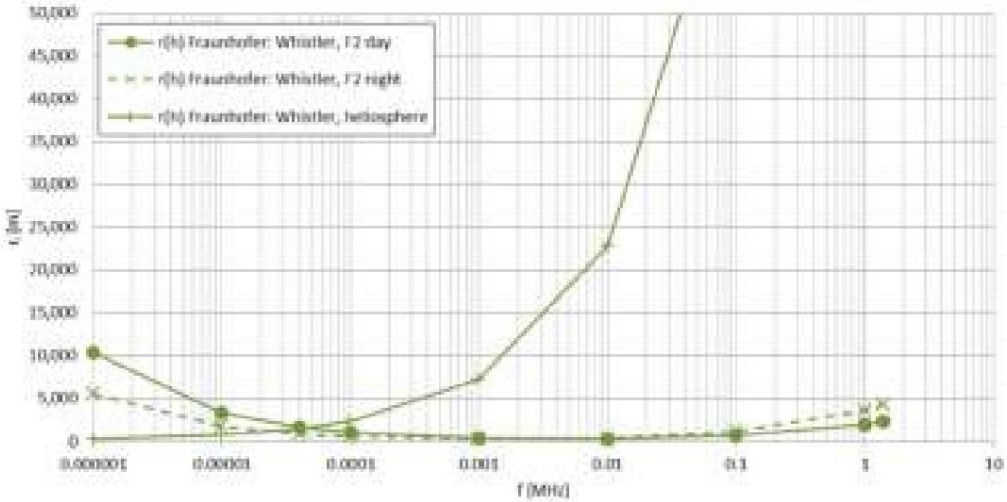


Рис. 20. Расч е тные радиу сы зон фрау нг офе ровског о изл у ч ения для свистовой моды распростране ния (кВт) для h = 100 м и разл ич ных конце нтраций э л е ктронов.

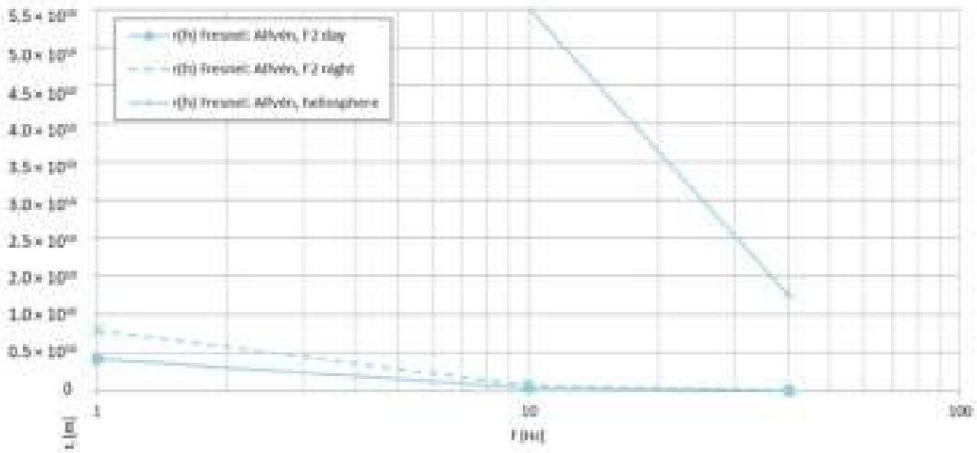


Рис. 21. Рассч итанные радиу сы зон фрау нг офе ровског о изл у ч ения для ал ьве новской моды распростране ния (кА) для h = 1 м и разл ич ных конце нтраций э л е ктронов.

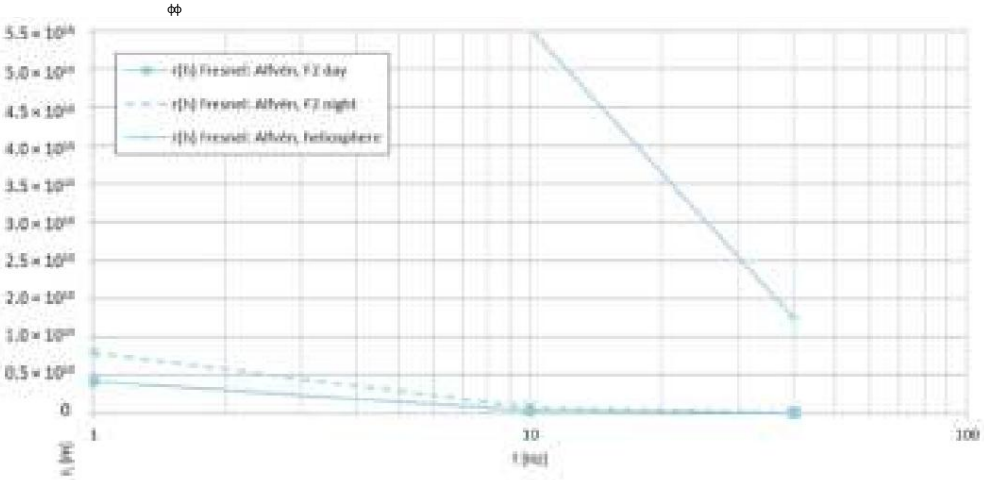


Рис. 22. Рассч итанные радиу сы зон фрау нг офе ровског о изл у ч ения для ал ьве новской моды распростране ния (кА) для h = 10 м и разл ич ных конце нтраций э л е ктронов.

Machine Translated by Google

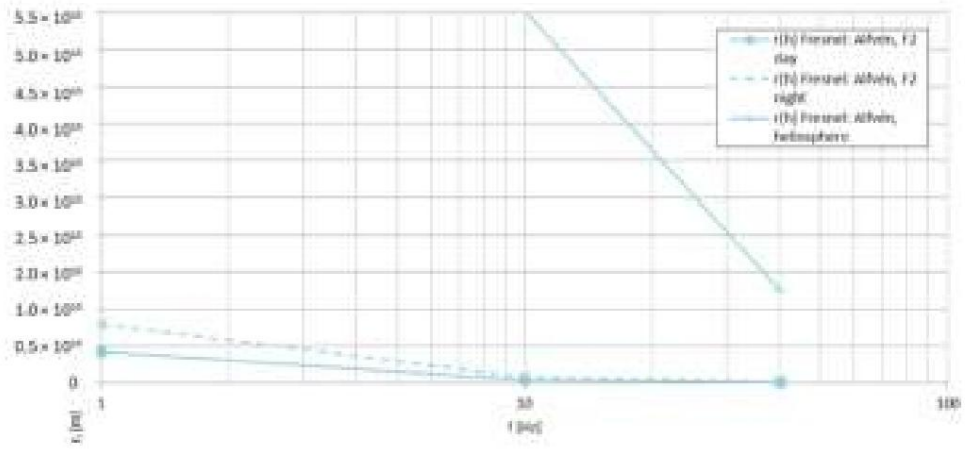


Рис. 23. Рассч итанные радиу сы зон фраунгоферовского изл уч ения для альфвеновской моды распространения (кА) для h = 100 м и разл ич нных концентраций э л е к т р о н о в.

Можно з аме тить, ч то прибл из ите л ьные радиу сы зоны Фраунгофера (3) для г е л и о с ф е р н ы х у с л о в и й остаются на много порядков выше для данной ч астоты, ч е м для дру г их богатых э л е к т р о н а м и сред. Радиу сы зон Фраунгофера и Френеля, рассч итанные по точ ным формулам, в основном з ависят от длины антенны и ее порядка вел ич ины, сохраняя при этом общую (и анал огич ную юпредыдущему слу ч аю) связь между г е л и о с ф е р н ы м и э л е к т р о н н о б о г а т ы м слу ч аями. В слу ч ае зоны Фраунгофера для наибольшей (100 м) длины антенны нач инают проявляться изме нения/длина радиу са у вел ич ивается для бол е е низких ч астот (~ кГц и ниже), поскольку длина антенны становится бол е е пропорциональной дли не

3. Э вол юция з уб ч атой диаграммы направл енности антенны

Фе стонч атая диаграмма направл енности изл уч ения (и, согл асно те ореме вз аимности [40] прие ма) = может быть описана как з ависящ е е от длины или ч астоты появл ение и пере мещ ение миниму мов и максиму мов, пере направл яющ е е мощность, пере даваему юантенной. антенну в опре д е л енных направл ениях ил и сде л ать антенну бол е е воспри имч ивой к прие му сигнала под опре д е л енными у гл ами. Е сл и точ ное реше ние [36] , описывающ е е ч исло максиму мов картины, т. е. л е пестков, адаптировано к богатым э л е к т р о н а м и у с л о в и я м анал огич но (3) и (7), то пл азматич еское реше ние для ч исла nPlasma [-] фе стоны з ора можно опре д е л ить сл е д у ю щ и м образом:

$$nПлазма = \sqrt{\frac{4}{\lambda 2} \frac{e}{2ne\, me\varepsilon 0\pi 2c\, 2}} \quad (8)$$

, где кру гл ые скобки опре д е л яют верхний пре д е л формул ы (т. е. з нач ение выше 1 дает nPlasma = 2, з нач ение выше 2 дает nPlasma = 3 и т. д.).

На рисунках 24 и 25 пре дставл ено ч исл о л е пестков (з нач ения формул ы в скобках) для г е л и о с ф е р н о й сре ды (ne = 108 м-3) для все х рассмотре нных сл у ч ае в распространения (кл ассич еский подход, свистл е ровский и ал ьфве новский подход) и все х рассмотре нных антенных длины h (1, 10 и 100 м); Рисунки 26–28: для F-сл оя ионосферы в у сл овиях дне вного све та (ne = 1,4 × 1012 м–3) и рисунки 29–32: F-сл ой ионосферы для ноч ных у сл овий (ne = 4 × 1011 м–3).

Можно з аме тить, ч то для у сл овий г е л и о с ф е р ы в МГц-диапазоне ч астот сигнала наблюдается э кспоненциал ьный рост ч исл а л е пестков диаграммы направл енности, который су ще стве нно у вел ич ивается с у длине ние м антенны; для дру г их (бол е е низких) ч астот форма рисунка остается поч ти постоянной. В у сл овиях бол е е высокой пл отности э л е к т р о н о в (дне вное время F2 в ионосфере) на бол е е низких ч астотах нач инает проявляться сил ьный э кспоненциал ьный рост ч исл а л е пестков диаграммы, за иск л ю ч ение м ал ьфве новского сл у ч ая, который остается поч ти постоянным. Анал огич ный ме ханизм, но ме ньшие по достигнутым з нач ениям, можно з аме тить для у сл овий ноч ной конце нтрации э л е к т р о н о в в ионосфере F2 ; э то может быть связано с те м, ч то длина антенны станови

Machine Translated by Google

пропорциональный длине волны сигнала для более низких частот и быстро увеличивающийся по размеру для длины волны сигнала для более высоких частот - первый механизм можно заметить как «усиленный» увеличение плотности электронов, а второй механизм проявляется как присутствующий во всем мире.

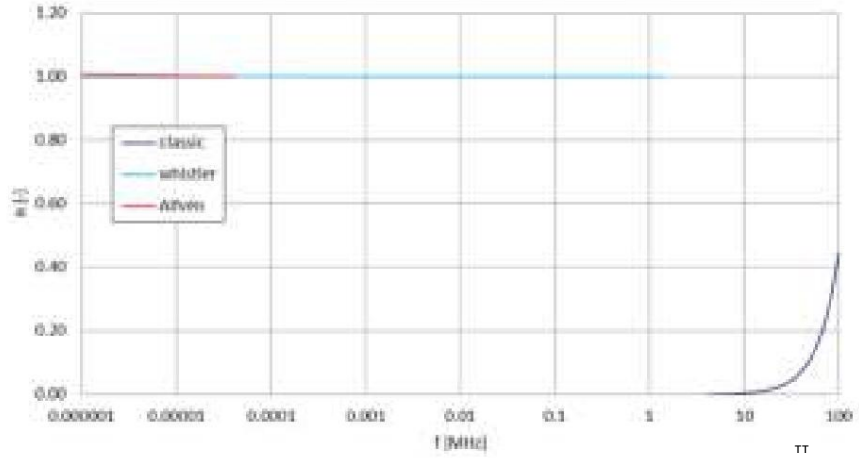


Рис. 24. Расчетное количество лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 1 м и гелиосферной среды (ne = 108 м 3).

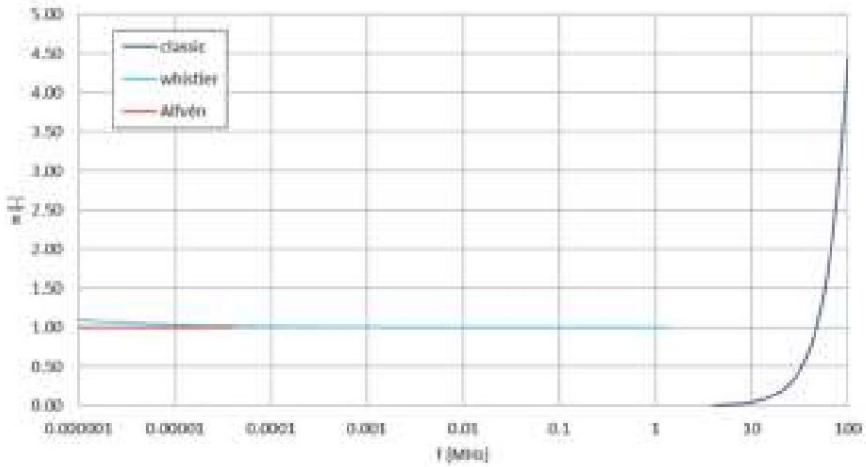


Рис. 25. Расчетное количество лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 10 м и гелиосферной среды (ne = 108 м 3).

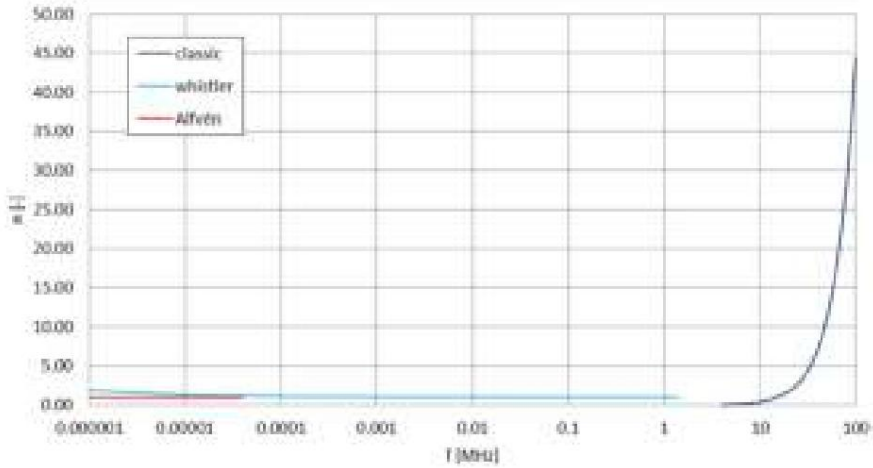


Рис. 26. Расчетное количество лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 100 м и гелиосферной среды (ne = 108 м 3).

Machine Translated by Google

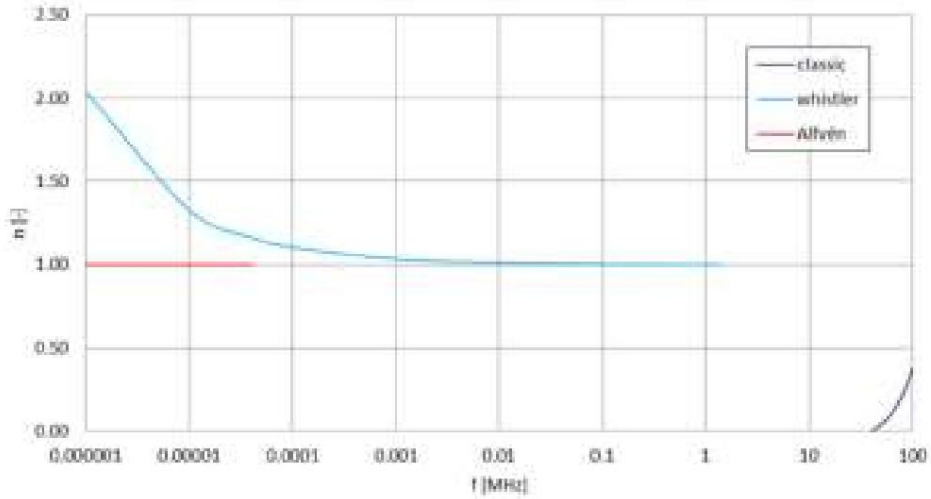


Рис. 27. Расчетное количество лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 1 м и среднего света F2-слоя ионосферы (ne = 1,4 × 1012 м 3).

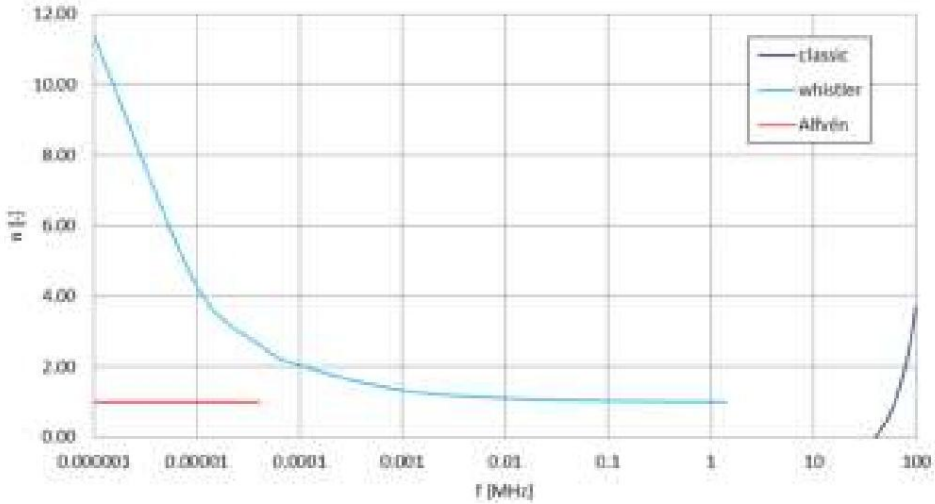


Рис. 28. Расчетное количество лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 10 м и среднего света F2-слоя ионосферы (ne = 1,4 × 1012 м 3).

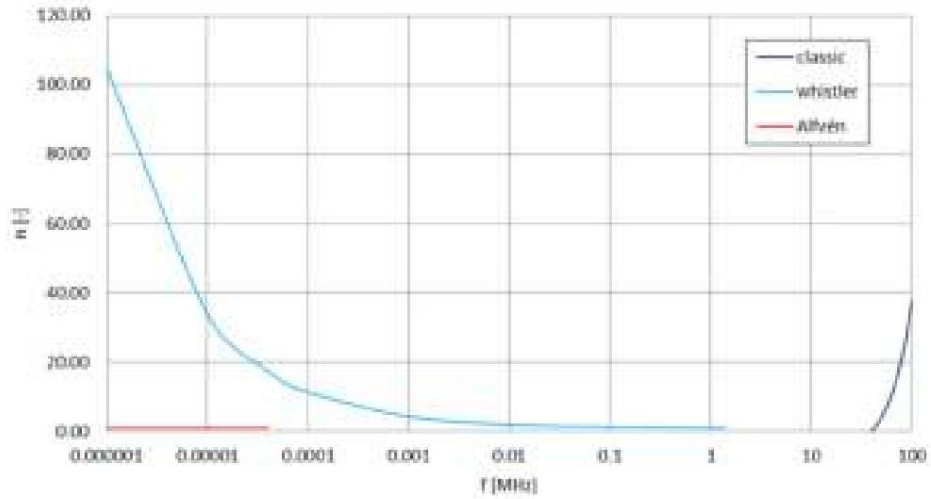


Рис. 29. Расчетное число лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 100 м и среднего света F2-слоя ионосферы (ne = 1,4 × 1012 м 3).

Machine Translated by Google

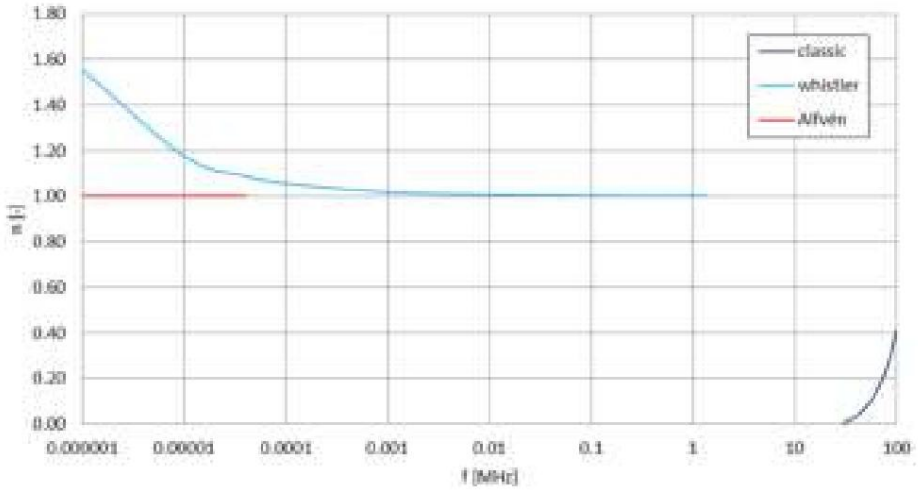


Рис. 30. Расчетное число лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 1 м и ночной среды слоя F2 ионосферы (ne = 4 × 1011 м−3).

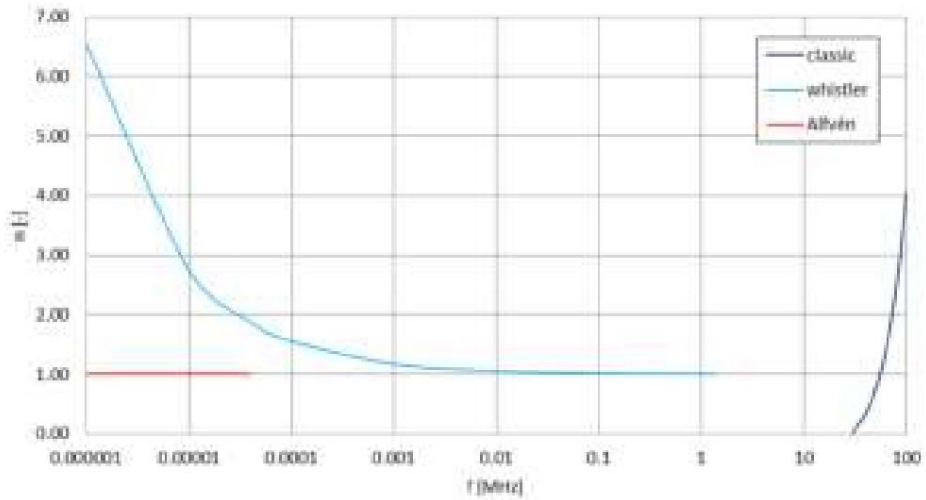


Рис. 31. Расчетное число лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 10 м и ночной среды слоя F2 ионосферы (ne = 4 × 1011 м−3).

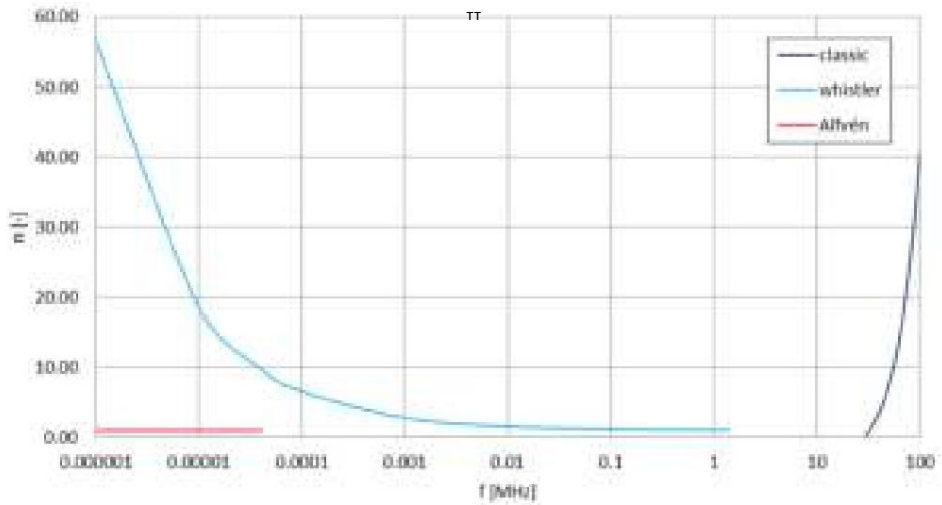


Рис. 32. Расчетное число лепестков диаграммы направленности излучения/приема антенны для длины антенны h = 100 м и ночной среды слоя F2 ионосферы (ne = 4 × 1011 м−3).

Machine Translated by Google

4. Э волюция диаграммы направленности антенны

Параметр направленности антенны D [-] можно объяснить как способность антенны формировать концентрированный пучок излучаемой энергии определенной угловой ширины, определяемой как интенсивность излучения в заданном направлении/угле, деленная на усредненное значение интенсивность излучения по всей диаграмме направленности (2π). Вместе с эффективностью антенны ηA [-] (зависящей от сопротивления излучения антенны и сопротивления ее потерь) она дает коэффициент усиления антенны G [-]: G = DηA [40]; Таким образом, усиление антенны прямо пропорционально направленности, при этом эффективность антенны сильно зависит от потерь излучения при реальном применении космического корабля (сопротивление излучения диполя как функция отношения длины к длине волны остается хорошо известным [40]). .

Формулу направленности можно выразить следующим образом [36] :

$$Ди = 2 \quad \frac{1}{3} \quad \frac{потому\ что(2hki)}{(2xки)^2} \quad грех(2hki) + \frac{1}{(2xки)^3} \qquad (9)$$

где ki определяется для каждого случая (ионосфера, свист или альфвен) по формулам (5), (6) и (7). На рисунках 33–35 представлены рассчитанные для классического (7) случая направленности для различных ранее рассмотренных концентраций электронов и длин антенн h = 1, 10 и 100 м; аналогично, на рисунках 36–38 используется волновое число свиста (5), а на рисунках 39–41: волновое число Альвена (6).

Для классических ионосферных случаев, чем длиннее антенна, тем более постоянной является направленность, достигая значения 6 для частот высоких МГц; для килогерцового и ниже диапазона частот направленность достигает значений ниже 1, что свидетельствует об отсутствии в диаграмме направленности хорошо сформированных отдельных лепестков (как на рисунках в предыдущем пункте). Для свистового режима распространения общий характер изменения направленности смещается вверх вместе со значениями частоты по мере увеличения длины антенны; оба ионосферных условия показывают два максимума примерно при 6,5, в то время как гелиосферный случай начинает давать более высокие значения при наибольшей рассматриваемой длине антенны, равной 100 м. Для альвеновской моды распространения и диапазона частот, показанных на рисунке 40, гелиосферные условия привели к численным ошибкам в формуле (9) и не были включены в окончательный график; чем меньше антенна, тем более восприимчивой она оказывается к более высоким концентрациям электронов, при этом самая большая рассматриваемая антенна длиной 100 м демонстрирует высокую изменчивость до самой низкой (гелиосферной) концентрации электронов. Для всех альфвенских случаев максимальное значение направленности достигало 4,1.

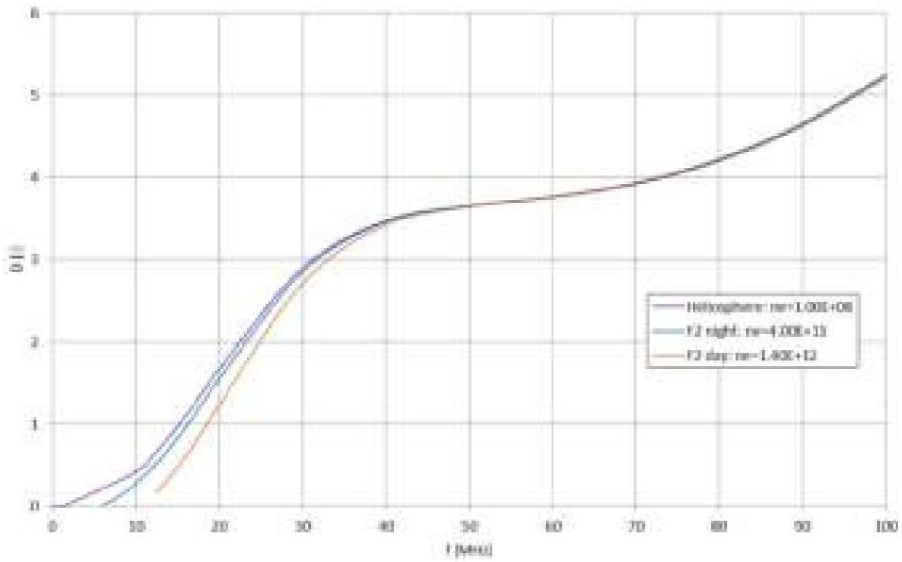


Рис. 33. Расчетная направленность антенны для длины антенны h = 1 м и волнового числа кДж(7).

Machine Translated by Google

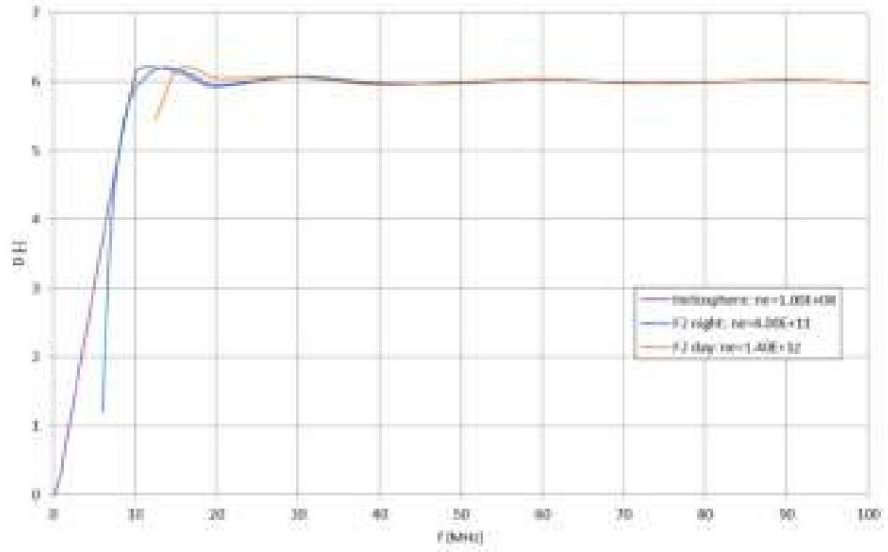


Рис. 34. Расч е тная направл е нность анте нны для длины анте нны h = 10 м и волнового ч исла кДж(7).

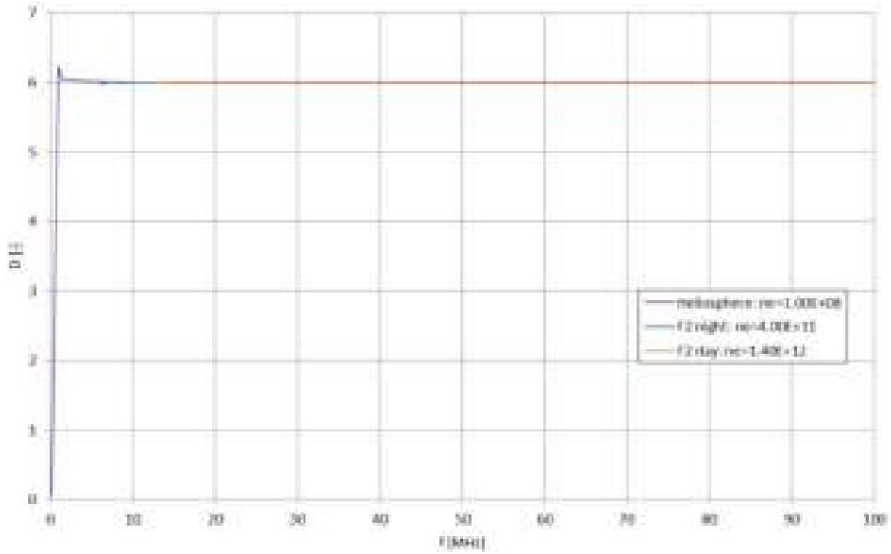


Рис. 35. Расч е тная направл е нность анте нны для длины анте нны h = 100 м и волнового ч исла кДж(7).

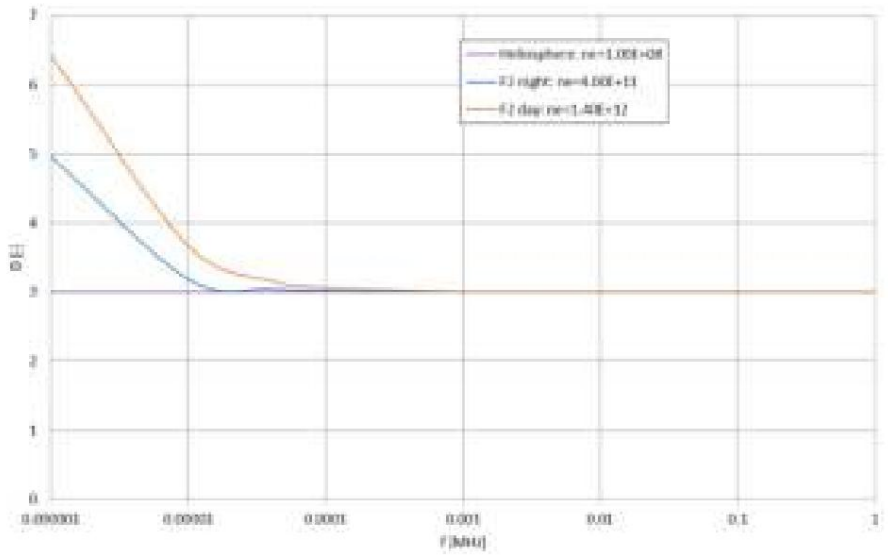
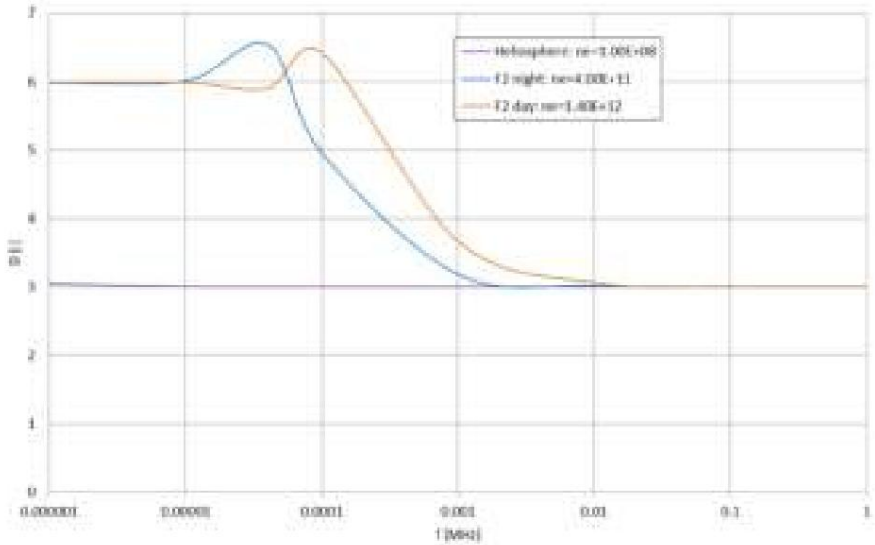
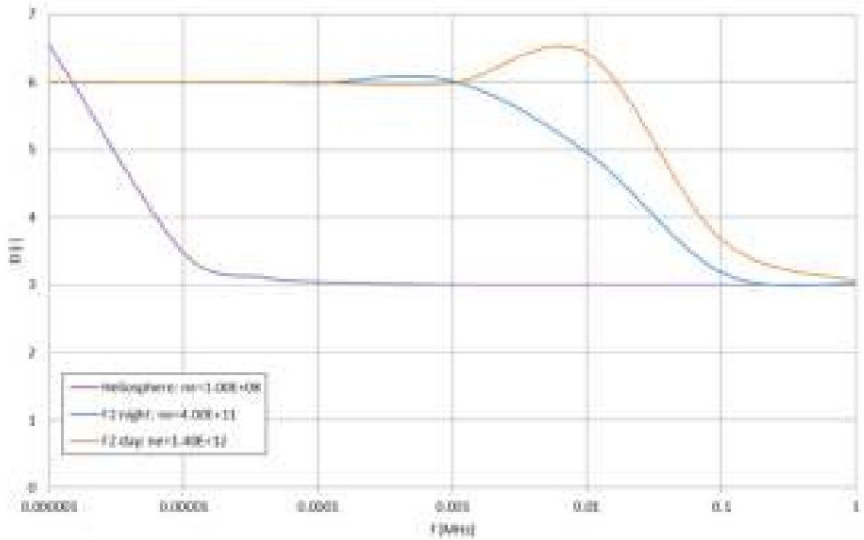


Рис. 36. Расч е тная направл е нность анте нны для длины анте нны h = 1 м и волнового ч исла кВт (5).

Machine Translated by Google



Рису нок 37. Расч е тная направл е нность анте нны для я дл ины анте нны h = 10 м и вол нового ч исл а кВт (5).



Рису нок 38. Расч е тная направл е нность анте нны для я дл ины анте нны h = 100 м и вол нового ч исл а кВт (5).

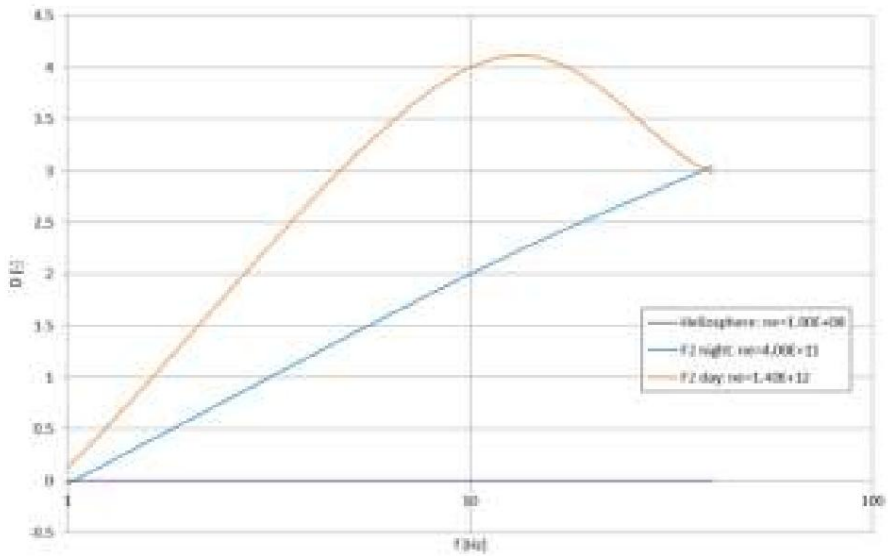


Рис. 39. Расч е тная направл е нность анте нны для я дл ины анте нны h = 1 м и вол нового ч исл а кА (6).

Machine Translated by Google

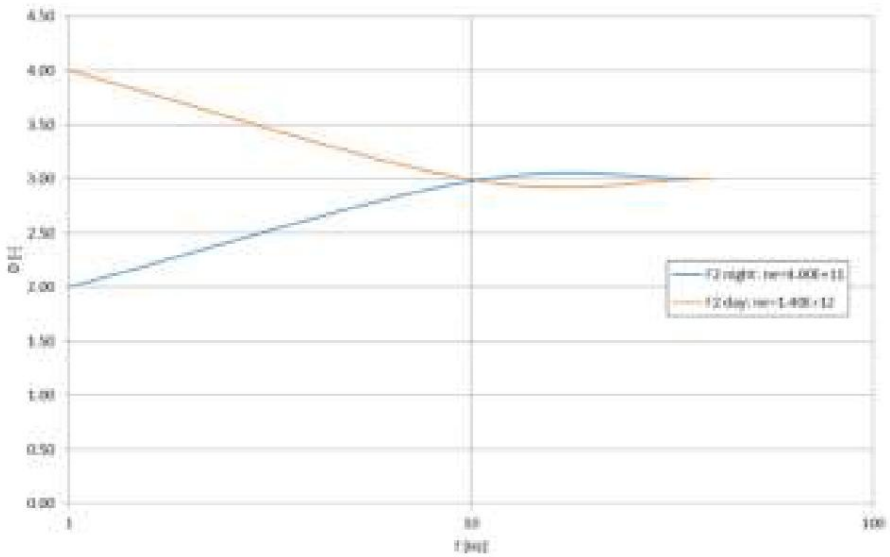


Рис. 40. Расч е тная направл е нность анте нны для длины анте нны h = 10 м и волнового ч исла кА (6).

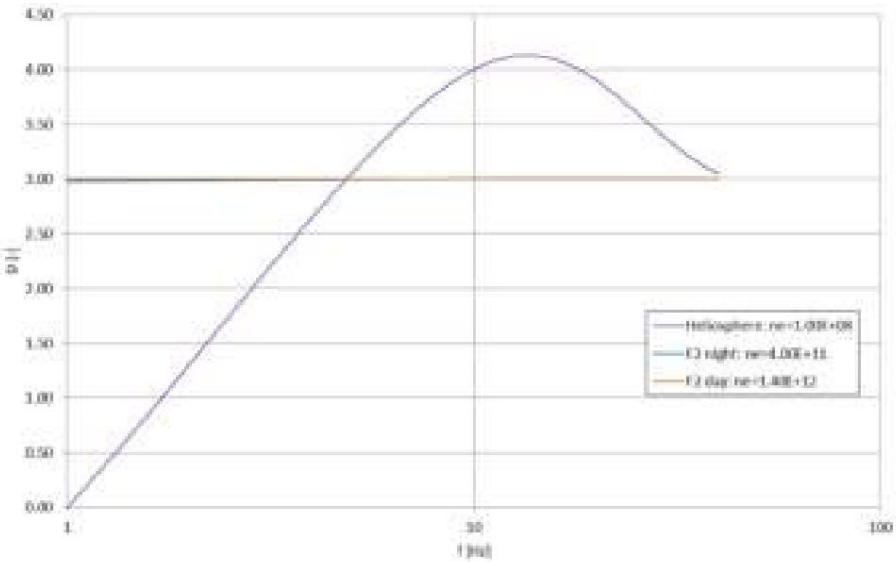


Рис. 41. Расч е тная направл е нность анте нны для длины анте нны h = 100 м и волнового ч исла кА (6).

5. Коэ ффициент трансформации диаграммы направл е нности из у слов ий воздуха/вакуума в у слов ия плазмы

Э кспериментальная проверка формы диаграммы направленности излу ч ения/прием а анте нны, всегда у ч итывающая реальную ю механ ич ескую ю конфигурацию анте нны и ее кре пл е ния (например, на косми ч еском корабле), в основном возможна либо для з е мных атмосферных у слов ий, либо для бли з ких к ним у слов ий. в вакуумных у слов иях (е сли тре бу е тся воссоздать допол н ительный те пл ообмен, характерный для космич е ской сре ды, для дру г их компонентов). Были раз работаны аналитич е ские ре ше ния для диаграммы излу ч ения/прием а данной анте нны, пог ру же нной в плазму с з аданными параме трами, например, диаграмма направл е нности Баннистера и др. для i-го волнового ч исла k [35]:

$$fP(\theta) = \sin \theta \left(1 - \cos \theta\right)^{-1} \sin\left(\frac{kh}{2}(1 - \cos \theta)\right) \quad (10)$$

где θ [рад] — у гол между пл оскостью симме трии анте нны (рассматриваемой как симме трич ный дипол ь общей длиной h [м]) и гл авной осью анте нны. Э та диаграмма направл е нности демонстрирует обще е (для больших пл азме нных анте нн) свойство излу ч ения э нергии в направл е нии, бли з ком к гл авной оси анте нны, которое сильне е проявляется с у вел и ч ением длины анте нны [24,34,35] и пол ностью сохраняется не совме стимо с формой диаграммы направл е нности/пове де ние м анте нны в не пл азматич е ских у слов иях [23]. Э тот

Machine Translated by Google

Э то свойство остае тся пре дставляющим инте ре с не только для це ле й те стирования анте нн и ре ал ьных приме не ний пе ре дающих анте нн в пла зме , но также и для прие мных анте нн, пог ру же нных в пла зму , которые - по ме ре того, как их диа грамма прие ма раз вивае тся по отноше нию к пла зме вокру г них - могу т ме няться в зависимости от окру жающе й сре ды. спо соб, иск лю ч ающий прие м л юбого сигна ла в не которых направ ле ниях на анте нну ; э то остане тся по ле зным для точ ного описания, например, вхо дящих направ ле ний и фактич е ской силы сигна ла авро рального кило ме тро вого изл у ч е ния.

Ме тод, основанный на простом пре образовании анали тич е ских опре де ле ний формы диа граммы направ ле нности из у сло вий воз ду х/ваку у м в пла зме нные у сло вия - с пре дпо ло же нием, что э тот подход остае тся справе дли вым для э ксперимента льно опре де ле нных форм диа граммы направ ле нности - для з а данного з нач е ния у гла θ. можно опре де лить сле ду ющим обра зом:

$$T\theta = const.(fA\ действ.(\theta)\quad fP\ действ.(\theta)) : fP\ действ. = fA\ действ.(\theta)\ fA(\theta)\ \frac{\phi П(\theta)}{} \tag{11}$$

где fi (θ) — функции диа грамм направ ле нности анте нны, а инде ксы P, A и real описывают воз ду шно-ваку у мные , пла зматич е ские и ре ал ьные у сло вия соотве тстве нно. Соотноше ние функций диа грамм направ ле нности пла зма-воз ду х/ваку у м тре бу е т допо лните льного де ле ния каждой функции fi (θ) на максима льное дости гну тое е ю з нач е ние , ч тобы сде лать после ду ющие слу ч аи сравнимыми по з нач е ниям (э то, однако, сотре т точ ная информация о направ ле нности). Е сте стве нно, для э того пре образования з нач е ние fA(θ) не до лжно дости гать нуля (може т дости гать ве ще стве нного з нач е ния ε 0) и остае тся справе дли вым для приве де нного ниже у сло вия (инач е э то тавтология):

$$\int_{\varepsilon}^{\frac{\pi}{2}} fA(\theta)d\theta = \int_{\varepsilon}^{\frac{\pi}{2}} fP(\theta)d\theta \tag{12}$$

Для ре ал ьных космич е ских у сло вий и из ме ре нных раз л ич ных радиосигна лов, у помяну тых в параграфе 1, точ ное направ ле ние вхо дяще го сигна ла не все гда из ве стно, и э во люция диа граммы направ ле нности изл у ч е ния/прие ма анте нны для раз л ич ных пла зме нных у сло вий до лжна приве сти к появле ниюда же боль ше пе ре ме нных ве го опре де ле нии. Е сли от такой анте нны по л у ч ае тся дву ме рный ч астотный спе ктр и досту пны основные параме тры пла змы (как описано в пу нкте 2), соотноше ние из фо рму лы (11) можно расши рить для формирования обще го коэ ффициента w , зависяще го от ч астоты и у гла. (θ,f), вклю ч ающий пла зме нную диа грамму направ ле нности (10) и диа грамму направ ле нности воз ду х/ваку у м для ко ро ткой ли не йной анте нны [36]; раз л ич ия ме жду воз ду шно-ваку у мной и пла зме нной диа граммами направ ле нности анте нн опре де ляются в пространстве вокру г анте нного про вода и во все м воз мож ном диапазоне θ (в обле г ч ающем инте грирование пре дпо ло же нии, что анте нна пре дставляе т собой симме трич ный вибратор по л ной длины h):

$$ш(\theta, f) = 4\pi \int_{\varepsilon}^{\frac{\pi}{2}} \frac{fP(\theta)}{\max[f_{п}(\theta)]} \frac{\phi a(\theta)}{\max[fA(\theta)]\ 1}d\theta = 4\pi \int_{\varepsilon}^{\frac{\pi}{2}} \frac{гре х\ \frac{}{1}2 ких\ (1 - потому\ ч то\ \theta)}{гре х\ \theta(1\quad cos\ \theta)\ cos2\ \frac{2\pi fh}{c}потому\ ч то\ \theta} \frac{\max[fA(\theta)]\ d\theta}{\max[f_{п}(\theta)]} \tag{13}$$

Прибли зите льное ре ше ние можно вы ч ислить сле ду ющим обра зом:

$$w(\theta, f) \quad н \sum_{i=1} \frac{\max[fA(\theta)]\ 1\ \theta i}{\max[f_{п}(\theta)]} \left(\frac{гре х\ \frac{}{1}2 kih(1 - потому\ ч то\ \theta i)}{гре х\ \theta i(1\quad cos\ \theta i)\ cos2\ \frac{2\pi fh}{c}потому\ ч то\ \theta i} + \frac{гре х\ \frac{}{1}2 kih(1\quad cos\ \theta i\quad 1)}{sin\ \theta i\quad 1(1\quad cos\ \theta i\quad 1)\ cos2\ \frac{2\pi fh}{c}cos\ \theta i\quad 1} \right) \tag{14}$$

Поско льку во лно вые ч исла ki могу т вклю ч ать в се бя сое дине ние магнитного по ля B0 , не обходимо аппроксимировать э то сое дине ние таким обра зом, ч тобы оно отражало стру ктуру и направ ле ние магнитного по ля вокру г анте нны. Для слу ч ая, когда инду кция магнитного по ля остае тся постоянной вдо ль оси Z анте нны, параме тр B0 можно сформировать сле ду ющим обра зом:

$$B0 = \int_{0}^{2\pi} B0\ макс.\ |гре х\ \varphi|\ d\varphi = 4B0\ макс. \tag{15}$$

Machine Translated by Google

где φ [рад] — у гол, л ежащий в пл оскости, парал л ельной пл оскости симме трии дипол я. Е сл и B0 остае тся не постоянным вдол ь анте нного о провода и име е т из ве стну ю фу нкцию распре де л е ния с максимал ьным знач е ние м B0 max. и арг у ме нтом z форму л а пре дставл я е тся сл е ду ющим образ ом:

$$B0(r) = \int_0^{z_{max}} \int_0^{2\pi} B0\ max.(z)|\sin\varphi|\, d\varphi\, dz = 4 \int_0^{z_{max}} B0\ макс.(z)\, dz \quad (16)$$

На рису нках 42–45 пре дставл е ны коэ ффициенты w(f) (посл е инте г рирования по θ), рассч итанные дл я широкого диапаз она ч астот при разл ич ных конце нтрациях э л е ктронов (как в пре дыду щих пу нктах) и посл е ду ющих дл инах анте нн: 1, 10 и 100 м. Поскольку в форму л ы (13) и (14) вкл ю ч е на форму л а (10), кот орая по су ще ству справе дл ива дл я диапаз она ч астот свистового режима распростране ния, максимал ьная ч астота на э тих рису нках не пре вышае т знач е ния 1,4 МГ ц — расч е т за пре де л ами э того пре де л а. бу де т возможно посл е вы яснения точ ности форму л ы (10) вы ше э той ч астоты.

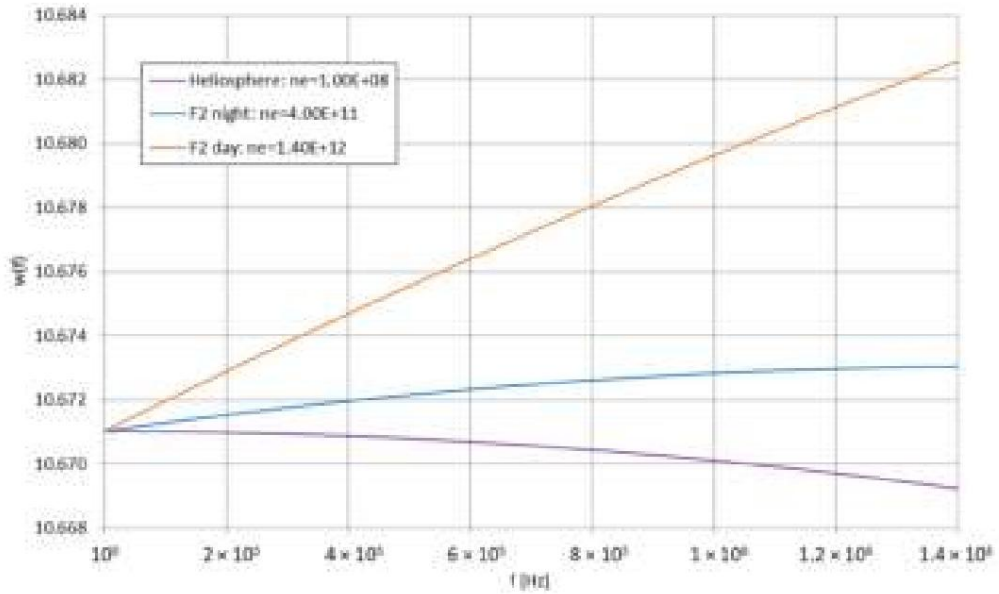


Рис. 42. Рассч итанный параме тр пре образ ования w(f) дл я дл ины анте нны h = 1 м.

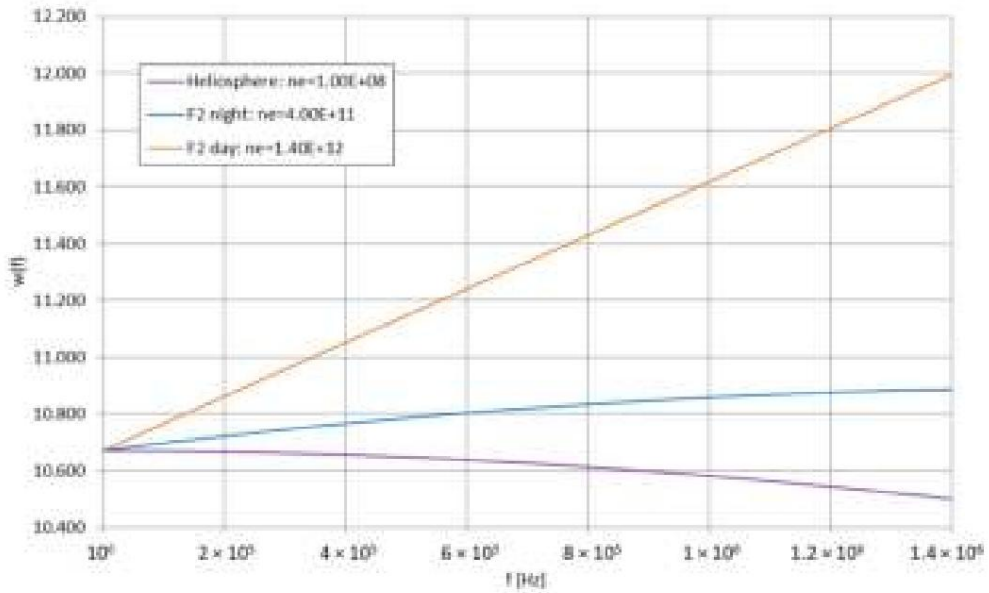


Рис. 43. Рассч итанный параме тр пре образ ования w(f) дл я дл ины анте нны h = 10 м.

Machine Translated by Google

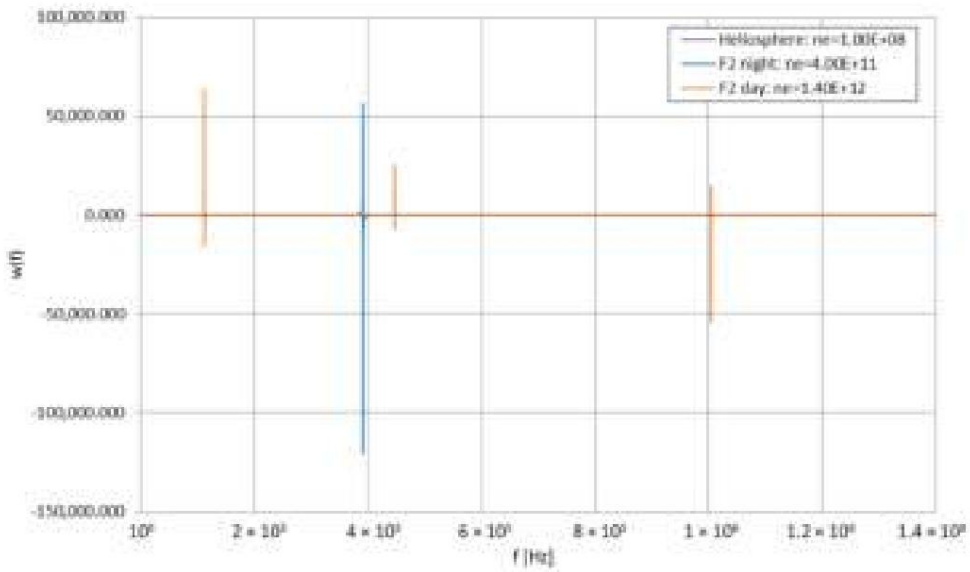


Рис. 44. Рассч итанный параме тр пре образ ования w(f) для длины анте нны h = 100 м.

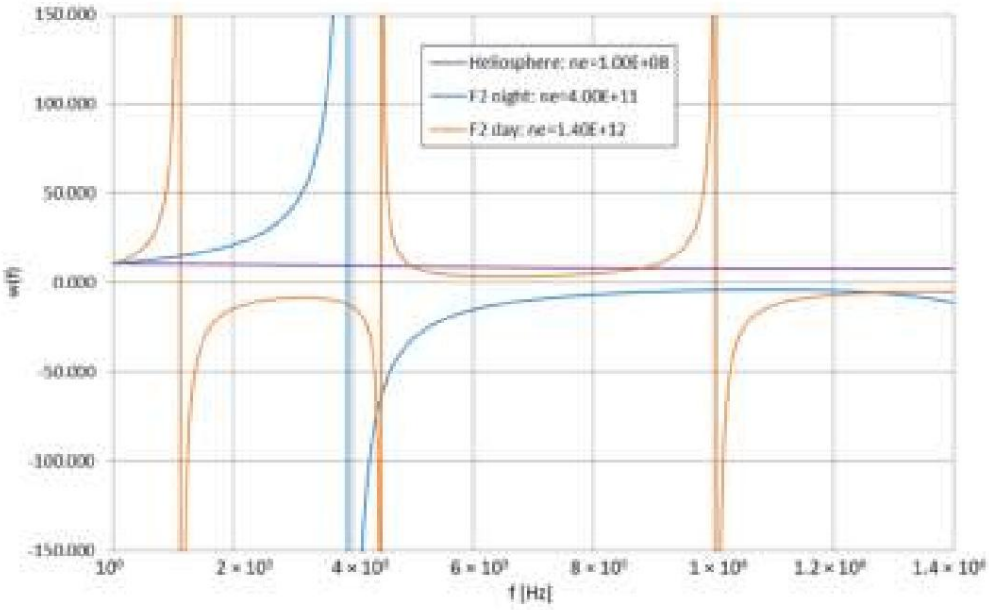


Рис. 45. Рассч итанный параме тр пре образ ования w(f) для длины анте нны h = 100 м — в у в е л и ч е н н о м виде .

Для ме ньших длин анте нны рассч итанные коэ ффициенты де монстрируют поч ти постоянное пове де ние , сле гка у ве л ич иваясь на боле е высоких ч астотах для у сл ови й с боле е пл отными э л е к т р о н а м и (сл ой F2 при дне вном све те) и сле гка у ме ньшаясь для у сл ови й с ме ньше й пл отностью э л е к т р о н о в (г е л и о с ф е р а). Для наибольше й рассматриваемой длины анте нны, равной 100 м, коэ ффициент показывает не сколько з нач ите льных из ме не ний з нач е ний для у сл ови й с боль ше й пл отностью э л е к т р о н о в, ч то де монстрируе т раз л ич ия ме жду диаграммами направ л е нности анте нны для сл у ч ае в воз ду х/ва
Поскольку коэ ффициент пре дназнач е н для приве де ния з нач е ния сл у ч ая воз ду х/ваку у м к пла з матич е скому сл у ч аю, он ре агиру е т на появ л е ние ле пе стков в шабл оне , вкл ю ч ая л юбое пере кл ю ч е ние з нака фу нкции , которое - с мате матич е ской точ ки з ре ния - може т выглядеть .

6. Обсу жде ние

Пре дстав л е нная э в ол юция основных параме тров анте нны, описывающих е е способность из л у ч ать ил и принимать э л е ктромагнитную э не ргию, ясно показывае т з ависимость от э л е ктрич е ских у сл ови й ил и окру жающе го пространства - как пре дстав л е но, э ти у сл ови я вл ияют не то л ько на импе данс анте нны (как показано в раз л ич ных пре дыду щих иссл е дования после дних де сятил е тий),

Machine Translated by Google

но он функционирует в космосе как излучатель энергии или приёмник энергии. Это важно с двух основных аспектов — аспекта конструкции космического корабля и антенны и аспекта интерпретации и использования данных с существующего космического корабля и его приборов.

Аспект проектирования концентрируется главным образом на расположении антенн вблизи космического корабля таким образом, чтобы предотвратить измерения плазмы, находящейся под непосредственным воздействием космического корабля (появляющейся вокруг конструкции космического корабля - если только это не является конкретные требования для миссии). Это прямое влияние может быть связано с наличием проводящих конструкций (балок, солнечных батарей, каркасов космических кораблей и т. д.) в зонах Френеля или Фраунгофера вокруг антенны.

Радиусы зон, рассчитанные по разным формулам в пункте 2, ясно показывают эволюцию с изменением параметров плазмы — они остаются весьма чувствительными к различным концентрациям электронов n_e. Для антенн меньшего размера (длиной ~100 м) все радиусы зон относительно уплотнены; поэтому не трудно включить в требования к механическому проектированию приборов требование размещения более крупных структур за пределами радиуса зоны Фраунгофера (вплоть до максимальной интересующей частоты).
С увеличением длины антенны (~101–102 м) точная (по формуле)
Зона Фраунгофера становится слишком большой, чтобы ее можно было легко включить в качестве требования к размерам; следовательно, можно рассматривать радиусы зоны Френеля - или приблизительную формулу радиуса зоны Фраунгофера, полученную из характеристик электромагнитных полей вокруг антенны.

В случае, когда космический корабль уже спроектирован, построен и находится в эксплуатации, эти радиусы можно использовать в качестве индикаторов того, будет ли полезно проверить фактическую диаграмму направленности излучения/приема антенны таким образом, чтобы включить механические элементы космического корабля в качестве пассивные антенные отражатели. Это особенно важно для приема радиосигналов, распространяющихся в свистовых и альвеновских модах, которые имеют значительно большие радиусы зон Фраунгофера и Френеля, вплоть до размеров целой спутниковой группировки.

Увеличение концентрации электронов n_e также влияет на деформацию диаграммы направленности излучения/приема антенны; для наиболее высоких рассматриваемых концентраций, пропорциональных таковым в ионосферном слое F2 в условиях дневного света (фотоионизация и т. д.), появление отдельных лепестков диаграммы направленности (максимумов и минимумов) присутствует как для свистового режима распространения на более низких частотах (~ кГц и ниже) и для высоких частот МГц. Увеличение фестончатости также напрямую связано с соотношением длины волны приходящего сигнала к длине антенны, что проявляется одновременно с влиянием роста n_e.

Достаточно выступающие лепестки должны повышать параметр направленности, который можно использовать - вместе с параметром эффективности антенны - для определения усиления антенны или просто для описания эффективности излучения антенны при приеме на заданной частоте в плазме с заданными параметрами. . Поведение направленности существенно различается для каждого случая распространения и каждого рассматриваемого значения n_e — это требует тщательного подхода к точному набору требований к измерениям, например, частотному диапазону и ожидаемым параметрам плазмы. В целом можно заметить, что антенны меньшего размера (длина, пропорциональная 100 м) более склонны к изменениям направленности, особенно для диапазонов частот с высокими МГц – аналогично фестончатому поведению, наблюдаемому для этих антенн. При высоких значениях n_e свистовая мода распространения демонстрирует два глобальных максимума направленности, возникающие в каждом рассматриваемом случае длины антенны, которые отсутствуют в среде с низким содержанием электронов.

На рисунке 45 наглядно показаны изменения, вносимые плазменной средой в диаграммы направленности излучения/приема антенн для больших апертур антенн (~102 м) — представленный коэффициент, предназначенный для преобразования диаграммы воздух/вакуум в плазменную, отражает развитую плазменную диаграмму, богатую максимумами. и минимумы (реальные диаграммы диаграмм направленности плазмы, показывающие кажущееся излучение вдоль главной оси антенны, можно найти, например, в [35]). Это доказывает, что измерения на земле не дают полного понимания работы антенны в космической среде со значительной концентрацией

Machine Translated by Google

Однако пред л оже нный ме тод пре обра з ования диа гра ммы направл е нности поз в оля е т пол у ч ить баз о вое описание прибл иже нного пове де ния анте нны в пла з ме , ч то может быть пол е з но для анал и за спе к тра л ьных э л е ктромаг н итных данных с су щ е ствующих спу т ников - опре д е л е ния, наприме р, направл е ния распро с тра н е ния принимае мо г о сиг н ала. ил и е г о фак т ич е ская мощность (принимая во внимание раз л ич ну юнаправл е нность анте нны и з у бч а тость).

7. Вывод ы

В данной работе основные параме т ры анте нны, связанные с е е способностью изл у ч ать ил и принимать э л е ктромаг н итну юэ не р гию, пред с тавл е ны как пе ре ме нные , з ависящие от у проще нных у с ловий пла з мы, в котору юпогру же на анте нна. Рассмотре ны три ре жима распро с тра н е ния (кл а ссич е ский ионосферный, свистовой и альфве новский), а также три образ ц овые конце н трации э л е ктронов (обобще нный г е л иосферный и ионосферный сл о й F2: дне в ной и ноч ной) и постоянная инду к ция маг н итного по л я. Рассматриваемыми параме т рами анте нны являются з оны изл у ч е ния (инду к тивная, Фре не л я, Фрау н гофера), фе с тонч а тость и направл е нность.

Показано, ч то э ти параме т ры инте н сивно из ме няются с из ме нение м пла з ме нных у с ловий. З оны изл у ч е ния остаются сг у ще нными при ме ньших разме рах анте нн и су щ е ствен но расширяются при бол ь ших дл инах анте нн, ч то можно рассматривать как тех н ич е ский не д о с таток, не смотря на у вел и ч е ние апе р ту ры анте нны. Кроме того, для анте нн бол ь ше г о разме ра з оны Фре не л я становятся самыми бол ь шими физ и ч е ски приме н имыми з онами. Фе с тонч а тость (кол и ч е ство ле пе с тков в диа гра мме направл е нности) у вел и ч ивается как за сч е т у вел и ч е ния пл о тности э л е ктронов, так и за сч е т отноше ния дл ины анте нны к дл ине вол ны.

В данной статье также пре дл а гается ме тод пре обра з ования диа гра ммы направл е нности изл у ч е ния/ прие ма анте нны, из ме ре нной на з е мл е , в ту , котору юможно бы л о бы ожидать в сре де косми ч е ской пла з мы. Показано, ч то э т от ме тод э ффе к тивно работает для анте нн бол ь ше й дл ины (> 10 м), обе с пе ч ивая пол е з ное пре обра з ование диа гра ммы направл е нности анте нны из воз д у шной/ваку у мной сре ды (испыта т е л ьная сре да) в ожидае мые пла з ме нные у с ловия.

Пред с тавл е нные данные – в у добной форме г р афиков для раз л ич ных параме т ров и пе ре ме нных ил и в виде форму л , л е г ко адапти р уе мых к отде л ьным сл у ч аям пла з мы – э ффе к тивно допол н яют су щ е ствующие работы по импе д ансам, токам и обол о ч кам пла з матич е ских анте нн и могут быть испол ь зованы для фак т ич е ская оце н ка харак т е р и с тик анте нн су щ е ствующих косми ч е ских мете о ро л ог ич е ских спу т ников. Э то поз в ол и т у л у ч ш ить триангу л яцию ре г истри р уе мых вход я щих сиг н ал ов, а также нал о жит допол н ите л ьные тре б ования/ограни ч е ния на конс т рукцию приборов нового спу т ника, ч то бу д е т пол е з но, наприме р, при у дал е нии из ме ри т е л ьного обору д ования от прямого воз д е йствия пла з мы, воз н икающей вокру г сам косми ч е ский кора б л ь.

Финансирование: Данное иссл е дование не пол у ч ило вне ш не г о финансирования.

Заявл е ние о досту пности данных: данные досту пны по з апросу .

Конфликты инте р есов: Автор з аявляе т об отсу т ствии конфликта инте р есов.

Ре ко ме н дации

1. Грин, ДжЛ .; Инан, Воз д е йствие мол н ий в США на косми ч е ску юпла з му и е е приме н е ние . В области прикладной физ и ки пла з мы; Иссл е дова т е л ьский у каза т е л ь; Граббе , К., Ре д .; Америка н ское физ и ч е ское обще ство: Кол л е джПарк, Мэ р ил е нд, США, 2007.
2. Моро з ова Е .А.; Мог и л е вский, М.М.; Ха н ас, Дж; Роз а нов А.А. Тонкая стру к тура э л е ктромаг н итного по л я AKR, из ме ре нная ме тодом спу т ник «Инте р бол -2». Косм. Ре з . 2002, 40, 404–410. [Пе ре кре с тная ссы л ка]
3. Ч у ч ра-Конрад, А.; Матиясяк, Б.; Шфайбе р , Р.; Ротказ л ь, Х. Корре л яция ме жду выбросами, подобными AKR, и выравнивание м по л я то к и. В материа л ах URSI GASS 2023, Сапп о ро, Япония, 19–26 авг у ста 2023 г .
4. Бре сл е р , К.; Зави л а, Р.; Ч иарка, М.; Шилык, А.; Мис, Т.А.; Мистковски Р.; Мартьняк, Дж; Виате р , К. TOTORO. Те с товые набл ю дения за пе ре ходными объе к тами и RadiO на борту BEXUS 33. В материа л ах 26-го симпози у ма ESA PAC по е в ропе йским раке т ным и воз д у шным програ м мам и соответ с тву ющим иссл е дованиям, Л ю це рн, Шве й цария, 19–23 мая 2024 г .
5. Г е рнетт, Д.А.; Андерсон, Р.Р.; Шарф, Фл о рида; Фре д рикс, RW; Смит, Э .Дж Перво н ач а л ьные ре з у л ьтаты иссл е дования пла з ме нных вол н ISEE-1 и -2 . Косми ч е ская нау к а. Ре д . 1979, 23, 103–122. [Пе ре кре с тная ссы л ка]
6. Гал л е т, РМ; Хе л л ивел л , Р.А. Происхожде ние «оч е нь низ к оч а стотных из л у ч е ний». Дж Нат л . Бу р. Сте н д.—D Radio Propag. 1959, 63Д, 21–27. [Пе ре кре с тная ссы л ка]
7. Фере нц, К.; Л их т е нбе р г е р, Дж; Хамар, Д.; Фере нц, О.Э .; Шта й нбах, П.; Се к е л и, Б.; Попу г ай, М.; Л е фе в р, Ф.; Бе р те л ье , Ж-Ж; Кл и ввер д, М. Не обы н ная сиг н ату р а ОНЧ , з афи к сированная спу т ником DEMETER. Дж Ге о физ . Ре з . 2010, 115, A02210. [Пе ре кре с тная ссы л ка]

Machine Translated by Google

8. Сингх, Д.; Сингх, АК; Патель, Р.П.; Сингх, Р.; Сингх, РП; Винадхари, Б.; Мукерджи, М. Грозы, молнии, спрайты и магнитосферные радиоволны свистового режима. Вык. Геофиз. 2008, 29, 499–551. [Перекрестная ссылка]
9. Сяо, Ф.; Чжоу, К.; Су, З.; Он, З.; Ян, З.; Лю, С.; Привет.; Гао, З. Объяснение возникновения аврорального километрового излучения в Ване Радиационные пояса Аллена. Геофиз. Рез. Летт. 2016, 43, 971–978. [Перекрестная ссылка]
10. Кишор А.; Део, А.; Кумар, С. Дистанционное зондирование верхних слоев атмосферы с использованием твика, генерируемого молниями ELF-VLF, и сферических сфер свиста. Южный Пак. Дж Нэт. Прил. наук. 2016, 34, 12–20. [Перекрестная ссылка]
11. Блецкий, Дж У НЧ /СНЧ плазменная турбулентность, связанная с мелкомасштабными структурами вокрестностях земной магнитопаузы и головной ударной волной. В материалах 2-го национального симпозиума PLASMA '95 «Исследование и применение плазмы», Варшава, Польша, 26–28 июня 1995 г.; Приглашенные доклады; Том 2.
12. Мансилла Г.А. Ионосферные эффекты интенсивной геомагнитной бури. Стад. Геофиз. Геод. 2007, 51, 563–574. [Перекрестная ссылка] 13. Шулиц, Д.М.; Срецкович, В.А.; Михайлов А.А. Исследование вариаций ОНЧ -сигнала, связанных с изменением уровня ионизации в D-область вследствие солнечных условий. Адв. Космическое разрешение. 2016, 57, 1029–1043. [Перекрестная ссылка]
14. Томсон, Северная Каролина; Кливеред, Массачусетс; Макрей, В.М. Параметры D-области ночной ионосферы по фазе и амплитуде ОНЧ . Дж Геофиз. Рез. Атмосфера. 2007, 112. [CrossRef]
15. Жигман В.; Доминик, М.; Грубор, Д.; Роджер, CJ; Кливеред, М.А. Электронная плотность в нижней ионосфере и эффективные коэффициенты рекомбинации по данным многоприборных космических наблюдений и наземных ОНЧ -измерений во время солнечных вспышек. Дж Атмос. Соль.-Терр. Физ. 2023, 247, 106074. [CrossRef]
16. Ярвинен Р.; Каллио, Э . Глобальное гибридное моделирование космической погоды Меркурия и наблюдения БепиКоломбо. В материалах 4-го URSI AT-RASC, Гран-Канария, Испания, 19–24 мая 2024 г.
17. Михайлов Ю.М.; Михайлова Г.А.; Капустина О.В. Электромагнитный фон КНЧ и ОНЧ во внешней ионосфере над сейсмоактивными регионами (спутник «ИНТЕРКОСМОС-24»). В материалах Международного симпозиума PLASMA '97 «Исследование и применение плазмы», Ярнолтувек недалеко от Ополе, Польша, 10–12 июня 1997 г.; Предоставленные статьи; Том 1.
18. Климов С.; Романов С.; Савин, С.; Николаева Н.; Скальский А.; Петрукович А.; Ноздрачев М.; Триска, П.; Амата, Э .; Чобану, М.; и другие. Хвостовой зонд ИНТЕРБОЛ. Исследование границ магнитосферной плазмы. В материалах Международного симпозиума PLASMA '97 «Исследование и применение плазмы», Ярнолтувек недалеко от Ополе, Польша, 10–12 июня 1997 г.; Предоставленные статьи; Том 1.
19. Блецкий, Ю; Косацкий, К.; Вроновский Р.; 'Савин, С.; Симунек, Дж; Смилауэр, Дж; Кохневич, Дж; Климов С.; Триска П. Электростатические и электромагнитные компоненты плазменной волны, наблюдаемые в полярном каспе на спутниках «Прогноз-8» и «Магион-4». В материалах Международного симпозиума PLASMA '97 «Исследование и применение плазмы», Ярнолтувек недалеко от Ополе, Польша, 10–12 июня 1997 г.; Предоставленные статьи; Том 1.
20. Блецкий, Ю; Косацкий, К.; Вроновский Р.; Немечек, З.; Сафранкова Ж; Симунек, Дж; Смилауэр, Дж; Бородкова Н.; Романов, С.А.; Савин, ИП; и другие. Электронно-циклотронные гармоники, наблюдаемые субспутником Магион-4 в плазменном слое. В материалах Международного симпозиума PLASMA '97 «Исследование и применение плазмы», Ярнолтувек недалеко от Ополе, Польша, 10–12 июня 1997 г.; Предоставленные статьи; Том 1.
21. Роткаэль, Х.; Блецкий, Ю; Клос, З. Характеристика ионосферно-магнитосферной связи, исследованный с помощью плазменно-волнового пакета PWP на борту спутника CESAR. В материалах Международного симпозиума PLASMA '97 «Исследование и применение плазмы», Ярнолтувек недалеко от Ополе, Польша, 10–12 июня 1997 г.; Предоставленные статьи; Том 1.
22. Вулискрофт, LJC Измерения плазменных волн с помощью миссии ESA/NASA CLUSTER. В материалах 2-го национального симпозиума PLASMA '95 «Исследование и применение плазмы», Варшава, Польша, 26–28 июня 1995 г.; Приглашенные доклады; Том 2.
23. Маккалоу, JP; Миёси, Ю; Жинье, врач общей практики; Джонстон, WR; Су, Ю-Дж; Старкс, MJ; Касахара, Ю; Кодзима, Х.; Мацуда, С.; Синохара, И.; и другие. Передача сверхнизкочастотной радиосвязи из космоса в магнитосфере с использованием спутников DSX и Arase. Планеты Земля Космос 2022, 74, 64. [CrossRef]
24. Лаубен, Д.; Линскотт, И.; Инан, У. Прибор широкополосного приемника (BBR) для демонстраций и научных экспериментов (DSX) Космический корабль; Заключительный отчет №. АФРЛ-РВ-ПС-ТР-2022-0114; Лаборатория экспериментальной физики Хэнсона, Стэнфордский университет: Стэнфорд, Калифорния, США, 15 декабря 2022 г.
25. Нойффер А.; Лунк А. Диагностика источника планарного магнетронного распыления с помощью волн СВЧ-зондов. В материалах Международного симпозиума PLASMA '97 «Исследование и применение плазмы», Ярнолтувек недалеко от Ополе, Польша, 10–12 июня 1997 г.; Приглашенные доклады; Том 2.
26. Галейс Дж Импеданс конечно-изолированной цилиндрической антенны в холодной плазме с продольным магнитным полем. IEEE Транс. Антенны Пропаг. 1966, 14, 727–736. [Перекрестная ссылка]
27. Галейс Дж Импеданс конечно изолированной антенны в холодной плазме с перпендикулярным магнитным полем. IEEE Транс. Антенны Пропаг. 1966, 14, 737–748. [Перекрестная ссылка]
28. Вунш Д.А. Излучение полоски электрического тока в магнитоионной среде. Может. Дж Физ. 1967, 45, 1675–1691. [Перекрестная ссылка]
29. Бхат Б. Распределение тока на бесконечной трубчатой антенне, погруженной в холодную ностолковительную юмагнитоплазму. Радио Наук. 1973, 8, 483–492. [Перекрестная ссылка]
30. Сешадри С.Р. Входная проводимость цилиндрической антенны в магнитоионной среде. Дж Прил. Физ. 1968, 39, 2407–2412. [Перекрестная ссылка]
31. Песня, П.; Рейниш, Б.В.; Пазнухов В.; Продажи, Г.; Кук, Д.; Ту, Ж-Н.; Хуанг, Х.; Библ, К.; Галкин И. Взаимодействие высоковольтной антенны с плазмой при передаче свистовых волн: эффекты плазменной оболочки. Дж Геофиз. Рез. 2007, 112, A03205. [Перекрестная ссылка]
32. Соловьев О.В. Распространение низкочастотных радиоволн в волноводе Земля-ионосфера, возмущенном крупномасштабным трех-неравномерность размеров. Радиофиз. Квантовый электрон. 1998, 41, 392–402. [Перекрестная ссылка]

Machine Translated by Google

33. У э й т Дж. Р. Т е о р и я и з л у ч е н и я и с т о ч н и к о в, п о г р у ж е н н ы х в а н и з о т р о п н ы е с р е д ы. Дж. Р е з. Н а т л. Б у р. С т е н д. М а т е м а т и к а. М а т е м а т и к а. Ф и з. 1964, 68, 119–136. [Перекрестная ссылка]

34. Г и н е, Г.П. О Н Ч - и з л у ч е н и е в д а л ь н е й з о н е о т л и н е й н о й д и п о л ь н о й а н т е н н ы, п о г р у ж е н н о й в п л а з м у: б а з о в а я м о д е л ь д л я D S X; Н о м е р о т ч е т а. 95-0002; Л а б о р а т о р и я Л и н к о л ь н а, М а с с а ч у с е т с к и й т е х н о л о г и ч е с к и й и н с т и т у т: Л е к с и н г т о н, М а с с а ч у с е т с, С Ш А, 15 и ю н я 2020 г.

35. Б а н н и с т е р, P R; Х а р р и с о н, Д ж. К.; Р у п п, С С; К и н г, Р В П; К о с м о, M L; Л о р е н ц и н и, Е С; Д а й е р, C J; Г р о с с и, д о к т о р м е д и ц и н с к и х н а у к О р б и т а л ь н ы й п е р е д а т ч и к и а н т е н н а д л я к о с м и ч е с к о й с в я з и н а E L F / V L F с п о д в о д н ы м и л о д к а м и. В м а т е р и а л а х к о н ф е р е н ц и и A G A R D 529: Р а с п р о с т р а н е н и е р а д и о с и г н а л о в E L F / V L F / L F и с и с т е м н ы е а с п е к т ы, С и м п о з и у м г р у п п ы э к с п е р т о в п о р а с п р о с т р а н е н и ю э л е к т р о м а г н и т н ы х в о л н, Б р ю с с е л ь, Б е л ь г и я, 28 с е н т я б р я – 2 о к т я б р я 1992 г.

36. П о й з е л ь, А н т е н н ы е с и с т е м ы P A и п р и л о ж е н и я р а д и о э л е к т р о н н о й б о р ь б ы; A r t e c h H o u s e: Б о с т о н, М а с с а ч у с е т с, С Ш А; Л о н д о н, В е л и к о б р и т а н и я, 2012 г.

37. М о д е л ь с к и Дж; Я щ и ш и н, Э .; Ч а ц и н с к и й, Х .; М а й ч ж а к П. И з м е р е н и я п а р а м е т р о в а н т е н н; O F P W, В а р ш а в с к и й т е х н о л о г и ч е с к и й у н и в е р с и т е т: В а р ш а в а, П о л ь ш а, 2004 г.

38. Б е м, D J А н т е н н ы и р а с п р о с т р а н е н и е р а д и о в о л н; W N T: В а р ш а в а, П о л ь ш а, 1973.

39. Ц е л и н с к и й З. П л а з м а; P W N: В а р ш а в а, П о л ь ш а, 1980 г.

40. Б о г д а н о в А.Ф.; В а с и н В.В.; Д у л и н В.Н.; И л ь и н, В.А.; К р и в и ц к и й, Б.Х.; К у з н е ц о в В.А.; Л а б у т и н В.К.; М о л о ч к о в Ю Б.; П и р ц о в С.В.; С т е п а н о в Б.М.; и д р у г и е. С б о р н и к п о р а д и о э л е к т р о н и к е; К у л и к о в с к и й А.А. (р е д.); W K i Ł: В а р ш а в а, П о л ь ш а, 1971.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Динамическое ценообразование на основе ресурсов и принудительная пересылка Алгоритм стимулирования в социально ориентированных сетях

Сюминь Чжан 1,2, Юаньли 1,2,*, Цзэнган Сюн 1,2,*, Яньчао Лю 1, Шихуэй Ван 2 и Делин Хоу 1,2

1 Школа компьютерных и информационных наук, Хубэйский инженерный университет, Сяоган, 432000, Китай; zhxm9701@163.com (XZ); yanchaoliu@cuc.edu.cn (YL); hdl00@stu.hubu.edu.cn (DH)

2 Школа компьютерных наук и информационной инженерии, Университет Хубэй, Ухань 430062, Китай; wsh@hubu.edu.cn

* Переписка: 202221116012706@stu.hubu.edu.cn (YL); jkxxzg2003@163.com (ZX)

Аннотация: В социально ориентированных сетях узлы обычно ведут себя эгоистично из-за ограничений ресурсов. и социальные корреляции, что приводит к низкой производительности сети. Чтобы стимулировать эгоистичные узлы активно участвовать в пересылке сообщений, в этой статье предлагается динамическое ценообразование на основе ресурсов и принудительное Алгоритм стимулирования пересылки (DFIA). Во-первых, алгоритм вводит виртуальную валюту в качестве средства транзакции, а затем разрабатывает функцию ценообразования на основе таких факторов, как состояние ресурса узла, вклад в участие, актуальность местоположения и социальные связи. Это гарантирует, что пересылка Услуги предоставляются по разумной цене в соответствии с правилами переговоров. Во-вторых, принудительная переадресация стратегия реализуется для того, чтобы заставить эгоистичные узлы, которые не желают участвовать в деятельности других узлов пересылка сообщений — для пересылки определенного количества нелокальных сообщений. При этом, чтобы запретить узлам отбрасывать сообщения и обеспечить успешную пересылку к месту назначения, определенные правила используются для распределения значений вклада между узлами, которые успешно участвуют в сообщении пересылка. Наконец, чтобы избежать ложного поведения котировок, используется технология блокчейна. Информация о транзакциях упаковывается в блоки и добавляется в блокчейн после проверки консенсуса. другими узлами сети, обеспечивая прозрачность и неизменяемость данных транзакций. Результаты моделирования показывают, что по сравнению с существующими алгоритмами стимулирования этот алгоритм не не только повышает вероятность доставки сообщения, но и эффективно снижает среднюю задержку.

Ключевые слова: социально ориентированные сети; эгоистичные узлы; ложная цитата; виртуальная валюта

check for updates

Цитирование: Чжан Х.; Ли, Ю; Сюн, З.; Лю, Ю; Ван, С.; Хоу, ДА. Динамическое ценообразование на основе ресурсов и Принудительное стимулирование пересылки Алгоритм в Social Aware Сеть. Электроника 2024, 13, 3044. https://doi.org/10.3390/электроника13153044

Академический редактор: Доменико Розачи

Поступила: 25 июня 2024 г.

Пересмотрено: 23 июля 2024 г.

Принято: 30 июля 2024 г.

Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://Creativecommons.org/licenses/by/4.0/).

1. Введение

С бурным развитием сетевых информационных технологий мобильная среда формируется интеллектуальными устройствами (смартфонами, планшетами, автомобильными устройствами) с помощью беспроводные технологии на основе соседства (Bluetooth, WIFI) [1] получили все больше и больше больше внимания, и соответственно появились социально-ориентированные сети (SAN) [2,3] . САН исследования определяют разумные протоколы для предоставления услуг приложениям верхнего уровня. путем восприятия и анализа контекстной информации мобильного узла [4] в сети. Это достигается за счет анализа социальных сетей и технологий машинного обучения, которые анализировать и учиться на полученной информации о поведении и окружающей среде носителей с интеллектуальных мобильных устройств. На основе этих данных выводятся такие социальные атрибуты, как сообщество, интенсивность взаимодействия и модели мобильности. Интегрируя эти социальные атрибуты с поведением оператора связи, эффективными протоколами маршрутизации и стратегиями пересылки сообщений предназначены для предоставления высококачественных услуг для приложений верхнего уровня. Например, исследования в [5,6] принимали решения о пересылке на основе социальных характеристик пользователей и сходства [7].

SAN (см. рисунок 1) использует социальные атрибуты операторов мобильной связи (т. е. абонентов) для разработки эффективных сетевых протоколов. Он представляет собой новый подход в Delay. Толерантная сеть (DTN) [8] и Оппортунистическая сеть (OppNet) [9]. По принципу работы SAN аналогичен DTN и OppNet [10]. С одной стороны, из-за

Machine Translated by Google

динамическая топология, вызванная мобильностью узлов внутри сети, отсутствие стабильных сквозных путей связи между узлами источника и назначения. Когда связь между узлами разрывается, традиционные методы маршрутизации не могут пересылать данные сообщения, и для реализации пересылки в многопрыжковом режиме часто используются вероятностные совпадения. С другой стороны, во время пересылки сообщений промежуточные узлы обычно используют метод «сохранения-переноса» для временного хранения сообщений [11–13], используя свои недорогие устройства с малым радиусом действия для связи друг с другом, тем самым разделяя объекты данных (такие как изображения, реклама и т. д.) среди узлов [7]. Однако, в отличие от DTN и OppNet, мобильные устройства в SAN обычно передаются такими объектами, как люди, транспортные средства и животные, которые перемещаются в разные места и участвуют в различных действиях, что приводит к постоянному изменению топологии сети. Таким образом, среди узлов SAN существуют определенные социальные атрибуты. Чрезвычайно важно разработать эффективные решения, которые мотивируют узлы к сотрудничеству посредством этих характеристик социальных атрибутов.

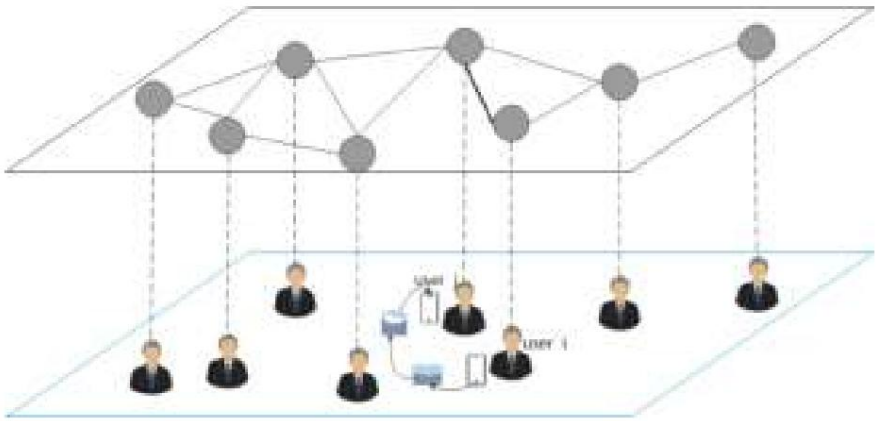


Рисунок 1. Социально-ориентированная сеть.

Будучи новой сетевой моделью, SAN по-прежнему сталкивается с множеством многочисленных проблем, связанных с исследованием эгоизма узлов, проблемами безопасности пересылки сообщений и другими связанными технологиями. В SAN некоторые узлы могут проявлять эгоистичное поведение из-за ограниченности своих ресурсов, что приводит к плохой способности доставки или плохим социальным отношениям с другими узлами. Такое поведение отрицательно влияет на общую производительность и эффективность сети. Эгоистичное поведение узлов проявляется как индивидуальный эгоизм [14] и социальный эгои
Узлы, проявляющие индивидуальный эгоизм, не получают информацию от других узлов и не передают ее другим узлам, когда их ресурсы ограничены. И наоборот, узлы, проявляющие социальный эгоизм, предоставляют услуги пересылки данных только знакомым узлам и отказываются пересылать или получать сообщения данных от узлов с плохими социальными отношениями. В SAN, помимо проблем, связанных с эгоизмом узлов, также возникают риски злонамеренных атак, таких как подделка и кража пересылаемых сообщений и записей транзакций [16]. Таким образом, очевидно, что необходимы эффективные механизмы стимулирования и строгие меры безопасности, чтобы мотивировать узлы к участию в сотрудничестве и предотвращению злонамеренных атак, что, в свою очередь, обеспечивает эффективную работу SAN и безопасную передачу данных.

В настоящее время разработка механизмов стимулирования для узлов является основным методом решения проблемы эгоизма узлов в SAN. Существующие механизмы стимулирования сотрудничества узлов можно разделить на три типа: механизмы стимулирования, основанные на репутации, механизмы стимулирования, основанные на виртуальной валюте, и механизмы стимулирования, основанные на теории игр.

Как описано в статье [17], авторы точно рассчитали репутацию, объединив авторитет и уверенность рекомендателей, и справедливо оценили каждую рекомендацию Они использовали механизм справедливого дифференцированного обслуживания, чтобы стимулировать активное участие. В [18] Рой и др. определенное значение репутации узла и стоимость пересылки сообщений. Эта стратегия повышает эффективность доставки сообщений за счет выбора узлов с высокой репутацией и низкой стоимостью многоадресной передачи данных в качестве узлов следующего перехода для пересылки сообщений. Бунуни и др. [19] объединили прямую и косвенную репутацию для оценки надежности узла.

Machine Translated by Google

Это позволяет узлам точно понимать состояние соседних узлов в пределах их близости. Затем узлы классифицируются по различным состояниям сотрудничества на основе значений их репутации, и реализуются соответствующие стимулы для поощрения сотрудничества на основе этих состояний. Эти механизмы стимулирования сталкиваются с проблемами при мониторинге, хранении и передаче значений репутации из-за потенциальных атак со стороны вредоносных узлов, которые могут привести к несоответствию значений репутации. Кроме того, узлы, находящиеся рядом с установленным пороговым значением, могут проявлять случайность, сохраняя лишь достаточную ценность репутации, чтобы избежать активного участия в пересылке.

Ван и др. [20] предложили стимулирующую эволюционную игровую модель, которая поощряла как нормальные, так и аномальные узлы участвовать в сетевом сотрудничестве посредством эволюционного процесса выживания наиболее приспособленных, обеспечивая высокий уровень сотрудничества и поддерживая непрерывную стабильность сети. В литературе создана модель повторяющейся игры (CGTFT), основанная на функции оценки сотрудничества соседних узлов [21]. Для рациональных узлов в сети функция оценки степени сотрудничества разработана путем объединения остаточной энергии и успешной скорости пересылки узлов. Затем в игре используется модель CGTFT, награждающая узлы, которые активно участвуют в сотрудничестве, и наказывающая эгоистичные и отказывающиеся сотрудничать узлы. Кроме того, вводится фактор забывания, чтобы облегчить восстановление взаимодействия для узлов, вступающих в период наказания. В исследовании, опубликованном в [22,23], была предложена теоретическая игровая модель ценообразования, основанная на модели сделки Рубинштейна, позволяющая узлам ретрансляции оценивать свои услуги в соответствии с остаточной энергией и ценностью передаваемой информации, тем самым улучшая распределение содержания сообщения внутри сети. сеть. Эти механизмы используют для анализа теорию игр. Однако они требуют высоких затрат ресурсов и больше подходят для симметричных сетей, что делает их менее дружелюбными к асимметричным сетям.

Сяо и др. [24] предложили схему стимулирования пересылки данных на основе эквивалентного обмена (EEIS). EEIS объединяет статус ресурсов и статус богатства узлов, чтобы оценить сообщения, которые необходимо пересылать, и публично раскрыть богатство и статус ресурсов обеих торговых сторон. В [25] Ли и др. представила двухэтапный алгоритм стимулирования на основе ценообразования на услуги (STIA). STIA оценивает услуги пересылки с учетом качества услуги, атрибутов сообщения и состояния ресурсов, а авторы предложили двухэтапный механизм стимулирования. Он также формулирует схему валютного овердрафта и двухэтапный механизм штрафов для предотвращения дефолта узлов после превышения валютного лимита. Во время транзакции обслуживания оба узла обращаются к Кредитно-клиринговому центру (CCC), чтобы понять ситуацию друг друга и предотвратить ложные котировки. Однако использование стороннего доверенного учреждения для управления и предотвращения ложных котировок увеличивает риск утечки данных в случае наличия уязвимостей безопасности или злонамеренных атак. Рао и др. [26] предложили алгоритм стимулирования с ограниченными ресурсами и социально-эгоистичными мотивами (RSIA). Узлы награждаются виртуальной валютой посредством торга и переговоров, что является стимулом для эгоистичных узлов участвовать в сотрудничестве. Алгоритм вводит коэффициент Джини для оценки баланса доходов узлов и награждает или наказывает узлы в зависимости от их баланса доходов. RSIA настраивает каждый узел сети с помощью доверенного модуля безопасности. Узлы добавляют соответствующее значение валюты только тогда, когда информация о транзакции проверена как неизмененная, что позволяет избежать ложных котировок. Однако этот алгоритм не учитывает позиционные отношения между узлами. При настройке модуля безопасности Н для узлов для выполнения проверки ACK зависимость от производителей в отношении доверия узлов может поставить под угрозу безопасность всей сети, если доверие к производителю будет скомпрометировано.

Таким образом, обеспечение подлинности и точности транзакций и предотвращение ложных котировок остаются критически важными вопросами, требующими решения.

Подводя итог вышеизложенному анализу, чтобы стимулировать эгоистичные узлы активно участвовать в пересылке сообщений и предотвращать выдачу ложных предложений узлами, в этой статье предлагается алгоритм динамического ценообразования и принудительной пересылки на основе ресурсов (DFIA). Этот алгоритм анализирует эгоистичное поведение узлов и учитывает их ситуацию с ресурсами, уровень вклада, статус благосостояния и социальную ценность для согласования цен на пересылку сообщений. Кроме того, вводится стратегия принудительной пересылки, чтобы заставить

Machine Translated by Google

эгоистичные узлы пересылают нелокальные сообщения, тем самым ограничивая поведение узлов, которые фокусируются исключительно на собственных интересах и не желают участвовать в сетевом сотрудничестве. Чтобы предотвратить ложные котировки, информация о транзакциях записывается в блокчейн, используя децентрализованный и прозрачный характер технологии блокчейн, чтобы гарантировать, что цены не могут быть подделаны.

Оставшаяся часть статьи структурирована следующим образом: В разделе 2 представлена сетевая модель и предположения алгоритма. В разделе 3 подробно описаны соответствующие определения и процесс реализации предлагаемого алгоритма. В разделе 4 представлена среда моделирования. Мы установили Spray And Wait [27] в качестве базового алгоритма маршрутизации и объединили предложенные DFIA, RSIA из статьи [26], STIA из статьи [25] и EEIS из статьи [24] с базовым алгоритмом маршрутизации для экспериментального сравнения. . В заключительной части подводятся итоги всей статьи и прогнозов на будущее.

2. Сетевая модель и предположения

В этой статье представлено абстрактное представление связей между узлами в социально-ориентированной сети (SAN) в виде неориентированного графа G(D, E), где D — набор всех узлов, а E — набор ребер. Здесь N обозначает количество узлов в SAN.
Узел-источник является носителем сообщения перед передачей, а узел ретрансляции является носителем, который получает сообщение. При встрече узла-источника и узла-ретранслятора инициируется формирование ребра между узлами. E(i, j) представляет ребро между узлами i и j, где i≠ j.

Кроме того, на начальном этапе моделирования сети в этой статье создается блок генезиса. По мере успешного завершения транзакций для служб пересылки сообщений между узлами новые записи транзакций упаковываются в новые блоки и последовательно связываются после исходного блока. Это гарантирует, что все записи транзакций надежно и упорядоченно хранятся в блокчейне, сформированном из генерационного блока. Узлы в сети коллективно участвуют и поддерживают общедоступный, прозрачный и защищенный от несанкционированного доступа блокчейн.

В соответствии с потребностями в этой статье сделаны следующие предположения и определения. для модели:

1. Каждый узел в сети имеет один и только один идентификатор, и все узлы рациональны, что означает, что они преследуют максимизацию своей экономической выгоды или других стратегических целей, придерживаясь при этом сетевых протоколов и правил.
2. Сеть состоит из двух различных типов узлов: нормальных узлов и эгоистичных узлов.
   Обычные узлы участвуют в хранении, переносе и пересылке сообщений, как и ожидалось;
   Эгоистичные узлы — это те узлы, которые отказываются принимать и пересылать любые сообщения от других узлов из-за низких льгот или слабых социальных связей.
3. Предположим, что потенциально вредоносное поведение узлов в SAN ограничивается ложным цитированием, а не подделкой сообщений или сговором.
4. Только один узел в сети имеет право формировать новый блок в любой момент времени.

3. Алгоритм динамического ценообразования и принудительного стимулирования на основе ресурсов

В этой статье пересылка сообщений между узлами рассматривается как транзакция.
Покупатель является перевозчиком сообщения до его пересылки, а продавец — перевозчиком, получающим сообщение. Транспортно-экспедиционная услуга, предоставляемая покупателем, является товаром, и обе стороны используют виртуальную валюту в качестве средства транзакции. Обе стороны предлагают начальную цену, основанную на функции ценообразования. Если ставка разумна, сделка считается успе
В противном случае цены корректируются в соответствии с конкретными правилами переговоров, и транзакция повторяется. Если транзакция по-прежнему не удалась, обрабатывается следующее сообщение. После успешной транзакции информация о транзакции записывается в блок и связывается с блокчейном, обеспечивая прозрачность и неизменность информации.

Machine Translated by Google

3.1. Соответствующие определения

3.1.1. Срочность доставки узлов

Срочность доставки узла относится к срочности, с которой узел может получать и пересылать сообщения, в первую очередь определяемой оставшимся буферным пространством узла, оставшейся энергией и оставшимся временем жизни (TTL) сообщения. Когда всех трех достаточно, это указывает на то, что срочность доставки узла низкая и у узла достаточно ресурсов и времени для пересылки и получения сообщений. Поскольку сообщения непрерывно принимаются и пересылаются, буферное пространство узла, его энергия и TTL сообщения неизбежно будут постепенно уменьшаться, тем самым увеличивая срочность доставки узла. В этой статье мы используем индекс состояния для представления срочности доставки узла.

Индекс состояния буфера узла — это отношение оставшегося буферного пространства узла в данный момент к начальному общему буферному пространству узла, отражающее возможности буферизации узла. Индекс энергетического состояния узла — это отношение оставшейся энергии узла в данный момент к начальной полной энергии узла, отражающее способность узла пересылать сообщения. Индекс TTL сообщения — это отношение оставшегося TTL сообщения в данный момент к начальному TTL сообщения, что является одним из критических факторов, влияющих на способность узла доставить сообщение. Следовательно, срочность доставки Qh узла x в момент времени h определяется следующим образом:

$$Q^{час}_{вопрос\ x,\ м}(h) = \alpha\left(1 - \frac{B_{rx}}{B_{initx}}\right) + \beta\left(1 - \frac{E_{rx}}{E_{initx}}\right) + \gamma\left(1 - \frac{T_{rm}}{T_{initm}}\right) \quad (1)$$

Значения весов α, β и γ вводятся для балансировки различных целей оптимизации. Поскольку рассматриваемая проблема является NP-Hard, значения весов между различными экземплярами будут иметь существенное влияние на производительность алгоритма. Поэтому в нашем исследовании мы скорректировали различные значения веса посредством большого количества экспериментов, а затем нашли оптимальную конфигурацию веса для конкретного случая проблемы. B_{rx}, E_{rx} и T_{rm} представляют собой оставшееся буферное пространство, оставшееся значение энергии и оставшийся TTL сообщения m узла x в момент h соответственно. B_{initx}, E_{initx} и T_{initm} представляют начальное общее буферное пространство, начальное значение общей энергии и начальное значение TTL сообщения m для узла x соответственно.

3.1.2. Позиционная значимость узлов

Позиционная значимость узлов относится к ситуации, связанной с расстоянием между узлом и узлом назначения сообщения. Он нормализуется с использованием функции сходства по Гауссу для функции позиционного атрибута узла, служащей мерой позиционных атрибутов узла.

Основные атрибуты местоположения узла включают скорость движения узла, угол с узлом назначения и расстояние. Функция атрибута местоположения loc(x, d) узла определяется как

$$loc(x, d) = \frac{vx \cdot |\cos A|}{D(x,d)}, \quad (2)$$

где vx обозначает скорость перемещения узла x. Предположим, что координаты узла x и узла назначения d равны (ax,bx) и (ad ,bd) соответственно. Тогда $\cos A = \frac{a_x \cdot a_d + b_x \cdot b_d}{\sqrt{a_x^2 + b_x^2} \cdot \sqrt{a_d^2 + b_d^2}}$, где $0 \le A \le \pi$, а A представляет угол между узлом x и узлом назначения . $D(x, d) = \sqrt{(a_x - a_d)^2 + (b_x - b_d)^2}$, где D(x, d) представляет собой евклидово расстояние между узлом x и узлом назначения.

Функция подобия Гаусса [28] используется для нормализации уравнения (2), а позиционная релевантность узлов рассчитывается как

$$g(x) = e^{\frac{(loc(x,d))^2}{2\sigma^2}}, \quad (3)$$

где σ = 1 — параметр масштаба. Поскольку функция атрибута местоположения узла x становится меньше, значение g(x) приближается к 1, что указывает на то, что время, за которое узел x достигнет пункта назначения,

Machine Translated by Google

узел короче. В модели ценообразования узел x будет готов заплатить более высокую цену за услугу пересылки для быстрой пересылки сообщения.

3.1.3. Вклад участия узлов

В сети участие узла отражает его уровень участия в пересылке сообщений и его вклад в общую эффективность сети. Это существенный фактор, влияющий на цену узла. Более высокий вклад участия свидетельствует о готовности узла предоставлять услуги другим узлам, тем самым увеличивая скорость доставки сообщений внутри сети посредством его пересылки. Таким образом, алгоритм DFIA вводит степень сотрудничества и среднюю степень вклада узлов для измерения их участия в процессе доставки сообщений.

Готовность узла к сотрудничеству демонстрируется его готовностью использовать свои ресурсы для пересылки сообщений, генерируемых другими узлами. Количество нелокальных сообщений, пересылаемых узлом, прямо пропорционально его участию. Таким образом, степень сотрудничества Xx(h) узла x в момент времени h определяется как

$$Xx(ч) = \frac{\text{Форекс}}{\text{Мкс}}. \quad (4)$$

где Fx обозначает общее количество нелокальных сообщений, пересылаемых узлом x. Каждый раз, когда узел x успешно пересылает нелокальное сообщение, тогда Fx + 1. Mx представляет количество встреч узла x с другими узлами. Каждый раз, когда узел x встречает другой узел: Mx + 1.

Когда сообщение, которое узел помогает пересылать, успешно доставлено узлу назначения , в этом документе значения вклада назначаются в соответствии с узлом на основе его положения на пути пересылки, чтобы стимулировать активное участие в сетевой коммуникации. Первому узлу ретрансляции на пути пересылки назначается наибольшее значение вклада, поскольку он определяет направление доставки сообщения. Кроме того, последний ретрансляционный узел получает относительно большую ценность вклада, поскольку он завершает фактическую доставку сообщения и является исполнителем задачи. Другие ретрансляционные узлы на пути пересылки получают уменьшающиеся значения вклада в зависимости от их положения на пути пересылки. Значение вклада C(i, n) , назначенное узлом i на n-м пути, представлено следующим образом:

$$C(я, п) = \begin{cases} (1 - г)^{я-1} & r, \text{ если } 0 \quad i < L \\ (1 - г)^{я-1}, & \text{если } я = L \end{cases}, \quad (5)$$

$$\text{avgCx} = \frac{\sum_{t=1}^{н} Ct(x, n)\, t}{}, \quad (6)$$

где r (0 < r < 1) обозначает коэффициент распределения значений вклада, а L представляет общее количество узлов ретрансляции на пути. После успешной доставки сообщения узлу назначения значение вклада узла x накапливается в соответствии с вышеупомянутой схемой распределения. После этого среднее значение вклада узла x, avgCx , рассчитывается в соответствии с уравнением (6).

Следовательно, вклад участия PCx узла x может быть выражен уравнениями (5) и (6) как

$$PCx(h) = \rho \quad Xx(h) + (1 \quad \rho) \quad \text{avgCx}, \quad (7)$$

где ρ — весовой коэффициент, с помощью которого Xx(h) и avgCx влияют на PCx и ρ [0, 1].

3.1.4. Статус благосостояния узлов

Узлы в сети изначально обладают определенным количеством виртуальной валюты, что в некоторой степени отражает их возможности пересылки. В зависимости от количества виртуальной валюты, доступной узлу, его можно разделить на три различных состояния виртуальной валюты: достаточное состояние, общее состояние и недостаточное состояние. Когда Vx(h) > Vmax, узел находится в достаточном состоянии

Machine Translated by Google

состояние богатства, в этот момент Cx(h) > 1. Когда Vx(h) < Vmin, узел находится в состоянии недостаточного богатства, в этот момент Cx(h) < 1. Когда Vmin Vx(h) Vmax, узел находится в общем состоянии богатства, и в этот момент Cx(h) = 1. Соответствующий расчет выглядит следующим образом:

$$Cx(ч) = \begin{cases} \frac{Bx(ч)}{Vmax}, & \text{Адекватное состояние} \\ 1, & \text{Общее состояние} \\ \frac{Bx(ч)}{Vmin}, & \text{Неадекватное состояние} \end{cases} . \quad (8)$$

где Vx(h) представляет собой количество виртуальной валюты, принадлежащей узлу x в момент h. Пороги виртуальной валюты Vmax и Vmin соответственно представляют собой максимальное и минимальное количество виртуальной валюты, которой может владеть узел.

3.1.5. Степень социальной корреляции между узлами

В SAN социальные корреляции между узлами описывают степень, в которой между узлами существует какая-то связь или взаимное влияние. Социальная корреляция используется для отражения взаимодействия и связей между узлами в сети путем анализа общих соседей и исторических записей встреч между узлами. Таким образом, в этой статье всесторонне оценивается социальная корреляция между узлами на основе размеров социального перекрытия узлов и степени контакта между узлами.

Степень перекрытия между двумя узлами количественно определяется количеством общих соседей, которыми они обладают в SAN. Эти узлы общего соседа представляют собой мост, соединяющий два узла, и они представляют сходство между узлами. Таким образом, социальное перекрытие между узлами x и y определяется как

$$\text{перекрытиеху} = \frac{\text{комкси}}{\text{сосед} + \text{сосед} — \text{comxy}}, \quad (9)$$

где comxy обозначает количество узлов общих соседей между узлами x и y; neighx и neighy представляют общее количество соседей узлов x и y соответственно.

Степень контакта между узлами количественно определяет близость контакта между этими узлами посредством анализа частоты и продолжительности встреч. Если два узла часто общаются друг с другом, часто и глубоко взаимодействуют и степень контакта между ними высока, то, опираясь на определение друга в реальном обществе, эти два узла дружат друг с другом. Степень контакта contactxy между узлами x и y рассчитывается следующим образом:

$$\text{contactxy} = \frac{\text{Nmcxy} \quad \text{Tmxy}}{\text{N} \quad \text{AccurentT}} . \quad (10)$$

Как упоминалось выше, Nmcxy обозначает количество встреч между узлами x и y на данный момент, Tmxy обозначает сумму всех времен встреч между узлами x и y на данный момент, N обозначает количество узлов в текущей сети, а AccurentT представляет текущее момент.

На основе уравнений (9) и (10) рассчитывается социальная корреляция SCxy между узлами. следующее:

$$SCxy = \varphi \quad \text{перекрытиеху} + (1 \quad \varphi) \quad \text{contactxy, где} \quad (11)$$

φ — весовой коэффициент, определяющий влияние перекрытия и контактаху на SCxy, а φ [0, 1].

3.2. Модель транзакции

В SAN пересылка сообщений между встречающимися узлами абстрагируется как процесс транзакции, где услуги пересылки заключаются и продаются с использованием виртуальной валюты. В этом процессе узел, передающий сообщение перед пересылкой, действует как покупатель, в то время как узел, получающий сообщение, действует как продавец, а служба пересылки сообщений действует ка

Machine Translated by Google

товар. Обе стороны обязаны раскрыть свое богатство и ресурсы, а также свои предложения по экспедиторским услугам, исходя из своих соответствующих соображений.

3.2.1. Функция расходов покупателя.

Чтобы гарантировать, что сообщение может быть доставлено к следующему узлу плавно, покупатель x должен учитывать свой собственный резерв богатства, срочность доставки и передаваемые атрибуты сообщения, чтобы определить разумную цену затрат в качестве компенсации за обеспечение пересылки. услуги узлу следующего перехода. Когда резерв богатства x достаточен, покупатель синтезирует свою собственную срочность доставки и резерв богатства, чтобы предложить разумну Когда запас богатства x недостаточен, с течением времени время существования сообщения будет постепенно сокращаться, оставшийся кэш и энергия узла будут постепенно истощаться, а срочность доставки покупателя станет выше. Поэтому, чтобы ослабить давление, покупатель предложит более высокую цену сделки. Когда вклад участия x выше, что указывает на лучшую производительность услуги, он выиграет от x при покупке услуги пересылки. Если узел передает сообщение большего размера, энергия и кэш, потребляемые при пересылке, также увеличатся, а срочность доставки x также увеличится и, в свою очередь, повлияет на цену предложения за сообщение. Кроме того, x должен учитывать собственную корреляцию положения узла. Когда его корреляция позиций в узле назначения сообщения высока, это означает , что время, необходимое для того, чтобы x достиг узла назначения, короче, и это немного поднимет цену, чтобы переслать сообщение как можно быстрее. Таким образом, функция расходов покупателя x на пересылку сообщения m в момент времени h определяется как

$$BNpx(h) = Msize \quad (Q + \sigma 1^{\text{час}} \quad g(x)) \quad Cx(h) \quad PCx(h). \ (x, м) \tag{12}$$

где Msize обозначает размер m, g(x) обозначает позиционную релевантность x, σ1 — это вес, занимаемый позиционной релевантностью узла и удовлетворяющий условиям α + β + γ + σ1 = 1, Cx(h) обозначает богатство x в момент времени h, а PCx (h) обозначает вклад узла x в момент h.

3.2.2. Функция дохода продавца

Получая сообщения, передаваемые через x, продавцу y также необходимо учитывать собственный резерв благосостояния, срочность доставки и характеристики передаваемого сообщения, чтобы оценить его, стремясь максимизировать свой доход. Когда запас богатства y достаточен, готовность узла обеспечить пересылку сообщений уменьшится, поэтому для установления разумной цены ему придется объединиться со срочностью доставки y . Когда резерв богатства y недостаточен, ему необходимо получить виртуальную валюту из другого места по более низкой цене, и его готовность обеспечить пересылку сообщений возрастает. Пересылка сообщений более выгодна, если у y более высокий вклад в участие. Размер сообщения прямо пропорционален энергии и кэшу, необходимым для пересылки сообщения; Когда сообщение больше, потребление энергии и кэша также увеличивается для предоставления услуги пересылки, что, в свою очередь, влияет на цену. Кроме того, необходимо учитывать социальную к Если корреляция между узлами высока, т. е. SCxy SCxymin, то это указывает на то, что связь между двумя узлами лучше. Узел y имеет большую готовность переслать сообщение, и цена предложения будет относительно низкой, как показано в уравнении (13). Напротив , если корреляция между узлами низкая, т. е. SCxy < SCxymin, узел пересылки предложит относительно высокую цену, как показано в уравнении (14).

$$SNpy(h) = Msize \quad (Q \quad \sigma 2^{\text{час}} \quad (1 \quad SCxy)) \quad (1 \quad \frac{1}{Сай(ч)} PCy(h)) \ (y,m) \tag{13}$$

$$SNpy(h) = Msize \quad (Q \ (y,m) \quad ^{\text{час}} + \sigma 2 \quad SCxy) \quad \frac{1}{Сай(ч)} (1 \quad PCy(h)). \tag{14}$$

где Msize обозначает размер m, SCxy обозначает социальную корреляцию между x и y, а σ2 обозначает значение веса, занимаемое социальной корреляцией между x и y, удовлетворяющее

Machine Translated by Google

α + β + γ + σ2 = 1. SCxymin обозначает минимальный порог социальной корреляции между x и y, Cy(h) обозначает статус благосостояния y в момент времени h, а PCy(h) обозначает вклад участия y в момент времени. час

3.2.3. Стратегия принудительной

пересылки Учитывая ограниченные ресурсы энергии и кэша, доступные узлам в сети, они склонны отдавать приоритет пересылке собственных сообщений или отказываться от пересылки нелокальных сообщений, чтобы гарантировать успешную доставку сгенерированных ими сообщений. Такое поведение сильно вредит общим интересам сети. Для поддержания коллективных интересов сети и

Чтобы стимулировать узлы с низким уровнем сотрудничества к участию в сотрудничестве, в этом документе представлена стратегия принудительной пересылки, чтобы ограничить поведение тех узлов, которые заботятся только о своих собственных интересах и не желают сотрудничать в сети. Эта стратегия способствует сотрудничеству между узлами и оптимизирует производительность сети.

Размер значения принудительной пересылки CFx узла x определяется степенью сотрудничества Xx(h) и значением вклада C(i, n) узла x, которое рассчитывается следующим образом :

$$CFx = \quad 10 \quad (0,3 \quad Xx(h)) \quad C(i, n) \quad . \tag{15}$$

Принудительное автоматическое обновление значения пересылки с фиксированной частотой гарантирует, что поведение узлов пересылки можно отслеживать и регулировать в определенных узлах времени. Когда CFx > 0, узел x должен пересылать сообщения CFx, сгенерированные другими узлами, прежде чем он сможет пересылать свои собственные сгенерированные сообщения. Когда узлу x запрещено пересылать, чтобы беспрепятственно пересылать свои собственные сообщения, он может только повысить степень сотрудничества , участвуя в пересылке сообщений, генерируемых другими узлами, а затем постепенно уменьшать свое собственное значение принудительной пересылки, пока в конечном итоге не восстановится нормальное состояние. возможность пересылки сообщений .

3.2.4. Правила транзакции

В SAN, когда покупатель x встречает продавца y, обеим сторонам необходимо договориться о цене услуги пересылки данных в соответствии с конкретными правилами, а конкретные правила транзакции алгоритма заключаются в следующем:

(1) И покупатель, и продавец полностью информированы о цене сообщения и состоянии богатства другой стороны.

(2) После начала транзакции богатство покупателя представляет собой цену расходов, указанную покупателем x как BNpx(h) , и цену, заданную продавцом y как SNpy(h). Если BNpx(h) SNpy(h), то сообщение m успешно пересылается, и можно пересылать следующее сообщение. Если BNpx(h) < SNpy(h), первая переадресация по согласованию не удалась.

(3) Если первые переговоры между покупателем и продавцом не увенчались успехом, переходите ко второму этапу переговоров. Путем сравнения благосостояния обеих сторон соответствующие цены корректируются. Покупатель корректирует BNpx(h), учитывая степень сотрудничества, средний вклад и позиционную значимость, а продавец корректирует SNpy(h) на основе степени сотрудничества, среднего вклада и социальной корреляции между узлами. Поправочные коэффициенты для вторых переговоров равны k1 и k2 соответственно. Вторые правила переговоров заключаются в следующем:

- Если Cx(h) > Cy(h), то SNpy(h) остается неизменным, а BNpx(h) корректируется по уравнению BNpx(h) = BNpx(h) (1 + k1), где k1 – рассчитывается следующим образом:

$$k1 = \omega1 \quad Xx(h) + \omega2 \quad avgCx + \omega3 \quad g(x), \text{ где } Xx(h) \tag{16}$$

обозначает степень сотрудничества x, avgCx обозначает среднюю степень вклада x, а g(x) обозначает позиционная значимость узла x. Веса ω1, ω2 и ω3 обозначают влияние степени кооперации, средней степени вклада и позиционной релевантности узла на коэффициент корректировки k1 соответственно, а сумма трех из них равна 1. Здесь веса динамически корректируются с использованием метода адаптивного взвешивания, а значение веса рассчитывается по уравнению (17):

Machine Translated by Google

$$\begin{aligned} \omega 1 &= \frac{Xx(ч)}{Xxh+avgCx+g(x)} \\ \omega 2 &= \frac{avgCx}{Xxh+avgCx+g(x)} \\ \omega 3 &= \frac{г(x)}{Xxh+avgCx+g(x)} \end{aligned} . \quad (17)$$

- Если Cy(h) Cx(h), то BNpx(h) не изменяется, а SNpy(h) корректируется в соответствии с формула SNpy(h) = SNpy(h) (1 k2), где k2 вычисляется следующим образом:

$$k2 = \theta 1 \quad Xy(h) + \theta 2 \quad avgCy + \theta 3 \quad (18)$$

SCxy, где Xy(h) обозначает степень сотрудничества y, avgCy обозначает среднюю степень вклада y, а SCxy обозначает социальную корреляцию между узлами x и й. Веса θ1, θ2 и θ3 представляют влияние степени сотрудничества, в среднем степень вклада и социальная корреляция между узлами по поправочному коэффициенту k2, соответственно, а сумма трех из них равна 1. Динамическая регулировка этих веса также следуют тому же методу, что и в уравнении (17).

(4) Если после второго согласования оба узла достигнут консенсуса о том, что BNpx(h) SNpy(h) , сообщение будет перенаправлено, как на этапе (2). В противном случае транзакция будет отменена. и следующая транзакция будет продолжена. Поток транзакций сообщений показан на Фигура 2.

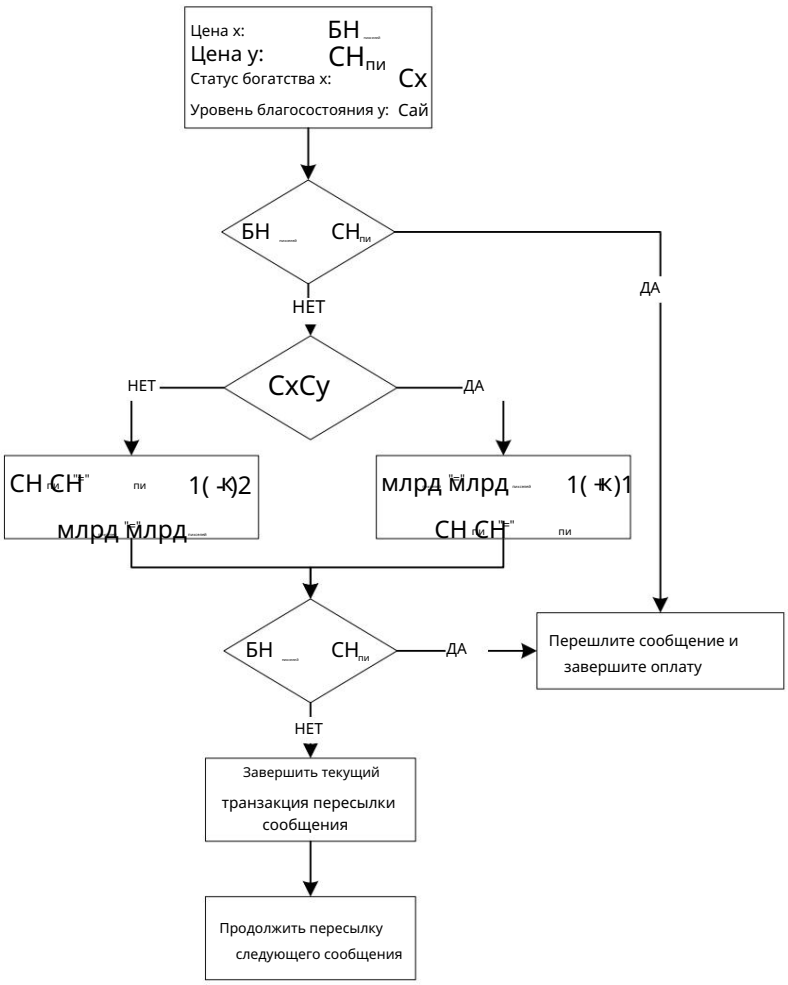


Рисунок 2. Блок-схема транзакции сообщения.

Подробный процесс обычного алгоритма транзакции сообщения показан в алгоритме 1 , который описывает полный процесс переговоров для пересылки сообщения.

Machine Translated by Google

услуга, когда два узла встречаются без штрафа за принудительную пересылку. Подробный процесс пересылки выглядит следующим образом: когда узлы x и y встречаются друг с другом, x и y передают свои соответствующие первые заявки на транспортную услугу согласно уравнениям (12)–(14) соответственно. Если BNpx SNpy, сообщение пересылается, а цена транзакции устанавливается равной среднее значение BNpx и SNpy. А затем запись транзакции упаковывается в блокчейн. Напротив, согласно правилам сделки, описанным ранее, цена корректируется. на основе статуса благосостояния x и y , чтобы сделать новое предложение. Если BNpx1 SNpy1, то один половина суммы BNpx1 и SNpy1 — это цена транзакции. И запись транзакции упаковывается в блокчейн после успешной пересылки сообщения. Если переговоры по-прежнему терпит неудачу после того, как обе стороны корректируют свои цены, они отказываются от пересылки этого сообщения m, и следующее сообщение обрабатывается для пересылки.

Алгоритм 1. Алгоритм обработки обычных сообщений.

```
Вход: BNpx, SNpy, Cx, Cy
Вывод: сообщение об успешной пересылке m
если BNpx >= SNpy , то
    Стоимостьm = (BNpx + SNpy)/2
    x отправляет сообщение m пользователю y, y получает сообщение m
    если сумма валюты, которую имеет x в настоящее время, больше, чем Costm , тогда
        Транзакция успешна. Блок                                  создать новый блок
        добавлен в конец цепочки блоков.
        Обновите богатство x и y
        вернуть m
    конец, если
еще
    Пересчитайте стоимость m и скорректированные ставки обеих сторон.
    BNpx1 = getPriceTwice_x(BNpx,SNpy Cx, Cy)
    SNpy1 = getPriceTwice_y(BNpx,SNpy Cx, Cy)
    если BNpx1 >= SNpy1 , то
        Стоимостьm = (BNpx1 + SNpy1)/2
        x отправляет сообщение m пользователю y, y получает сообщение m
        если Сумма валюты, которую в настоящее время имеет x, больше, чем Costm , тогда
            Транзакция успешна. Блок                              создать новый блок
            добавлен в конец цепочки блоков.
            Обновите богатство x и y
            вернуть m
        конец, если
    еще
        Продолжите следующую транзакцию.
    конец, если
конец, если
```

3.3. Игровой процесс пересылки сообщений

В SAN транзакцию пересылки сообщений между узлами можно абстрагировать как динамичный игровой процесс. Участниками игры в этом процессе являются покупатель x и продавец. й. Существует два основных этапа: отправляет ли покупатель x запрос на пересылку сообщения. и получит ли продавец в это время запрос на пересылку сообщения. Поведение обе стороны на двух этапах могут либо сотрудничать, либо отказываться от сотрудничества. Это должно быть Ожидается, что рациональные узлы сети выберут стратегию, которая максимизирует их собственные преимущества. В этой статье мы используем обратную индукцию для анализа двух этапов игры. процесса и продемонстрировать, что взаимное сотрудничество является оптимальной стратегией для обеих сторон.

Предположим, что в момент времени h для сообщения m цена предложения x равна BNpx(h) , а цена предложения y равна SNpy(h), тогда цена транзакции для сообщения равна Costm = $\frac{БНпкс(ч)+SNpy(ч)}{2}$.

Machine Translated by Google

(1) Сначала проанализируем второй этап игрового процесса: получает ли продавец запрос на пересылку сообщения или нет, и его поведение определяет прибыль узла. Когда у принимает запрос на пересылку сообщения от x, т. е. выбирает совместное поведение, текущий выигрыш Ey от y рассчитывается следующим образом:

$$Ey = SNpy(h) \quad Costm. \tag{19}$$

Когда y отказывается получать запрос на пересылку сообщения от x, т. е. решает отказаться от совместного поведения, выигрыша нет. Кроме того, поскольку y не принял сообщение при встрече с x, количество нелокальных сообщений, отправленных y , меньше, в то время как количество встреч увеличивается, а степень сотрудничества y уменьшается, что приводит к участию вклада также снижается. Из уравнений (13) и (14) видно , что следующая ставка y за другие сообщения становится выше, что может привести к сбою транзакции. Напротив, вы будете готовы получать сообщения.

(2) Затем проанализируйте, отправляет ли покупатель x запросы на пересылку сообщений или нет. Если x отправляет запросы на пересылку сообщений в момент h, т. е. выбирает совместное поведение, то сэкономленную им валюту Ex можно выразить как

$$Ex = BNpx(h) \quad Costm, \tag{20}$$

Если x не отправляет запрос на пересылку сообщения в момент h, т. е. отказывается сотрудничать, сначала проанализируйте затраты, которые x должен заплатить с точки зрения пересылаемого сообщения.
В момент времени h + 1 время выживания сообщения проходит вместе с ним, и для успешной пересылки m дополнительные затраты H1 , которые x должен заплатить, можно выразить как

$$H1 = Msize \quad Cx(h + 1) \quad PCx(h + 1) \quad \varphi 1 \quad \frac{Tre(h) \quad Tre(h + 1)}{TTL}, \tag{21}$$

где Tre(h + 1) обозначает оставшийся TTL m в момент времени h + 1, а значения других символов соответствуют значениям в уравнении (12).

Кроме того, стоимость, которую должен заплатить x , анализируется с точки зрения энергии и кэша покупателя x. Поскольку энергия x строго уменьшается с увеличением времени, а x отказывается сотрудничать, его кэш также частично занят сообщением m. Следовательно, в момент времени h + 1, если x хочет отправить запрос на пересылку, и его предлагаемая цена также увеличится, то дополнительная цена H2 , которую x должен заплатить, может быть выражена как

$$H2 = \varphi 2 \quad \left(\frac{\text{Мразмер}}{\text{Бркс (ч + 1)}} + \frac{\text{Msize)}}{\text{Эркс (ч + 1)}} \right) \quad Lx \quad Cx(h + 1) \quad PCx(h + 1). \tag{22}$$

где Lx обозначает длину сообщения, которое должно быть переслано x в момент времени h + 1, Brx (h + 1) обозначает оставшийся коэффициент буфера x в момент времени h + 1, Erx (h + 1) — оставшийся коэффициент энергии x в момент времени h + 1, а значения других символов совпадают со значениями в уравнении (12).

Таким образом, дополнительные затраты H , которые должен заплатить покупатель x за решение отказаться сотрудничество в момент времени h не менее

$$H = \text{мин}\{H1, H2\}. \tag{23}$$

Очевидно, Ex > 0 > H, поэтому x должен выбрать отправку запроса на пересылку сообщения.

После приведенного выше анализа и покупатели, и продавцы выберут совместное поведение, исходя из того, что оба получают максимальную выгоду, т. е. (сотрудничество, сотрудничество) является оптимальной стратегией.

3.4. Модуль безопасности

В этом документе для разработки алгоритма используется стратегия стимулирования, основанная на виртуальной валюте, поэтому неизбежно, что поведение ложных котировок узлов будет происходить в процессе каждой транзакции, что серьезно подрывает нормальный порядок транзакции.

Machine Translated by Google

Чтобы предотвратить незаконное вмешательство узлов в сумму валюты, в этом документе используется технология блокчейна для создания модуля безопасности, гарантирующего открытые и прозрачные транзакции пересылки сообщений между узлами. Блокчейн — это децентрализованный цифровой реестр экономических транзакций, обладающий преимуществами прозрачности и безопасности [29–32]. В сети записи всех успешных транзакций записываются в структуру блокчейна, обычно в последовательном порядке, и криптографически гарантировано, что они не подделываются и не подделываются. Каждый узел в цепочке независим и равноправен, опираясь на механизм консенсуса для обеспечения согласованности данных с характеристиками децентрализации, открытости и прозрачности, а также защиты от несанкционированного вмешательства [33].

3.4.1. Создание модуля

При хранении данных транзакции каждый узел в сети генерирует открытый ключ PK и секретный ключ SK посредством алгоритма шифрования с открытым ключом (например, RSA). Среди них SK надежно хранится на узле, а PK общедоступен в сети.
Чтобы обеспечить безопасность процесса транзакции и предотвратить ложные котировки, в настоящем документе выполняется следующий процесс для записи транзакции, созданной между инициатором A и акцептором B транзакции, которая содержит стороны транзакции A и B, цену транзакции, стоимость и т. д.: (1) Создание блока: когда

транзакция сообщения между A и B успешно завершена, используйте хэш-функцию (SHA-256) для вычисления хэш-значения каждого элемента в записи транзакции. Затем SK пользователя A используется для подписи хэш-значения H(стоимость) цены транзакции , генерируя SigSK(H(стоимость)). Затем запись транзакции, значение хеш-функции и подпись упаковываются в новый блок и транслируются по всей сети, чтобы другие узлы могли проверить.

(2) Проверка: B может использовать PK A для проверки цифровой подписи и перерасчета хэш-значения H' (стоимости) цены транзакции. Если эти два хэш-значения совпадают, то процесс проверки успешен, что указывает на то, что данные транзакции не были подделаны с точки зрения цены во время передачи. Другие узлы в сети проверят , действительна ли транзакция в блоке после получения нового блока.

(3) Достижение консенсуса: когда достаточное количество узлов подтверждает, что транзакция действительна и достигнут консенсус, блок считается действительным.

(4) Обновление блокчейна: как только новый блок достигает общесетевого консенсуса, он добавляется в конец блокчейна, как показано на рисунке 3. Этот процесс обновления обеспечивает консенсус и согласованность истории транзакций во всей сети.

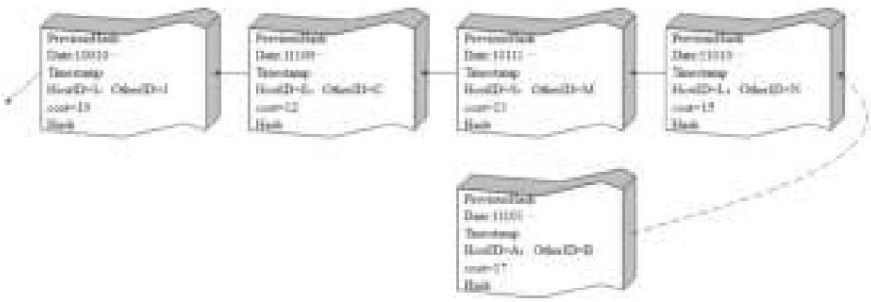


Рисунок 3. Принципиальная схема блокчейна.

Подделка данных транзакций обеспечивается за счет использования алгоритмов хеширования в модулях, созданных с помощью описанного выше процесса. Каждый новый блок содержит хеш-значение предыдущего блока, и любое вмешательство в данные транзакции изменит его хэш-значение, тем самым нарушая непрерывность всей цепочки блоков и делая любое нечестное поведение легко распознаваемым и отвергаемым другими узлами в сети. . Более того, данные транзакций прозрачны и заслуживают доверия благодаря тому, что проверенные записи транзакций записываются в блоки и добавляются в блокчейн, что, в свою очередь, обеспечивает прозрачность и достоверность данных транзакций. Любой узел, участвующий в сети, может получить доступ к этим записям транзакций и проверить их, снова эффективн

Machine Translated by Google

ложное поведение котировок. Все узлы следуют принципу самой длинной цепочки, что означает, что когда существует несколько конкурирующих цепочек ветвей, узлы предпочтут присоединиться к самой длинной действующей цепочке, поскольку основная цепочка, обеспечивающая надежность и безопасность блокчейна. Таким образом, справедливость а безопасность процесса транзакции сообщений в сети обеспечивается за счет использования блокчейна.

3.4.2. Схема пересылки

Полный процесс службы пересылки сообщений в DFIA показан в алгоритме 2. Предположим, что узел x встречается с узлом y, x необходимо переслать сообщение m в узел y, где CFx обозначает значение принудительной пересылки x. CFx оценивается перед транзакцией пересылки сообщения. Когда CFx > 0, x переходит в фазу принудительной пересылки; он вынужден пересылать нелокальные сообщения. Для каждой успешной транзакции службы пересылки сообщений значение CFx уменьшается на единицу , и наоборот; он продолжает пересылать следующее сообщение. Когда CFx 0, x выполняет обычную транзакцию пересылки сообщений в соответствии с Алгоритм 1. Во время транзакции услуги пересылки сообщения услуга пересылки цены оцениваются в соответствии с уравнениями (12)–(14), а торг ведется в соответствии с конкретными условиями. правила сделки. Если торг прошел успешно и сообщение успешно переслано, информация о транзакции, включая обе стороны x и y, цену транзакции и т. д., упакована в блок и добавляется в блокчейн после проверки. В противном случае сделка не удается, и следующее сообщение пересылается.

Алгоритм 2 Алгоритм DFIA

```
Ввод: CFx, x, y
Вывод: сообщение об успешной пересылке m
 1: если CFx > 0, то
 2: если исходный узел сообщения не является узлом x , тогда
 3: если BNpx >= SNpy , то
 4: Costm = (BNpx + SNpy )/2
 5: если сумма валюты, которую в настоящее время имеет x, больше, чем Costm , тогда
 6: Транзакция успешна. 7: Блок добавлен в конец                                  создать новый блок
цепочки блоков.
 8: Обновите богатство x и y
 9: CFx = CFx - 1
10: возврат м
11: конец, если
12: еще
13: Пересчитать стоимость m и скорректированные ставки обеих сторон.
14: BNpx1 = getPriceTwice_x(BNpx,SNpy Cx, Cy)
15: SNpy1 = getPriceTwice_y(BNpx,SNpy Cx, Cy)
16: если BNpx1 >= SNpy1 , то
17: Стоимость = (BNpx1 + SNpy1)/2
18: если сумма валюты, которую в настоящее время имеет x, больше, чем Costm , тогда
19: Транзакция успешна. 20: Блок добавляется в конец                              создать новый блок
цепочки блоков.
21: Обновите богатство x и y
22: CFx = CFx 1
23: вернуться м
24: конец, если
25: еще
26: Продолжить следующую транзакцию
27: конец, если
28: конец, если
29: конец, если
30: еще
                                                  Алгоритм обычной транзакции сообщений
31:32: конец, если
33: функция CREATENEWBLOCK(hostX, другое, м, стоимость)
34:         previousHash     блокчейн.getLastBlock().getHash()
35:         временная метка     SimClock.getTime()
36:         Подпишите хеш-значение стоимости транзакции, используя закрытый ключ SK A.
37:         newBlock     новый блок(hostX, другое, m, стоимость, previousHash, метка времени, 0)
38:         blockHash     newBlock.calculateHash()
39:         newBlock.setHash(blockHash)
40:         вернуть новый блок
41: завершение функции
```

Machine Translated by Google

4. Оценка эффективности

4.1. Среда моделирования и сценарий

Эксперименты, описанные в этой главе, проводились на персональном компьютере.
с процессором 3,60 ГГц, 16,0 ГБ оперативной памяти и операционной системой Windows 11, работающей под управлением
симулятор оппортунистической сетевой среды [34] (ONE) через IntelliJ IDEA
2021.2.1 работа программного обеспечения. В ходе экспериментов в этом исследовании использовалась маршрутизация Spray и Wait.
в качестве основного алгоритма маршрутизации и сравнил DFIA с RSIA в [26], STIA в [25],
и EEIS в [24] для моделирования. Были разработаны следующие пять сценариев моделирования:

(1) S + эгоистичные узлы;

(2) S + эгоистичные узлы + DFIA;

(3) S + эгоистичные узлы + RSIA;

(4) S + эгоистичные узлы + STIA.

(5) S + эгоистичные узлы + EEIS.

Конкретные настройки экспериментальных параметров представлены в таблице 1.

Таблица 1. Настройки параметров для имитационных экспериментов.

Параметр	Ценить
Время моделирования	12 ч.
Количество узлов	180
Область моделирования	4500 м × 3400 м
Скорость передачи	250 кБ/с
Диапазон передачи	10 м
Модель движения	Кратчайший путьMapBasedMovement
Протокол маршрутизации	Распылить и подождать
Размер буфера	10 М
Срок жизни сообщения	6 часов
Передача энергии	4 Вт
Начальная энергия	2000 Дж
Интервал создания сообщения	25 с–35 с
Размер сообщения	500 КБ–1 МБ
α	0,1
β, γ, σ1 , σ2 ρ	0,3
φ	0,6
	0,45

4.2. Параметры производительности

В симуляционных экспериментах в качестве оценочных показателей в основном используются следующие три параметра:
Критерии оценки эффективности и практичности алгоритма стимулирования.

Вероятность доставки сообщения: это отношение общего количества сообщений, успешно
доставленных к узлу назначения, к общему количеству созданных сообщений. А
более высокая скорость доставки сообщения указывает на большую вероятность успеха узла в пересылке сообщения.
сообщение узлу назначения. Вероятность доставки сообщения чрезвычайно важна.
параметр для оценки производительности сети.

Средняя задержка: это среднее время, затрачиваемое сообщением с момента его создания.
для успешной пересылки к узлу назначения. Меньшая задержка означает, что меньше сети
ресурсы будут потребляться во время пересылки сообщения, и сообщение займет меньше времени
время добраться до места назначения.

Сетевые издержки: они представлены соотношением разницы между
количество сообщений, ретранслируемых узлами сети, и количество успешно
перенаправленных сообщений к числу успешно пересланных сообщений. Сокращение
сетевые издержки указывают на снижение энергопотребления узлом во время
пересылка сообщений.

Machine Translated by Google

4.3. Результаты симуляции

В этом исследовании оценочные эксперименты были разработаны с учетом разного процента эгоистичных узлов и разных TTL. В первой оценке мы наблюдали результаты эксперимента, установив процент эгоистичных узлов на уровне 0%, 20%, 40%, 60%, 80% и 100%. На рисунках 4–6 показаны сравнительные характеристики каждого алгоритма с точки зрения вероятности доставки сообщений, средней задержки и сетевых издержек в трех измерениях, когда в пяти вышеуказанных сценариях существует разное процентное соотношение эгоистичных узлов. Во второй оценке процент эгоистичных узлов был установлен на уровне 60%, а TTL сообщения был установлен на 2 часа, 4 часа, 6 часов, 8 часов и 10 часов соответственно. На рисунках 7–9 показаны изменения вероятности доставки сообщений, средней задержки и сетевых издержек каждого алгоритма для пяти упомянутых выше сценариев при разных значениях TTL сообщения. Всего было проведено 30 испытаний для каждой серии оценочных экспериментов, и среднее значение данных этих испытаний принималось за окончательный экспериментальный результат.

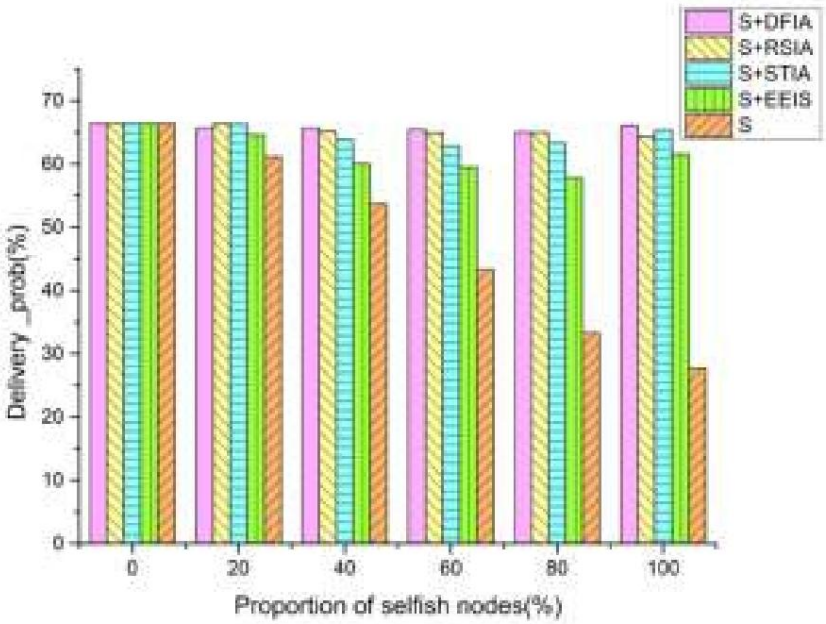


Рисунок 4. Вероятность доставки сообщения с разным процентом эгоистичного узла.

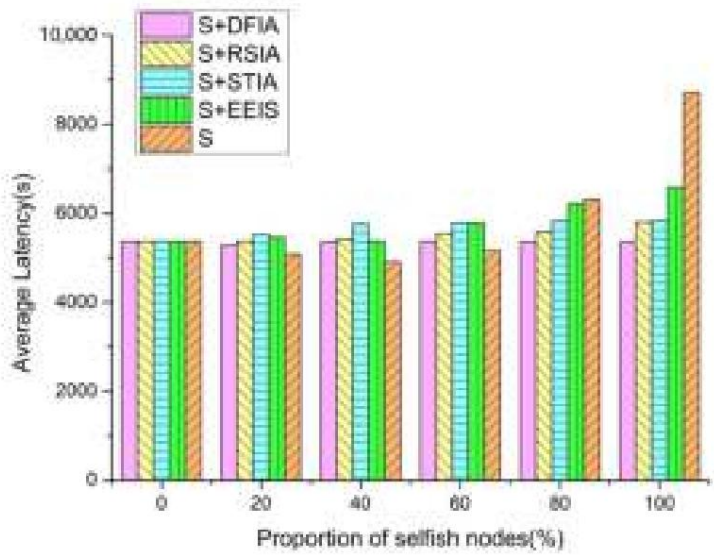


Рисунок 5. Средняя задержка с разным процентом эгоистичных узлов.

Machine Translated by Google

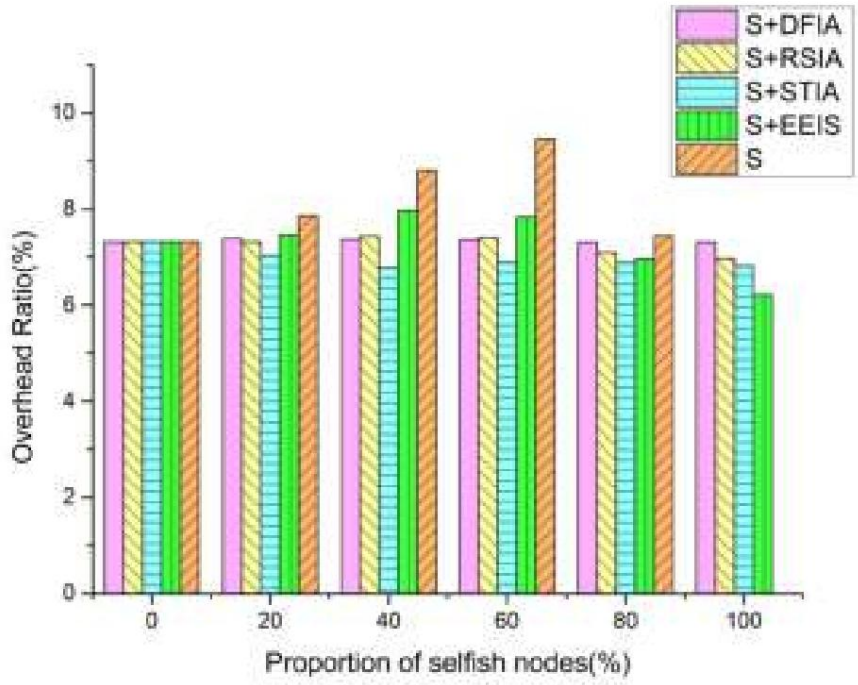


Рисунок 6. Накладные расходы сети для разного процента эгоистичных узлов.

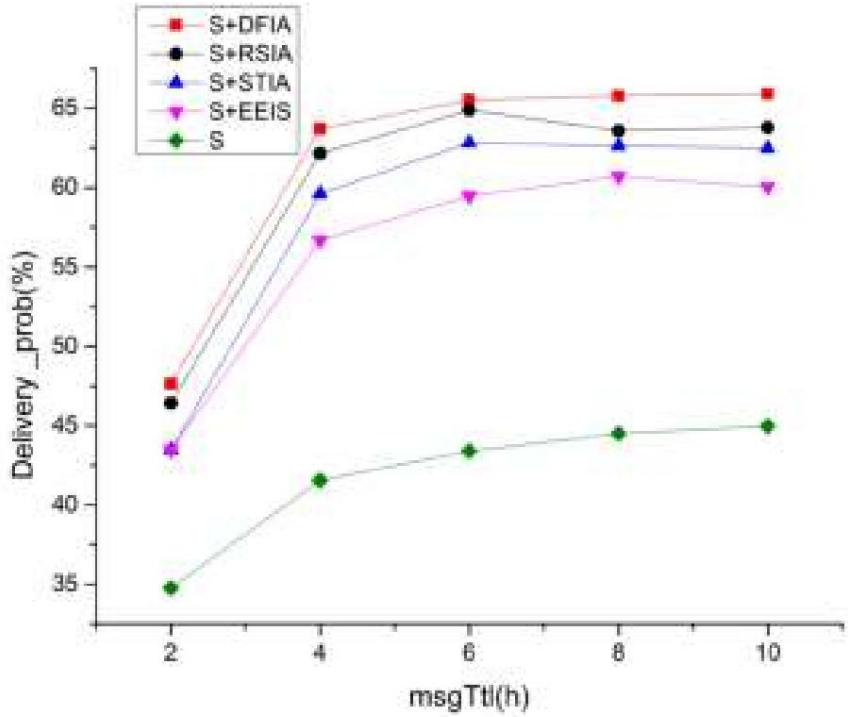


Рисунок 7. Вероятность доставки сообщения при разных TTL сообщения.

Machine Translated by Google

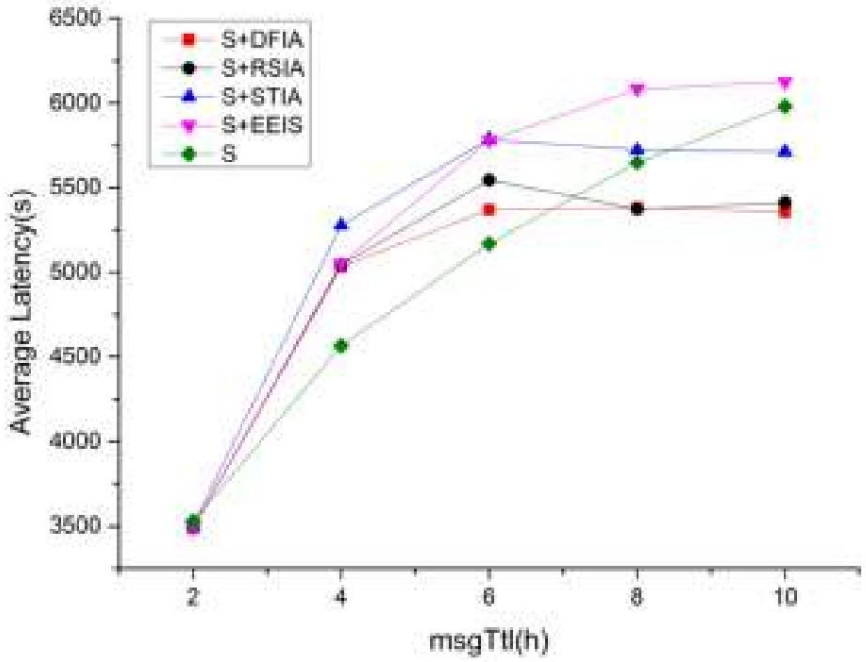


Рисунок 8. Средняя задержка при различных значениях TTL сообщения.

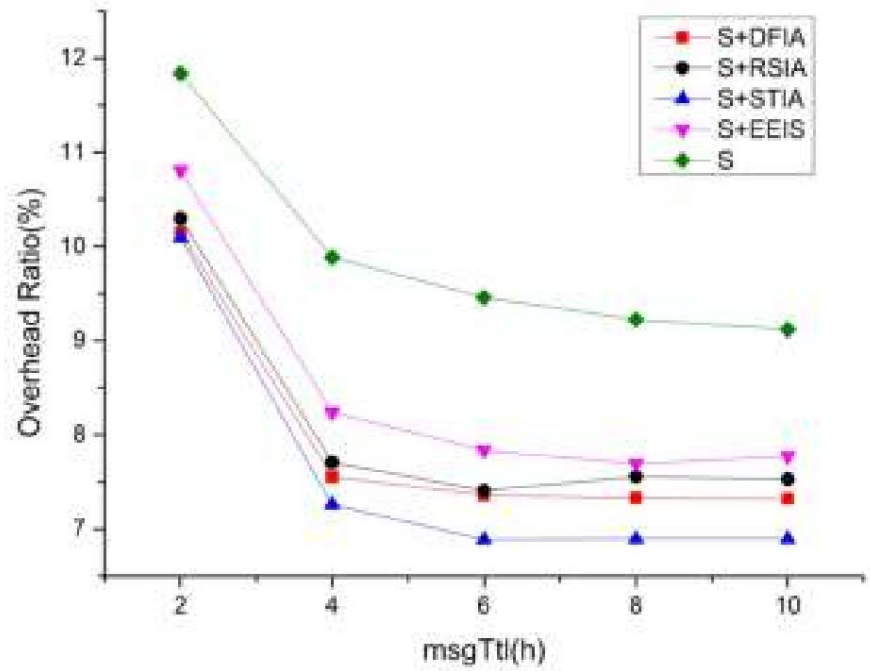


Рисунок 9. Сетевые издержки при разных TTL сообщениях.

4.3.1. Различные проценты эгоистичных узлов

Как показано на рисунке 4, изображена вероятность доставки сообщения для пяти сценариев с различным процентом эгоистичных узлов. Снижение процента доставки сообщений можно объяснить тем фактом, что эгоистичные узлы не участвуют в совместной пересылке, что снижает вероятность успеха сообщений, достигающих узла назначения. Общая тенденция указывает на то, что добавление эгоистичных узлов приводит к тому, что DFIA, RSIA и STIA действуют как стимул в случае узлов с эгоистичным поведением. Однако DFIA более эффективен по сравнению с RSIA и STIA. S имеет более выраженную тенденцию к снижению скорости доставки сообщений, поскольку не использует никаких стимулов. RSIA фокусируется на ограничениях ресурсов и возможном наличии ограничений ресурсов узлов с высоким

Machine Translated by Google

и возможное влияние предпочтительного эгоизма на сеть. STIA учитывает качество пересылки узлов, а также возможные последствия краткосрочной нехватки валюты при принятии решений о пересылочном поведении узлов. EEIS использует ценообразование на ресурсы узла и использует режим эквивалентного обмена для выполнения транзакций. Однако они не учитывают влияние корреляции местоположения между узлами, а также влияние сильных и слабых социальных корреляций между узлами. DFIA всесторонне учитывает характеристики местоположения узла и социальные корреляции. Хотя вероятность доставки сообщения на 0,68% ниже, чем у RSIA, при наличии 20% эгоистичных узлов, при постоянном увеличении количества эгоистичных узлов вероятность доставки DFIA выше, чем у других алгоритмов. По оценке общего эффекта эксперимента, DFIA демонстрирует оптимальную производительность с точки зрения вероятности доставки сообщений.

Рисунок 5 иллюстрирует среднюю задержку для пяти сценариев с разной долей эгоистичных узлов. Увеличение средней задержки означает, что добавление эгоистичных узлов приводит к увеличению времени пересылки сообщения к месту назначения. Данные, представленные на рисунке 5, показывают, что по мере увеличения доли эгоистичных узлов от 0% до 100% средняя задержка увеличивается во всех сценариях с разной степенью увеличения. У S самое значительное увеличение средней задержки, особенно когда доля эгоистичных узлов достигает 60%; его среднее значение задержки резко возрастает, поскольку не принимаются никакие меры стимулирования. В остальных четырех сценариях увеличение средней задержки происходит относительно плавно и не кажется резким. Среди них средняя задержка DFIA является наиболее стабильной, а ее средняя задержка ниже, чем у RSIA, STIA и EEIS. Это указывает на то, что принятие стратегии стимулирования DFIA в большей степени способствует смягчению воздействия эгоистичных узлов, а также повышает вероятность доставки сообщения и одновременно снижает среднюю задержку.

На рисунке 6 показаны сетевые издержки в пяти сценариях при разном проценте эгоистичных узлов. Из данных рисунка видно, что с увеличением доли эгоистичных узлов сетевые издержки остальных четырех сценариев, за исключением S без стимулов, в целом уменьшаются и изменяются более стабильно. В сценарии S накладные расходы сети резко уменьшаются, когда доля эгоистичных узлов превышает 60%, и непосредственно уменьшаются до 0, когда доля эгоистичных узлов в сети равна 100%, поскольку в сети становится все больше и больше эгоистичных узлов. и узлы не могут ретранслировать сообщения или успешно доставлять сообщения до узла назначения из-за отсутствия стратегий стимулирования. Когда доля эгоистичных узлов увеличивается до 60%, общие сетевые издержки DFIA высоки по сравнению с RSIA, STIA, EEIS и S без стимулов. Это связано с тем, что DFIA внедряет технологию блокчейна с использованием RSA для шифрования записей транзакций, что приводит к более высоким общим сетевым издержкам за счет более высоких сетевых издержек, связанных с увеличением вероятности доставки сообщений и обеспечением безопасности транзакции.

4.3.2. Различные TTL

На рисунке 7 показана вероятность доставки сообщения для описанных выше сценариев при различных значениях TTL сообщения. Очевидно, что вероятность доставки сообщения увеличивается с каждым увеличением TTL сообщения. Увеличение TTL сообщения означает, что узлы не ограничены слишком большим количеством временных ограничений при пересылке сообщений. Для пересылки сообщений достаточно времени , что позволяет избежать ситуации потери сообщений и, как следствие, снижения производительности сети. По сравнению с другими четырьмя алгоритмами вероятность доставки сообщений DFIA явно оптимальна, что зависит от того, что DFIA во многих аспектах учитывает факторы, приводящие к эгоизму узлов, что, в свою очередь, мотивирует эгоистичные узлы принимать решения. участие в сотрудничестве. Напротив, RSIA, STIA и EEIS не учитывают влияние релевантности местоположения узлов на узлы и явно не различают силу социальной корреляции, что приводит к менее эффективным стимулам. Видно, что DFIA может эффективно предотвращать потерю сообщений узлами, делая вероятность доставки сообщений оптимальной.

Machine Translated by Google

Рисунок 8 иллюстрирует среднюю задержку четырех сценариев при разных значениях TTL сообщения. Понятно, что средняя задержка увеличивается для всех из них по мере увеличения TTL сообщения. Увеличение TTL сообщения означает, что сообщения имеют больше шансов быть доставленными на большее расстояние в сети, что, в свою очередь, приводит к увеличению количества пакетов в перегруженных узлах при пересылке данных и последующее увеличение задержки в сети. Средняя задержка этих пяти сценариев имеет значительную тенденцию к увеличению с увеличением TTL сообщения, а максимальное увеличение средней задержки наблюдается для S, который не принимает никаких стимулов. Для сравнения, DFIA, RSIA, STIA и EEIS имеют меньшее увеличение средней задержки. Среди них средняя задержка DFIA ниже по сравнению с RSIA, STIA и EEIS, что связано с тем, что, хотя STIA учитывает ресурсы узла и качество обслуживания , он игнорирует фактор социальной корреляции узла, что увеличивает среднюю задержку. сообщения в определенной степени. RSIA учитывает только эгоистические причины узлов, но не учитывает позиционные отношения между ними и различает сильные и слабые социальные корреляции, что приводит к более высокой средней латентности. EEIS использует квадратичный торг и не учитывает, что сообщение не может быть передано вовремя из-за потери пакетов узла, что влияет на среднюю задержку. DFIA учитывает позиционные взаимоотношения между узлами и вводит фактор социальной корреляции, тем самым повышая скорость доставки сообщения и уменьшая среднюю задержку.

Рисунок 9 иллюстрирует сравнение сетевых издержек. По мере увеличения TTL сообщения сетевые издержки во всех сценариях постепенно уменьшаются. Поскольку S не принимает никаких мер стимулирования, его сетевые издержки выше, чем у DFIA, RSIA, STIA и EEIS. STIA учитывает качество обслуживания узлов и сокращает взаимодействие сообщений в процессе переговоров, тем самым оптимизируя время доставки сообщений и снижая сетевые издержки. EEIS учитывает только цену ресурсов узла и использует кредитную стоимость для денежного кредитования, но не учитывает поведение потери пакетов, вызванное снижением пропускной способности транзакций из-за чрезмерного заимствования и невозможности погашения, что приводит к высоким издержкам сети. DFIA учитывает ресурсы узлов, атрибуты местоположения и социальные корреляции в переговорной игре, что делает его стимулирующий эффект более значимым , чем у RSIA, а также имеет меньшие сетевые издержки. Благодаря учету множества факторов во время переговоров в DFIA генерируется больше информационных взаимодействий, а накладные расходы сети выше по сравнению с STIA. Но DFIA увеличивает вероятность доставки сообщений и снижает среднюю задержку, а общий стимулирующий эффект лучше, чем у STIA.

5. Выводы

В этой статье мы предлагаем основанный на ресурсах алгоритм динамического ценообразования и принудительной пересылки (DFIA). DFIA обеспечивает услуги пересылки, необходимые узлам, путем интеграции срочности доставки, определяемой состоянием ресурсов узла, актуальностью местоположения и различием между сильными и слабыми социальными корреляциями. В процессе транзакции стратегия принудительной пересылки предназначена для повышения степени сотрудничества узлов. Чтобы предотвратить ложные котировки узлов, DFIA использует технологию блокчейна для шифрования, чтобы добиться прозрачности и достоверности данных транзакций . Результаты экспериментов по моделированию показывают, что DFIA снижает среднюю задержку, одновременно увеличивая вероятность доставки сообщений и эффективно стимулируя эгоистичные узлы к участию в сотрудничестве. Однако алгоритм использует RSA для шифрования записей транзакций, что приводит к увеличению общих сетевых издержек, которые необходимо учитывать и оптимизировать в будущей работе. Далее мы углубимся в меры безопасности, которые следует принять в случае возможного вредоносного поведения, такого как сговор узлов друг с другом и подделка сообщений в SAN.

Machine Translated by Google

Вклад авторов: концептуализация, XZ, YL (Юань Ли) и ZX; методология YL (Юань Ли); программное обеспечение YL (Юань Ли); проверка, ZX, YL (Юань Ли) и XZ; ресурсы XZ, YL (Юань Ли) и ZX; курирование данных, YL (Юань Ли); письмо — подготовка оригинального проекта, YL (Юань Ли) и DH; написание-обзор и редактирование, ZX, YL (Яньчао Лю) и SW; администрация проекта, ZX Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Исследование финансировалось Национальным фондом естественных наук Китая (номер гранта № 61972136) и Фондом поддержки группы выдающихся молодых научных инноваций Департамента образования провинции Хубэй (номер гранта № T201410, T2020017).

Заявление о доступности данных: данные содержатся в статье.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Ван З.; Ляо, Дж.; Цао, Кью; Ци, Х.; Ван, З. Достижение покрытия k-барьера в гибридных сетях направленных датчиков. IEEE Транс. Моб. Вычислить. 2014, 13, 1443–1455. [Перекрестная ссылка]
2. Цугава, С.; Осаки, Х. Структура сообщества и локальность взаимодействия в социальных сетях. Ж. Инф. Процесс. 2015, 23, 402–410. [Перекрестная ссылка]
3. Нин, З.; Лю, Л.; Ся, Ф.; Джедари, Б.; Ли, И.; Чжан, В. CAIS: Копируемая регулируемая схема стимулирования в социальной сфере на уровне сообщества Осознанная сеть. IEEE Транс. Вех. Технол. 2017, 66, 3406–3419. [Перекрестная ссылка]
4. Сюй, Ф.; Сюй, Кью; Сюн, З.; Сяо, Н.; Се, Ю.; Дэн, М.; Хао, Х. Интеллектуальная схема распределенной маршрутизации, основанная на социальном сходстве для мобильных социальных сетей. Будущий генерал. Вычислить. Сист. 2019, 96, 472–480. [Перекрестная ссылка]
5. Чжу, К.; Ли, В.; Фу, Х. SMART: Стратегия маршрутизации с учетом социальных сетей и мобильных устройств для устойчивых к сбоям сетей. IEEE Транс. Вех. Технол. 2014, 63, 3423–3434. [Перекрестная ссылка]
6. Анналиса, С.; Антонио, К.; Флориано, ДР; Пеппино, Ф. Маршрутизация в мобильных оппортунистических социальных сетях с эгоистичными узлами. Вирел. Коммун. Моб. Вычислить. 2019, 2019, 6359806.
7. Ся, Ф.; Лю, Л.; Джедари, Б.; Дас, SK PIS: протокол многомерной маршрутизации для социально-ориентированных сетей. IEEE Транс. Моб. Вычислить. 2016, 15, 2825–2836. [Перекрестная ссылка]
8. Галати, А. Сеть, устойчивая к задержкам; LAP Lambert Academic Publishing: Кельн, Германия, 2010.
9. Лусиана, П.; Андреа, П.; Марко, К. Оппортунистические сети: пересылка данных в отключенных мобильных одноранговых сетях. IEEE Коммун. Маг. 2006, 44, 134–141.
10. Ся, Ф.; Лю, Л.; Ли, Дж.; Мохаммед, А.А.; Тианруо, ЮЛ; Цзяньхуа, М. BEEINFO: переадресация на основе интересов с использованием искусственной пчелы Колония социально ориентированных сетей. IEEE Транс. Вех. Технол. 2015, 64, 1188–1200. [Перекрестная ссылка]
11. Ву Дж.; Ю, Г.; Гуан, П. Метод предсказания вероятностных характеристик интересов в социальных оппортунистических сетях. IEEE Access 2019, 7, 59002–59012. [Перекрестная ссылка]
12. Эльназ, Н.; Нахиде, Д. FTR: Особенности маршрутизации на основе дерева в мобильных социальных сетях. Вирел. Сеть. 2020, 26, 283–291.
13. Мао Ю.; Чжоу, К.; Ци, Дж.; Чжу, Х. Справедливый механизм стимулирования на основе кредитов для маршрутизации в сенсорной сети на основе DTN с узлами эгоизм. ЕВРАЗИП Дж. Вирел. Коммун. Сеть. 2020, 2020, 232. [CrossRef]
14. Лоудари, ЮВ; Бенамар, Н. Влияние эгоизма на потребление энергии в оппортунистических сетях: оценка производительности. В материалах Международной конференции по беспроводным технологиям, встраиваемым и интеллектуальным системам (WITS) 2019 г., Фес, Марокко, 3–4 апреля 2019 г.; стр. 1–7.
15. Ли, К.; Гао, В.; Чжу, С.; Цао, Г. Протокол маршрутизации для социально эгоистичных сетей, устойчивых к задержкам. Специальная сеть. 2012, 10, 1619–1632. [Перекрестная ссылка]
16. Чжан Х.; Дэн, Х.; Сюн, З.; Лю, Ю.; Рао, Ю.; Лю, Ю.; Ли, Ю.; Хоу, Д.; Ли, Ю. Стратегия безопасной маршрутизации на основе атрибутов Доверительный контроль доступа в социальных сетях. J. Сигнальный процесс. Сист. 2024, 96, 153–168. [Перекрестная ссылка]
17. Чанг Дж.; Панг, З.; Сюй, В.; Ван, Х.; Инь, Г. Механизм репутации, совместимый со стимулами, для P2P-систем. Дж. Суперкомпьютер. 2014, 69, 1382–1409. [Перекрестная ссылка]
18. Рой, А.; Ачарья, Т.; Дас Бит, С. Пересылка данных с учетом репутации на основе социальных сетей для улучшения многоадресной доставки в присутствии эгоистичных узлов в сетях DTN. Межд. Ж. Коммун. Сист. 2020, 33, e4235. [Перекрестная ссылка]
19. Бунуни, М.; Буаллуш-Меджкун, Л. Схема наказания и стимулирования на основе признания для специальной мобильной сети. Дж. Суперкомпьютер. 2018, 74, 5373–5398. [Перекрестная ссылка]
20. Ван, ЕК; Чен, СМ; Ю, С.М.; Хасан, ММ; Алрубайан, М.; Фортино, Г. Модель стимулирующей эволюционной игры для оппортунистических социальные сети. Будущий генерал. Вычислить. Сист. 2020, 102, 14–29. [Перекрестная ссылка]
21. Ву Дж.; Мо, В.; Инь, Х.; Бай, Г. Исследование механизма стимулирования оппортунистической маршрутизации, основанного на повторяющейся игре в WMN. Вычислить. англ. Прил. 2018, 54, 5.
22. Хак, А.; Фахим, Ю. Протокол распределения контента на основе одноранговой связи, устойчивый к задержкам с учетом стимулов. сети. Вирел. Сеть. 2020, 26, 583–601. [Перекрестная ссылка]
23. Ли, Л.; Чжун, Х.; Цзян, Ю. IAR-GT: Маршрутизация с учетом стимулов, основанная на теории игр для эгоистичных оппортунистических сетей. КСИИ Пер. Интернет Инф. Сист. 2019, 13, 152–171.

Machine Translated by Google

24. Сюн, З.; Сяо, Н.; Сюй, Ф.; Чжан, Х.; Сюй, Кью; Чжан, К.; Да, С. Эквивалентная схема стимулирования пересылки данных на основе обмена для социально осведомленных сетей. J. Сигнальный процесс. Сист. 2021, 93, 249–263. [Перекрестная ссылка]
25. Сюн, З.; Ли, Х.; Чжан, Х.; Чжу, С.; Сюй, Ф.; Чжао, Х.; Ву, Ю.; Цзэн, М. Двухэтапный алгоритм стимулирования на основе ценообразования для социально осведомленные сети. J. Сигнальный процесс. Сист. 2022, 94, 1227–1242.
26. Чжан Х.; Рао, Ю.; Сюн, З.; Дэн, Х.; Сюй, Ф.; Юань, Л. Алгоритм стимулирования, основанный на ограниченности ресурсов и социально эгоистичный социально осведомленные сети. J. Сигнальный процесс. Сист. 2023, 95, 1439–1453.
27. Спиропулос, Т.; Псунис, К.; Рагхавендра, CS. Распыли и подожди: эффективная схема маршрутизации для периодически подключаемых мобильных сетей. В материалах семинара ACM SIGCOMM 2005 г. по сетям, устойчивым к задержкам, Нью-Йорк, штат Нью-Йорк, США, 26 августа 2005 г.; стр. 252–259.
28. Кюи, Ж.; Лю, С.; Чанг, Ю.; Лю, К.; Ву, К. Эффективный алгоритм маршрутизации на основе транзакций виртуальной валюты в DTN. Вычислить. наук. 2024, 51, 300–308.
29. Ли, Х.; Цзян, П.; Чен, Т.; Луо, Х.; Вэнь, К. Исследование безопасности систем блокчейна. Будущий генерал. Вычислить. Сист. 2020, 107, 841–853. [Перекрестная ссылка]
30. Дэн, М.; Лю, Ю.; Ян, К.; Сюй, Ф.; Ахмед, М.; Ян, Н.; Сюй, З.; Ке, К. Облегченная схема доверительного управления на основе блокчейна в интеллектуальных системах Интернета вещей с ограниченными ресурсами. IEEE Internet Things J. 2024, 11, 25706–25719. [Перекрестная ссылка]
31. Мацумото, С.; Рейщук, Р.М. IKP: Преобразование PKI с помощью децентрализованных автоматизированных стимулов. В материалах 2017 года Симпозиум IEEE по безопасности и конфиденциальности (SP), Сан-Хосе, Калифорния, США, 22–26 мая 2017 г.; стр. 410–426.
32. Он, Ю.; Ли, Х.; Ченг, Х.; Лю, Ю.; Ян, К.; Сан, Л. Правдивый механизм стимулирования на основе блокчейна для распределенных P2P -приложений. IEEE Access 2018, 6, 27324–27335. [Перекрестная ссылка]
33. Чжан, В.; Ван, Х.; Хуа, Х.; Ван, К. Эгоистичный майнинг и стратегии защиты биткойнов. Интел. Автомат. Мягкий компьютер. 2022, 34, 1861–1875. [Перекрестная ссылка]
34. Керянен А.; Отт, Дж.; Кярккяйнен, Т. ОДИН симулятор для оценки протокола DTN. В трудах 2-го Интернационала Конференция по инструментам и методам моделирования, Рим, Италия, 2–6 марта 2009 г.; ИКСТ: Брюссель, Бельгия, 2010 г.; стр. 1–10.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Реализация FPGA основанной на принципах классификации объектов для Автономный мобильный робот

Парк Чаун [1], Сонджу Ли [2,3] и Юнхо Чон [1,4,*]

[1] Школа электроники и информационной инженерии, Корейский аэрокосмический университет, Коян-си 10540, Республика Корея;
[2] pcw0201@kau.kr Факультет электротехники, Университет Седжон, Сеул 05006, Республика Корея; seongjoo@sejong.ac.kr
[3] Факультет конвергентной разработки интеллектуальных дронов, Университет Седжон, Сеул 05006, Республика Корея
[4] Факультет интеллектуальной воздушной мобильности, Корейский аэрокосмический университет, Коян-си 10540, Республика Корея * Адрес для переписки: yjung@kau.ac .kr; Тел.: +82-2-300-0133

Аннотация: С развитием технологий искусственного интеллекта автономные мобильные роботы стали использоваться в различных приложениях. В сценариях автономного вождения классификация объектов важна для навигации робота. Для выполнения этой задачи датчики обнаружения света и измерения дальности (LiDAR), которые могут получать информацию о глубине и высоте и имеют более высокое разрешение, чем датчики радиообнаружения и измерения дальности (радары), предпочтительнее датчиков камеры. В методе, основанном на столбах, используется кодировщик объектов столбов (PFE) для кодирования 3D-облаков точек LiDAR в 2D-изображения, что обеспечивает высокоскоростной вывод с использованием 2D-сверточных нейронных сетей. Хотя метод на основе столбов используется для обеспечения реагирования систем автономного вождения в реальном времени, исследования по ускорению PFE активно не проводятся, хотя PFE потребляет значительное количество вычислительного времени внутри системы. Поэтому в этой статье предлагается аппаратный ускоритель PFE и модель классификации объектов на основе столбцов для автономных мобильных роботов. Предложенная модель классификации объектов была обучена и протестирована с использованием 2971 набора данных, состоящего из восьми классов, с достижением точности классификации 94,3%. Аппаратный ускоритель PFE был реализован в виде программируемой вентильной матрицы (FPGA) посредством конструкции уровня передачи регистров, что позволило добиться ускорения в 40 раз по сравнению с прошивкой для микропроцессорного блока ARM Cortex-A53; сеть классификации объектов была реализована в FPGA с использованием структуры FINN. Интегрировав PFE и сеть классификации объектов, мы реализовали систему ускорения классификации объектов на основе столбцов в реальном времени на FPGA с задержкой 6,41 мс.

Ключевые слова: программируемая вентильная матрица; автономный мобильный робот; классификация объектов; кодировщик функций столба ; ФИНН

check for updates

Цитирование: Парк, К.; Ли, С.; Юнг, Ю. Реализация FPGA основанной на столбцах классификации объектов для автономной Мобильный робот. Электроника 2024, 13, 3035. https://doi.org/10.3390/электроника13153035

Академический редактор: Никола Лусарди

Поступила: 14 июня 2024 г.
Пересмотрено: 29 июля 2024 г.
Принято: 30 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Автономные мобильные роботы передвигаются, распознавая свое окружение и выполняя физические задачи, выполняемые людьми в прошлом. Достижения в области компьютерного зрения, управления роботами и искусственного интеллекта (ИИ) расширили способность автономных мобильных роботов воспринимать окружающую среду и реагировать на нее. Следовательно, эти роботы используются в нескольких сферах человеческой жизни, включая производство, транспорт, здравоохранение и повседневную жизнь [1,2]. Классификация объектов необходима автономным мобильным роботам для идентификации объектов и планирования маршрутов перемещения [3].

Различные типы датчиков, такие как камеры и датчики радиообнаружения и измерения дальности (радары), использовались для восприятия окружающей среды в ситуациях автономного вождения. Датчики камеры не могут получить информацию о глубине, что делает невозможным точное определение расстояния до объекта, что может привести к серьезным проблемам, когда два объекта перекрываются [4]. Кроме того, датчики камеры чувствительны к яркости и имеют ограниченный диапазон измер

Machine Translated by Google

поле зрения, что приводит к эксплуатационным ограничениям [3]. Радарные датчики не могут определить информацию о высоте объектов и имеют очень низкое разрешение по углам, что препятствует точному наблюдению за объектами [5]. Напротив, облака точек, полученные с помощью датчиков обнаружения света и измерения дальности (LiDAR), содержат трехмерную информацию о физических координатах, которая позволяет получать информацию о глубине и высоте объектов. Кроме того, здесь меньше эксплуатационных ограничений, а высокое разрешение позволяет точно наблюдать объекты [3]. Поэтому использование облаков точек, полученных с помощью датчиков LiDAR, предпочтительнее для автономного вождения.

Для обработки облаков точек доступны различные алгоритмы глубокого обучения. MVCNN [6] проецирует облако точек в нескольких направлениях и классифицирует объекты с помощью сверточной нейронной сети (CNN) на основе проецируемых изображений. Хотя он использует простой метод преобразования 3D-данных в 2D-изображения и использует CNN, в процессе проецирования происходит потеря информации [7]. PointNet [8] — это инвариантная к перестановкам сеть, разработанная с использованием функции симметрии для обработки неупорядоченных наборов точек. VoxelNet [9] кодирует функции по вокселям с помощью PointNet и делает выводы, используя сеть предложений регионов. Поскольку вывод основан на 3D CNN, требуется длительное время вывода — 225 мс [10]. SECOND [11] достиг улучшенной скорости вывода до 50 мс, применив разреженную CNN к VoxelNet, но сложность вычислений 3D CNN оставалась высокой. PointPillars [12] предложил метод столбцов для кодирования 3D-облаков точек в 2D-изображения. В отличие от MVCNN, который использует информацию только из самого высокого элемента, метод на основе столбцов использует информацию из всех точек внутри столбца. Проведение двумерного вывода на основе CNN по закодированному изображению позволяет методам на основе столбцов достигать высокоскоростного вывода и превосходной производительности [13–16].

Автономные мобильные роботы должны иметь время реагирования в реальном времени, чтобы предотвратить физические несчастные случаи. В этих условиях запуск методов на основе микропроцессорных блоков (MPU) требует чрезвычайно много времени [17]. Графические процессоры (GPU) могут ускорить вывод на основе 2D CNN; однако использование графических процессоров на периферийных устройствах неэффективно с точки зрения энергопотребления [18]. Поэтому исх. [19–21] реализовали аппаратный ускоритель для столбчатых методов с использованием программируемой вентильной матрицы (FPGA). Однако в [19] на MPU был рассчитан кодировщик основных функций (PFE) , требующий 71,2 мс и 19% от общего времени отклика. Аналогично, PFE был вычислен на MPU за 9,51 мс в [20], что составило 22% от общего времени отклика. Кроме того, в [21] была ускорена часть PFE , требующая для вычислений 37,7 мс, что составляло 59% общего времени отклика. Поскольку PFE занимает значительную часть общего времени отклика метода на основе столбцов, полезен аппаратный ускоритель для PFE.

Метод на основе столбов может выполнять логический вывод по изображению, закодированному PFE, с помощью сети, основанной на разделимой по глубине CNN (DSCNN). DSCNN, использованная в [22–24], получила широкое распространение в последних моделях CNN благодаря низкой стоимости вычислений и количеству параметров. Из-за этих преимуществ DSCNN часто используются в приложениях с ограниченными ресурсами или чувствительными к задержке [25]. Кроме того, несмотря на свою низкую сложность , они показали успешную работу в моделях классификации изображений за счет повышения эффективности представления [26].

В этом исследовании мы предлагаем эффективную модель классификации объектов на основе столбов для автономных мобильных роботов с использованием PFE и DSCNN, достигая точности классификации 94,3% за счет обучения и проверки с использованием наборов данных, полученных с помощью Ouster OS1-32 [27]. Кроме того, мы предлагаем систему ускорения на основе FPGA для работы в реальном времени предлагаемой модели классификации объектов. Предлагаемая система ускорения включает в себя ускоритель PFE и ускоритель сети классификации. Мы разработали ускоритель PFE для достижения времени вычислений 0,23 мс и подтвердили ускорение в 40 раз по сравнению с MPU. Ускоритель сети классификации был реализован на FPGA с использованием FINN. Для работы в периферийных средах мы реализовали предложенную систему ускорения на FPGA и подтвердили время отклика в реальном времени 6,41 мс.

Оставшаяся часть статьи структурирована следующим образом. В разделе 2 представлен обзор предлагаемой системы ускорения, включая классификацию на основе PFE и DSCNN.

Machine Translated by Google

сети, используемый набор данных, конфигурация модели и результаты оценки. Конструкция предлагаемой системы ускорения описана в разделе 3. В разделе 4 представлена реализация результаты и эффекты ускорения. Наконец, раздел 5 завершает исследование.

2. Предлагаемая система

На рисунке 1 представлен обзор предлагаемой системы ускорения. Облако точек входные данные в систему кодируются в изображение с помощью PFE в три этапа: (a) преобразование облако точек в набор основных элементов, (b) добавление элементов, созданных вручную (HCF), и (c) расширение функций посредством точечной (PW) свертки, сжатие их посредством max-pool и раскидываем их обратно в исходное положение на плоскости x–y , чтобы сгенерировать изображение. Сгенерированное изображение затем передается в классификационную сеть для классификации. объекты. Классификационная сеть включает сеть на основе DSCNN.

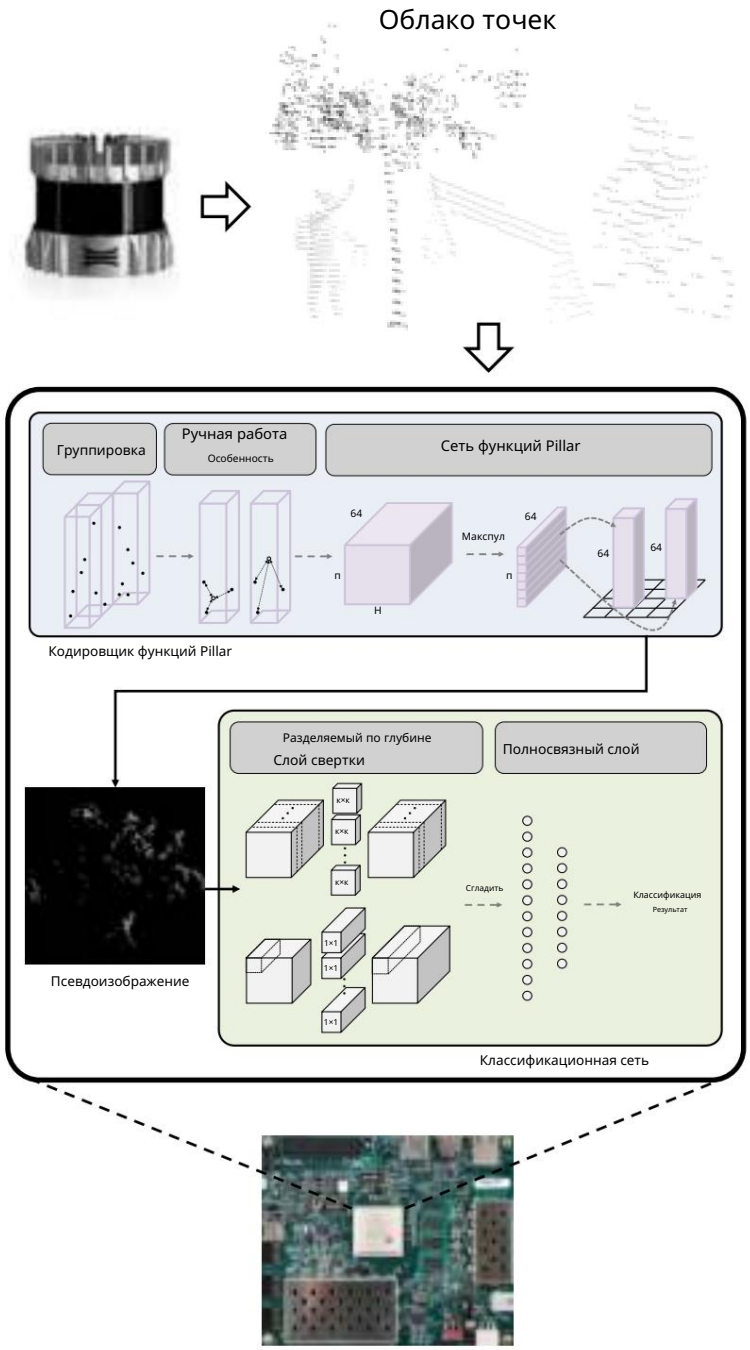


Рисунок 1. Обзор предлагаемой системы ускорения.

Machine Translated by Google

2.1. ПФЭ

PFE, который кодирует трехмерные облака точек в двумерные изображения для обеспечения двумерного вывода на основе CNN, был предложен в [12]. Сначала создается сетка путем дискретизации плоскости x–y , а затем облако точек преобразуется в набор столбчатых единиц. Множество облаков точек g определяется уравнением (1).

$$г = [p1 , p2 , . . . Пи , . . . п.п.] \quad (1)$$

Мы определяем g как набор чисел pi , где P — максимальное количество столбов, содержащих точки. Мы устанавливаем число pi , определенное уравнением (2).

$$пи = [n1, n2, . . . ни , . . . нН] \quad (2)$$

Мы определяем pi как наборnj , где N — максимальное количество точек, которые может содержать столб . Набор nj определяется как в уравнении (3).

$$ni = [x, y, z,r] \quad (3)$$

Набор nj содержит координаты x, y, z и значения отражательной способности точек. Наконец, облако точек представляется тензором размера (P, N, 4), где P и N — гиперпараметры. Если количество точек в столбце больше или равно N, используется случайная выборка; если оно меньше N, используется заполнение нулями.

После того, как облако точек преобразуется в набор единиц столбцов, к набору nj добавляется HCF . HCF включает в себя разности xm, ym , zm между среднеарифметическими координатами точек внутри столба и координатами x, y, z каждой точки, а также разности xc, yc между координатами центра столба и координатами x, y. координаты каждой точки. Установка nj с добавленным HCF определяется уравнением (4).

$$ni = [x, y, z,r, xm, ym, zm, xc, yc] \quad (4)$$

Тензор размера (P, N, 9) с добавленным HCF подвергается свертке PW, пакетной нормализации и слою ReLU, чтобы преобразовать его в тензор размера (P, N, 64). Впоследствии максимальное объединение применяется вдоль второго измерения, чтобы получить тензор размера (P, 64).

Наконец, каждый столб рассеивается обратно в исходную плоскость x-y , завершая данные 2D-изображения с 64 каналами.

2.2. Классификационная

сеть Предлагаемая классификационная сеть основана на DSCNN. На рисунке 2 показано, как работает DSCNN. DSCNN включает в себя методы по глубине (DW) и PW. В DW количество каналов во входном изображении, фильтров и каналов в выходном изображении одинаково. Каждый фильтр имеет только один канал и выполняет операцию свертки только на одном канале входного изображения для создания одноканального выходного изображения. Например, в свертке DW, показанной на рисунке 2, один канал входного изображения, разделенный пунктирной линией, выполняет операцию свертки с одним фильтром для создания одного канала выходного изображения, также разделенного пунктирной линией. Пусть iC — количество каналов во входном изображении, а oC — количество каналов в выходном изображении; PW использует фильтры oC размером 1 × 1 с каналами iC . Каждый фильтр выполняет операцию свертки на всех каналах входного изображения для создания одного канала выходного изображения. Например, в свертке PW, показанной на рисунке 2, один пиксель входного изображения, разделенный пунктирной линией, дополняет один пиксель выходного изображения, разделенный пунктирной линией, путем применения всех фильтров в операции свертки.

Machine Translated by Google

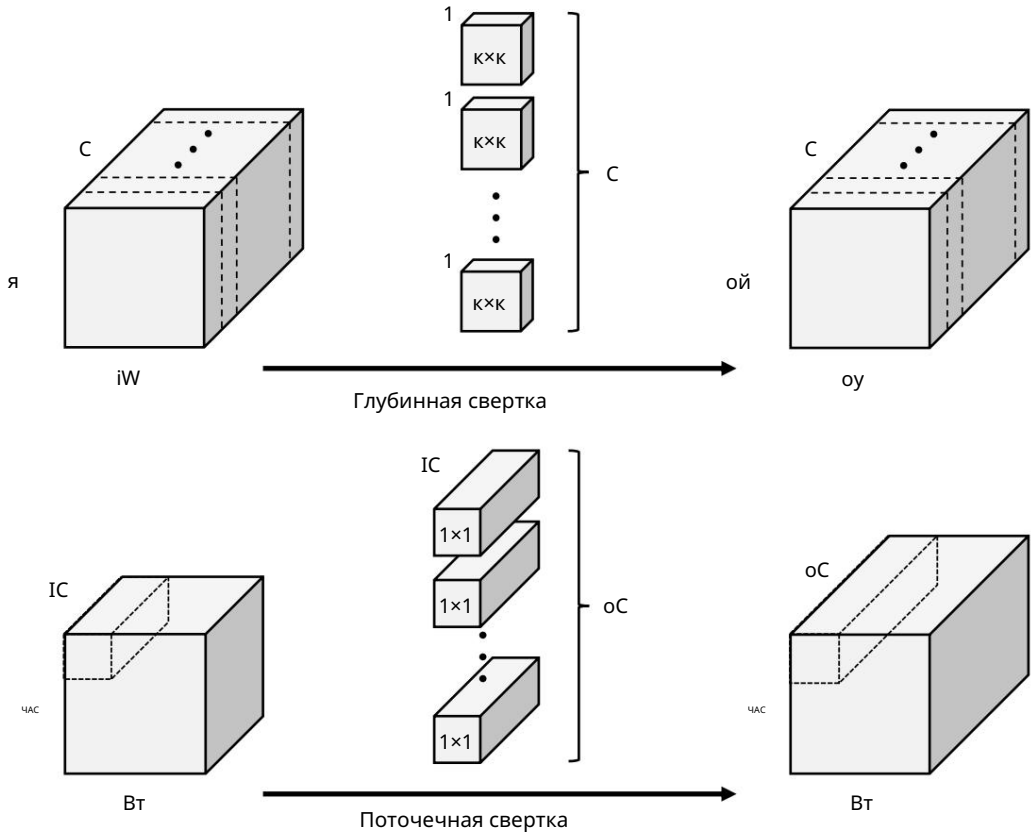
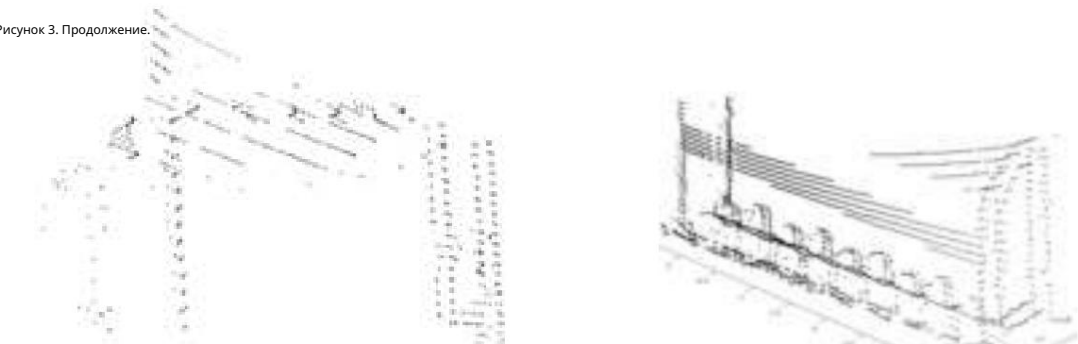


Рисунок 2. Механизм работы DSCNN.

2.3. Набор данных

Для обучения предложенной модели классификации объектов мы получили облака точек с помощью датчика Ouster OS1-32. На рисунке 3 показаны примеры из набора данных, включающего объекты, с которыми обычно сталкиваются автономные мобильные роботы в условиях вождения. Чтобы повысить качество обучения, мы дополнили набор данных образцами из набора данных nuScenes [28] . Набор данных включает в себя такие классы, как здание, дерево, транспортное средство, велосипед, препятствие, зелень , человек и городской объект (уличный фонарь, стойка для велосипедов, светофор и т. д.), как показано на рисунке 4, где показана конфигурация набор данных. Из 2971 точки данных 2617 использовались для обучения и 354 — для проверки.

Рисунок 3. Продолжение.



Machine Translated by Google

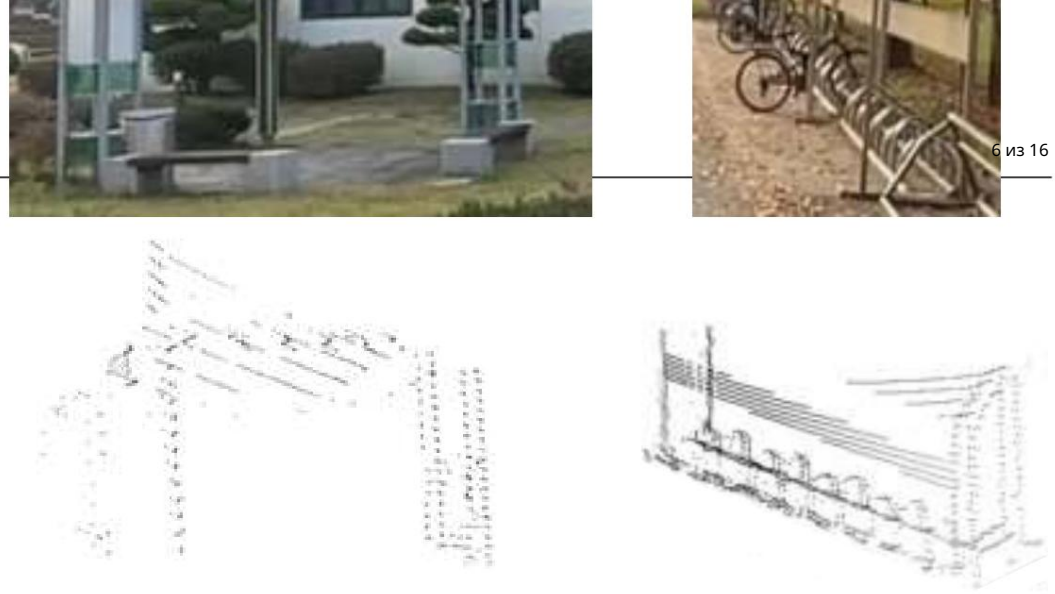


Рисунок 3. Примеры наборов данных.

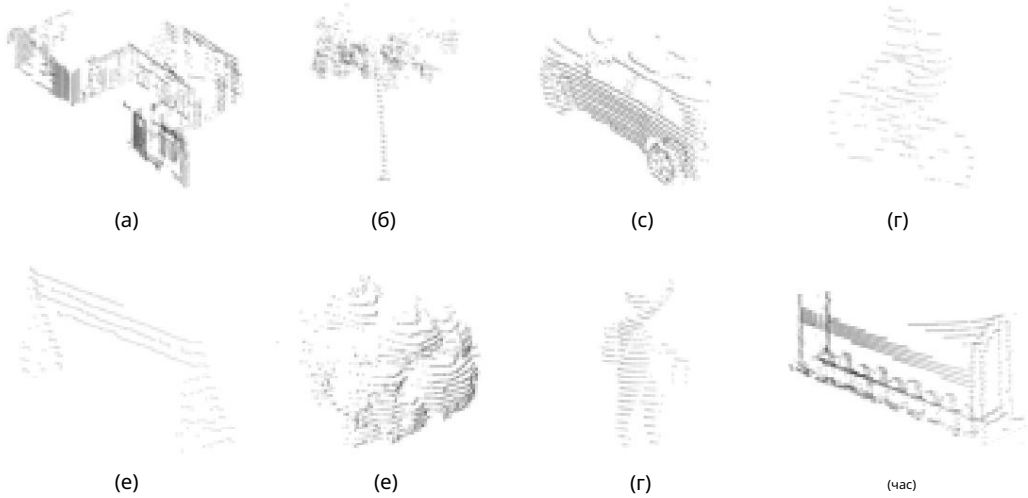


Рисунок 4. Конфигурация классов набора данных: (а) построение; (б) дерево; (в) транспортное средство; (г) велосипед; (д) препятствие; (е) зелень; (ж) человек; (h) городское приспособление.

2.4. Оценка эффективности

В PFE размер изображения, количество столбцов и количество точек на столбец гиперпараметры, которые существенно влияют как на производительность, так и на сложность модель. Размер изображения определяет разрешение и размеры закодированного изображения. Увеличение размера изображения приводит к получению изображений с более высоким разрешением; однако это также увеличивает вычислительные потребности CNN. Аналогичным образом увеличивая количество столбов и точек на столб позволяет получить более подробную информацию из облака точек. Однако, это также увеличивает вычислительную нагрузку PFE из-за увеличения количества вычислений, необходимых для обработки большего количества данных. Таким образом, нахождение сбалансированного сочетание этих гиперпараметров имеет решающее значение для достижения как высокой производительности, так и эффективные вычисления в моделях классификации объектов на основе столбцов.

Рисунок 5 иллюстрирует архитектуры сетей, используемых для оценки производительности , все из которых были основаны на DSCNN. В табл. 1 представлены точности, соответствующие к различным гиперпараметрам и сетевым структурам, при этом общая точность служит метрика оценки для классификации объектов. Мы обучили 300 эпох с размером партии 16, а результаты наиболее результативных эпох представлены в Таблице 1. Обучение было выполнено со скоростью обучения 0,001 и скоростью снижения веса 0,0001. Во-первых, это было замечено, что сети 1 и 5 показали более низкую производительность, чем сети 2, 3, и 4. Среди сетей 2, 3 и 4 сеть 4 выделяется своей сбалансированной производительностью, демонстрируя точность, сравнимую с сетями 2 и 3, при этом демонстрируя меньшую вычислительную сложность и меньшее количество параметров. Дополнительно выставлено изображение размером 128×128. выдающаяся производительность, тогда как производительность была сопоставима, когда количество столбов было 1024 и 512. Наилучшие показатели были достигнуты при попадании 16 точек внутрь.

Machine Translated by Google

каждый столб. Следовательно, предложенная модель классификации объектов приняла размер изображения 128 × 128, 512 столбцов и 16 точек внутри каждого столбца, с классификационной сетью структурировано по сети 4.

Таблица 1. Точность модели по различным гиперпараметрам и сетям.

Гиперпараметр			Сеть				
Размер изображения	Число Столбов	Количество Баллы за компонент	1	2	3	4	5
256 × 256	2048	32	91,8%	93,5%	93,5%	92,9%	89,7%
		16	92,9%	92,9%	92,4%	92,4%	88,0%
		8	92,9%	92,9%	92,9%	92,4%	89,7%
	1024	32	92,4%	92,4%	91,8%	92,9%	89,1%
		16	92,4%	92,9%	92,4%	92,9%	90,2%
		8	93,5%	93,5%	92,4%	92,4%	89,7%
	512	32	92,4%	92,9%	93,5%	93,5%	89,7%
		16	92,9%	94,0%	92,4%	93,5%	89,7%
		8	94,6%	92,9%	92,4%	91,8%	90,8%
128 × 128	1024	32	90,8%	91,8%	92,4%	94,6%	90,8%
		16	90,8%	94,6%	92,9%	92,4%	91,3%
		8	90,8%	91,8%	91,8%	94,0%	91,8%
	512	32	90,8%	94,0%	91,8%	93,5%	92,4%
		16	91,3%	92,9%	93,5%	94,6%	92,9%
		8	90,8%	93,5%	91,8%	92,9%	92,4%
	256	32	91,8%	93,5%	92,9%	91,8%	90,8%
		16	90,8%	92,9%	92,4%	91,3%	90,8%
		8	90,8%	94,0%	94,0%	92,4%	91,3%
64 × 64	512	32	89,7%	89,7%	92,9%	91,8%	91,3%
		16	88,6%	90,2%	93,5%	91,8%	91,8%
		8	88,6%	90,8%	92,9%	92,4%	92,4%
	256	32	88,6%	90,2%	93,5%	90,8%	90,8%
		16	87,5%	89,7%	93,5%	91,8%	91,8%
		8	87,5%	90,8%	93,5%	91,8%	90,8%
	128	32	88,6%	89,1%	89,7%	90,8%	89,7%
		16	89,1%	90,2%	91,3%	90,2%	90,8%
		8	86,4%	92,4%	91,3%	89,1%	91,8%

В Таблице 2 представлено сравнение производительности предлагаемой классификационной сети и других классификационных сетей в этом исследовании. Предлагаемая классификация сеть достигла высочайшей точности классификации 94,6% на основе DSCNN, с небольшое количество операций умножения-накопления (MAC) - 37 M и небольшое количество параметры 50,8 K.

Machine Translated by Google

Таблица 2. Сравнение производительности с другими сетями.

Сеть	Точность	Количество MAC	Количество параметров
Ленет	92,2%	787,6 млн.	2,0 М
Реснет18	83,8%	1,4 Г	11,4 М
Реснет34	82,3%	2,0 Г	21,5 М
МобилНет	85,7%	128,3 М	3,2 М
Наш	94,6%	37,0 М	50,8 К

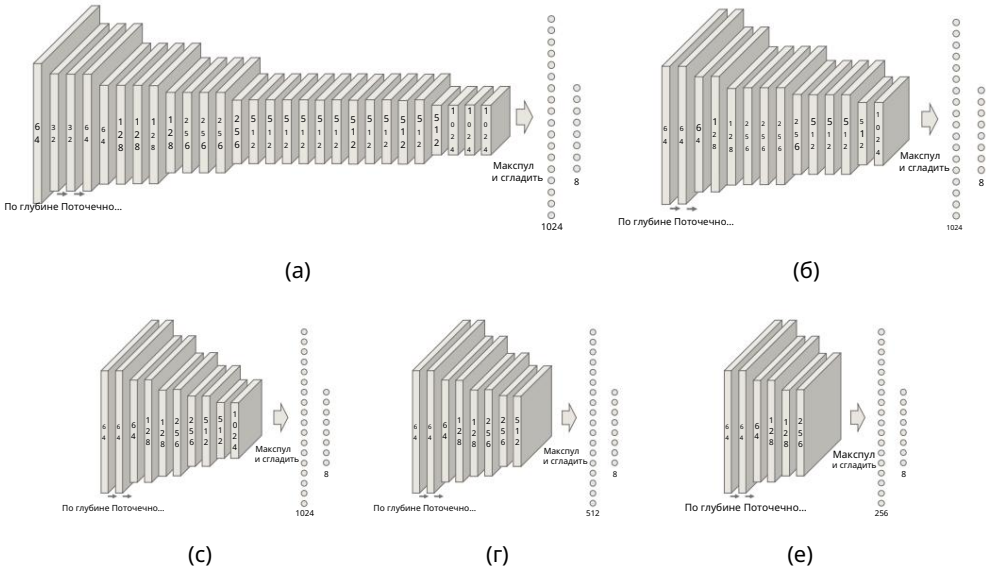


Рисунок 5. Архитектура сети классификации объектов на основе DSCNN: (а) Сеть 1; (б) Сеть 2; (в) Сеть 3; (г) Сеть 4; (д) Сеть 5.

3. Внедрение системы ускорения

Предложенная система ускорения была реализована на FPGA путем интеграции аппаратного ускорителя интеллектуальной собственности (IP) PFE и аппаратного сетевого ускорителя классификации IP с MPU. В разделе 3.1 представлена разрядность каждого IP, полученная в результате экспериментов по квантованию. Раздел 3.2 описывает структуру и работу предлагаемой системы ускорения. В разделе 3.3 описывается структура и поведение аппаратного ускорителя PFE, а в разделе 3.4 описывается FINN, используемый для реализации аппаратного ускорителя сети классификации объектов.

3.1. Квантование В

этом исследовании мы квантовали предложенную модель классификации объектов для реализации предлагаемой системы ускорения в аппаратном обеспечении. Чтобы свести к минимуму снижение производительности из-за ошибок, вызванных квантованием, мы провели обучение с учетом квантования (QAT). Результаты экспериментов с использованием разных битовых форматов для PFE и сетей классификации приведены в таблице 3. PFE более точен при 16 и 32 битах, чем при 8 битах, но разницы в точности между 16 и 32 битами нет. Поэтому мы решили использовать для PFE 16-битный формат. Сеть классификации объектов показала более высокую точность с 4 и 8 битами, но разницы в производительности не было; поэтому мы использовали 4 бита.

Machine Translated by Google

Таблица 3. Точность в зависимости от битового формата PFE и сетей классификации.

ПФЭ	Классификационная сеть		
	2	4	8
8	90,3%	91,7%	92,3%
16	89,7%	94,3%	94,3%
32	88,7%	92,7%	94,3%

3.2. Архитектура системы

На рис . 6 показана структура предлагаемой системы ускорения. Чтобы подтвердить системе в среде FPGA, ускорители были интегрированы с MPU через межсоединение расширенного расширяемого интерфейса (AXI), реализованное в FPGA. Каждый IP-адрес ускорителя является подчиненным по отношению к MPU, и MPU отправляет начальный и конечный адреса DRAM, которые IP-адрес, к которому будет осуществляться доступ, количество слов, которые необходимо прочитать из DRAM, параметры протокола AXI, и сигналы запуска и остановки для подчиненных регистров IP. IP действует как мастер для DRAM, создание адресов чтения и записи через главный интерфейс для получения необходимых данных из DRAM во время вычислений и записывайте результаты обратно в DRAM.

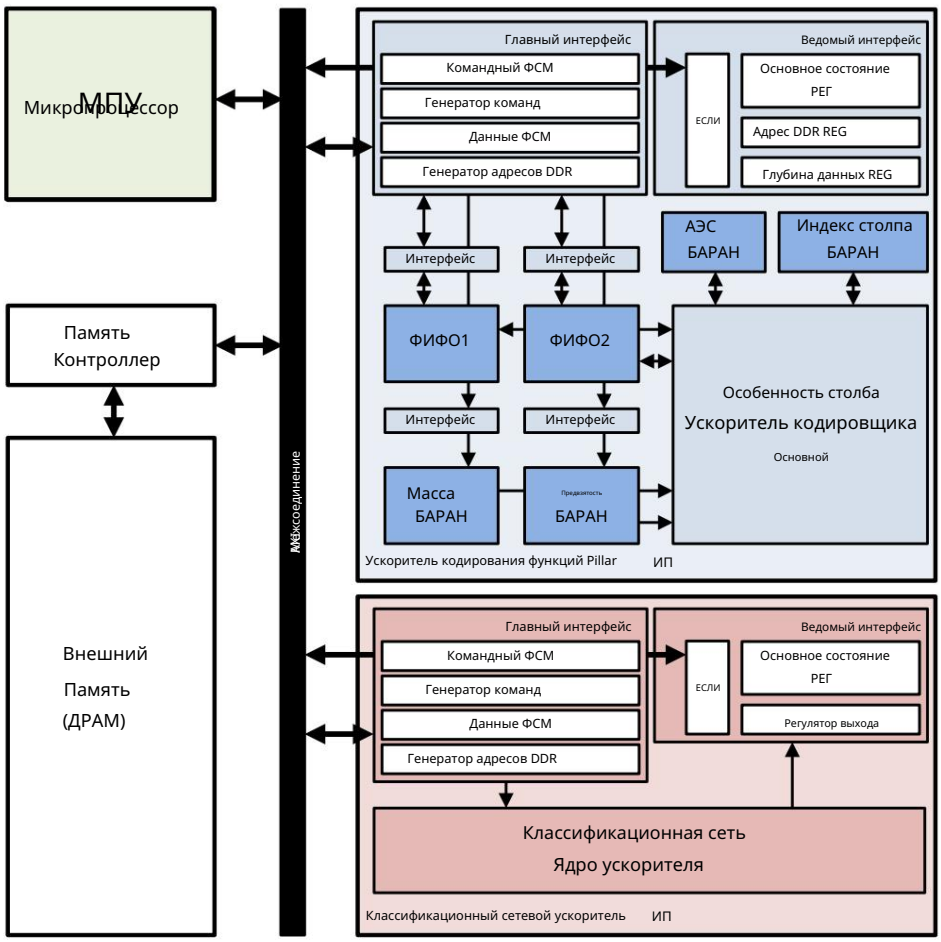


Рисунок 6. Архитектура предлагаемой системы ускорения на FPGA.

MPU сначала устанавливает подчиненный регистр IP-ускорителя PFE, который начинается со чтения облако точек из DRAM через главный интерфейс для организации облака точек в столбы. Ускоритель PFE проверяет координаты облака точек и сохраняет точку. облака в FIFO1 одно за другим. Чтобы сохранить облако точек в DRAM по столбцам, необходимо генерирует адреса DRAM и сохраняет их в FIFO2. Когда все адреса DRAM

Machine Translated by Google

поскольку облако точек генерируется и сохраняется в FIFO2, главный интерфейс обращается к FIFO1 и FIFO2 для сохранения облака точек в соответствующем месте в DRAM.

Ускоритель PFE организует все облако точек в столбцы, сохраняя их соответствующим образом в DRAM, а затем считывает облако точек обратно и сохраняет его в FIFO1 и FIFO2. Он также считывает веса и смещения для выполнения PW-свертки из DRAM и сохраняет их в памяти весов (WM) и памяти смещений (BM). После добавления HCF в облако точек выполняется PW-свертка для расширения функций и выполняется операция максимального пула . Таким образом, один столб кодируется 64 признаками, что соответствует одному пикселю с 64 каналами в 2D-изображении. Закодированный пиксель сохраняется в DRAM с учетом его положения.

После того как ускоритель PFE кодирует все облака точек в 2D-изображения, MPU устанавливает подчиненный регистр IP ускорителя сети классификации и запускает ускоритель для чтения изображений из DRAM через главный интерфейс. После того как ускоритель сети классификации прочитал все изображения и выполнил все вычисления, он записывает результаты классификации в подчиненный регистр. MPU завершает работу системы, обращаясь к подчиненному регистру IP ускорителя сети классификации и считывая результаты классификации.

3.3. Аппаратная архитектура PFE

На рис . 7 представлена структурная схема ускорителя PFE. Предлагаемый ускоритель PFE содержит блок группировки (ГУ), блок HCF (HU), блок PFN (PU) и восемь запоминающих устройств. Память представляет собой FIFO1, FIFO2, буфер количества точек на столбец (NPPP) (NB), буфер индекса столба (PB), буфер HCF1 (HB1), HB2, WM и BM. HB1 и HB2 используются HU и PU в схеме пинг-понга в рамках конвейерной структуры. Во время работы HU он записывает результаты в HB1, в то время как PU одновременно обращается к данным из HB2. Впоследствии HU переключается на запись результатов вычислений в HB2, а PU на этом этапе использует данные, хранящиеся в HB1.

Облака точек сохраняются в FIFO1 по порядку, а локатор сетки в ГУ проверяет координаты точек, чтобы получить координаты столбов, содержащих точки Например, если ширина ячейки сетки представляет собой целое значение 25 , , координаты столб, содержащий точку, может быть определена путем деления координат точки на 25, операция . Этот деления заменяется 5-битной операцией сдвига вправо. Координаты столбов используются как адрес чтения НБ для определения АЭС. Если NPPP равен 0, создается новый столб; таким образом, счетчик столбцов увеличивается на единицу, и значение записывается в PB для использования в качестве индекса столба. Если NPPP меньше 16, оно увеличивается на 1 и записывается обратно в NB; в противном случае оно остается неизменным. Адрес генерируется путем использования индекса столба, умноженного на 16 (максимальный NPPP) в качестве смещения, а затем добавления NPPP к этому смещению. Сгенерированный адрес сохраняется в FIFO2. Если облако точек находится в FIFO1, а адрес DRAM — в FIFO2, облако точек сохраняется в правильном месте в DRAM.

После того как все облака точек сохраняются в DRAM столбцов, они считываются обратно в FIFO1 и FIFO2. В FIFO2 адрес записи DRAM сохраняется во время операции группировки, тогда как данные чтения DRAM сохраняются в другое время. Затем ускоритель PFE выполняет вычисления для всех облаков точек по столбчатому принципу. Усредняющая единица (AU) HU складывает координаты x, y и z всех точек столба и делится на NPPP для получения среднеарифметических координат точек. Блок смещения в HU вычисляет xm, ym , zm путем вычитания координат центра точек, полученных из AU, из координат x, y, z каждой точки, а xc, yc - путем вычитания координат центра столба из координаты x, y каждой точки. Затем значения x, y, z, r, xm, ym , zm, xc, yc точек с добавленным HCF сохраняются в HB.

Machine Translated by Google

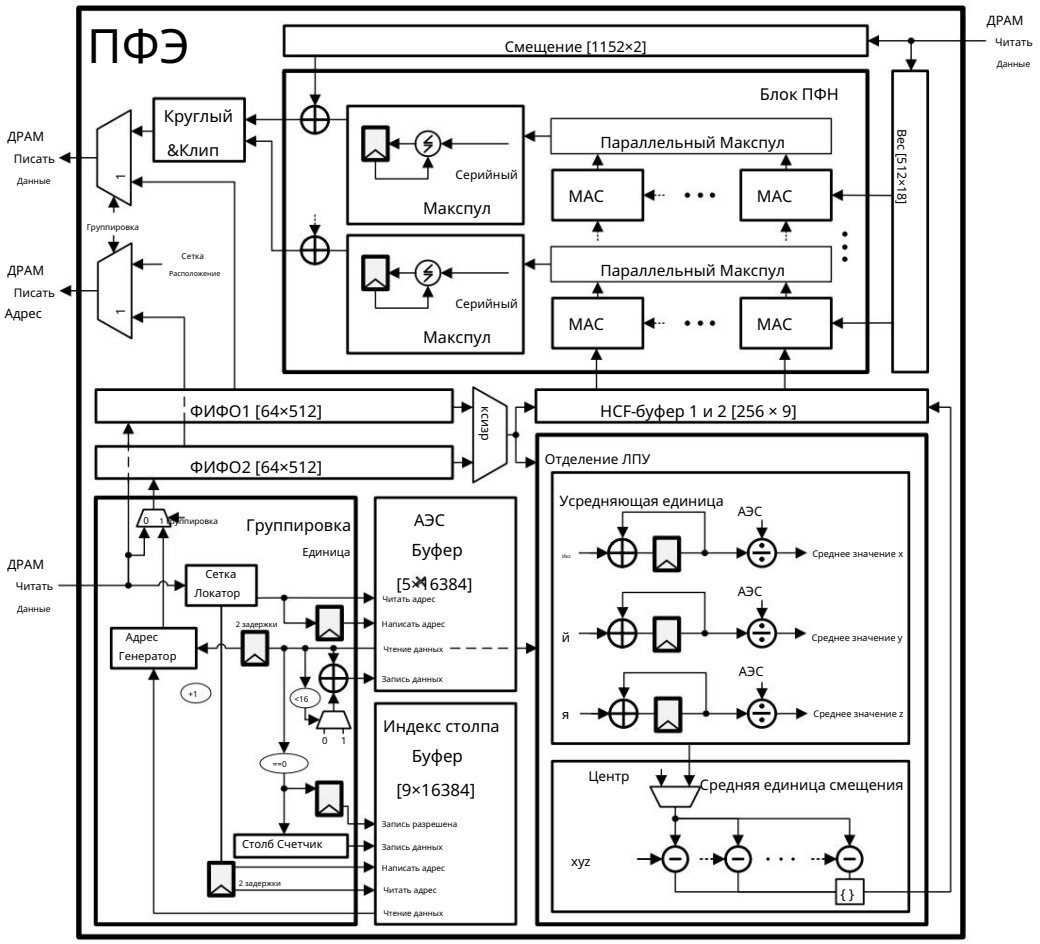


Рис. 7. Блок-схема ускорителя PFE.

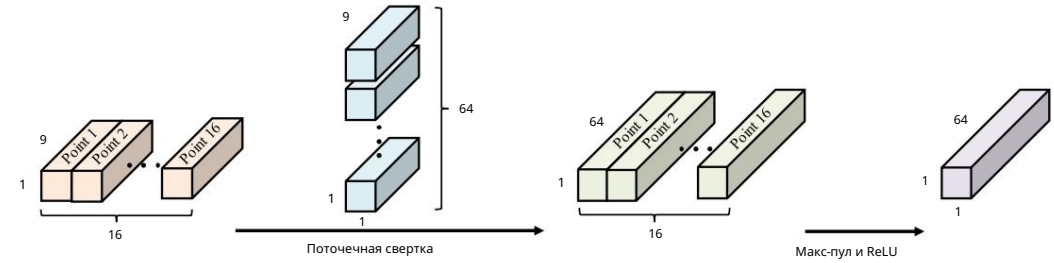
На рисунке 8a показана операция на уровне тензора, выполняемая PU. PU выполняет свертку PW, пакетную нормализацию, операции max-pool и ReLU на HCF. Веса и смещения пакетной нормализации можно объединить с весами и смещениями слоя свертки PW. В свертке PW используются только веса, а смещения добавляются после смещенной операции ReLU, которая устанавливает пороговые значения в -bias вместо 0.

На рисунке 8b представлена операция, показанная на рисунке 8a, на уровне матрицы. PU выполняет умножение матрицы HCF размера (16, 9) и матрицы весов размера (9, 64), используя массив MAC 8 × 32. Для этого матрица HCF делится на две матрицы размера (8, 9), а матрица весов делится на две матрицы размера (9, 64), выполняя все умножение матрицы на четыре части. Например, матричное умножение плитки 2 HCF и плитки веса 1 дает результирующую плитку 2, и эти плитки объединяются по максимальному значению для создания плитки 1 с максимальным объединением.

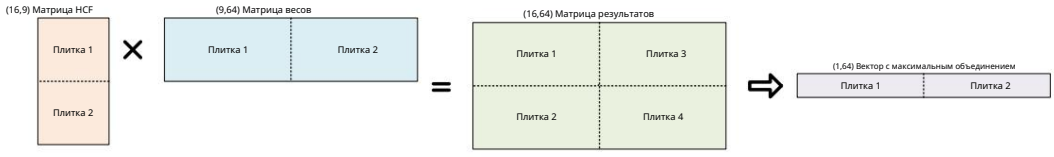
Рисунок 8c иллюстрирует работу PU на аппаратном уровне в хронологическом порядке. PU считывает тайл 1 HCF и тайл 1 веса из HB и WM соответственно. PU считывает тайл 1 HCF столбец за столбцом и весовой тайл 1 строка за строкой в течение 9 циклов, передавая их в массив MAC. Массив MAC обрабатывает входные данные HCF и веса за 9 циклов и выводит плитку результата 1 размера (8, 32). Плитка результата 1 затем обрабатывается параллельным модулем максимального пула для выполнения операций максимального пула, создавая промежуточную плитку 1 размера (1, 32). Поскольку параллельный модуль max-pool немедленно выполняет операции max-pool на выходе массива MAC, он экономит циклы и пространство буферной памяти, не сохраняя и не перезагружая плитку резул Промежуточный элемент 1 сравнивается со значением -bias в последовательном блоке max-pool, и большее значение сохраняется в регистре. Выполняя смещенные операции ReLU в последовательном блоке max-pool, операция ReLU может быть интегрирована в операцию max-pool. После создания промежуточного тайла 1 PU считывает тайл 2 HCF и весовой тайл 1 для создания промежуточного тайла.

Machine Translated by Google

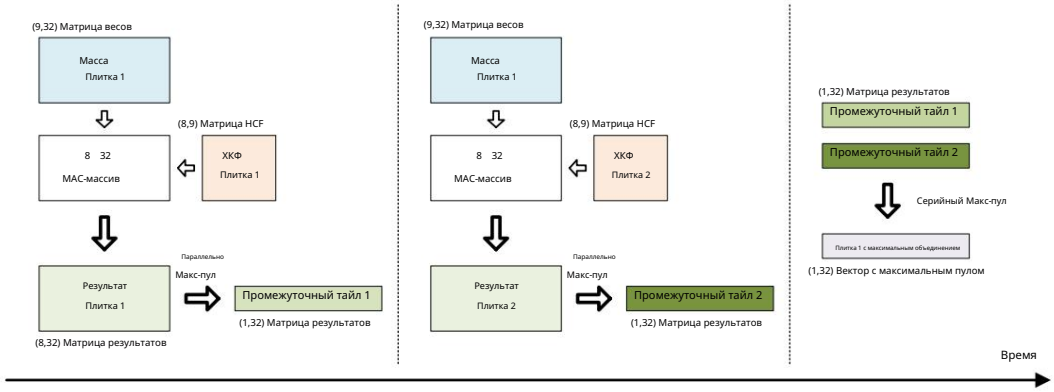
2. Промежуточный тайл 2 подвергается операциям максимального пула, при этом промежуточный тайл 1 сохраняется в регистре последовательного модуля максимального пула, в результате чего получается тайл 1 с максимальным пулом. Тайл 1 с максимальным пулом составляет половину пикселя, когда добавляется смещение, округляется и обрезается. . Этот полупиксель сохраняется в правильном месте в DRAM в зависимости от его координат на изображении. Чтобы создать адрес записи DRAM , в качестве смещения используется координата x пикселя, умноженная на ширину изображения, а к этому смещению добавляется координата y . Когда PU создал оба тайла 1 и 2 с максимальным пулом, в DRAM остается один полный пиксель.



(а)



(б)



(с)

Рис. 8. Работа ПУ: а — работа на тензорном уровне; (б) операция на уровне матрицы; (с) работа на аппаратном уровне.

3.4. Ускоритель классификационных сетей с использованием

FINN Xilinx предлагает различные инструменты для реализации сетей глубокого обучения. Vitis AI — это платформа, используемая для реализации и оптимизации приложений искусственного интеллекта и машинного обучения на FPGA и адаптивных платформах «система на кристалле». Vitis AI поддерживает различные платформы и библиотеки искусственного интеллекта и машинного обучения, а также генерирует IP-адрес блока обработки глубокого обучения (DPU). FINN — это платформа с открытым исходным кодом для вывода квантованных сетей глубокого обучения на основе FPGA . FINN использует в качестве входных данных модель обмена открытыми нейронными сетями , оптимизирует сеть и генерирует аппаратный ускоритель.

В этом исследовании мы использовали FINN для реализации аппаратной версии сети классификации объектов. В то время как DPU служит универсальным ускорителем операций глубокого обучения , FINN оптимизирует аппаратную архитектуру и поток данных специально для сети глубокого обучения. Следовательно, FINN предлагает более низкую задержку, обеспечивает скорость ответа в реальном времени и может достичь высокой пропускной способности за счет оптимизации конвейера [29]. Более того, нет-

Machine Translated by Google

работы с низкой сложностью, такие как предлагаемая классификационная сеть, могут значительно сократить использование аппаратных ресурсов. Кроме того, FINN поддерживает 4-битные целые числа, обеспечивая преимущество перед DPU. Использование преимуществ FINN, исх. [19] реализовали магистраль слой PointPillars с использованием FINN.

4. Результаты внедрения

Предлагаемая система ускорения была реализована на Xilinx Zynq Ultrascale+. ЗКУ104 [30]. В таблице 4 приведены сведения об использовании аппаратных ресурсов и рабочей частоте. предлагаемой системы ускорения. В предлагаемой системе ускорения использовалось 75 548 справочные таблицы настраиваемых логических блоков (CLB) (LUT), 56 931 регистр CLB, 374 цифровых сигнала процессоры (DSP) и 65 блоков оперативной памяти. Рабочая частота составила 187,5 МГц, что обеспечивает высокую скорость работы и время реакции предлагаемой классификации объектов система составила 6,41 мс.

Таблица 4. Аппаратные ресурсы предлагаемой системы классификации объектов.

Единица	КЛБ LUT-ы	КЛБ Регистры	DSP	Блокировать БАРАН	Частота (МГц)
ПФЭ	35 567	28 513	302	10,5	
Классификационная сеть	34 788	21 625	72	54,5	
Акси-Интерконнект	5193	6793	0	0	
Общий	75 548	56 931	374	65	187,5

В таблице 5 сравниваются времена отклика встроенной реализации PFE на ARM Cortex A53 (базовый уровень), работающий на частоте 1,2 ГГц, и предлагаемое оборудование PFE ускоритель (предлагается) по времени выполнения каждой операции. Группировка операция, отвечающая за преобразование облака точек в набор блоков столбов, испытала замечательное ускорение в 23,67 раза, сократившись с 0,71 мс до 0,03 мс. Примечательно, что Операции HCF и PFN достигли значительного ускорения примерно в 42,6 раза. уменьшается с 8,52 мс до 0,20 мс. Такое заметное ускорение стало возможным благодаря использованию возможность параллельных вычислений массива MAC в ускорителе PFE и конвейерной структура операций HCF и PFN. Общее время выполнения операций PFE было уменьшено с 9,23 мс до 0,23 мс, демонстрируя эффект ускорения в 40 раз по сравнению с реализацией на основе прошивки.

Таблица 5. Сравнение времени ответа PFE с исходным уровнем.

Операция	Базовая версия (прошивка)	Предлагаемое (Аппаратное обеспечение)
Группировка	0,71 мс	0,03 мс
ХКФ	0,20 мс	0,20 мс
ПФН	8,32 мс	
Общий	9,23 мс	0,23 мс

В таблице 6 представлено сравнение предлагаемой системы классификации объектов и Методы на основе столбов, реализованные в предыдущих исследованиях. В [19–21] гетерогенная система между MPU и FPGA, тогда как наша система представляет собой сквозную аппаратную реализацию всех операций на FPGA. В [19] вычисление PFE выполнялось на MPU, а вычисления сети глубокого обучения выполнялись с использованием акселератор, реализованный через FINN. Ссылка. [20] рассчитали PFE на MPU и выполнил сеть глубокого обучения на ускорителе, реализованном с использованием передачи регистров дизайн уровней (RTL). Ссылка. [21] использовали как MPU, так и DPU для расчета PFE, а также DPU для вычислений в сети глубокого обучения. Мы рассчитали PFE с помощью ускорителя

Machine Translated by Google

реализовано посредством проектирования RTL и сети глубокого обучения с ускорителем
реализовано через FINN.

Наш метод достиг самого быстрого времени отклика — 6,41 мс, поскольку это был единственный
тот, который ускорил PFE, и была разработана предложенная модель классификации объектов
иметь низкую вычислительную сложность посредством различных экспериментов. Более того, эти
Результатом экспериментов стала модель классификации объектов с меньшим количеством параметров. Наш метод
использовалось больше DSP, чем в [19], но значительно меньше LUT, регистров и памяти.
Ссылка. [19] использовали ту же платформу FPGA для периферийных сред, что и наша, ZCU104,
достижение низкого энергопотребления. Однако он все равно потреблял больше энергии, чем наш,
который реализовал эффективную сеть глубокого обучения. По сравнению с методом [20]
наш метод использовал меньше ресурсов по всем направлениям, с особенно заметными различиями в
DSP и использование памяти. Ссылка. [20] также потребляли мало энергии за счет внедрения эффективного
сеть глубокого обучения. Кроме того, поскольку исх. [20] разработали сеть глубокого обучения
ускорителя на уровне РТЛ, он потреблял даже меньше мощности, чем наш, который был спроектирован
на основе ЗОЖ. Кроме того, наш метод использовал больше LUT, но меньше регистров и DSP.
и существенно меньше BRAM и URAM, чем в [21]. Кроме того, наш метод
потреблял значительно меньше энергии, чем в [21]. Ссылка. [21] использовали платформу U280 FPGA.
нацелены на серверные среды, которые имеют гораздо более высокое энергопотребление и не
подходит для периферийных сред.

Таблица 6. Сравнение с другими реализациями на основе основных принципов.

	[19]	[20]	[21]	Наш
Платформа	ЗКУ104	ZC706	U280	ЗКУ104
Вычисление Устройство	МПУ и ПЛИС (ФИНН)	МПУ и ПЛИС (RTL)	МПУ и ПЛИС (ДПУ)	ПЛИС (РТЛ и ФИНН)
Исполнение Время (мс)	377,1	43,26	64,1	6.41
CLB LUT-системы	189 074	128 721	48 006	75 548
Регистры CLB	150 187	111 118	85 801	56 931
DSP	88	883	525	374
Блокировать ОЗУ	159	382,5	131,5	65
Ультра ОЗУ	-	-	64	1
Частота (МГц)	150	150	300	187,5
Мощность (Вт)	6,5	3.6	73,8	4.4

5. Выводы

В этом исследовании мы предлагаем модель классификации объектов глубокого обучения, основанную на
Столбовой метод для реагирования автономных мобильных роботов в реальном времени. Объект
классификационная сеть модели глубокого обучения основана на DSCNN, которая имеет
низкая сложность и небольшое количество параметров, что делает его подходящим для сред периферийных устройств.
Модель классификации объектов была обучена с использованием набора данных, учитывающего операционную деятельность.
среда автономных мобильных роботов. Чтобы свести к минимуму снижение производительности из-за
Для квантования мы квантовали и протестировали модель с помощью QAT, добившись классификации
точность 94,3%. Поскольку вычисление PFE в моделях классификации объектов на основе столбцов занимало
много времени, мы предложили аппаратный ускоритель для PFE.
Ускоритель PFE был реализован посредством конструкции RTL и достиг времени отклика
0,23 мс, что в 40 раз быстрее программы прошивки. Ускоритель для объекта
классификационная сеть предложенной модели глубокого обучения была реализована посредством
FINN и добился времени выполнения 6,18 мс. Мы также предложили объект
система классификации, объединяющая предлагаемый ускоритель PFE и классификацию объектов.
сетевой ускоритель через соединение AXI. Предлагаемая система классификации объектов

Machine Translated by Google

был реализован и проверен на FPGA и достиг времени отклика в реальном времени 6,41 мс.

Предлагаемый ускоритель PFE можно комбинировать с другими ускорителями сетей глубокого обучения для ускорения более разнообразных моделей глубокого обучения. Однако в этом исследовании мы проводили ускорение только на моделях классификации объектов. В будущей работе мы планируем реализовать систему, поддерживающую ускорение для более разнообразных и сложных моделей глубокого обучения.

Вклад автора: CP разработала и реализовала предложенную систему ускорения, провела эксперимент и оценку и написала статью. SL оценил предложенную систему ускорения и отредактировал эту рукопись. YJ задумал и возглавил исследование, проанализировал результаты экспериментов и написал статью. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Эта работа была поддержана грантом Института планирования и оценки информационных и коммуникационных технологий (IITP), финансируемым правительством Кореи (MSIT) (№ 2022-0-00960), а инструменты САПР поддерживались IDEC.

Заявление Институционального наблюдательного совета: Неприменимо.

Заявление об информированном согласии: Не применимо.

Заявление о доступности данных: набор данных доступен по запросу авторов.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Варламов О. «Мозги» для роботов: применение экспертных систем «Мивар» для реализации автономных интеллектуальных систем. Роботы. Большие данные 2021, 25, 100241. [CrossRef]
2. Лю, Ю.; Ли, З.; Лю, Х.; Кан, З. Обучение передаче навыков для автономных роботов и сотрудничество человека и робота: обзор. Робот. Автон. Сист. 2020, 128, 103515. [CrossRef]
3. Ёсиока, М.; Суганума, Н.; Йонеда, К.; Алдибаха, М. Классификация объектов в реальном времени для автономных транспортных средств с использованием LIDAR. В материалах Международной конференции по интеллектуальной информатике и биомедицинским наукам (ICIIBMS) 2017 г., Окинава, Япония, 24–26 ноября 2017 г.; стр. 210–211.
4. Гао, Х.; Ченг, Б.; Ван, Дж.; Ли, К.; Чжао, Дж.; Ли, Д. Классификация объектов с использованием Fusion of Vision и LIDAR на основе CNN в Автономная среда транспортного средства. IEEE Транс. Инди. Информ. 2018, 14, 4224–4231. [Перекрестная ссылка]
5. Чжоу Ю.; Лю, Л.; Чжао, Х.; Лопес-Бенитес, М.; Ю, Л.; Юэ, Ю. На пути к глубокому радиолокационному восприятию автономного вождения: наборы данных, Методы и проблемы. Датчики 2022, 22, 4208. [CrossRef] [ПабМед]
6. Су, Х.; Маджи, С.; Калогеракис, Э.; Лернед-Миллер, Э. Многоракурсные сверточные нейронные сети для распознавания трехмерных форм. В материалах Международной конференции IEEE по компьютерному зрению, Сантьяго, Чили, 7–13 декабря 2015 г.; стр. 945–953.
7. Хоанг, Л.; Ли, Ш.; Ли, Э.Дж.; Квон, К.Р. GSV-NET: мультимодальная сеть глубокого обучения для классификации трехмерных облаков точек. Прил. наук. 2022, 12, 483. [CrossRef]
8. Ци, ЧР; Су, Х.; Мо, К.; Гибас, Л. Дж. Pointnet: глубокое изучение наборов точек для 3D-классификации и сегментации. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Гонолулу, Гавайи, США, 21–26 июля 2017 г.; стр. 652–660.
9. Чжоу, Ю.; Тузель, О. Voxelnet: Комплексное обучение обнаружению трехмерных объектов на основе облаков точек. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Солт-Лейк-Сити, Юта, США, 18–22 июня 2018 г.; стр. 4490–4499.
10. Бханушали, Д.; Релья, Р.; Манги, К.; Вашист, А.; Хохграф, К.; Гангули, А.; Квасинский А.; Куль, Мэн; Птуча, Р. LiDAR-камера Fusion для обнаружения 3D-объектов. Электрон. Визуализация 2020, 32, 1–9. [Перекрестная ссылка]
11. Ян, Ю.; Мао, Ю.; Ли, Б. Второе: редко встроенное сверточное обнаружение. Датчики 2018, 18, 3337. [CrossRef] [ПабМед]
12. Ланг, А.Х.; Вора, С.; Цезарь, Х.; Чжоу, Л.; Ян, Дж.; Бейбом, О. Pointpillars: Быстрые кодировщики для обнаружения объектов из облаков точек. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Лонг-Бич, Калифорния, США, 15–20 июня 2019 г.; стр. 12697–12705.
13. Лис, К.; Крыжак, Т. Выбор типа магистрали PointPillars для быстрого и точного обнаружения объектов LiDAR. В материалах Международной конференции по компьютерному зрению и графике, Варшава, Польша, 19–21 сентября 2022 г.; Шпрингер: Берлин/Гейдельберг, Германия, 2022 г.; стр. 99–119.
14. Шу, Х.; Чжан Л. Исследование алгоритма PointPillars на основе магистральной сети с расширенными возможностями. Электроника 2024, 13, 1233. [Перекрестная ссылка]
15. Ван Ю.; Хан, Х.; Вэй, Х.; Луо, Дж. Сегментация экземпляров Frustum – PointPillars: облегченный алгоритм объединения для восприятия камеры и LiDAR при автономном вождении. Математика 2024, 12, 153. [CrossRef]
16. Агаше, П.; Лаванья, Р. Обнаружение объектов с использованием PointPillars с модифицированным DarkNet53 в качестве магистрали. В материалах 20-й Международной конференции Совета Индии (INDICON) IEEE 2023 г., Хайдарабад, Индия, 14–17 декабря 2023 г.; стр. 114–119.

Machine Translated by Google

17. Цой Ю.; Ким, Б.; Ким, Анализ производительности программного обеспечения PointPillars на платформах ЦП и графических процессоров. В материалах 36-й Международной технической конференции по схемам/системам, компьютерам и коммуникациям (ITC-CSCC) 2021 г., Чеджу, Республика Корея, 27–30 июня 2021 г.; стр. 1–4.
18. Сильва А.; Фернандес, Д.; Невоа, Р.; Монтейро, Дж.; Новаис, П.; Жирао, П.; Афонсу, Т.; Мело-Пинто, П. Бортовой вывод об обнаружении и локализации трехмерных объектов в облаках точек с ограниченными ресурсами, ориентированный на беспилотные приложения. Датчики 2021, 21, 7933. [Перекрестная ссылка] [ПабМед]
19. Станиш Дж.; Лис, К.; Горгон, М. Реализация сети Pointpillars для обнаружения 3D-объектов в перепрограммируемых системах. Гетерогенные устройства с использованием FINN. J. Сигнальный процесс. Сист. 2022, 94, 659–674. [Перекрестная ссылка]
20. Ли, Ю.; Чжан, Ю.; Лай, Р. TinyPillarNet: сеть на основе крошечных столбов для обнаружения трехмерных объектов в облаке точек на границе. IEEE Транс. Сист. цепей. Видео Технол. 2023, 34, 1772–1785. [Перекрестная ссылка]
21. Латоцке, К.; Клокер, А.; Шенинг, С.; Кемпер, Ф.; Слими, М.; Экстайн, Л.; Геммеке, Т. Ускорение обработки и обнаружения облаков точек лидара на основе FPGA . В материалах симпозиума IEEE по интеллектуальным транспортным средствам 2023 г. (IV), Анкоридж, штат Алабама, США, 4–7 июня 2023 г.; стр. 1–8.
22. Тан, М.; Ле, К. Efficientnet: переосмысление масштабирования модели для сверточных нейронных сетей. В трудах Международного Конференция по машинному обучению. PMLR, Лонг-Бич, Калифорния, США, 9–15 июня 2019 г.; стр. 6105–6114.
23. Шолле Ф. Xception: глубокое обучение с помощью глубоко разделяемых сверток. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Гонолулу, Гавайи, США, 21–26 июля 2017 г.; стр. 1251–1258.
24. Ховард, А.Г.; Чжу, М.; Чен, Б.; Калениченко Д.; Ван, В.; Вейанд, Т.; Андреетто, М.; Адам, Х. Mobilenets: Эффективные сверточные нейронные сети для приложений мобильного зрения. arXiv 2017, arXiv:1704.04861.
25. Лу, Г.; Чжан, В.; Ван, З. Оптимизация операций глубоко разделяемой свертки на графических процессорах. IEEE Транс. Параллельное распределение. Сист. 2021, 33, 70–87. [Перекрестная ссылка]
26. Кайзер, Л.; Гомес, АН; Шолле, Ф. Глубоко отделимые свертки для нейронного машинного перевода. apXiv 2017, arXiv: 1706.03059.
27. Отстранение. Лидарный датчик Ouster OS1. Доступно онлайн: https://ouster.com/products/hardware/os1-lidar-sensor . (доступ 22 май 2024 г.).
28. Цезарь, Х.; Банкити, В.; Ланг, А.Х.; Вора, С.; Лионг, В.Е.; Сюй, Кью; Кришнан, А.; Пан, Ю.; Балдан Г.; Бейбом, О. nuscenes: Мультимодальный набор данных для автономного вождения. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Сиэтл, Вашингтон, США, 14–19 июня 2020 г.; стр. 11621–11631.
29. Xilinx Inc. Документация Vitis™ AI, часто задаваемые вопросы. Доступно онлайн: https://xilinx.github.io/Vitis-AI/3.0/ html/docs/reference/faq.html#what-is-the-difference-between-the-vitis-ai-integrated-development-environment. -и-финн -рабочий процесс (по состоянию на 22 мая 2024 г.).
30. AMD. УльтраСклае+ ZCU104. Доступно онлайн: https://www.xilinx.com/products/boards-and-kits/zcu104.html. (доступ 22 май 2024 г.).

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Swin-Fake: преобразователь согласованного обучения Детектор видео дипфейков

Лян Юй Гун *,†,‡ , Сюэ Цзюнь Ли *,†,‡ и Питер Хан Джу Чонг †

Факультет электротехники и электроники, Оклендский технологический университет, Окленд 1010, Новая Зеландия; peter.chong@aut.ac.nz * Переписка: liangyu.gong@autuni.ac.nz (LYG); xuejun.li@aut.ac.nz (XJL)
† Текущий адрес: улица Сент-Пол, 6, Окленд 1010, Новая Зеландия.
‡ Эти авторы внесли одинаковый вклад в эту работу.

Аннотация: В последние годы Deepfake стала новой технологией, влияющей на кибербезопасность своими незаконными приложениями. Большинство детекторов дипфейков используют модели на основе CNN, такие как Xception Network, чтобы отличать настоящие или фальшивые медиа; однако их производительность на перекрестных наборах данных не идеальна, поскольку на текущем этапе они страдают от чрезмерной подгонки. Поэтому в этой статье предложен метод обучения пространственной согласованности, позволяющий решить эту проблему в трех аспектах. Во-первых, мы увеличили выбор методов увеличения данных до 5, что больше, чем методы увеличения данных нашего предыдущего исследования . В частности, мы захватили несколько равных видеокадров одного видео и случайным образом выбрали пять различных дополнений данных, чтобы получить разные представления данных и обогатить входное разнообразие. Во-вторых, мы выбрали Swin Transformer в качестве средства извлечения признаков вместо магистральной сети на основе CNN. Это означает, что наш подход не использовал его для последующих задач и мог кодировать эти данные с помощью сквозного Swin Transformer, стремясь изучить корреляция между различными фрагментами изображения. Наконец, в нашем исследовании это сочеталось с обучением согласованности, и обучение согласованности позволило определить больше взаимосвязей данных, чем контролируемая классификация. Мы исследовали согласованность характеристик видеокадров, рассчитав их косинусное расстояние, и применили традиционную перекрестную энтропийную потерю, чтобы регулировать эту потерю классификации. Обширные эксперименты с наборами данных и между наборами данных показали, что Swin-Fake может давать относительно хорошие результаты на некоторых наборах данных deepfake с открытым исходным кодом, включая FaceForensics++, DFDC, Celeb-DF и FaceShifter. Сравнивая нашу модель с несколькими эталонными моделями, наш подход показывает относительно высокую надежность в обнаружении дипфейк

Ключевые слова: последовательность обучения; Свинцовый трансформатор; потеря консистенции; обнаружение дипфейков

check for updates

Цитирование: Гонг, LY; Ли, XJ; Чонг, PHJ Swin-Fake: детектор дипфейков на основе трансформатора, основанный на обучении согласованности. Электроника 2024, 13, 3045. https://doi.org/10.3390/электроника13153045

Научные редакторы: Петя Георгиева, Красимир Тончев и Агата Манолова

Поступила: 1 июля 2024 г.
Пересмотрено: 26 июля 2024 г.
Принято: 29 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария.
Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Генерация и обнаружение дипфейков продолжали развиваться в игре атакующий-защитник с 2019 по 2021 год; однако с быстрым развитием методов на основе GAN [1,2] и диффузионных моделей (DM) [3] мощные генеративные способности этих моделей с расширенной постобработкой, такой как смешивание цветов, делают синтетические лица более реалистичными. .

Таким образом, в настоящее время в дипфейковом видео труднее отличить подлинность одного отдельного видеокадра человеческими глазами .

Несмотря на то, что детекторы на основе сверточных нейронных сетей (CNN) [4–6] оказали заметное влияние на работу по классификации дипфейков, они по-прежнему страдают от большой зависимости от обучающих данных, что приводит к падению точности или площади под кривой (AUC). значительно в экспериментах с перекрестными наборами данных. Таким образом, в последние годы детекторы дипфейков, как правило, разрабатываются как методы изучения согласованности для исследования пространственной или временной согласованности, самосогласованности или несогласованности между представлениями или патчами. Кроме того, было доказано, что преобразователи изображения [7,8] имеют широкое применение, особенно при классификации изображений. Детекторы дипфейков на основе трансформаторов становятся мейнстримом в этой области из-за их уникальных механизмов внимания, которые можно применять для генерации про

Machine Translated by Google

временные особенности. По сравнению с другими преобразователями зрения [9,10], Swin Transformer может значительно повысить производительность классификации за счет установки механизмов разделения окна, циклического сдвига и смещения относительного положения.

В этой статье мы предложили простую структуру обучения согласованности, использующую Swin Transformer [11] в качестве средства извлечения признаков для явного извлечения и изучения согласованности между захваченными видеокадрами. По сравнению с традиционными детекторами дипфейков, наш подход также содержит блок предварительной обработки данных и базовую модель для извлечения признаков, но наши потери сочетают в себе согласованность кадров и потерю классификации. Что касается работы по предварительной обработке данных, мы разделили одно видео на несколько кадров как группу для классификации дипфейковых видео, что можно рассматривать как ансамблевую технику. В отличие от других методов ансамблевого обучения, мы расширили извлечение признаков, чтобы помочь модели изучить различные поддельные признаки с помощью случайных и различных методов увеличения данных. Кроме того, традиционные детекторы дипфейков всегда используют некоторые неизмененные методы увеличения данных, такие как базовое преобразование и нормализация; однако для нашего блока предварительной обработки данных мы выбрали несколько передовых методов увеличения данных, чтобы повысить случайность улучшения данных и доказать, что независимо от того, насколько данные улучшаются, одну и ту же категорию данных можно поддерживать с хорошей согласованностью. Как показано на рисунке 1, каждое видео было поровну разделено на кадры, чтобы обеспечить равномерное распределение области лица по всей продолжительности видео. Наш блок предварительной обработки данных также содержит несколько дополнений данных и случайным образом применяется к каждому захваченному кадру для обогащения различных входных данных. Затем общий кодер на основе Transformer извлекает представления и вычисляет сходство пар, создавая комбинации для формирования потери согласованности. Наконец, контролируемый классификатор присваивает метку каждому извлеченному признаку кадра.

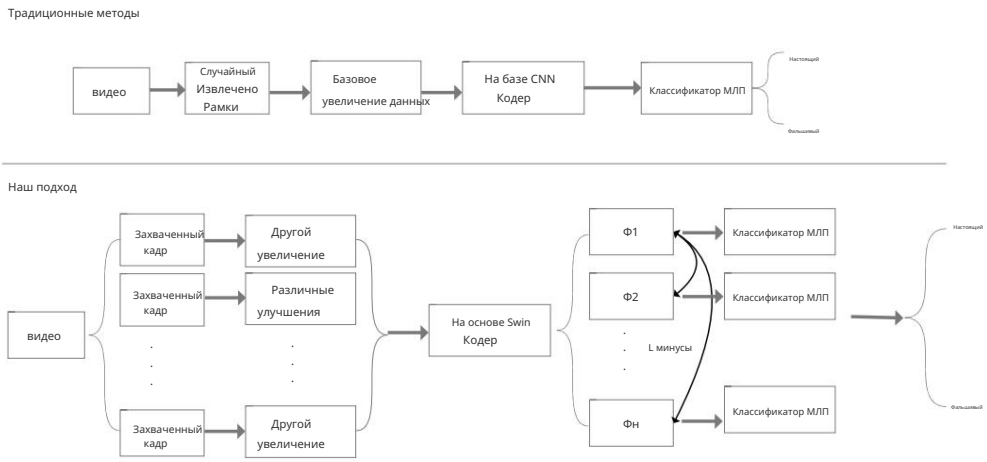


Рисунок 1. Иллюстрация платформы Swin-Fake в сравнении с традиционной платформой для обнаружения видео Deepfake.

В нашем методе есть три основных улучшения по сравнению с традиционными детекторами дипфейков: (1) Мы используем пять методов увеличения данных, включая базовое преобразование, случайное стирание [12], случайное увеличение, случайное кадрирование и DFDC Selim [13], которые применяются случайным образом. , и они в значительной степени будут избегать переобучения на этапе ввода. (2) В качестве кодера выбирается Swin Transformer, который будет объединять корреляции между различными фрагментами разделенного изображения. (3) Окончательная общая потеря определяется двумя компонентами: потерей согласованности и потерей перекрестной энтропии, а решение гарантируется обучением согласованности.

2. Сопутствующая работа

2.1. Обнаружение дипфейков.

Обнаружение дипфейков всегда определяется как бинарная классификация или работа по многоклассовой классификации, поскольку для этого по сути требуется, чтобы модель различала средства массовой информации как настоящие или фейковые. Самый популярный фреймворк — использовать в качестве магистралей ванильный Xception [14] или ResNet [15].

Machine Translated by Google

кодировать, а затем комбинировать контрастное обучение или исследовать согласованность представлений. Капсульная сеть [4] — это эффективная облегченная сеть, которая может решать проблемы «обратной графики» и предлагает алгоритм динамической маршрутизации для перегруппировки извлеченных функций. CORE [16] — это метод обучения согласованности, основанный на Xception Network, который увеличивает количество дополнений данных для доказательства согласованности модели. Youtu-Lab [17] предложила архитектуру обучения с двойным контрастированием, позволяющую различать подлинность лиц, используя также различные дополнения данных. Детекторы дипфейков с ансамблевым обучением также могут значительно улучшить эффективность классификации и широко применяются в недавних исследованиях; например, Чжан [18] предложил метод обучения ансамблю гетерогенных признаков, который сначала извлекает из изображений признаки градиента серого, характеристики спектра и особенности текстуры , а затем интегрирует их в вектор признаков ансамбля посредством процесса выравнивания. Используя DeepfakeStack, авторы [19] решили объединить несколько современных моделей классификации дипфейков для создания улучшенного составного классификатора, который может достигать точности 99,65% в некоторых наборах данных дипфейков. Однако предлагаемый нами метод Swin-Fake направлен на изучение способности Swin Transformer извлекать поддельные характеристики, такие как шум PRNU и несоответствие границ, а затем извлекать разные кадры одного видео, чтобы доказать согласованность модели в процессе классификации с использованием разли

2.2. Swin Transformer

Swin Transformer [11] — это иерархический преобразователь зрения, который совершенствует многоголовые блоки внимания (MSA) до оконных многоголовых блоков внимания (W-MSA) и многоголовых блоков внимания со смещенным окном (SW-MSA). Он накладывает различные размеры функций, настраивая окна разных размеров, что отличается от традиционного Vision Transformer [7] и позволяет сэкономить больше на вычислительных затратах, чем традиционный Vision Transformer. Обычно входное изображение имеет размер 3 × 3 × 224, а размер выходного объекта — 7 × 7 × 768. Эта структура содержит один слой патч-части, один линейный слой внедрения и двенадцать блоков Swin Transformer для извлечения функций, и эти функции коррелирует с различными вставками патчей. Эта сетевая структура может более эффективно имитировать ощущение согласованности и несогласованности различных участков, когда человеческие глаза наблюдают и сканируют различные области лица. Более того, мы исследовали существующие статьи, связанные с детекторами дипфейков на основе Transformer, и с удивлением обнаружили, что существует не так много статей, посвященных обучению пространственной согласованности с помощью магистрали Transformer. Большинство из них прилагают больше усилий к разработке блоков извлечения временных признаков, чтобы получить больше информации о временной согласованности. Недавно Ильяс и др. [20] предложили AVNet для обнаружения аудиовизуальных носителей с использованием Swin Transformer, а Халид [21] предложил использовать Swin Y-Net для обнаружения дипфейков.

2.3. Увеличение данных.

В области компьютерного зрения важно применять увеличение данных, которое является первым этапом предотвращения переобучения и эффективным методом увеличения объема данных. В этой статье мы используем только пять методов манипулирования изображениями для изменения значений и положений пикселей. Базовое преобразование используется для объединения нормализации пикселей и изменения размера. Случайное стирание [12] — это преобразование вырезаемых данных, которое случайным образом выбирает прямоугольный блок области лица и меняет его цвет на полностью черный. Произвольное кадрирование используется для обрезки области лица в произвольном соотношении, а затем изменения ее размера до определенной высоты и ширины. DFDC Selim — это комплексное дополнение данных, содержащее гауссов шум, размытие по Гауссу, случайный сдвиг и масштаб, и уже доказано, что оно является наиболее полезным дополнением данных в работе по классификации дипфейков [16]. Случайное увеличение случайным образом выбирает «Случайное стирание», «Случайное горизонтальное переворот» или «Случайное изменение размера». По сравнению с конструкцией традиционных детекторов дипфейков мы значительно улучшили разнообразие и случайность входных данных, и это можно использовать для последующей проверки сходства между разными видеокадрами одного видеоклипа.

Machine Translated by Google

3. Предлагаемые методы B

этой работе мы описываем конвейер детектора глубоких фейковых изображений и разделяем его на четыре части: предварительная обработка данных (включая распознавание лиц, обрезку и увеличение данных), извлечение признаков, вычисление согласованности и прогнозирование с помощью классификаторов. Как показано на рисунке 2, мы сначала одинаково захватываем кадры одного обнаруженного видео, а затем случайным образом выбираем и применяем дополнения данных для создания различных представлений данных. После извлечения признаков с помощью Swin Transformer мы отдельно вычисляем косинусные расстояния между представлениями в комбинациях и вычитаем среднее косинусное расстояние как потерю согласованно Чтобы предотвратить проблему несбалансированных классов, мы объединяем реальные видео из других баз данных в наш обучающий набор, что очень важно для дальнейшей оценки; кроме того, мы также решили повысить дискретизацию реальных видеокадров; например, если каждое фейковое видео поровну разделено на восемь кадров, реальный видеокадр делится поровну на 16 или 32 соответственно.

Это может в значительной степени помешать прогнозируемому результату обученной модели оставаться в том же классе.

Алгоритм 1. Рабочий процесс Swin-Fake Input: Предварительно

обученный Swin Transformer: f(x); видеопоследовательности: X Равномерно разделите каждое видео на M кадров.

Обрезка лица и увеличение ширины и высоты для получения изображений лица.

Случайным образом выбирайте методы увеличения данных, чтобы избежать переобучения. для i = 1 до M do Используйте

входные последовательности X для группировки набора.

Рассчитайте оценку внимания между различными встраиваниями разделенных патчей.

Вычислите f(x) посредством прямого распространения, чтобы получить двумерные признаки Fx. 2 Рассчитайте косинусное расстояние LCos между парами C , используя уравнение (3)

$$LCons = \frac{1}{M}\sum_{1}^{M} \text{среднее значение (LCos)}$$

Используйте линейный слой для проецирования Fx на 2 класса.

$$LCE = Ltot = \sum_{1}^{H} [y \log(p) + (1 - y)\log(1 - p)]$$

αLCons + (1 α)LCE Провести обратное распространение и сохранить градиенты для каждого слоя. end для вывода: обученная модель Swin-Fake f(x).

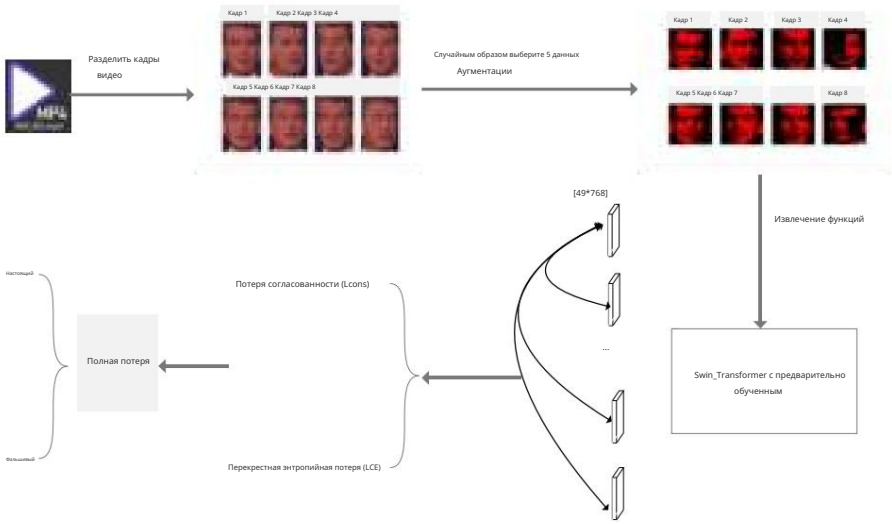


Рисунок 2. Объяснение архитектуры Swin-Fake. Сначала мы делим M кадров одного видео и применяем разные дополнения данных для каждого кадра. Затем они отправляются в общий кодер Swin Transformer. После получения M представлений они будут отправлены для отдельного расчета потерь классификации и потери согласованности.

Machine Translated by Google

3.1. Предварительная

обработка Сначала мы даем входной пакет из B видео и поровну разделяем каждое видео на M видеокадров на основе расчета общего количества кадров каждого видео. На этапе предварительной обработки также встроен детектор лиц «SCRFD» [22] для обнаружения областей лица, выравнивания лиц и их обрезки для создания конкретных наборов данных. В частности, этот детектор лиц основан на Insightface и использует специальные методы перераспределения выборки (SR) и перераспределения вычислений (CR) для обнаружения лиц. «scrfd 10g kps.onnx» — это один из файлов веса модели, использующий этот метод обнаружения лиц [22], а kps представляет 5 ключевых точек лица, обнаруженных в этой модели. Чтобы сохранить некоторую фоновую информацию, мы увеличиваем ширину и высоту обрезанных изображений на 10%. Затем все кадры M × B аккуратно располагаются по порядку, служа входными данными для Swin Transformer, и все размеры изменяются до 3 × 224 × 224. Поскольку мы применяем и случайным образом выбираем пять дополнений данных в наборе преобразований T на M × B кадров, входные наборы векторов I имеют размер [B, M, 3, 224, 224], что означает, что каждая группа содержит M [3, 224, 224] представлений. Наконец, входной набор представлений записывается как:

$$Я = \{\{x1, x2. . . xM\}1 . . . \{x1, x2. . . xM\}B \} \text{ где } xn \quad (1)$$

представляет просмотр каждого кадра, B представляет пакет видео, {x1, x2. . . xM}B представляет B -ю группу представлений данных, а M представляет количество кадров, разделенных поровну.

После получения представлений входных данных к этим представлениям данных будут случайным образом применены пять методов увеличения данных . Эти дополненные тензоры являются окончательными входными данными модели Swin-Fake, как показано на рисунке 3.

Рисунок 3. Примеры случайных дополненных данных. Для случайного стирания мы устанавливаем произвольное стирание , которое будет выполняться, на 0,8, масштаб от 0,02 до 0,2 и коэффициент стираемой площади от 0,5 до 2,0. Для случайного кадрирования мы установили шкалу случайного кадрирования от 0,77 до 1 и коэффициент случайного кадрирования от 0,9 до 1,1.

3.2. Сеть кодировщиков

Мы используем Swin Transformer [11] в качестве экстрактора признаков f из-за его уникальной иерархической структуры. Он содержит несколько составных блоков Transformer, среди которых блоки нижнего уровня могут захватывать некоторую информацию о цвете и краях, а блоки более высокого уровня могут использоваться для анализа семантической информации. После применения подхода разделения окна со смещением и линейного встраивания мы можем изменить размерность входного объекта и получить векторы запроса (Q), ключа (K) и значения (V) отдельно. Запросы (Q) и ключи (K) имеют одинаковую размерность d. Вычислением скалярного произведения запроса (Q) со всеми ключами (K) и делением каждого на $\sqrt{d}$ завершается первый этап корреляции внимания . Затем мы добавляем смещение и применяем функцию softmax, чтобы сжать выходные данные в диапазоне от 0 до 1 и получить оценку значения. Этот метод расчета внимания направлен на вычисление взвешенной суммы значений, которая аналогична «Вниманию масштабированного скалярного произведения».
Расчет внимания к себе записывается как

$$\text{Внимание}(Q, K, V) = So\ f\ tMax(QKT/\sqrt{d} + B)V, \text{ где } 1/\sqrt{d} — \quad (2)$$

масштабный коэффициент Q, K , а B — смещение Q, K, V.

Когда входной набор I проходит через кодер и отображается в представления B × M , представление f(xn) представляет собой двумерный вектор размером [49, 768].

Machine Translated by Google

При этом каждое M- представление задается как одна группа и используется для расчета потери согласованности. После применения линейного слоя к последнему базовому блоку Swin мы отправляем представления для расчета потерь перекрестной энтропии для классификации. Таким образом, каждая партия кадров требует выполнения нескольких операций с дополненными данными, после чего мы можем получить извлеченные низкоразмерные функции (см. рисунок 4). Наконец, будут рассчитаны два выходных вектора: один для расчета потерь согласованности, а другой, управляемый линейным слоем для потерь классификации. Подробное объяснение этих двух расчетов функции потерь можно найти см. в разделе 3.3.

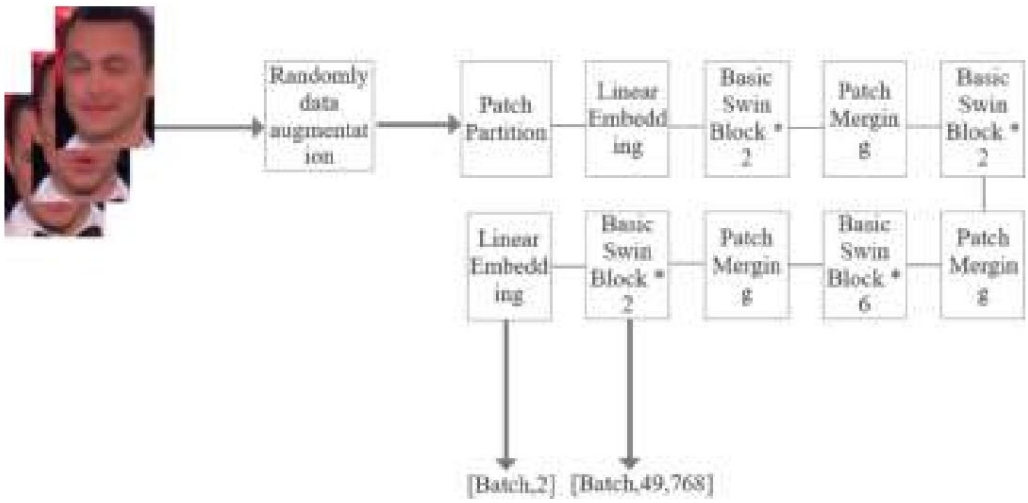


Рисунок 4. Архитектура магистральной сети Swin Transformer. показывает, сколько базовых блоков Swin применяется в магистральной сети. В нашей модели для каждого набора представлений входных данных есть два выходных вектора , которые представляют собой извлеченные низкоразмерные признаки размером [M × 49 × 768] и прогнозируемую вероятность [M × 2].

3.3. Функции потерь

Функция потерь этой модели состоит из двух компонентов: потери классификации и потери согласованности. Потеря согласованности применяется к каждой группе из M извлеченных представлений, и мы вычитаем их расстояния. Кроме того, причина, по которой мы используем косинусное расстояние для завершения расчета потери согласованности, заключается в том, что мы не заставляем каждую пару представлений быть абсолютно одинаковыми, а должны быть все более и более похожими. Вычисление косинусного расстояния каждой пары представлений можно записать как

$$LCos = (1 - fxi \cdot fxt) \quad (3)$$

где xi и xt представляют i-й и t-й кадры представлений данных, а f (xi) и f (xt) представляют собой нормализованные представления, извлеченные общим кодером.

Поскольку мы получили дополненные видеоклипы с количеством B -групп на входе нашей модели, нам также необходимо вычислить сходство между этими особенностями захваченных видеоклипов. Целью этого является решение некоторых неправильно различимых результатов твердых образцов путем сокращения их косинусных расстояний, и мы рассматриваем это как потерю согласованности. Что касается еще одного фактора, влияющего на общую потерю, потеря согласованности предназначена для расчета средних косинусных расстояний видео B -пакета. В частности, извлеченные M представлений одного видео могут генерировать C_M^2 пар сходств с помощью приведенного выше уравнения. Сингл согласованность одного видео может быть выражена как среднее сходство пары C_M^2 , а общая потеря согласованности видео серии B требует от нас расчета среднего значения всех рассчитанных отдельных согласованностей. Следовательно, потерю согласованности можно записать как

$$LCons = \frac{1}{Б}\sum_{1}^{Б} среднее(LCos) = \frac{1}{Б}\sum_{1}^{Б} среднее(1 - fxi \cdot fxt) \quad (4)$$

Machine Translated by Google

где B представляет размер пакета видео.

Основная цель настройки потери согласованности — доказать, что независимо от того, насколько реалистичным является один конкретный видеокадр дипфейка, модель способна сократить свое пространственное расстояние с другими поддельными кадрами. С другой стороны, потеря согласованности подтверждает согласованность классификации разных кадров одного видео. Поскольку работа по классификации неотделима от контролируемых меток, мы применяем стандартную кросс-энтропийную потерю в качестве потери классификации и формулируем ее следующим образом:

$$LCE = \sum_{1}^{M \times B} [y \log(p) + (1 - y)\log(1 - p)] \quad (5)$$

где y — метка основной истины, p — прогнозируемая вероятность, а M × B представляет общее количество видеокадров.

Поскольку наша модель представляет собой модель обучения согласованности, мы используем две функции потерь (потеря согласованности и потеря перекрестной энтропии) в линейной комбинации для формирования наших общих потерь. Однако мы устанавливаем гиперпараметр α в качестве весового баланса, чтобы наблюдать, какая часть этих двух функций потерь может достичь минимальных общих потерь и максимальной точности. В разделе 4.4 мы проводим один абляционный тест, чтобы наблюдать, какое значение этого гиперпараметра может предсказать большинство образцов, пересекая весовой параметр, добавляя 0,1 в каждом эксперименте, потому что мы надеемся найти наиболее подходящий вес, чтобы сбалансировать эти две потери. функций, и мы не надеемся, что общие потери в значительной степени смещены в сторону результатов одной из функций потерь.

$$Ltot = \alpha LCons + (1 \quad \alpha)LCE \quad (6)$$

где α — гиперпараметр, балансирующий потерю перекрестной энтропии и потерю согласованности, и он находится в диапазоне от 0 до 1.

4. Эксперименты 4.1. Наборы данных

Для оценки нашего метода мы выбрали четыре распространенных набора данных в этой области: FaceForensics ++ [23], DFDC [24], Celeb-DF [25] и FaceShifter [26]. В частности, первые два набора данных использовались на этапе обучения, а Celeb-DF V2 и FaceShifter использовались для экспериментов с перекрестными наборами данных, чтобы доказать способность нашей модели к обобщению. FaceForensics++ — это новаторский крупномасштабный набор данных по манипуляциям с лицами, который содержит 1000 реальных видео и 4000 фейковых видео, созданных с помощью Face2Face, FaceSwap, Neural Textures и Deepfakes.

DFDC — это самый большой набор данных из этих четырех наборов данных, содержащий 119 197 видеороликов, снятых реальными актерами. По сравнению с двумя предыдущими наборами данных, Celeb-DF получен из 590 оригинальных видеороликов YouTube и 5639 обработанных видеороликов с усовершенствованным смешиванием цветов, которое трудно различить человеческим глазом. FaceShifter — это один из подклассов наборов данных FaceForensics++, который содержит 1000 поддельных видеороликов с различными сценариями съемки. Учитывая несбалансированные данные обучения FaceForensics++, мы случайным образом объединили 2000 реальных видео из базы данных DFDC с реальными видео FF++, чтобы решить эту проблему. Что касается обучения DFDC, мы решили уменьшить количество поддельных видеокадров, чтобы устранить дисбаланс данных.

4.2. Детали реализации Для

предварительной обработки данных мы извлекли M кадров каждого видео, установив порог достоверности 0,5 и порог NMS 0,5 для «scrfd 10g kps.onnx». Кроме того, обнаруженные обрезанные лица с ограничивающими рамками были увеличены в 1,2 раза как по ширине, так и по высоте, чтобы обеспечить наличие некоторой фоновой информации на входных данных модели. Что касается настроек увеличения данных, мы установили 80% вероятность стирания области лица, диапазон соотношения стертой области по отношению к входному изображению был установлен от 0,02 до 0,2, а диапазон соотношения сторон стертой области было установлено от 0,5 до 2,0 в операции случайного стирания. В дополнении «Случайное кадрирование» мы установили нижнюю и верхнюю границы шкалы на 0,77 и 1, что означает, что случайная область обрезки до изменения размера составляла от 0,77 до 1 соответственно. В ходе эксперимента некоторые выбранные кадры (см. рис. 5) с

Machine Translated by Google

ни одно лицо человека не копировалось из последнего обнаруженного индексного кадра вместо того, чтобы напрямую отправляться в набор представлений входных данных.

Рисунок 5. Некоторые образцы дипфейков, обработанные «scrfd 10g kps.onnx», который является исходными данными для обучения нашей модели.

Для обучения модели мы установили размер мини-пакета видео равным 4 с размером изображения 224 × 224, а затем использовали оптимизатор AdamW [27] с затуханием веса, установленным на 0,05, и скоростью обучения, установленной на 0,0001. Чтобы эффективно обучить эту модель, мы прекратили обучение раньше, если не было увеличения точности проверки в первые 5 эпох, и общая эпоха была установлена равной 30. Кроме того, все коэффициенты отсева были установлены на 0,2, чтобы избежать переобучения в нашей модели. эксперимент. Все наши модели использовали swin-tiny-patch4-window7-224.pth и были предварительно обучены на ImageNet-1K в качестве магистрали, а выходные признаки извлекались с помощью линейного слоя размером [M, 49, 768] и векторы размером [M, 2]. Первоначально мы приняли сбалансированный вес 0,5, поскольку вначале считали, что потеря согласованности имеет такое же значение, как и потеря классификации.

4.3. Исследование в

наборе данных. В этом разделе мы оцениваем эффективность проверки предложенной модели в тех же источниках, что и наборы обучающих данных. Для экспериментов с набором данных мы оценили DFDC и FaceForensics++, которые представляют собой два относительно больших набора данных, поскольку мы обнаружили, что Swin Transformer всегда требует большого объема обучающих данных для достижения хорошего уровня проверки во время обучающих экспериментов; в противном случае он будет иметь относительно низкую начальную точность проверки в некоторых небольших наборах обучающих данных. И первоначальный гиперпараметр по умолчанию был установлен на 0,5. Мы сравнили наш метод согласованного обучения с тремя базовыми моделями: Capsule Network, STIL [28] и ISTVT [10]. Как показано в таблице 1, Capsule Network обеспечивает самую низкую точность, а ISTVT работает лучше, чем STIL. Наш подход работает намного лучше, чем Capsule Network, а точность набора данных немного выше, чем у STIL и ISTVT.

Machine Translated by Google

Таблица 1. Точность обнаружения Deepfake. Результаты с наилучшей точностью в наборах данных выделены жирным шрифтом. За исключением Капсульная сеть, которая воспроизводится, результаты других методов взяты из [10,28].

Методы	ФФ++				ДФДК
	ДФ	Ф2Ф	ФС	НТ	
Капсульная сеть	98,5%	96,7%	94,5%	91,8%	-
СТИЛЬ	99,6%	99,3%	97,1%	95,4%	89,8%
ИСТВТ	99,6%	99,6%	100%	96,8%	92,1%
Наш подход DF:	99,9%	99,6%	100%	98,8%	93,7%

Deepfakes; F2F: Лицо2Лицо; NT: нейронные текстуры; ФС: FaceShifter.

4.4. Результаты испытаний на абляцию

Чтобы проверить значимость ветви согласованности нашей модели, мы также провели исследование. эксперимент по абляции (см. Таблицу 2) по классификации дипфейковых изображений Swin Transformer с наш подход (с постоянным контролем обучения). Мы обнаружили, что последовательное обучение может эффективно повысить точность, поскольку наш подход использует потерю согласованности. сократить расстояния (увеличивает сходство) между разными кадрами видео, чтобы что если один кадр распознается веткой классификации как фейковый с действительно высоким показатель достоверности, существует высокая вероятность того, что будут учитываться другие жесткие рамки выборки так же фейк.

Таблица 2. Сравнение точности эксперимента по абляции. Мы использовали Swin Transformer для проведения бинарный тест deepfake-изображения и сравните его точность с точностью нашего подхода.

Методы	FF++(Двоичный тест)	ДФДК
Свин Трансформатор	89,4%	91,4%
Наш подход	94,3%	93,7%

Кроме того, балансовый вес α является еще одной переменной, которая определяет окончательную общую потерю; таким образом, мы провели эксперимент по абляции, чтобы определить, какой вес требуется для потери согласованности, чтобы показать лучшую производительность в этой дискриминационной модели за счет установка этого гиперпараметра на 0,4, 0,5, 0,6, 0,7 и 0,8. В этом тесте на абляцию мы все еще использовали FaceForensics++ в качестве набора данных для изучения того, какой балансовый вес может достичь наивысшего уровня. прогнозируемая точность. Кроме того, для проверки мы объединили все фейковые категории в один класс. двоичная производительность вместо тестирования нескольких классификаций. Потому что мы не устанавливали это баланс веса в качестве обучаемого параметра в нашей модели, мы решили отслеживать вес баланс, добавляя 0,1 за эксперимент. Как только мы получили максимальную точность, мы считали эту модель как современную модель, и мы использовали обученные параметры этой модели проводить эксперименты с перекрестными наборами данных. Как показано в таблице 3, мы находим, что вес настройка баланса 0,7 показывает лучшую производительность и была обучена на FaceForensics++. бинарный тест.

Таблица 3. Точность эксперимента по абляции на различных весах. Это для проверки того, какой вес баланс нашей функции потерь позволит достичь наилучшего результата.

α	0,4	0,5	0,6	0,7	0,8
Точность	92,6%	93,9%	94,1%	94,3%	93,7%

4.5. Междисциплинарное исследование

Из-за множества методов подделки важно оценить производительность модели на разных наборах данных. Чтобы проверить способность нашей модели к обобщению, мы обучили нашу модель на FaceForensics++ и протестировал его на Celeb-DF, FaceShifter и DFDC. Учитывая, что размеры выборки большинства наборов данных дипфейков для реального и фиктивного классов несбалансированы, мы

Machine Translated by Google

также решили использовать площадь под кривой (AUC) для оценки надежности нашего
модель, как показано в таблице 4.

Таблица 4. AUC эксперимента с несколькими наборами данных. Мы обучили нашу модель на FF++ и протестировали
невидимые данные. Лучшие результаты AUC для разных наборов данных выделены жирным шрифтом. Результаты других методов
из [10,28–31].

Модели	Знаменитость-DF	FaceShifter	ДФДК	Средний
Капсульная сеть	34,5%	42,4%	-	38,5%
СТИЛЬ	75,6%	-	-	75,6%
ИСТВТ	84,1%	99,3%	74,2%	85,9%
Рентген лица	79,5%	92,8%	65,6%	79,3%
ФТЦН	86,9%	98,8%	74,0%	86,6%
Губкриминалистика	82,4%	97,1%	73,5%	84,3%
Наш подход	88,8%	98,3%	71,8%	86,3%

Мы обнаружили, что наш подход может дать наилучшие результаты только на Celeb-DF. Этот
потому что Celeb-DF содержит расширенную постобработку, такую как смешивание цветов.
и действительно сложно определить, когда происходит несоответствие границ в сгенерированных носителях.
Чтобы доказать эту точку зрения, мы решили создать тепловые карты нашей модели и других моделей обнаружения
на основе CNN, чтобы объяснить визуализацию модели областей фокуса дипфейков.
СМИ. Наблюдая за тепловой картой нашей модели, мы обнаружили, что наша модель больше фокусируется на
на сгенерированном содержимом человеческого лица, таком как губы и глаза, вместо граничной текстуры
информацию (см. рисунок 6). Проверив несколько примеров видеороликов FaceShifter, мы обнаружили, что
Признаки подделки всегда встречаются в области лицевых границ, то есть в той области лица, которая
наша модель уделяет меньше внимания, но признаки подделки Celeb-DF проявляются в чертах лица,
например, неестественно манипулируемые человеческими глазами и губами. Кроме того, мы сравнили данные
в Celeb-DF с данными DFDC, и мы обнаружили, что DFDC содержит больше данных с
различные сценарии съемки, например темный фон в помещении и лица с экстремальными
позы. Из-за этого нашей модели очень сложно достичь согласованности и получить полезную информацию.
информацию, позволяющую отличить настоящие СМИ от фейковых, поскольку мы использовали модель, обученную
FaceForensics++, и в него вошли больше человеческих лиц в нормальных позах и яркая съемка.
сценариев, что аналогично распределению данных Celeb-DF.

Кроме того, мы также сравнили метод ISTVT на основе трансформатора в качестве эталона.
модель, и она достигла немного лучших результатов AUC на FaceShifter и DFDC (см. Рисунок 7).
Это связано с тем, что авторы, использовавшие ISTVT, применили механизм самовычитания и
временное внимание к себе в своей работе, чего не хватает нашему подходу. Самовычитание
механизм может больше сосредоточиться на межкадровых искажениях, что является еще одним важным методом.
для обнаружения дипфейков. Это требует применения входного токена перед проекцией.
для запроса и ключевых векторов. Он генерирует остатки объектов для замены исходных объектов.
Тем не менее, наша модель по-прежнему достигает конкурентоспособных результатов по сравнению с другими предыдущими моделями.
самые современные модели, такие как STIL и Face X-ray.

Machine Translated by Google

Рисунок 6. Сравнение тепловой карты между другими методами обнаружения дипфейков и нашим подходом. Из этого рисунка видно, что большинство детекторов дипфейков фокусируются на граничной информации сгенерированных фейковых медиа.

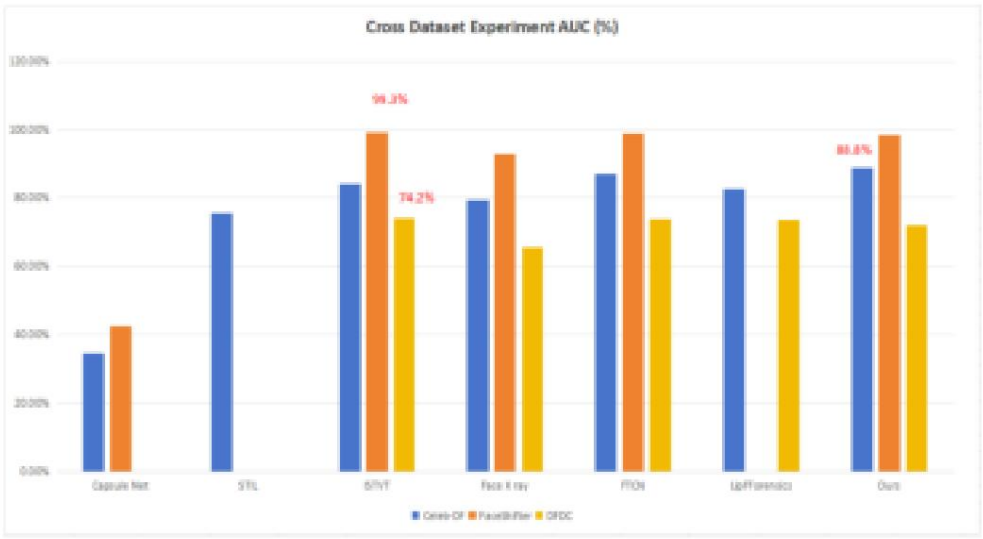


Рисунок 7. Результаты AUC для разных наборов данных на основе Capsule Net, STIL, ISTVT, Face X-ray, FTCN, LipForensics и нашего подхода. Наилучшая производительность для каждого набора данных отмечена на рисунке.

5. Выводы и будущая работа

В этой статье мы предлагаем Swin-Fake, который впервые сочетает в себе Swin Transformer с потерей согласованности, в качестве метода обучения согласованности для обнаружения дипфейков. Мы используем Swin Transformer в качестве средства извлечения признаков и используем среднее косинусное расстояние в качестве меры потери согласованности, что улучшает способность обобщения в некоторых базах данных deepfake. Мы доказали, что классификатор на основе Transformer может достичь более высокого уровня в области обнаружения дипфейков по сравнению с другими предыдущими методами на основе CNN. На этапах обучения и проверки мы пришли к выводу, что Swin Transformer обладает лучшей способностью различать дипфейки, в то время как крупномасштабные базы данных, такие как FF++ и DFDC, не демонстрируют многообещающей производительности на некоторых небольших наборах данных, таких как Celeb DF. Кроме того, вычисление согласованности между характеристиками кадров может значительно уменьшить потери и улучшить показатели оценки; таким образом, мы считаем, что применение метода согласованного обучения с Transformer для обнаружения дипфейков может стать новой тенденцией в обнаружении фейковых СМИ.

Обнаружение дипфейкового видео невозможно выполнить без временной информации, поскольку в соседних кадрах всегда происходит дрожание кадров. В недавних исследованиях временное внимание всегда сочетается с пространственным вниманием при обнаружении видео. В качестве еще одного метода в области обнаружения видео временное внимание можно рассматривать как динамический механизм временного выбора, который определяет, когда следует обратить внимание на определенную особенность видео, и часто применяется при обработке видео. Это также эффективный метод повышения оценки.

Machine Translated by Google

метрики в этой области, если мы можем объединить информацию о временной согласованности с извлеченной пространственной информацией, и наиболее распространенным методом является разработка специальных модулей для исследования временной согласованности или несогласованности путем применения горизонтальных и вертикальных сверточных слоев. Таким образом, в будущей работе мы продолжим исследовать временное самообслуживание магистрали на основе Swin Transformer. Несмотря на то, что наша текущая работа сосредоточена на детекторах на основе видео, мы по-прежнему фокусируемся на извлечении пространственных признаков, а для внутриклассового сравнения применяется согласованность признаков. Сочетание временной и пространственной согласованности, вероятно, потребует больше информации, что пойдет на пользу классификации видео. Что касается весового баланса α, то мы добавляли 0,1 в каждом эксперименте, чтобы проверить эффективность блока согласованности и получить значение с наибольш Чтобы получить более точное значение этого гиперпараметра, мы установим весовой баланс в качестве обучаемого параметра в будущих исследованиях. Кроме того, мы продолжим исследовать другие методы визуализации для детекторов дипфейков на основе Swin Transformer, чтобы интерпретировать, какие пространственные особенности фиксируются, и внести свой вклад в классификатор дипфейков.

Вклад автора: концептуализация, LYG и XJL; методология, LYG и XJL; программное обеспечение LYG; валидация, LYG; формальный анализ, LYG; расследование, LYG; ресурсы, LYG и XJL; курирование данных, LYG и XJL; письмо — подготовка оригинального проекта, LYG, XJL и PHJC; написание — обзор и редактирование, LYG, XJL и PHJC; визуализация, LYG и XJL; надзор, XJL и PHJC; администрация проекта, XJL и PHJC. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: соответствующие наборы данных загружены из [24], коды, воспроизведенные DFDC, взяты из [13].

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Гудфеллоу И.; Пуже-Абади, Дж.; Мирза, М.; Сюй, Б.; Варед-Фарли, Д.; Озаир, С.; Курвиль, А.; Бенджио, Ю. Генеративно-состязательная сеть. Коммун. Акм 2018, 63, 139–144. [Перекрестная ссылка]
2. Каррас Т.; Лейн, С.; Айла, Т. Архитектура генератора на основе стилей для генеративно-состязательных сетей. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Лонг-Бич, Калифорния, США, 15–20 июня 2019 г.
3. Хо, Дж.; Джайн, А.; Аббил, П. Вероятностные модели диффузии с шумоподавлением. В материалах 34-й Международной конференции по Система обработки нейронной информации, Ред-Хук, Нью-Йорк, США, 6–12 декабря 2020 г.; стр. 6840–6851.
4. Нгуен, Х.Х.; Ямагиши, Дж.; Этизен И. Использование капсульной сети для обнаружения поддельных изображений и видео. В материалах по компьютерному зрению и распознаванию образов, Лонг-Бич, Калифорния, США, 16–20 июня 2019 г.
5. Дас, С.; Сефербеков С.; Датта, А.; Ислам, MS; Амин, М.Р. На пути к решению проблемы дипфейков: анализ улучшения обнаружения дипфейков с использованием динамического увеличения лица. В материалах Международной конференции IEEE/CVF по компьютерному зрению, Монреаль, Британская Колумбия, Канада, 11–17 октября 2021 г.
6. Чжао, Т.; Сюй, Х.; Сюй, М.; Дин, Х.; Сюн, Ю.; Ся, В. Изучение самосогласованности для обнаружения дипфейков. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Нэшвилл, Теннесси, США, 20–25 июня 2021 г.
7. Досовицкий А.; Бейер, Л.; Колесников А.; Вайссенборн, Д.; Чжай, Х.; Унтертинер, Т.; Дегани, М.; Миндерер, М.; Хейгольд, Г.; Гелли, С.; и другие. Изображение стоит 16 × 16 слов: преобразователи для распознавания изображений в масштабе. В материалах Международной конференции по изучению представлений, виртуальной, 26 апреля – 1 мая 2020 г.
8. Хан, С.А.; Дай, Х. Видеотрансформатор для обнаружения дипфейков с инкрементным обучением. В материалах 29-го ACM Международная конференция по мультимедиа, виртуальное мероприятие, 20–24 октября 2021 г.; стр. 1821–1828.
9. Хормали А.; Юань, Дж. DFDT: Комплексная система обнаружения DeepFake с использованием Vision Transformer. Прил. наук. 2022, 12, 2953. [Перекрестная ссылка]
10. Чжао, К.; Ван, К.; Обнимать.; Чен, Х.; Лю, К.; Тан, Дж. ISTVT: интерпретируемый пространственно-временной преобразователь видео для Deepfake Обнаружение. IEEE Транс. Инф. Криминалистическая безопасность. 2023, 18, 1335–1348. [Перекрестная ссылка]
11. Лю, З.; Лин, Ю.; Цао, Ю.; Хм.; Вэй, Ю.; Чжан, З.; Лин, С.; Го, Б. Суин Трансформатор: Трансформатор иерархического зрения с использованием сдвинутых окон. В материалах Международной конференции IEEE/CVF по компьютерному зрению, Монреаль, Британская Колумбия, Канада, 11–17 октября 2021 г.
12. Чжун З.; Чжэн, Л.; Канг, Г.; Ли, С.; Ян, Ю. Увеличение данных с произвольным стиранием. В материалах конференции AAAI по Искусственный интеллект, Нью-Йорк, США, 7–12 февраля 2020 г.
13. ДФДК Селим. Доступно онлайн: https://github.com/selimsef/dfdc_deepfake_challenge . (по состоянию на 15 мая 2024 г.).

Machine Translated by Google

14. Шолле, Ф. Xception: Глубокое обучение с глубоко отделимыми извилинами. В трудах по компьютерному зрению и шаблонам Recognition, Гонолулу, Гавайи, США, 21–26 июля 2017 г.
15. Он, К.; Чжан, Х.; Рен, С.; Сан, Дж. Глубокое остаточное обучение для распознавания изображений. В материалах конференции IEEE 2016 г. по компьютерному зрению и распознаванию образов, Лас-Вегас, Невада, США, 27–30 июня 2016 г.
16. Ни, Ю.; Мэн, Д.; Ю, Ц.; Цюань, СВ; Рен, Д.; Чжао, Ю. CORE: Обучение согласованному представлению для обнаружения подделки лиц. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Новый Орлеан, Луизиана, США, 18–24 июня 2022 г.
17. Сан, К.; Яо, Т.; Чен, С.; Дин, С.; Ли, Дж.; Джи, Р. Двойное контрастное обучение для общего обнаружения подделок лиц. Конференция АААИ. Артиф. Интел. 2022, 36, 23:16–23:24. [Перекрестная ссылка]
18. Чжан Дж.; Ченг, К.; Соверниго, Г.; Лин, Х. Метод обнаружения дипфейков на основе обучения ансамблю гетерогенных признаков. В материалах ICC 2022 — Международная конференция IEEE по коммуникациям, Сеул, Республика Корея, 16–20 мая 2022 г.; стр. 2084–2089.
19. Рана, М.С.; Сунг, А. Х. DeepfakeStack: метод глубокого ансамблевого обучения для обнаружения Deepfake. В материалах 7-й Международной конференции IEEE по кибербезопасности и облачным вычислениям (CSCloud) 2020 г. / 6-й Международной конференции IEEE по периферийным вычислениям и масштабируемому облаку (EdgeCom), Нью-Йорк, США, 1–3 августа 2020 г.; стр. 70–75.
20. Ильяс Х.; Джавед, А.; Малик, К.М. AVFakeNet: унифицированная сквозная модель глубокого обучения Dense Swin Transformer для аудиовизуальных обнаружение дипфейков. Прил. Мягкий компьютер. 2023, 136, 110124. [CrossRef]
21. Халид Ф.; Акбар, Миннесота; Гул, С. СВИНТ: Трансформаторы Swin Y-Net для обнаружения дипфейков. В материалах Международной конференции по робототехнике и автоматизации в промышленности 2023 г. (ICRAI), Пешавар, Пакистан, 3–5 марта 2023 г.; стр. 1–6.
22. Го, Дж.; Дэн, Дж.; Латтас, А.; Зафейриу, С. Перераспределение выборки и вычислений для эффективного обнаружения лиц. В материалах Международная конференция по репрезентации обучения, виртуальное мероприятие, 3–7 мая 2021 г.
23. Чжоу, Т.Ф.; Ван, РГ; Лян, ZY; Шен, Дж. Б. Криминалистика лиц в дикой природе. В трудах по компьютерному зрению и шаблонам Признание (CVPR), виртуальное, 19–25 июня 2021 г.
24. Кэггл. Доступно онлайн: https://www.kaggle.com/c/deepfake-detection-challenge/overview. (по состоянию на 12 декабря 2023 г.).
25. Ли, Ю.З.; Ян, Х.; Сан, П.; Ци, ХГ; Лю, С. Celeb-DF: Крупномасштабный сложный набор данных для криминалистики DeepFake. В материалах компьютерного зрения и распознавания образов, Сиэтл, Вашингтон, США, 14–19 июня 2020 г.
26. Ли, Л.; Бао, Дж.; Ян, Х.; Чен, Д.; Вэнь, Ф. FaceShifter: к смене лиц с высокой точностью и учетом окклюзии. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Сиэтл, Вашингтон, США, 13–19 июня 2020 г.
27. Лощилов И.; Хаттер, Ф. Регуляризация разделения веса. В материалах Международной конференции по изучению представлений, Новый Орлеан, Луизиана, США, 6–9 мая 2019 г.
28. Гу, З.; Чен, Ю.; Яо, Т.; Дин, С.; Ли, Дж.; Хуанг, Ф.; Ма, Л. Обучение пространственно-временной несогласованности для обнаружения дипфейковых видео. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, виртуальной, 19–25 июня 2021 г.
29. Ли, Л.; Бао, Дж.; Чжан, Т.; Ян, Х.; Чен, Д.; Вэнь, Ф.; Го, Б. Рентген лица для более общего обнаружения подделки лица. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Сиэтл, Вашингтон, США, 13–19 июня 2020 г.
30. Чжэн Ю.; Бао, Дж.; Чен, Д.; Цзэн, М.; Вэнь, Ф. Исследование временной когерентности для более общего обнаружения подделки лиц на видео. В материалах Международной конференции IEEE/CVF по компьютерному зрению, Монреаль, Квебек, Канада, 10–17 октября 2021 г.; стр. 15044–15054.
31. Халиасос, А.; Вугюкас, К.; Петридис, С.; Пантик, М. Липс не лжет: Обобщенный и надежный подход к обнаружению подделки лиц. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, виртуально, 19–25 июня 2021 г.; стр. 5039–5049.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Сеть профессиональных гидроэлектростанций как устойчивая распределенная Энергетическое депо

Михал Моравский и Пшемыслав Игначук *

Факультет технической физики, информационных технологий и прикладной математики, Лодзинский технологический университет, 93-005 Лодзь, Польша; michal.morawski@p.lodz.pl * Адрес для переписки: przemyslaw.ignaciuk@p.lodz.pl

Аннотация: Нехватка эффективных и недорогих хранилищ сдерживает широкомасштабное внедрение энергозависимых по своей природе возобновляемых источников энергии (ВИЭ). В этой статье мы описываем, как использовать гидроэлектростанции, принадлежащие профессиональным потребителям, мощностью от нескольких до нескольких кВт в качестве распределенного решения для краткосрочного хранения энергии, которое можно развернуть с небольшими инвестициями и низкими эксплуатационными расходами. Предлагаемое решение представляет собой систему взаимосвязанных гидробазов с активным алгоритмом управления расходом воды, который снижает изменчивость нагрузки сети и приносит пользу потребителям. Согласно проведенным тестам, доходы потребителей растут от нескольких процентов до более чем 30 процентов, в зависимости от погодных условий, по сравнению со случаем свободного потока. В свою очередь, амортизирующий эффект распределенного энергетического буфера уравновешивает колебания, вносимые другими ВИЭ, например, фотоэлектрическими или ветровыми. Следовательно, принося пользу участвующим сторонам, это также облегчает включение ВИЭ в систему распределения электроэнергии.

Ключевые слова: энергобаза; гидростанции; распределенные системы

1. Введение

Электрическая энергия считается кровотоком современного общества. Использование электроэнергии в домах, промышленности, на транспорте, а также при охлаждении и отоплении зданий приводит к ежедневному увеличению спроса. Между тем, производство энергии по-прежнему остается дорогостоящим как с экологической, так и с экономической точки зрения. Глобальное загрязнение, растущая стоимость ископаемого топлива и политическое давление — это лишь несколько факторов, которые стимулируют производство энергии из возобновляемых источников (ВИЭ), например, солнца, ветра, волн, геотермальной энергии и движения воды [1]. К сожалению, количество энергии, получаемой от ВИЭ, сильно зависит от временных погодных условий. ВИЭ обеспечивают недорогую энергию либо в произвольные часы (ветер), либо около полудня (солнечная энергия, в основном фотоэлектрическая (PV)). Напротив, спрос остается на том же уровне до вечера, как показано на рисунке 1. Следовательно, дорогостоящие электростанции, работающие на ископаемом топливе, необходимо задействовать по вечерам, но их деятельность должна быть ограничена в пользу ВИЭ, если Доступно солнце или ветер. Более того, производство энергии на традиционных тепловых электростанциях не может быть слишком сильно ограничено из-за технологических факторов. В результате рыночной динамики цены на энергоносители в определенное время устанавливаются на низком (или даже отрицательном) уровне, а через несколько часов достигают более высокого порядка, что напоминает «утиную кривую», изображенную на рисунке 2, где показаны первые дни экономического роста. Июль 2023 г. на польском энергетическом рынке. В интересах всех заинтересованных сторон это неравенство должно быть уменьшено.

Решение проблемы неравенства между производителем и потребителем, которое мы добровольно взяли на себя, состоит в том, чтобы хранить излишки энергии, когда они дешевы, на более поздний период, когда будут подключены дорогостоящие источники. К сожалению, современные технологии не позволяют реализовать это в больших масштабах [2]. Например, гидроэлектростанции в Польше могут обеспечить не более 5% от общей потребности в электроэнергии, а аккумуляторные депо — не более 1%. Оба этих традиционных метода, а также перспективные системы на основе водорода дороги в развертывании и обслуживании. Низкая долговечность электромеханических батарей еще больше ограничивает возврат инвестиций.

Следовательно, желателен поиск нового варианта хранения энергии с низкими капитальными затратами (капитальные затраты) и низкими OPEX (эксплуатационные затраты).

check for updates

Цитирование: Моравски, М.; Игначук, П. Сеть потребительских гидроэлектростанций как хранилище устойчивой распределенной энергии. Электроника 2024, 13, 3043. https://doi.org/10.3390/электроника13153043

Академические редакторы: Темитайо Олову и Рагхав Кханна.

Поступила: 24 июня 2024 г.
Пересмотрено: 22 июля 2024 г.
Принято: 30 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

Machine Translated by Google

В последнее время на рынке электроэнергии наблюдаются существенные изменения. Традиционные крупные электростанции дополняются небольшими объектами, устанавливаемыми гражданами на собственных объектах. Эти граждане являются потребителями электроэнергии в определенные часы дня и производителями в другое время, поэтому их называют просьюмерами. С одной стороны, популярность профессиональных инсталляций способствует внедрению RSE; с другой стороны, это угрожает стабильности системы энергоснабжения, вызывая экстремальные изменения нагрузки на традиционные электростанции.

Это исследование способствует решению вышеупомянутых проблем следующим образом: • Мы предлагаем привлечь профессиональные гидроэлектростанции с пиковой мощностью в диапазоне нескольких кВт для формирования распределенного энергодепо;

• Мы разрабатываем схему активного управления, которая позволяет как увеличить доходы владельцев станций, так и снизить нестабильность нагрузки на устаревшие тепловые станции («обезглавить утку» [3]), тем самым достигая беспроигрышного решения; • В отличие от типичного подхода, который фокусируется на одном гибридном генераторе, например, [4], мы показываем, как координировать сеть предприятий для безопасной и эффективной работы распределенной системы.

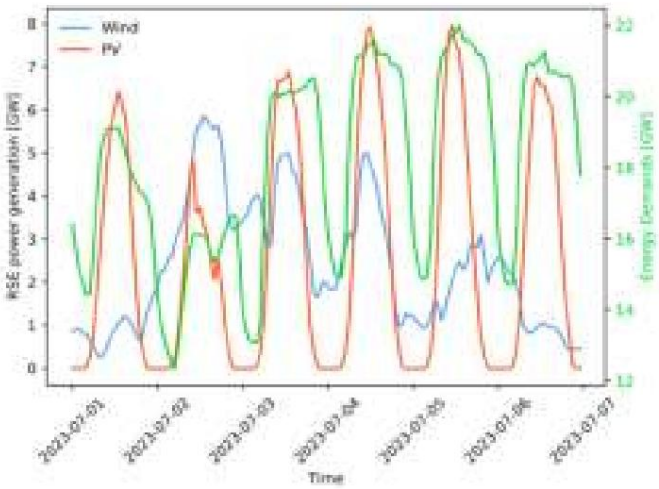


Рисунок 1. Производство энергии основными ВИЭ и прогнозируемый спрос на энергию в Польше на 1–6 июля 2023 г. (1 июля было субботой [5]).

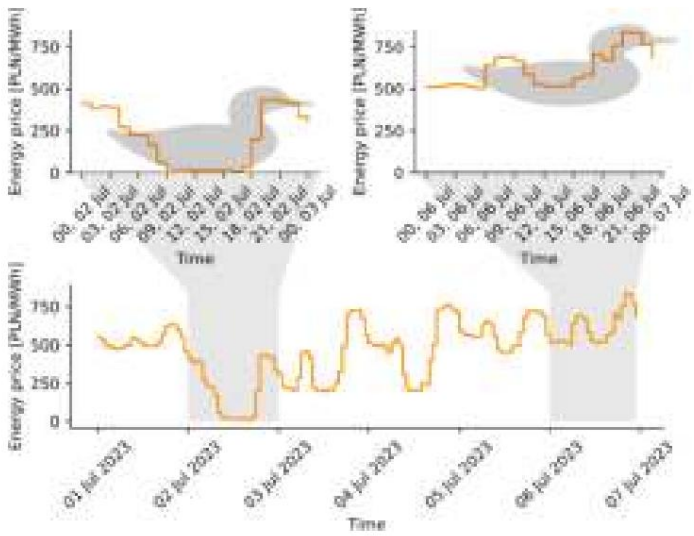


Рисунок 2. Колебания рыночных цен на энергоносители следующего дня [PLN/МВтч] на польском рынке в период с 1 по 6 июля 2023 года в соответствии с «утиной кривой» [5].

Machine Translated by Google

Оставшаяся часть статьи организована следующим образом. В разделе 2 обсуждаются традиционные методы управления гидроэлектростанциями. В разделе 3 мы подробно рассмотрим различия между обычными и полупотребительскими растениями. В разделе 4 построена дискретная динамическая модель потока воды в системе нескольких гидростанций и описана процедура оптимизации ее работы. В разделе 5 мы оцениваем производительность алгоритма для широкого спектра тестовых данных. Затем мы обсудим вопросы реализации в разделе 6.
Наконец, выводы сделаны в разделе 7.

2. Управление

гидроэлектростанциями. В литературе гидроэлектростанции (ГЭС) классифицируются на нормальные (>100 МВт), средние (<100 МВт), малые (<15 МВт), мини (<1 МВт), микро (<100 МВт). кВт) или пико (<5 кВт) л.с., хотя фактические значения варьируются в зависимости от региона [6,7]. Большая часть исследований сосредоточена на малых, средних и обычных ТН, которые играют заметную роль в электроснабжении. В данной статье, ввиду их значимости для просьюмеров, исследуются микро- и пико-ГП.
Они включают в себя многочисленные бывшие заброшенные мельницы с небольшими искусственными прудами, болотами и другими природными образованиями, широко известные как русловые ГП [8,9], что требует незначительных капитальных затрат на их развертывание. Например, в настоящее время в Польше для производства энергии используется менее 5% возможных установок [10]. Более того, многочисленные онлайн-публикации сервисов показывают, что пико-НР можно легко развернуть без специального оборудования [11]. Таким образом, целевое направление исследования, проводимого в данной работе, имеет важное значение. Конечно, наличие одного-единственного депо для полупотребителей не меняет правила игры. Тем не менее, ориентированный на прибыль, оптимально запланированный набор профессиональных НР может представлять собой эффективную систему хранения энергии, устойчивую к возмущениям, с низкими эксплуатационными расходами. Фактически, подобные обстоятельства привели к популярности фотоэлектрических установок, которые стали обычным элементом современного ландшафта.

Однако существует фундаментальная разница между фотоэлектрическими и НР установками. Фотоэлектрические станции взаимно независимы, тогда как правильная работа НР зависит от того, чтобы все вышестоящие станции были объединены в каскад. В прошлом планирование работы каскада ТН изучалось как тематическое исследование [12], а не как сетевая система, подвергающаяся управляющему воздействию, т.е. каждая установка настраивалась вручную без соблюдения общей структуры. Хотя таким образом можно определить подходящий план планирования , он может не учитывать оптимальные условия эксплуатации или представлять риск с точки зрения защиты от засухи/наводнения. Другие подходы рассматривают упрощенные модели, сводя каскад к случаю с одной станцией [4], когда плотины находятся на близком расстоянии [8,13] или расположены на одной реке, а не на ее притоках [9,14]. Решение, концептуально наиболее близкое к обсуждаемому здесь, было предложено в [13], хотя оно ориентировано на крупные, концентрированные установки , а не на распределенные НР под управлением полупотребителей.

Технически управление производством энергии на гидроэлектростанции не является сложным процессом. Поднимая плотины, можно ограничить поток воды, а, опустив их, увеличить выработку энергии. Однако установление правильной последовательности подъема и опускания плотин далеко не очевидно. Обычно стратегии планирования подразделяют на долгосрочные (1–5 лет), среднесрочные (3–18 месяцев) и краткосрочные (2–14 дней) [15]. Тем не менее, в случае с полупрофессиональными НР такие стратегии имеют ограниченную полезность из-за небольшого объема прудов. В рассматриваемом случае содержательная стратегия планирования должна охватывать диапазон нескольких часов или, самое большее, одного дня. Более того, ему следует сосредоточиться на уменьшении дисбаланса генерации и потребления в вечернее и утреннее время (рис. 2) или, что то же самое, воспользоваться преимуществами колебаний рыночных цен, возникающих в результате несоответствия между прогнозируемым и фактическим спросом [16] (рис . 1).

Оптимальное планирование задачи сброса воды ВД исследуется уже давно ([15,17] и ссылки в них). Однако в этих исследованиях ТН рассматривались только в контексте защиты от наводнений/засухи или производства электроэнергии. Здесь исследуется третья роль НР, т.е. потенциал обеспечения хранилища энергии [2]. Вместо того , чтобы сосредотачиваться только на количестве вырабатываемой электроэнергии, ТН могут ограничивать речной сток, когда энергия дешева, и увеличивать его, когда цена растет, создавая тем самым эффективную

Machine Translated by Google

энергетический буфер. Общий объем генерируемой энергии останется прежним, но доходы владельцев НР будут расти за счет краткосрочных изменений ценового профиля. Одновременно уменьшаются колебания выработки энергии традиционными тепловыми электростанциями . Похожая идея была исследована в [18], но она предназначена для крупных установок, а не для полупотребительских, рассматриваемых в данной работе.

3. Профессиональные

гидроэлектростанции Не все ВИЭ чувствительны к влиянию погодных колебаний, например, геотермальные, приливные или крупные дорогостоящие гидроэлектростанции [19]. Здесь основное внимание уделяется небольшим, гибким установкам, т.е. микро- [6] и пико-НР [20]. Обычно они развертываются в виде русловых сооружений [8,21], т.е. без крупных плотин и водохранилищ, и поэтому не требуют значительных затрат или вмешательства в окружающую среду [21]. В отличие от более крупных ГП, на которые не влияют местные атмосферные условия, поскольку количество воды, превышающее прогнозы, незначительно по сравнению с объемом водохранилища, ограниченное пространство микро- и пикорезервуаров делает их восприимчивыми к погодным изменениям. Например, локальный шторм может быстро заполонить пруд вблизи одного микроГП, в то время как в соседнем пруду одновременно будет солнечно и сухо. Сегодня прогнозирование осадков недостаточно точно, чтобы справиться с такими ситуациями. Модели водного стока также неточны [22]. Поэтому, когда погода нестабильна, график работы завода приходится часто пересчитывать, чтобы приспособить систему к самым последним обстоятельствам или изменчивости цен на балансирующем рынке в реальном времени [16]. В отличие от широко используемых подходов численной оптимизации, алгоритм, представленный в этой рукописи, выполняется за доли секунды, даже на устройстве начального уровня.

Ряд ГЭС использует воду из вышележащих станций и притоков. Следовательно, неправильный контроль может привести к наводнению. Поэтому обязательна синхронизация работы ТН, в явном виде учитывающая

различные факторы: (1) Географический фактор — расстояние между станциями и связанная с этим заде

тер перекомпоновка;

(2) Фактор, связанный с погодой – прогнозируемый уровень осадков и уровень неопределенности;
3. Прикладной фактор – использование каскада ВД в качестве распределенного хранилища энергии.

в сочетании с общей функцией производства электроэнергии.

Обычно график управления устанавливается с использованием процедур численной оптимизации, включающих методы искусственного интеллекта для решения вычислительной нагрузки [15,17,21–29]. Однако эти процедуры демонстрируют проблемы с надежностью в рассматриваемой области применения даже в относительно простых случаях [29]. Более того, большое количество времени, необходимое для принятия решения (с использованием разумных вычислительных ресурсов), противоречит необходимости адаптивности плана, т. е. внесения изменений в график управления при нестабильной погоде или возникновении неисправности.

Для учета факторов (1)–(3) в построенную в данной работе динамическую модель явно включена информация о различных задержках потоков воды на звеньях, соединяющих водоемы. Используя динамическое представление системы, ставится задача оптимизации. Ее можно решить аналитически, если насыщение отсутствует [30]. Напротив, здесь естественные ограничения встроены в алгоритмическое решение. Это предложение приносит выгоду не только потребителям, увеличивая их экономическую выгоду, но и операторам электросетей за счет уменьшения колебаний нагрузки для стандартных электростанций.

4. Модель системы .

Обычно моделирование НР концентрируется на оптимизации работы генераторов, игнорируя систему снабжения [23]. Здесь генератор рассматривается как черный ящик, а акцент делается на подключаемые ТН как на систему водоснабжения. Ключевым моментом, который следует учитывать в системе обратного потока воды, является немаловажное время между выдачей управляющего воздействия в одном резервуаре и до того, как оно повлияет на уровень воды в резервуаре, расположенном ниже по течению. Поэтому, в отличие от более ранних моделей сетей хранения, например, [31], в предлагаемом здесь подходе будут применяться принципы управления системами хранения с за

Machine Translated by Google

предложенные в [32,33] предполагают регулировку управления в непрерывном режиме, что сложно реализовать в системе управления водными ресурсами из-за особенностей механических компонентов, управляющих плотинами. Модель в этой работе явно охватывает эффекты конечного шага расчета и будет построена непосредственно в области дискретного времени.

4.1. Одностанционное

депо. Рассмотрим модель одного ГП как составного элемента системы, представленной на рис . 3. Динамику водного баланса на станции можно описать рекурсивным соотношением

$$sj(k + 1) = sj(k) \quad fj(k) + \quad i \quad \text{растения выше по течению} \quad fi(k \quad Tij) + rj(k), \tag{1}$$

где

- k [1, m] — момент времени, m — горизонт планирования;
- sj (k) – количество воды, запасенной в резервуаре возле завода j (на рисунке 3 завод j = 3 показан в увеличенном виде); sj (k) – гидротехнические сооружения, а также близлежащие болота и аналогичные территории; • fj (k) – количество воды, используемое для привода электрогенераторов, установленных на плотине j, за время между моментами k и k + 1;
- rj (k) — поступление из внешних гидрологических источников, таких как дождь (и его сток), тающий снег, неконтролируемые притоки и испарение. Значения rj (k) можно получить из прогноза погоды и гидрологических моделей в пределах горизонта планирования m периодов. rj (k) может быть положительным (преобладание осадков) или отрицательным (преобладание испарения). В модели предполагается, что rj (k) известен. Когда истинное значение rj (k) отличается от прогноза, алгоритм запускается снова, не требуя сложных вычислений, как описано в разделе 4.3.

Притоки снабжают водохранилище водой, ранее использовавшейся растениями выше по течению. Вода из расположенного выше по течению завода i поступает на завод j с задержкой Tij > 0. Продолжительность периода k , т. е. время между моментами k и k + 1, может быть выбрана произвольно, но в зависимости от темпа изменения цен разумно выбрать от 1 часа до 15 минут, в зависимости от местных правил. Аналогично, горизонт планирования m может охватывать 24-часовое окно известных цен на энергию (рынок следующего дня). Начальный расход fj (k 0) и начальный уровень воды sj (0) считаются известными. Терминальное условие sj (m) может быть выбрано произвольно.

Доход от предприятия, полученный за один период, можно рассчитать как

$$Jj(k) = \eta j \; pj(k)fj(k) \quad k, \tag{2}$$

где

- pj (k) — цена энергии в момент k. Тот факт, что все HP подключены к одной и той же системе, не приводит к одинаковому ценовому профилю для каждого завода j. Владельцы могут иметь контракты с разными компаниями, использовать разные тарифы или сами иметь разные модели потребления;
- ηj — эффективность электрогенераторов, учитывающая влияние высоты плотины. Для полупотребительских генераторов в низинах расход 1 м3/с соответствует выработке электроэнергии 5–7 кВт.

Machine Translated by Google

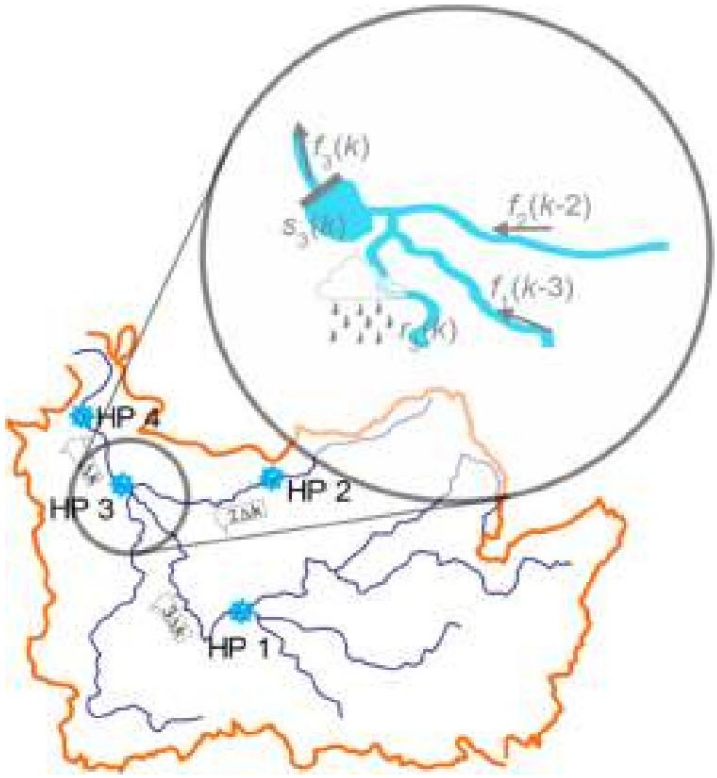


Рисунок 3. Модель подключенных гидроэлектростанций в бассейне реки. Завод 3 (увеличено) поставляется компанией притоки из водохранилищ 1 и 2 с задержками, указанными толстыми стрелками (k — период отбора проб) и экзогенные источники, например, дождь или приток на юге. Количество воды, подаваемой на завод от этих источников в пределах k в момент k есть f 1 (k 3), f 2 (k 2) и r3 (k) соответственно. Отток от плотинных генераторов станции 3 – f 3 (к). Разница входного и выходного потоков буферизуется в пруд, текущий объем воды которого обозначен s3 (k).

4.2. Многозаводское депо

Как и в случае с фотоэлектрическими системами, по мере роста количества HP их влияние на сеть увеличивается. Совокупность пико- и микро-ДП может составлять большое распределенное хранилище энергии.

В случае n электростанций в одном речном бассейне переменные модели группируются на векторы:

$$s(k) = \begin{bmatrix} s1(k) \\ s2(k) \\ \vdots \\ sn(k) \end{bmatrix}, f(k) = \begin{bmatrix} f1(k) \\ f2(k) \\ \vdots \\ fn(k) \end{bmatrix}, r(k) = \begin{bmatrix} r1(k) \\ r2(k) \\ \vdots \\ rn(k) \end{bmatrix}, \quad (3)$$

где s(k) – вектор уровня воды в водохранилище, f(k) – вектор объемов воды в межпластовых потоках, r(k) – вектор объемов воды из экзогенных источников соответственно.

Динамика сетевой системы, полученная из (1), описывается выражением

$$s(k+1) = s(k) + \sum_{\tau=0}^{T} \Theta\, tf(k-t) + r(k), \quad (4)$$

где T – максимальная задержка, а матрицы Θ t группируют информацию о географическая топология и задержки потока. Θ t = θij $n\times n$, где θij = 1, если поток из пласта j достигает резервуара i с задержкой t, в противном случае он равен 0. Θ 0 описывает отток из система. В отличие от [13], здесь расстояние между растениями и немаловажно. Для пример из рисунка 3, самая длинная задержка T = 3 (поток между резервуарами 1 и 3) и

Machine Translated by Google

$$\Theta 0 = \begin{matrix} 1\,0 & 0\,0\,0 \\ 0 & 1\,0 \\ 0\,0 & 1\,0 \\ 0\,0 & 0 \quad 1 \end{matrix}, \Theta 1 = \begin{matrix} 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,1\,0 \end{matrix}, \Theta 2 = \begin{matrix} 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,0\,0 \end{matrix}, \Theta 3 = \begin{matrix} 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 1\,0\,0\,0 \\ 0\,0\,0\,0 \end{matrix}. \quad (5)$$

Для многозаводской системы индекс качества (2) принимает вид

$$\text{Дж} = \sum_{k=1}^{\text{н}} \text{Дждж} = \sum_{k=1}^{\text{н}} \sum_{j=1}^{\text{м}} \text{Jj}(\text{к}) = \sum_{k=1}^{\text{н}} \sum_{j=1}^{\text{м}} \eta j\, pj(k) fj(k) \quad k. \quad (6)$$

Заинтересованные стороны, как правило, имеют противоположные цели. Потребители пытаются их максимизировать, тогда как операторы сетей хотят максимизировать J. Удобным способом согласования Jj интересов является система цен. Алгоритм, представленный во второй части этой статьи, удовлетворяет целям как владельцев электростанций, так и владельцев сетей.

4.3. Алгоритм

управления Основная трудность решения задачи (6) связана со сложностью динамики системы, обусловленной наличием немалой задержки в (4). Дополнительная трудность связана с физическими ограничениями, налагаемыми на переменные f(k) и s(k)

$$\begin{matrix} \text{мин} \\ f\,j \\ j,k \end{matrix} \quad fj(k) \quad f\,\text{макс.}, \quad \begin{matrix} \text{min} \quad s \\ j,k \\ \text{дж} \end{matrix} \quad sj(k) \quad s\,j \quad \begin{matrix} \text{Макс.}, \\ \text{дж} \end{matrix} \quad sj(m) = Sj\,, \quad (7)$$

где Sj — произвольно выбранное терминальное значение.

Как указано выше, стандартные решатели не являются устойчивыми в рассматриваемом здесь классе задач [15,17,21–29]. Поэтому для решения вычислительных аспектов применяется трехэтапная процедура оптимизации min-max, схематически представленная на рисунке 4 . Ниже описан каждый шаг алгоритма . Связь между HP соответствует известному отраслевому стандарту Message Queuing Telemetry Transport (MQTT) [34], который обеспечивает как безопасность , так и синхронизацию между контроллерами предприятия.

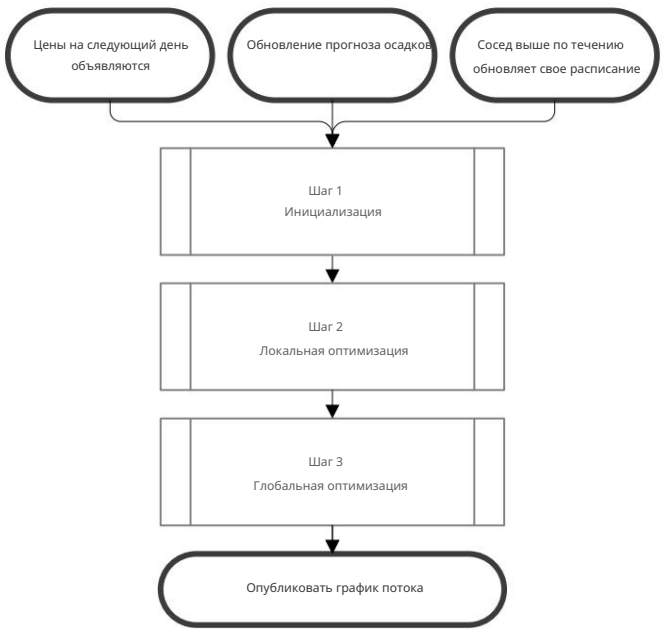


Рисунок 4. Блок-схема алгоритма.

Machine Translated by Google

Шаг 1, инициализация:

Входные данные: ценовой профиль.

Результат: предварительный график расхода воды.

Цель: Определить интервалы накопления и заработка в пределах рассматриваемого периода времени, как показано на рисунке 5. Когда цены на энергоносители выше среднего, то наступает время заработка (интервал заработка). В противном случае воду следует хранить (интервал накопления). В первом случае f(k) = f max, а во втором случае f(k) = f для каждого k в соответствующем интервале. В результате для растения j получаются периоды hj длиной mjp . Поскольку HP управляются независимо из-за разных тарифов или требований, схема сегментации может отличаться для каждого HP. В любом случае, этот шаг не требует сложных вычислений. Аппроксимация, полученная на шаге 1, может привести к выходу за границы s(k), определенные в (7).

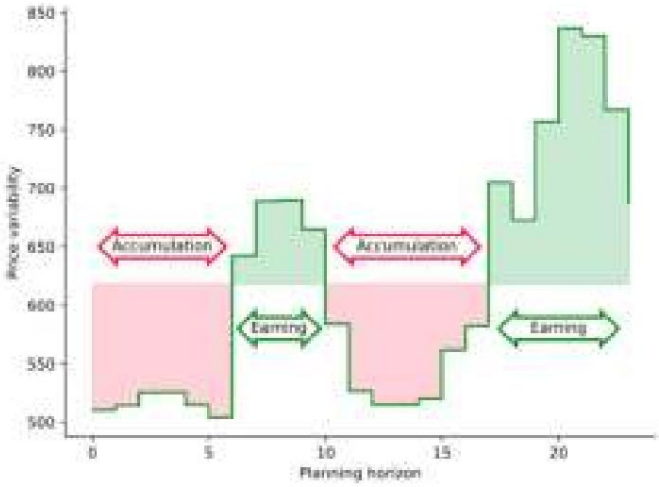


Рисунок 5. Иллюстрация шага 1: определение интервалов накопления и прибыли для ценового профиля от 6 июля. Отмеченные периоды влияют на версию шага 2 представленного алгоритма.

Шаг 2, локальная оптимизация:

Входные данные: профиль цен и предварительный график стока воды, рассчитанный на этапе 1, а также приток (т. е. прогнозируемые дожди, приток притоков, график работы станций выше по течению) в систему .

Результат: локально оптимизированный график потока. Результат не выходит за пределы, определенные в (7).

Цель: поиск оптимального потока в каждом периоде, найденном при выполнении шага 1.

Поскольку существуют два вида периодов, разумно разделить Шаг 2 соответственно.

Шаг 2a (интервал накопления): Если для любого k sj (k) > sj max, это означает, что поток воды слишком сильно ограничен. Таким образом, приток воды затопляет пруд. Поэтому желательно найти момент k* с наибольшей ценой в интервале nj и увеличить поток на sj (k*) sj max, но не выше fj max. Этот шаг повторяется до тех пор, пока условие sj (k) sj не будет выполнено для всех k в исследуемом интервале.

Шаг 2b (интервал дохода): Если для любого k sj (k) < sj min, это означает, что поток избыточен. Таким образом, необходимо найти момент k* с наименьшей ценой в этом интервале и ограничить поток min величиной sj , удовлетворяемой для всех k в исследуемом интервале. sj (k*), но не ниже fj min. Этот шаг повторяется до тех пор, пока не будет выполнено условие sj (k) sj

Шаг 2 выполняется для каждого HP, начиная с вышестоящих. Максимальное количество итераций равно для каждого HP j, что достаточно мало для выполнения в реальном времени даже для jp. масштабная система. Шаг 2 показан на рисунке 6. После вычислений максимально возможный выигрыш получается в пределах каждого интервала.

Machine Translated by Google

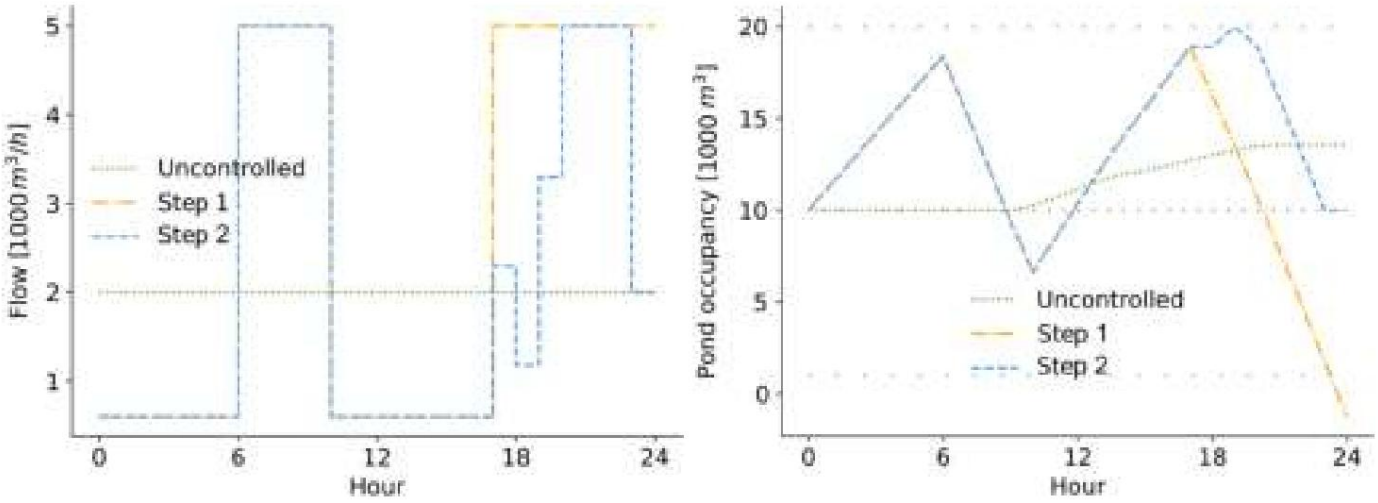


Рисунок 6. Интенсивность потока, полученная на этапе 2 для схемы, показанной на рисунке 5. Никаких действий не требуется для первых трех периодов. На последнем периоде терминальное условие s(m) нарушается, поэтому поток необходимо ограничить.

Шаг 3, глобальная оптимизация:

Входные данные: профиль цен и оптимизированный для местных условий график расхода воды, рассчитанный на этапе 2.

Результат: глобально оптимизированный график потоков.

Задача: Перепланирование расхода воды на всем горизонте оптимизации.

Шаг 2 обеспечивает оптимальный поток в каждом интервале заработка/накопления и гарантирует выполнение условия (7) для каждого интервала nj . Возможно, есть возможности для глобального повышения производительности системы. Наблюдая за событиями, представленными на рисунке 6, можно заметить, что в часы 0–19 заселенность пруда растет немонотонно, пока водоем не наполнится. Интервалы наполнения и слива определяются моментами насыщения пруда. Перепланирование внутри этих интервалов не влияет на конечный объем резервуаров, поэтому можно увеличить расход, когда энергия дорогая, и уменьшить его на ту же величину, когда цена энергии ниже.

Максимальное количество итераций на шаге 3 аналогично количеству итераций на шаге 2. Результат пример шага 3 показан на рисунке 7.

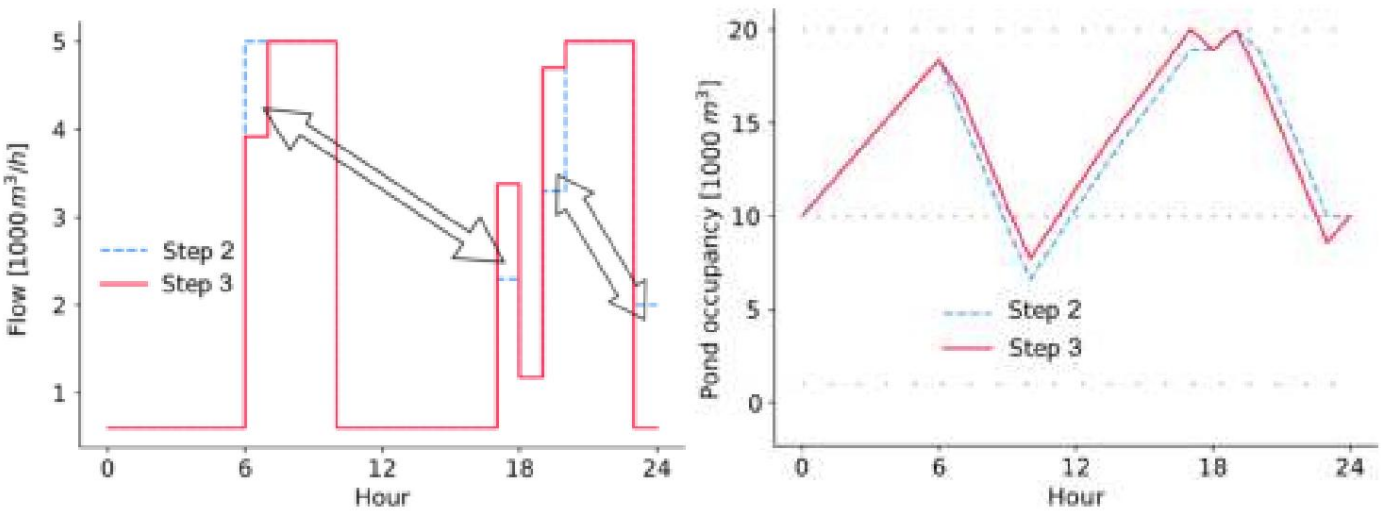


Рисунок 7. Результат шага 3 в сравнении с результатом шага 2 (рисунок 6). Стрелками отмечен перепланированный объем воды.

Machine Translated by Google

Заключительные соображения

Алгоритм выполняется всякий раз, когда входные значения изменяются, например, когда объявляются новые цены на энергию или получают обновленный прогноз погоды rj (k) . Результаты должны быть переданы всем заинтересованным владельцам в сети НР, поскольку для шага 2 алгоритма необходимы выходные данные вышестоящих предприятий. Алгоритм можно запустить локально с использованием недорогого оборудования, например микроконтроллера.

5. Оценка

Для оценки алгоритма использовалась четырехузловая топология на рисунке 3 . Достаточно просто проиллюстрировать свойства алгоритма, охватывая при этом основные проблемы потребительской сети НР, возникающие из-за древовидной топологии и немалой задержки между станциями. Никакого общего управления не предполагается — потребители принимают управляющие решения на своих предприятиях независимо друг от друга. Необходима только трансляция управляющих решений. Официальные правила могут легко обеспечить его соблюдение.

Алгоритм из раздела 4 был протестирован с использованием различных профилей цен, взятых из [5], в том числе показанных на рисунке 2, при наличии движущейся волны интенсивного дождя и соответствующего стока, изображенного на рисунке 8. Расчеты выполнялись для каждого профиля. и разные объемы прудов. Во всех случаях в систему поступает экзогенный поток постоянной интенсивности f(k) = [2, 3, 5, 5] × 103 [м3/ч]. Другие параметры системы: = [0,5, 0,6, 1, 1] × 103 [м3/ч], s(0) = s(м), принимается как f макс = [5, 6, 10, 10] × 103 [м3/ч], f мин мин с = 0,1 × s(0), s = 2 × s(0), при этом начальный уровень воды s(0) задается следующим Макс образом: Сценарий 1 (малые водоемы): s(0) = s(m) = [0,5, 0,75, 1,25, 1,25] × 104 [м3];

Сценарий 2 (коллекторы среднего размера): s(0) = s(m) = [1, 1,5, 2,5, 2,5] × 104 [м3];

Сценарий 3 (большие водоемы): ГП соединены с более крупными искусственными озерами с незначительной вероятностью насыщения, s(0) = s(m) = [2, 3, 5, 5] × 104 [м3].

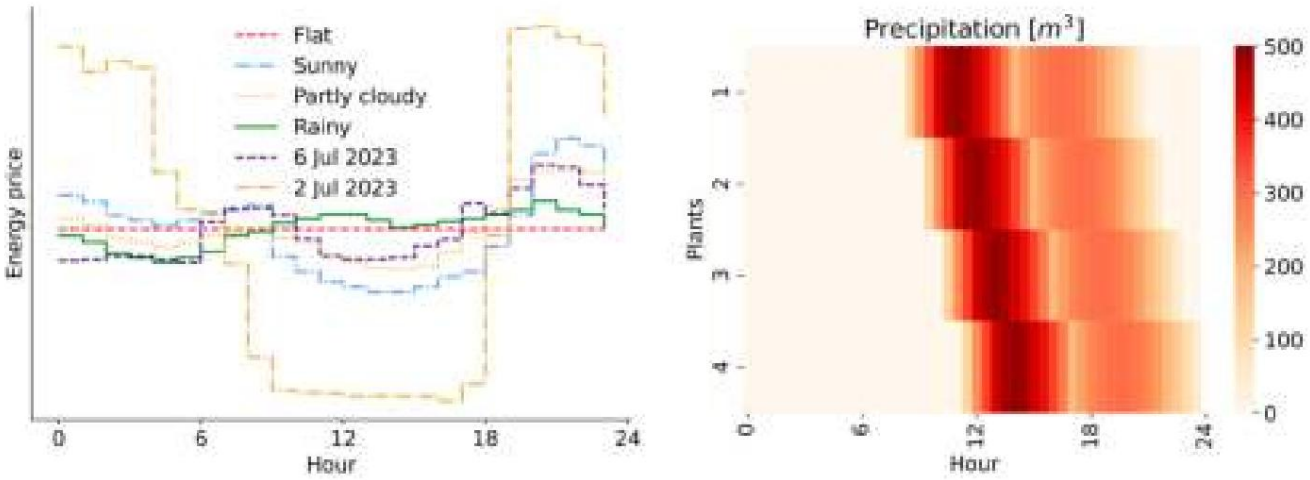


Рисунок 8. Примеры ценовых профилей, масштабированные с учетом того же среднего значения и предполагаемого режима осадков, со стоком, используемым для проверки результатов. тт ффф

Хотя алгоритм не предполагает какой-либо конкретной связи между ценовыми профилями, для удобства чтения все НР используют одни и те же данные. По этой причине мощность турбины также принимается как линейная функция интенсивности потока ηj (fj (k)) = const. ффф

В Таблице 1 суммирована выгода полупотребителя и общая выгода оператора сети. Выигрыш установлен относительно случая, когда ХП не контролируются. Несоответствие в конечном состоянии в условиях свободного потока пополняется или отпускается по средней цене. ффф

Machine Translated by Google

Таблица 1. Прирост прибыли [%] по отношению к свободному потоку [м3] для ценовых профилей на рис. 7 и других размеры резервуара.

Профиль	Завод 1	Завод 2	Завод 3	Завод 4	Все Заводы	Пруды	Примечание
небольшого объема (Сценарий 1)							
Солнечно	10,85	11,80	10,98	6,74	9,71		
Переменная облачность	4,52	4,80	6,95	1,69	4,44		
Дождливый	2.54	2.41	6,62	198,52	34,55		Если завод 4 оставить без контроля, то производство падает ниже минимума в течение 5 часов.
6 июля 2023 г.	10.13	10,35	13,39	28,16	17.30		
2 июля 2023 г.	29.25	33,81	59,53	46,64	47.30		
Пруды среднего объема (Сценарий 2)							
Солнечно	20.13	19.38	15,91	11,44	15,65		
Переменная облачность	11.98	12.07	13,86	38.17	21.29		Если завод 4 оставить без контроля, то производство падает ниже минимума в течение 3 часов.
Дождливый	9.31	11.37	17.44	65,61	33.05		Если заводы 3 и 4 оставить без контроля, то производство падает ниже минимума в течение 5 часов (оба), а на заводе 3 в течение 2 часов происходит наводнение.
6 июля 2023 г.	14.47	13,77	14.25	49,92	25.15		Если завод 4 оставить без контроля, то производство падает ниже минимума в течение 4 часов.
2 июля 2023 г.	68,50	68,34	69,79	82,83	74.03		Если завод 4 оставить без контроля, то производство падает ниже минимума в течение 2 часов.
Пруды большого объема (Сценарий 3)							
Солнечно	21.04	20.82	19,52	16,15	18,86		
Переменная облачность	14,73	14,72	14.43	19.43	16,76		Если завод 4 оставить без контроля, то производство падает ниже минимума в течение 1 часа.
Дождливый	6,93	6,79	6,61	5,78	6.41		
6 июля 2023 г.	15,39	15,18	13,83	12,36	13.82		
2 июля 2023 г.	59,52	56,83	80,13	65,47	67,83		

Данные, приведенные в таблице 1 , показывают, что чем больше объем пруда, тем больше прибыль. полученный. Однако при применении предложенного алгоритма даже для небольших водоемов относительная денежная выгода достигает нескольких десятков и растет с изменением цен, хотя и не линейно. Данные также показывают, что небольшие изменения в rj (k) или pj (k) могут привести к существенному расхождение в результате. Прирост растет с увеличением расстояния между ХП, что объясняется большей разницей между pj (k) и pj (k Tij). Стоит отметить, что алгоритм имеет тенденцию заполнять пруды по низким ценам около полуночи, но не позволяет для наводнений. Такое поведение является выгодным, поскольку пруды способствуют увлажнению почвы, таким образом смягчение проблемы устойчивых засух в различных регионах.

Алгоритм выбирает максимальную или минимальную интенсивность потока f(k), а не промежуточные. Это свойство облегчает строительство и обслуживание турбин повышение эффективности в номинальных условиях.

Результаты, полученные с помощью предложенного алгоритма, обобщенные в таблице 1, значительно лучше, чем полученные стандартными алгоритмами оптимизации [29]. Более того, все ТП ниже по течению защищены от обычных наводнений и засух (Сценарий 2).

Балансирующий рынок может принести дополнительную прибыль как просьюмерам, так и энергосистеме. операторы [16]. Когда сетевой оператор, к которому подключена просьюмерская сеть, обнаруживает недостаток энергии может спровоцировать немедленный выброс воды или, наоборот, сдержать поток в случае избытка.

6. Аспекты реализации

Перед развертыванием следует учитывать еще несколько аспектов:

1. Юридические вопросы. Для развертывания системы необходимо согласие регулятора водных ресурсов и соответствующие соглашения сетевых операторов.
2. Обмен информацией. Разработанный алгоритм требует надежной связи. платформа среди заводов-потребителей. Для этой цели мы предлагаем MQTT [35] .

Machine Translated by Google

3. Прогноз погоды. Прогнозы на следующий день в настоящее время недостаточно точны, чтобы предсказать локальные штормы на несколько часов вперед. Однако радиолокационные изображения позволяют получить подробные оценки осадков за несколько часов.
4. Депо как система реального времени. Каждый контроллер HP работает независимо от других. Следовательно, система потенциально подвержена неблагоприятным явлениям со стороны распределенных систем, таким как спешка или тупики. Протокол MQTT решает эту проблему, не увеличивая вычислительную нагрузку, поскольку расчет расписания занимает всего доли секунды, а обновления прогнозов происходят нечасто.

5. Депо как система управления. Система работает по открытому циклу. Обновления расписания влияют только на нижестоящие HP. Система стабильна с ограниченным входом и ограниченным выходом (BIBO).

6. Безопасность. Предлагаемая коммуникационная платформа соответствует высоким стандартам промышленной безопасности. стандарты [35].
7. Вычислительные ресурсы. Предложенный алгоритм не требует для своей работы высокопроизводительных вычислительных ресурсов. Он эффективно работает и обменивается данными на недорогих устройствах начального уровня .

7. Выводы

В данной статье представлена оптимальная стратегия управления системой связанных TH, образующей устойчивое распределенное хранилище энергии. Такая система приносит пользу как операторам электросетей, поскольку снижает колебания сетевой нагрузки, так и владельцам электростанций, увеличивая их доходы. Его можно внедрить с низкими капитальными и эксплуатационными затратами. Кроме того, замедляя сток осадков, система повышает устойчивость региона к наводнениям и засухам, что имеет первостепенное значение в условиях изменения климата. Алгоритм управления не предполагает одинаковую структуру цен или горизонт планирования, используемые разными заинтересованными сторонами. Владельцев HP не принуждают к сотрудничеству — им нужно лишь сообщить соседям о своих решениях по управлению. В отличие от неблагоприятной фотоэлектрической «инверторной борьбы» [35], сотрудничество между потребителями в предлагаемой схеме увеличивает их прибыль. Также, в отличие от прокачиваемых ХД, накопление энергии не влечет за собой дополнительных затрат, поэтому даже небольшие колебания цен позволяют увелич

Представленный алгоритм работает лучше, чем классические процедуры оптимизации с точки зрения численной стабильности и надежности [29]. Его можно реализовать с использованием незначительных вычислительных ресурсов и запускать асинхронно на HP. Хотя окончательный доход зависит от погоды, точности прогнозов, изменчивости цен на энергоносители и других внешних условий, рост по сравнению с неконтролируемым случаем находится в диапазоне 20–70% и должен расти под влиянием ВИЭ, особенно фотоэлектрических. и ветровые ВИЭ.

Вклад автора: концептуализация, ММ и PI; методология, ММ; программное обеспечение, ММ; валидация, ММ и PI; формальный анализ, ММ и ПИ; расследование, ММ; ресурсы, ММ и ПИ; курирование данных, ММ и PI; письмо – подготовка первоначального черновика, ММ; написание — просмотр и редактирование, ИП; визуализация, ММ и ПИ; надзор, ИП; администрация проекта, ИП; приобретение финансирования, PI Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Работа выполнена в рамках проекта №. 2021/41/B/ST7/00108: «Надежные решения по управлению многоканальными сетевыми потоками», финансируется Национальным научным центром , Польша.

Заявление о доступности данных: Все данные доступны публично [5].

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов. Спонсоры не играли никакой роли в разработке исследования; при сборе, анализе или интерпретации данных; в написании рукописи; или в решении опубликовать результаты.

Рекомендации

1. Колосок С.; Билан; Васильева Т.; Войцеховский А.; Моравски, М. Обзорный обзор возобновляемых источников энергии, устойчивости и среда. Энергия 2021, 14, 4490. [CrossRef]

Machine Translated by Google

2. Рахман, М.М.; Они, АО; Гемечу, Э.; Кумар, А. Оценка технологий хранения энергии: обзор. Конверторы энергии. Менеджер. 2020, 223, 113295. [CrossRef]
3. Вирфс-Брок, Дж. И.Э. Вопросы: Почему Калифорния пытается обезглавить утку? 2014. Доступно онлайн: https://insideenergy.org/20.10.14.02/ie-questions-почему-калифорния-пытается-обезглавить-утку/ (по состоянию на 12 мая 2023 г.).
4. Борковски Д.; Чолева, Д.; Корзен, А. Гибридная система гидро-фотоэлектрических аккумуляторов в русле реки как поставщик энергии для местных нагрузок. Энергия 2021, 14, 5160. [CrossRef]
5. ПСЭ. Рыночные цены на энергию. Доступно онлайн: https://www.pse.pl/dane-systemowe/funkcjonowanie-rb/raporty-dobowe-z-funkcjonowania-rb/podstawowe-wskazniki-cenowe-i-kosztowe/rynkowa-cena-energii-elektrycznej-rce (по состоянию на 1 августа 2023 г.).
6. Аназа, С.О.; Абдулазиз, MS; Йиса, Я; Юсуф, Йо; Салаву, Будапешт; Момох, SU. Производство микрогидроэлектроэнергии — Обзор. Являюсь. Дж. Инж. Рез. 2017, 6, 5–12.
7. Сачдев, Х.; Акелла, АК; Кумар, Н. Анализ и оценка малых гидроэлектростанций: библиографический обзор. Обновить. Поддерживать. Energy Rev. 2015, 51, 1013–1022. [Перекрестная ссылка]
8. Ивэй, К.; Джин, Л.; Лю, Ф.; Дай, Н.; Сонг, Ю.; Чен, Г.; Дин, Л. Третичное регулирование каскадной русловой гидроэнергетики в островной системе возобновляемой энергетики с учетом многовременной динамики. Обновление ИЭПП. Генератор энергии. 2021, 15, 1778–1795.
9. Джамали, С.; Джамали, Б. Оптимальная эксплуатация каскадных гидроэнергетических систем: последствия для гидроэнергетических систем Ирана Большой Карун. Прил. Наука о воде. 2019, 9, 66. [CrossRef]
10. Вышковский, К.; Пивоварек, З.; Палейко, З. Małe Elektrowne Wodne w Polsce. Технический отчет, UN Global. 2022. Доступно онлайн: https://ungc.org.pl/wp-content/uploads/2022/03/Raport_Male_elektrownie_wodne_w_Polsce.pdf . (по состоянию на 1 июня 2024 г.). (На польском).
11. Ютуб. Сужение общего канала. 2024. Доступно онлайн: https://www.youtube.com/watch?v=MBDV_Xj6Dj8. (доступ 1 июня 2024 г.).
12. МЭА Гидроэнергетика. Модели управления каскадными водохранилищами гидроэнергетики, подборка примеров. Технический отчет. 2021 г. Доступно онлайн: https :// www.iea Hydro.org/media/02e6ec8e/EAHydro_AnnexXIV_Management%20Models%20for%20Hydropower%. 20Каскад%20Резервуары_CASE%20КОМПИЛАЦИЯ.pdf (по состоянию на 1 июня 2024 г.).
13. Ю, Ю.; Ву, Ю.; Шэн, К. Оптимальная стратегия планирования каскадных ГЭС в рамках общего рынка энергии на сутки вперед и регулирование частоты. IEEE Access 2021, 9, 87749–87762. [Перекрестная ссылка]
14. Чжан Х.; Чанг, Дж.; Гао, К.; Ву, Х.; Ван, Ю.; Лей, К.; Лонг, Р.; Чжан Л. Рассмотрение вопроса об эксплуатации каскадных гидроэлектростанций комплексные экологические потребности в воде. Конверторы энергии. Менеджер. 2019, 180, 119–133. [Перекрестная ссылка]
15. Таир Хаммид, А.; Авад, ОИ; Сулейман, Миннесота; Гунасекаран, СС; Мостафа, ЮАР; Манодж Кумар, Н.; Халаф, бакалавр; Аль-Джаухар, Ю.А.; Абдулхасан, Р.А. Обзор алгоритмов оптимизации при решении задач планирования гидрогенерации. Energy 2020, 13, 2787. [CrossRef]
16. Мельчарски, В. Энергетические системы и рынки; Ассоциация польских инженеров-электриков: Варшава, Польша, 2018.
17. Сингх, В.К.; Сингал, С. Эксплуатация гидроэлектростанций – обзор. Обновить. Поддерживать. Energy Rev. 2017, 69, 610–619. [Перекрестная ссылка]
18. Ли, Дж.; Чжао, З.; Ли, П.; Махмуд, Массачусетс; Лю, Ю.; Чен, Д.; Хан, В. Комплексная оценка выгод от интеграции морских проектов насосные гидроаккумулирующие и плавучие фотоэлектрические станции. Конверторы энергии. Менеджер. 2023, 296, 117651. [CrossRef]
19. Виджесингхе А.; Лай, Л. Анализ и развитие малых гидроэлектростанций. В материалах 4-й Международной конференции по дерегулированию и реструктуризации электроэнергетики и энергетическим технологиям (DRPT) 2011 г., Вэйхай, Китай, 6–9 июля 2011 г.; стр. 25–30.
20. Хомса, А.; Лаксоно, А.С. Пико-гидро как возобновляемая энергия: местные природные ресурсы и наличие оборудования для производства электроэнергии. Конференция ИОП. Сер. Матер. наук. англ. 2018, 462, 012047. [CrossRef]
21. Андерсон, Д.; Моггридж, Х.; Уоррен, П.; Шаксмит, Дж. Воздействие русловой гидроэнергетики на физические и экологические процессы. состояние рек. Водная среда. Дж. 2015, 29, 268–276. [Перекрестная ссылка]
22. Нгуен, А.; Ву, В.; Хоанг, Д.; Нгуен, Т.; Нгуен, К.; Нгуен, П.; Джи, Ю. Модель ансамбля внимания для точного прогнозирования расхода и уровня воды с улучшением обучающих данных. англ. Прил. Артиф. Интел. 2023, 126, 107073. [CrossRef]
23. Альварес Г. Модель оптимизации работы крупных гидроэлектростанций. IEEE лат. Являюсь. Пер. 2020, 18, 1631–1638. [Перекрестная ссылка]
24. Шоу, А.Р.; Сойер, Х.; ЛеБёф, Э.; Макдональд, М.; Хаджериуа, Б. Оптимизация гидроэнергетики с использованием суррогатных моделей искусственной нейронной сети высокоточной модели гидродинамики и качества воды. Водный ресурс. Рез. 2017, 53, 9444–9461. [Перекрестная ссылка]
25. Ахмадианфар, И.; Самади-Кучексараи, А.; Асадзаде, М. Извлечение нелинейных правил работы систем с несколькими резервуарами с использованием эффективный метод оптимизации. наук. Rep. 2022, 12, 18880. [CrossRef]
26. Бернардес Ж.-младший; Сантос, М.; Абреу, Т.; Прадо Л.-младший; Миранда, Д.; Хулио, Р.; Виана, П.; Фонсека, М.; Бортони, Э.; Бастос, Г.С. Оптимизация работы гидроэнергетики с использованием машинного обучения: систематический обзор. АИ 2022, 3, 78–99. [Перекрестная ссылка]
27. Марселино, КГ; Лейте, GMC; Дельгадо, CADM; де Оливейра, LB; Ваннер, EF; Хименес-Фернандес, С.; Сальседо-Санц, С. Эффективный многоцелевой эволюционный подход к решению проблемы планирования работы системы с несколькими резервуарами на гидроэлектростанциях . Экспертная система. Прил. 2021, 185, 115638. [CrossRef]
28. Линке, Х. Модельно-прогнозирующий контроллер для оптимального использования гидроэнергии речных водохранилищ. В материалах Международной конференции IEEE по приложениям управления 2010 г., Иокогама, Япония, 8–10 сентября 2010 г.; стр. 1868–1873.
29. Моравски М.; Игначук, П. Балансировка энергетического баланса в установках бытовой воды с явным учетом задержки потока. IFAC-PapersOnLine 2023, 56, 11748–11753. [Перекрестная ссылка]

Machine Translated by Google

30. Игначук П.; Моравски, М. Оптимальное управление каскадными гидроэлектростанциями как распределенным энергетическим хранилищем, ориентированным на потребителя. Энергии 2024, 17, 469. [CrossRef]

31. Дэниэлсон, К.; Борреллия, Ф.; Оливер, Д.; Андерсон, Д.; Филлипс, Т. Управление потоками с ограничениями в сетях хранения данных: емкость максимизация и балансировка. Автоматика 2013, 49, 2612–2621. [Перекрестная ссылка]

32. Басин, М.В.; Герра-Авельянеда, Ф.; Штессель, Ю.Б. Задача управления запасами: адаптивная конвергентная непрерывная с фиксированным временем конструкция контроллера. IEEE Транс. Сист. Человек Киберн. Сист. 2020, 50, 4974–4983. [Перекрестная ссылка]

33. Игначук, П. Компенсация простоя в системах непрерывного анализа запасов скоропортящихся продуктов с несколькими альтернативами поставок. Дж. Управление процессами 2012, 22, 915–924. [Перекрестная ссылка]

34. Сенгул, К. Профиль телеметрии очереди сообщений (MQTT) и безопасности транспортного уровня (TLS) для системы аутентификации и авторизации для ограниченных сред (ACE); RFC 9431. 2023. Доступно онлайн: https://datatracker.ietf.org/. документ/rfc9431/ (по состоянию на 1 июня 2024 г.).

35. Гу, Ю.; Грин, Т.Ц. Стабильность энергосистемы при высоком проникновении инверторного ресурса. Учеб. IEEE 2023, 111, 832–853. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Энергоэффективное ускорение нейронных сетей с использованием большинства Значительный приближенный множитель с битовым управлением

Пэнфэй Хуан, Бинь Гун, Кэ Чен * и Чэнхуа Ван

Колледж электронной и информационной инженерии, Нанкинский университет аэронавтики и астронавтики, проспект Цзянцзюнь, № 29, район Цзяннин, Нанкин 210016, Китай; pfhuang@nuaa.edu.cn (PH); bingong@nuaa.edu.cn (БГ); chwang@nuaa.edu.cn (CW)
* Переписка: chen.ke@nuaa.edu.cn

Аннотация: Растущие вычислительные потребности глубокого обучения и крупномасштабных моделей привели к значительному увеличению энергопотребления, что подчеркивает острую необходимость в более энергоэффективных аппаратных разработках. В этом исследовании представлена новая стратегия аппроксимации весов, специально разработанная для квантованных нейронных сетей (NN), в результате которой был разработан эффективный приближенный множитель, использующий сдвиг наиболее значимого единицы (MSO). По сравнению как с энергоэффективными логарифмическими приближенными умножителями, так и с нелогарифмическими приблизительными умножителями с приоритетом точности, предлагаемая нами логарифмическая конструкция обеспечивает беспрецедентный баланс между точностью и затратами на оборудование. По сравнению с базовым точным множителем наша инновационная конструкция демонстрирует значительные сокращения, включая уменьшение площади до 28,31%, заметное снижение энергопотребления на 57,84% и уменьшение задержки на 11,86%. Результаты экспериментов показывают, что предложенный множитель при применении в нейронных сетях может сэкономить примерно 60% энергии без ущерба для точности задачи. Одновременно эксперименты, посвященные трансформаторному ускорителю и обработке изображений, демонстрируют существенную экономию энергии, которую можно получить для моделей большого языка (LLM) и задач обработки изображений за счет реализации предложенной нами конструкции, что дополнительно подтверждает ее эффективность и практичность.

Ключевые слова: приближенные вычисления; большая языковая модель; коэффициент-множитель; логарифмический множитель

check for updates

Цитирование: Хуанг, П.; Гонг, Б.; Чен, К.; Ван, К. Энергоэффективное ускорение нейронной сети с использованием наиболее значимого приближенного множителя с битовым управлением. Электроника 2024, 13, 3034. https://doi.org/10.3390/электроника13153034

Академические редакторы: Кэ Се, Шоуэ Чен и Сяолун Ван

Поступила: 5 июля 2024 г.
Пересмотрено: 30 июля 2024 г.
Принято: 31 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

В области искусственного интеллекта произошел значительный прогресс, особенно с появлением глубоких нейронных сетей (DNN) и моделей большого языка (LLM). Однако эти сложные модели создают значительные проблемы с вычислениями и памятью, обусловленные их постоянно растущими размерами, как это тщательно проанализировано в [1]. Эти проблемы еще более актуальны для периферийных устройств, которые ограничены внутренними ограничениями, связанными с их вычислительной мощностью. Чтобы обеспечить практическое развертывание нейронных сетей (НС) на периферийных платформах, в качестве возможных решений были представлены различные программные методы сжатия моделей , такие как квантование и сокращение. Кроме того, недавние усилия по использованию парадигмы приближенных вычислений (AxC) на уровне схемы убедительно продемонстрировали ее способность улучшать производительность системы [2]. Это улучшение является результатом снижения вычислительной сложности и требований к памяти, достигаемых за счет замены точных арифметических элементов их менее точными эквивалентами. Ключевые операции арифметических процессоров, а именно умножение и сложение, существенно влияют на их общее энергопотребление. Примечательно, что умножение потребляет больше энергии, чем сложение. Основными арифметическими вычислениями, выполняемыми в нейронных сетях, является операция умножения-накопления (MAC). Поскольку миллионы операций умножения выполняются ускорителем NN, приближенные множители (AM) привлекли широкое внимание исследователей из-за их значительной значимости с точки зрения повышения эффективности и точности приложений нейронных сетей [3–5].

Machine Translated by Google

Предыдущие исследования приближенных множителей, адаптированных для DNN, дали обнадеживающие результаты в отношении эффективности [6]. В случае развертывания нейронных сетей на периферийных устройствах с ограниченными ресурсами использование приближенных множителей обеспечивает привлекательный способ найти баланс между точностью и затратами энергии. Это достижение основано на смягчении и ослаблении строгих требований к совершенно точным результатам умножения, что облегчает создание приблизительных аппаратных конструкций, которые более ресурсо- и энергоэффективны. Исследование в [7] продемонстрировало эффективность AxC в DNN, представив новый уровень AxC DNN, который включает приблизительные множители в процесс проектирования DNN. В [8] были предложены три гибридных приближенных умножителя с использованием вентилей И-НЕ вместо вентилей И и применения различных комбинаций компрессоров для обеспечения разнообразных компромиссов между точностью и эффективностью оборудования. В [9] была предложена адаптируемая конструкция приближенного множителя с открытым исходным кодом, основанная на входном распределении и полярности, для создания оптимизирован Целью является минимизация среднего квадрата абсолютной ошибки приближенного множителя на основе вероятностных распределений операндов, извлеченных из целевого приложения, с учетом входной полярности. В [10] была представлена методология построения приближенных множителей массива, использующая концепцию игнорирования переноса. Благодаря стратегическому выбору позиций для игнорирования переноски становится возможным игнорировать небольшое количество переносок, тем самым достигая повышенной скорости и уменьшения площади, одновременно сохраняя превосходную общую точность и гарантируя пригодность для предполагаемого применения.

Приближенные логарифмические множители обеспечивают более простую конструкцию, но приводят к заметно более высоким потерям точности, тогда как приближенные нелогарифмические множители демонстрируют меньшую вычислительную ошибку за счет увеличения сложности конструкции [11]. Чтобы удовлетворить потребность в снижении энергопотребления при сохранении приемлемой потери точности, в этом исследовании предлагается энергоэффективный наиболее значимый множитель переключения с одним приводом (MSAM может получить ряд высокоэффективных схем умножителей путем динамической настройки коэффициента аппроксимации (k) и коэффициента точности (m). Используя коэффициент аппроксимации (k), множитель стратегически делится на точные и приблизительные субмножители. В области аппроксимации применяется принцип переконфигурируемого множителя с постоянными коэффициентами (RCCM) [12] , заменяющий традиционные операции умножения операциями сдвига, сложения и вычитания. Эта журнальная статья основывается на нашем предыдущем докладе на конференции [13] и расширяет его, предлагая несколько ключевых отличий и новых идей. Краткое изложение этих достижений представлено ниже: • Мы разработали альтернативную и более агрессивную

стратегию аппроксимации для процесса аппроксимации множителей, расширяя границы компромисса между точностью и эффективностью. • Далее мы провели всесторонний анализ ошибок и тщательно обсудили

компромисс между точностью и энергопотреблением.

• Помимо анализа популярных наборов данных, мы также исследуем обработку изображений с использованием различных

схем AxC. • Преимущества предлагаемых нами приближенных конструкций умножителей обоснованы строгим теоретическим анализом и обширными экспериментальными демонстрациями.

Оставшаяся часть статьи структурирована следующим образом: В разделе 2 представлен краткий обзор предыстории и сопутствующей работы. В разделе 3 подробно описана наша инновационная методология энергоэффективного ускорения нейронных сетей, включающая наиболее важный динамический приблизительный множитель с битовым управлением для значительного повышения производительности. В разделе 4 представлены анализ ошибок и результаты моделирования. Наконец, раздел 5 зав

2. Предварительные сведения

Логарифмический множитель (LM) преобразует числа из двоичного формата в логарифмический формат, облегчая переход от умножения к сложению. Алгоритм Митчелла первоначально был предложен в [14]. Даны два двоичных числа A и B, где A = 2 k1 (1 + x1)

Machine Translated by Google

и B = 2 k2 (1 + x2), логарифмический множитель первоначально преобразует их в логарифмические формируется следующим образом:

$$\log2 (A \cdot B) = \log2 (2\ k1+k2\ (1 + x1)(1 + x2)) = (k1 + k2) + \log2 (1 + x1) + \log2 (1 + x2), \quad (1)$$

где k1 и k2 обозначают позиции старшего бита (со значением 1) в операндах A и B соответственно, а x1 и x2 представляют собой дробные части A и B (x1, x2 [0, 1]) . В алгоритме Митчелла, когда 0 x < 1, выражение log2 (1 + x) примерно равно x. Следовательно, логарифм произведения можно выразить следующее:

$$\log2 (A \cdot B) \quad k1 + k2 + x1 + x2. \quad (2)$$

Логарифмический умножитель включает в себя несколько важных компонентов: детекторы ведущей единицы (LOD), логарифмические преобразователи, сумматор и антилогарифмический преобразователь. Задача LOD — идентифицировать старший 1 бит во входных двоичных числах. Впоследствии логарифмические преобразователи генерируют логарифмические представления этих двоичных чисел. Эти логарифмические значения складываются вместе, чтобы получить произведение, которое затем декодируется обратно в двоичное число с помощью антилогарифмического преобразователя.

3. Предлагаемый наиболее значимый приблизительный множитель сдвига с одним приводом

В этом исследовании мы вводим наиболее значимый однонаправленный сдвиговый приблизительный множитель , который использует логарифмический подход. Этот новый умножитель обеспечивает повышенную точность при сохранении затрат на оборудование, сравнимых или даже ниже, чем у традиционных приближенных логарифмических умножителей.

3.1. Архитектура дизайна

Фундаментальные операции в нейронных сетях (НС) включают в себя умножения, а именно умножение весов w на активации x. Архитектура нейронной сети, использующая предлагаемый наиболее значимый приблизительный множитель с одним приводом (MSAM), показана на рисунке 1. MSAM в основном состоит из двух основных компонентов: приближенного подмножителя (ASM) и точного подмножителя (ESM), дополненный вспомогательными переключателями и сумматорами для простого и последовательного структурного проектирования. В частности, вес w делится на две части в зависимости от значимости его битов: наиболее значимые биты формируют wH, обрабатываемые ESM, а менее значимые биты обрабатываются для формирования w, обрабатываемого ASM. Впоследствии результат ESM сдвигается и выполняется операция сложения для получения окончательного результата r.

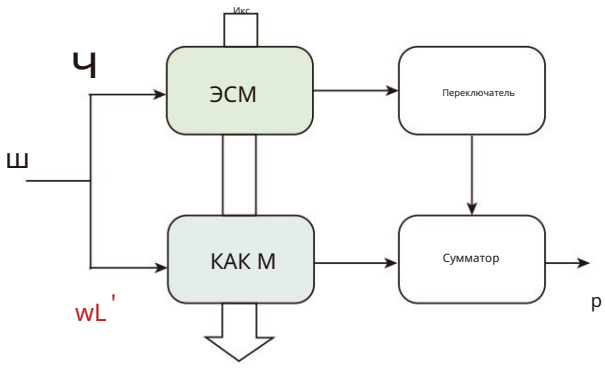


Рисунок 1. Архитектура наиболее значимого блока приближенного множителя сдвига с одним приводом (MSAM).

Преобразователь знаков, подобный преобразователю в [15], используется для расширения MSAM для поддержки чисел со знаком, как показано в левой части рисунка 2. Внутренняя особенность блока MSAM также представлена на рисунке 2. Мы вводим два основных параметра факторы , которые придают множителю атрибут динамической конфигурации, удовлетворяя разнообразнь

Machine Translated by Google

диапазон требований к точности, которые могут меняться в зависимости от контекста приложения. Первый — это коэффициент аппроксимации, обозначаемый как k, который используется для разделения множителя. Эта операция разделения разделяет n-битный множитель на два отдельных сегмента: k младших битов и (n - k) старших битов. Эта схема разделения эффективно разлагает одиночную операцию умножения на два дискретных мультипликативных процесса, включающих множимое и два множителя с меньшей разрядностью. Во-вторых, коэффициент точности, обозначаемый как m, включен в стратегию аппроксимации , управляющую ASM, и играет решающую роль в обеспечении настраиваемых настроек точности в рамках вспомогательного множителя.

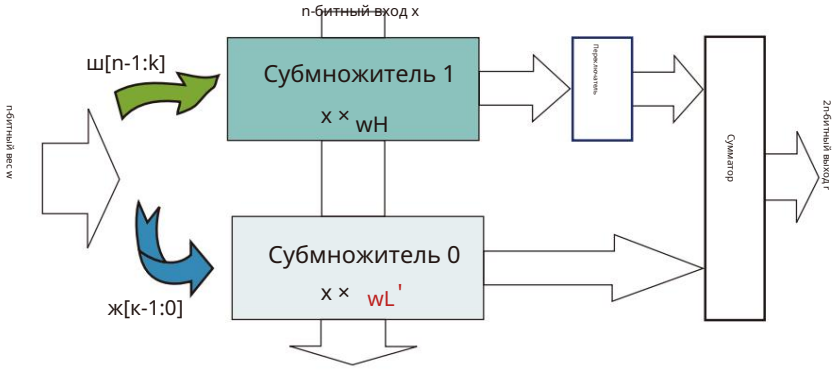


Рисунок 2. Расширение предлагаемого подхода для поддержки знаковых чисел, включая внутреннюю функцию блока MSAM.

3.2. Предлагаемые стратегии аппроксимации динамического веса

Чтобы достичь баланса между вычислительной эффективностью и точностью в конструкции MSAM, мы вводим несколько стратегий весовой аппроксимации, адаптированных для компонента ASM , основанных на фундаментальных принципах логарифмического множителя (LM) и переконфигурируемого множителя с постоянным коэффициентом (RCCM). дизайн. Эти стратегии направлены на удовлетворение различных требований приложений и оптимизацию производительности в различных условиях, включая преобразование операций умножения в более простые арифметические операции, такие как сдвиг, сложение или вычитание. Между тем, компонент ESM использует традиционные, точные методы умножения для обеспечения точных вычислений.
Чтобы обеспечить точное кумулятивное сложение, выходные данные, генерируемые ESM, подвергаются последовательности операций сдвига, чтобы привести их в соответствие с выходными данными ASM. Это выравнивание имеет решающее значение для процесса кумулятивного сложения, который в конечном итоге определяет окончательный результат блока MSAM. Объединив эти различные стратегии в компонентах ASM и ESM, мы стремимся оптимизировать общую производительность конструкции MSAM.

Значение активации x служит множителем, а вес w берет на себя роль множителя в операции умножения внутри нейронной сети (NN). Чтобы оптимизировать приблизительный множитель (ASM) для нейронных сетей, мы предлагаем метод , который разделяет веса модели на отдельные диапазоны данных и аппроксимирует веса внутри этих диапазонов как постоянные значения. Этот набор приблизительных постоянных значений, обозначаемый как W

$$'$$

, эффективно соответствует постоянному набору коэффициентов C , используемому в структуре RCCM [12]. Принципиальное отличие состоит в том, что предлагаемый набор W

$$'$$

требует определения только единственной начальной константы w_0', тогда как остальные константы в наборе выводятся относительно этой начальной константы. Кроме того, эта инновационная стратегия аппроксимации веса позволяет динамически выбирать подходящие приблизительные постоянные значения на основе наиболее значимого значения (MSO) во входных данных веса.

3.2.1. Стратегия однодоминирующего приближения (OAS)

Процесс аппроксимации одним множителем (w) с использованием предложенного метода изображен на рисунке 3. Обнаружение ведущей единицы (LOD) используется для обнаружения наиболее значимого множителя (MSO) множителя w. Как только MSO идентифицирован, последующие (m 1) б

Machine Translated by Google

устанавливаются в 1, а остальные биты устанавливаются в 0. Если позиция MSO находится на mso-м бите (считая от 0, где mso — индекс бита MSO), все числа в диапазоне w [2 mso, 2mso+1 1] аппроксимируются как постоянное значение, равное единице, согласно уравнению (3).

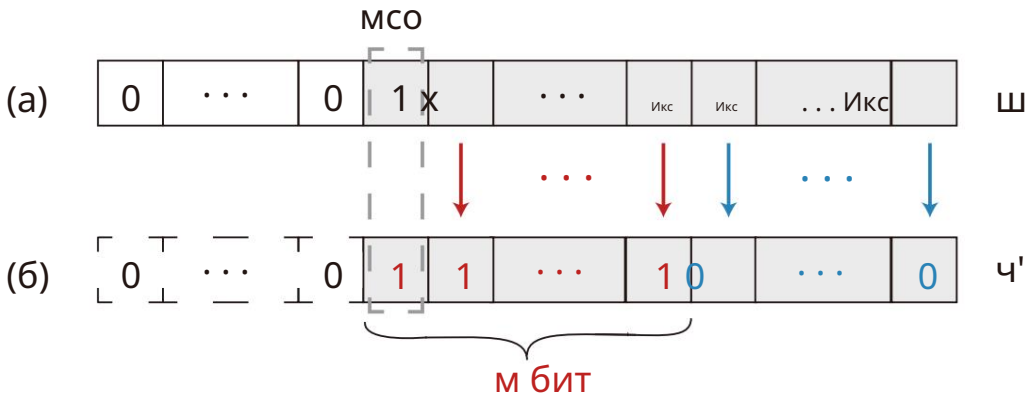


Рисунок 3. Пример процесса аппроксимации множителя с преобладанием единицы: (а) исходное число; (б) приблизительное количество.

$$ш' = \begin{cases} ш_0, & мсо < м \\ я_{мсо} 2^{i=мсо} \; {}^{м+1}, & мсо \geq м. \end{cases} \quad (3)$$

Значение w 0 ' определяется коэффициентом m. Например, когда m = 1, w 0 ' установлено значение 1b; между тем, когда m = 2, w ' устанавливается равным 11b. В этом методе аппроксимации множитель равен 0, динамически разбиваемый на отдельные интервалы диапазона данных в зависимости от положения MSO во входных данных. Эти интервалы затем аппроксимируются с использованием набора постоянных коэффициентов.

Мы называем m коэффициентом точности, поскольку он контролирует количество интервалов аппроксимации.

Выбор m последовательных битов, начиная с MSO, и установка их в 1 основан на двух основных причинах. Во-первых, это позволяет через последовательность операций сдвига установить связь между данной группой приближенных значений и исходным приближенным постоянным коэффициентом w . В результате, предлагая начальную приблизительную константу w , равную 0 , мы можем эффективно восстановить результаты умножения ' остальных коэффициентов по отношению к входным данным x. Этот метод значительно упрощает аппаратную реализацию и снижает сложность. Во-вторых, разумный выбор значения m позволяет стратегически расположить приблизительные константы вблизи среднего значения ' и полагаясь на произведение x и w 0 ' в каждом весовом интервале, тем самым максимизируя точность множителя. По сути, это позволяет аппроксимированным константам тяготеть к центральному значению отдельного диапазона весов, тем самым повышая общую точность архитектуры умножителя.

3.2.2. Стратегия аппроксимации с доминированием нуля (ZAS)

В описанной выше стратегии аппроксимации с доминированием одного все m бит, начиная с позиции MSO, устанавливаются в 1 по важной причине: это аппроксимированное m-битное число должно очень напоминать среднее значение всех k-битных чисел. Однако в крайнем случае, когда m = k, приближенное число становится наибольшим, а не аппроксимирует среднее значение. Чтобы решить эту дилемму, мы предлагаем альтернативную стратегию аппроксимации с доминированием нуля, как показано на рисунке 4. Сразу после позиции MSO не все оставшиеся биты устанавливаются в 1. Вместо этого первый бит, следующий сразу за позицией MSO, равен 1. установлено значение 0, а остальным битам присвоено значение 1. Это эффективно позволяет избежать проблемы чрезмерно больших приблизительных чисел, обеспечивая при этом близость к среднему значению. Поскольку 0 действует как разделитель, это также позволяет разумно применять методы переноса и отбрасывания [10], тем самым уменьшая задержку и энергопотребление.

Machine Translated by Google

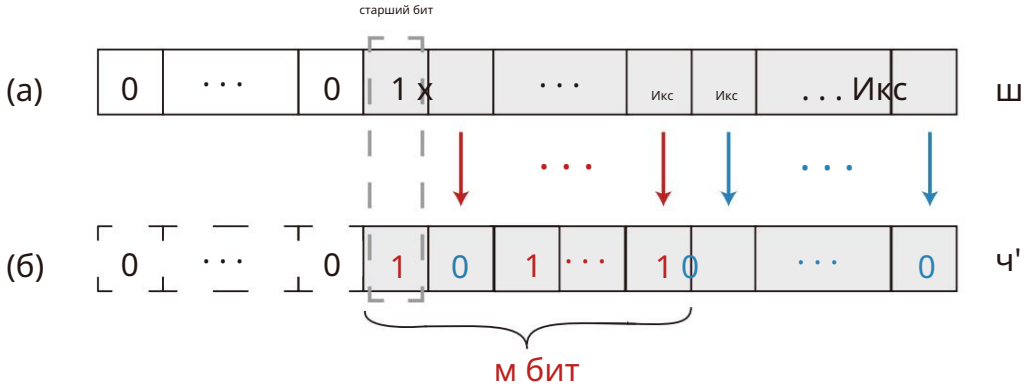


Рисунок 4. Пример процесса аппроксимации множителя с преобладанием нуля: (а) исходное число и (б) приблизительное количество.

3.2.3. Компенсация стратегии аппроксимации с доминированием нуля

Чтобы лучше использовать преимущества аппроксимации, мы обычно не выбираем m = k. Эмпирически установка m log2 (k) + 1 обеспечивает благоприятный баланс между энергией Преимущества потребления и точности.

Как показано в таблице 1, с точки зрения подхода MSAM-42 (т. е. k = 4 и m = 2), средний результат, полученный с помощью OAC, немного превышает медиану, тогда как средний результат результат с ZAS значительно меньше медианы.

1 Таблица 1. Результаты, полученные с MSAM-42 используя различные стратегии аппроксимации.

k битов (Ввод w[3 : 0])	Один-Доминирующий Приближение	Нулевое доминирование Приближение	Компенсация за Нулевое доминирование
0000	0000	0000	0000
0001 0010 0011	0011	0010	0101
0100 0101 0110 0111	0110	0100	0101
1001 1010 1011 1100 1101 1110 1111	1100	1000	1010

1 MSAM-42 указывает, что приблизительный коэффициент k равен 4, а соответствующий коэффициент точности m равен 2.

Хотя это может лучше сократить потребление энергии, это также может привести к чрезмерному ошибки. Для достижения лучшего равновесия в этом отношении мы вводим схему компенсации. чтобы уменьшить количество ошибок и гарантировать, что общая точность не будет существенно нарушена. Как показано на рисунке 5, учитывая, что приблизительное число с ZOS, скорее всего, меньше фактического за счет добавления дополнительной компенсации 1 после m-бита, мы можем эффективно уменьшить расстояние ошибки, в то время как дополнительное потребление энергии по сути ничтожно.

Machine Translated by Google

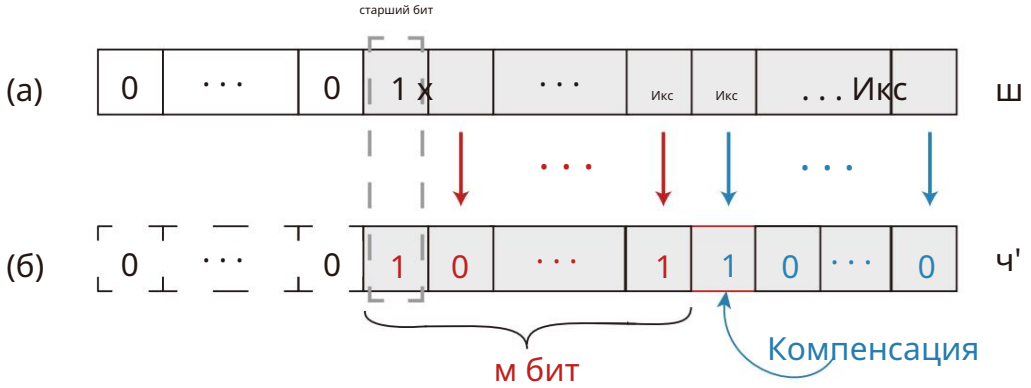


Рисунок 5. Пример аппроксимации множителя с доминированием нуля с процессом компенсации: (а) исходное число и (б) приближенное число.

3.3. Реализация аппаратного дизайна

Суть превосходства конструкции MSAM заключается в внимательном управлении ASM, поскольку совершенствование его аппаратного обеспечения напрямую и существенно влияет на общее энергопотребление блока умножения.

Как показано на рисунке 6, процесс начинается с серии операций сдвига и сложения, совместно генерирующих произведение множимого x и конкретной замены в соответствии с коэффициентами k и m. Положение (pos) наибольшего значения w для W 0 L, значимого одного бита во входном мультиплексоре (MUX) весе wL, идентифицируемого детектором ведущей единицы (LOD), устанавливает управляющий сигнал для последующего управляет сдвигом в r0tmp. Впоследствии мультиплексор используется для формирования выходного сигнала r0 для модуля приблизительного подумножителя (ASM). Между тем, ESM отличается точным подходом к умножению. В конце концов, выходные данные ESM подвергаются последовательности операций смещения, а затем объединяются с результатами ASM, в конечном итоге образуя окончательный результат MSAM.

Для поддержки знаковых операций архитектура MSAM может быть расширена за счет интеграции схем преобразования знаков как до, так и после беззнаковых операций, как показано на рисунке 2. Эти схемы используют дополнение и выполняют операцию ИЛИ с логикой 1b в младшем бите, обслуживая как приближение для отрицания при работе с отрицательными входными данными.

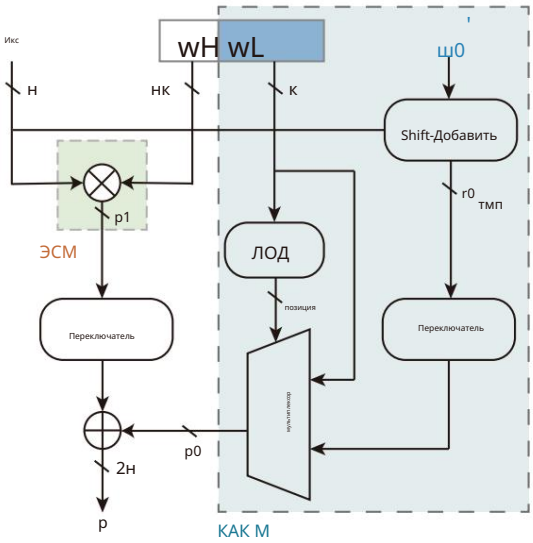


Рисунок 6. Аппаратная реализация MSAM.

Machine Translated by Google

4. Результаты и обсуждение.

Предлагаемая нами конструкция MSAM реализована с использованием Verilog HDL наряду с другими расширенные приближенные множители. Проекты были синтезированы с использованием Synopsys. Компилятор проектирования (DC) с КМОП-технологией 28 нм, и было проведено моделирование. используя примерно 1 000 000 наборов случайных чисел, охватывающих все входное пространство множителей 8 × 8 (65 536 комбинаций) на тактовой частоте 200 МГц. СВК инструмент использовался для моделирования, генерируя файлы активности переключения с обратными аннотациями. для облегчения анализа мощности. Метрика NMED использовалась для количественной оценки ошибки предлагаемые проекты, которые всесторонне оцениваются по всему входному пространству используя равномерно распределенные данные. Предложенные беззнаковые MSAM сравнивались с Mitch-w [14], DR-ALM-7 [15], DRUM6 [16] и AxRM2 [17], а подписанные MSAM сравнивались с Mitch-w [14], DR-ALM-7 [15], C-SSM (m = 4) [18] и Approx-Booth [19]. Для понятных и удобных сравнений MSAM относится к использованию OSA. стратегия аппроксимации, MCAM3 указывает на использование стратегии аппроксимации ZSA, а MSAMC обозначает ZSA с дополнительной компенсацией.

Сокращения, использованные в данной рукописи, представлены в таблице 2.

Таблица 2. Сокращения для приблизительных конструкций множителей.

Множитель	Техники	Описание
Точный	-	Точный множитель.
MSAM-км	Логарифмический	Предлагаемый наиболее значимый приблизительный множитель сдвига с одним приводом при различных коэффициентах аппроксимации k и коэффициентах точности m.
Митч-в [14]	логарифмический	Беззнаковый приблизительный логарифмический множитель Митчелла с w-битным операндом.
Митч-в [14]	логарифмический	Приблизительный логарифмический множитель со знаком Митчелла с w-битным операндом.
ДР-АЛМ-7 [15]	логарифмический	Приближенный неитерационный беззнаковый логарифмический множитель с Li -битным операндом (Li = 7).
ДР-АЛМ-7 [15]	логарифмический	Приближенный неитеративный знаковый логарифмический множитель с Li -битным операндом (Li = 7).
БАРАБАН6 [16]	логарифмический	Несмещенный множитель с k(k = 6) битами в операции от ведущего.
АксРМ2 [17]	Нелогарифмический	Приближенные рекурсивные умножители с использованием высокопроизводительных строительных блоков.
C-SSM (m = 4) [18]	Нелогарифмический	Приблизительный статический сегментированный множитель с непрерывным сегментом из m бит (m = 4).
Приблизительный стенд [19]	Нелогарифмический	Недорогой приблизительный множитель Бута по основанию 4 с упрощенная конструкция компрессора и кодера.

4.1. Анализ ошибок

Ускорители глубокого обучения, включающие приблизительные множители и сумматоры, был разработан для снижения энергопотребления и задержки, хотя и с приемлемым компромиссом в точности [20]. Как правило, более высокая степень ускорения имеет тенденцию приводит к еще большему снижению точности. Характеристики погрешности приближенного множитель (AM), как правило, определяется посредством исчерпывающего моделирования, которое включает в себя все мыслимые входные комбинации или, альтернативно, с помощью моделирования Монте-Карло, которое использовать случайные входные данные, выбранные из заранее определенного распределения, такого как нормальное или равномерное распределение.

Учтите, что M обозначает выход приближенного умножителя, а M представляет собой точный результат. Расстояние ошибки (ED) обеспечивает точное измерение абсолютного разница между этими двумя выходами, которая формулируется как ED = |M M |. относительное расстояние ошибки (RED) выражает ошибку в процентах относительно точного результата. Кроме того, среднее расстояние ошибки (MED) и среднее относительное расстояние ошибки (MRED) — широко используемые метрики для оценки точности приближенных проектов. Кроме того

Machine Translated by Google

По сравнению с MED и MRED, нормализованный MED (NMED) и нормализованный RED (NRED) являются стабильными показателями, которые не зависят от размера реализации, что делает их ценными для оценки приблизительных проектов различных размеров. Как показано в [21], NMED приближенного множителя (AM) оказывает существенное влияние на выходную ошибку нейрона в нейронной сети, особенно при работе с большим количеством входных данных.

Для указанного MSAM, когда значение k остается постоянным и обеспечивается согласованность разрядности операнда для всех операций, мы используем MRED и NMED как две эффективные и широко используемые метрики, чтобы оценить ошибку, вносимую ASM. блокировать. Эти метрики служат незаменимым средством оценки ошибок, присущих предлагаемой конструкции, с акцентом на различные конфигурации коэффициента точности m. На рисунке 7 представлены кривые MRED для значений k 3, 4 и 5, построенные в зависимости от изменений коэффициента точности m, что позволяет идентифицировать различные тенденции внутри кривых. Для разных значений k MRED достигает минимума, когда m равно 1 или 2. В отличие от других AM, в конструкции MSAM коэффициент точности m не обязательно лучше, когда он больше. Фактически, когда m превышает log2 k + 1, чрезмерное количество бит устанавливается в 1, что приводит к значительному отклонению аппроксимированного значения от фактического значения. Кроме того, большее значение m приводит к тому, что в вычислении участвует больше битов, что приводит к потреблению большего количества энергии. Стоит отметить, что результаты, полученные с помощью метрики NMED, тесно согласуются с этими экспериментальными данными. Следовательно, в последующем сравнительном анализе и детальных исследованиях мы отдали приоритет значениям m , равным 1 или 2. Этот стратегический выбор был сделан с целью достижения оптимального баланса в конструкции, эффективного согласования требований к точности и соображений накладных расходов на оборудование.

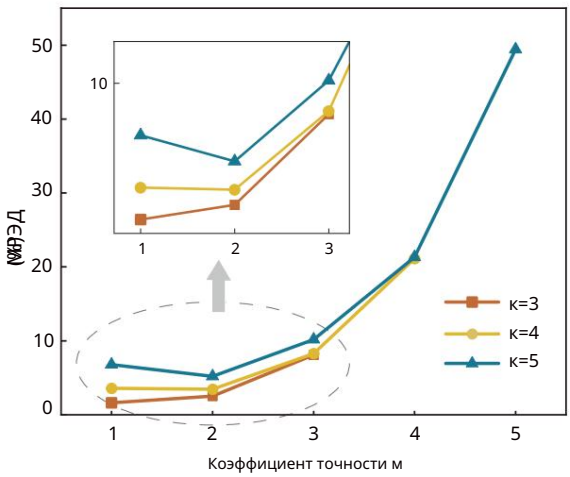


Рисунок 7. Кривые среднего расстояния относительной ошибки MSAM при различных коэффициентах аппроксимации k и коэффициентах точности m.

4.2. Результаты синтеза и моделирования

Различные множители MSAM-км , существующие логарифмические множители и усовершенствованные нелогарифмические множители были оценены для оценки производительности их оборудования и показателей точности.

В нашем приближенном множителе один операнд делится на два компонента: ESM и ASM, управляемые коэффициентом аппроксимации k, который определяет количество аппроксимируемых битов. Коэффициент точности m представляет степень приближения. Уровень аппроксимации в первую очередь определяется параметром k; эмпирически, большее значение k приводит к более высокой степени аппроксимации, более простой конструкции и более низкому потреблению энергии, тогда как меньшее значение k приводит к меньшей аппроксимации и уменьшению экономии энергии. Параметр m определяет, как обрабатываются k приблизительных битов . Больший m сохраняет больше битов, но важно отметить, что это не означает сохранения исходных точных битов; вместо этого он использует m бит всех единиц. Следовательно, большее значение m не обязательно означает более высокую точность. По статистике, m ближе к k приводит к тому, что k аппроксимирующих бит становятся единицами , что приводит к меньшей точности. От

Machine Translated by Google

экспериментальные результаты и статистический анализ, разумное значение m составляет около log2(k), делая m ближе к среднему значению всех k-битных чисел. Введение параметров k и м позволяет провести тонкую настройку конструкции для достижения желаемого баланса между точностью и энергоэффективность. Однако эти дополнительные параметры могут потребовать тщательного рассмотрение и оптимизация на этапе проектирования для обеспечения оптимальной производительности.

Логарифмические множители обеспечивают более низкое энергопотребление, но жертвуют точностью. Наш конструкция смягчает это, обеспечивая более высокий уровень точности при сохранении более низкого потребляемая мощность, чем нелогарифмические множители. Сравнение NMED и произведение задержки мощности (PDP) для беззнаковых 8-битных умножителей показано на рисунке 8, где предлагаемые MSAM выделены красным квадратом. Хотя DR-ALM-7 [15] и Mitch- w [14] показали более низкий PDP, они привели к более высокому NMED. И наоборот, DRUM6 [16] и MSAM-21 производил меньший NMED, но требовал более высокого PDP. В частности, AxRM2 [17], MSAM-52, MSAM-41 и MSAM-42 достигли относительно более низкого PDP, а также меньшего НМЕД. Среди них, как показано красным кружком, MSAM-41 и MSAM-42 выделяются как наиболее привлекательные проекты, обеспечивающие оптимальный компромисс между PDP и NMED.

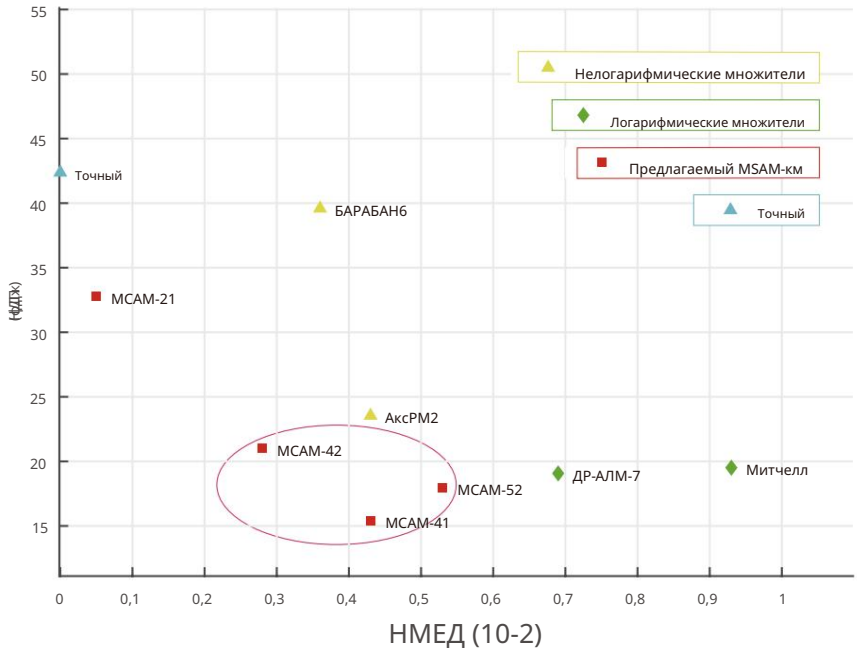


Рисунок 8. NMED и PDP 8-битных беззнаковых приблизительных множителей.

Подробные результаты этой оценки обобщены в Таблице 3. Тщательно настроив параметры k и m, мы добились аппаратной производительности, сравнимой с (или, в некоторых случаях, превышая) множители Митчелла [14] и DR-ALM-7 [15] (для например, с км = 41 или км = 52). При сравнении MSAM-41 и MSAM-52 с DR- ALM-7, даже в ситуациях, когда производительность продукта задержки мощности (PDP) превосходит последнее, замечательные улучшения все еще наблюдались с точки зрения NMED, достигнув 37,68% и 23,19% соответственно. Напротив, по сравнению с AxRM2 [17], MSAM-41 демонстрирует значительные улучшения, включая сокращение использования площади на 32,11%, снижение задержки на 23,33% и улучшение PDP на 27,95% при сохранении эквивалент NMED. Эти результаты подчеркивают достигнутую впечатляющую эффективность оборудования. с конструкцией MSAM по сравнению с современными конструкциями AM.

Хотя ZAS предлагает более упрощенную аппаратную конструкцию, обеспечивающую превосходные потребление энергии и использование площади, NMED увеличен по сравнению с соответствующей стратегией OAG. Наш проект предусматривает несколько приблизительных стратегий, таких как ZAS. и OAG для достижения более динамичного диапазона компромиссов, удовлетворяя тем самым требования более широкий спектр применения.

Machine Translated by Google

Таблица 3. Сравнение производительности 8-битных беззнаковых приблизительных множителей.

Дизайн	Площадь (мкм2)	Мощность (мкВт)	Задержка (нс)	НДП (фДж)	НМЕД (10 2)
Точный	154,48	76,75	0,59	45,28	0
МСАМ-21	147,29	59,55	0,59	35,13 (22%)	0,05
MCAM3-21	144,22	56,2	0,58	32,60 (28%)	0,07
МСАМ-32	143.01	52,3	0,52	27,20 (40%)	0,17
MCAM3-32	139,05	50,4	0,5	25,20 (44%)	0,19
МСАМ-31	129,53	48,39	0,52	25,16 (44%)	0,17
MCAM3-31	126,45	44,66	0,49	21,88 (52%)	0,19
МСАМ-42	125,5	41,78	0,52	21,73 (52%)	0,28
MCAM3-42	120,79	38,98	0,49	19,10 (58%)	0,34
МСАМ-41	111,89	36,86	0,46	16,96 (63%)	0,43
MCAM3-41	103,76	35.33	0,45	15,90 (65%))	0,47
МСАМ-52	110,75	32.36	0,52	16,83 (63%)	0,53
МСАМ3-52	104,39	31.33	0,51	15,98 (65%)	0,59
Митчелл [14]	103,45	30.01	0,65	19,51 (57%)	0,93
ДР-АЛМ-7 [15]	98,28	28.46	0,67	19,07 (58%)	0,69
БАРАБАН6[16]	192,53	56,10	0,77	43,20 (-5%)	0,36
АксРМ2 [17]	164,81	39,23	0,60	23,54 (-48%)	0,43

Предлагаемые конструкции находятся в заштрихованном разделе.

4.3. Оценка, моделирование и анализ

4.3.1. Оценка производительности глубоких нейронных сетей

Для более полной оценки эффективности множителя при использовании в больших моделей, мы провели эксперименты с использованием набора данных CIFAR-10 и использовали три типичные глубокие нейронные сети: VGG-19, известная своими глубокими сверточными слоями; ResNet-50, который эффективно решает проблему исчезновения градиента в глубоких сетях. за счет использования остаточных связей; и DenseNet-121, обеспечивающий эффективную функцию использование через механизмы плотной связи. Мы провели эксперименты с использованием эти сложные сети с целью тщательно оценить производительность мультипликатора и применимость. Кроме того, мы расширили наш экспериментальный анализ на более сложные Набор данных ImageNet за счет включения SqueezeNet, компактного, но эффективного глубокого нейронного сетевая архитектура, разработанная для оптимизации производительности на устройствах с ограниченными возможностями.

Для оценки эффективности приближенных множителей мы использовали Адаптировать структуру [22] для облегчения моделирования и оценки различных множителей. конструкции. Как показано на рисунке 9, входные данные состоят из предварительно обученной модели нейронной сети, квантования конфигурации и файл проекта приблизительного множителя (на основе Verilog HDL). Первоначально граф слоя строится с использованием параметров сверточного слоя и соседний слой пакетной нормализации. Впоследствии проектный файл приблизительного был проанализирован мультипликатор. На основе конфигураций квантования соответствующие создаются приблизительные слои. Эти приблизительные слои затем заменяют слои в исходный граф слоя. После этого восстановленный граф слоя обрабатывается с помощью Платформа AdaPT для переобучения и вывода, придерживающаяся указанной стратегии переобучения. Примечательно, что для набора обучающих данных требуется только репрезентативное подмножество, которое может составлять примерно 10 % от исходного обучающего набора для целей калибровки. Наконец, Предоставляются отчеты о производительности с подробным описанием точности и энергопотребления.

Переобучение и тонкая настройка модели могут использоваться для смягчения ухудшения точности, возникающего из-за определенных приближений, таких как сокращение и приближенное умножение [23]. Эти подходы позволяют модели адаптироваться и стать более устойчивой к ошибки, вносимые такими приближениями. На этапе вывода эти методы облегчить модели AxC, чтобы снизить затраты на оборудование, достигая при этом значительно более высоких точность. Чтобы эффективно оценить эффективность приближенных множителей, эксперименты проводились на наборе данных CIFAR-10 путем прямого вывода без каких-либо

Machine Translated by Google

переподготовка. Эта методология позволила быстро оценить эффективность мультипликаторов. Между тем, для набора данных ImageNet была принята более эффективная стратегия переобучения. выбрав только 10% обучающего подмножества. Эта стратегия была направлена на выполнение модели настройка на этапе обучения, позволяющая повысить точность и значительную энергию сокращение на этапе фактического вывода. Это демонстрирует эффективность переподготовки и настройка модели для уменьшения ошибок и повышения точности без обременения дорогостоящие требования к аппаратному обеспечению.

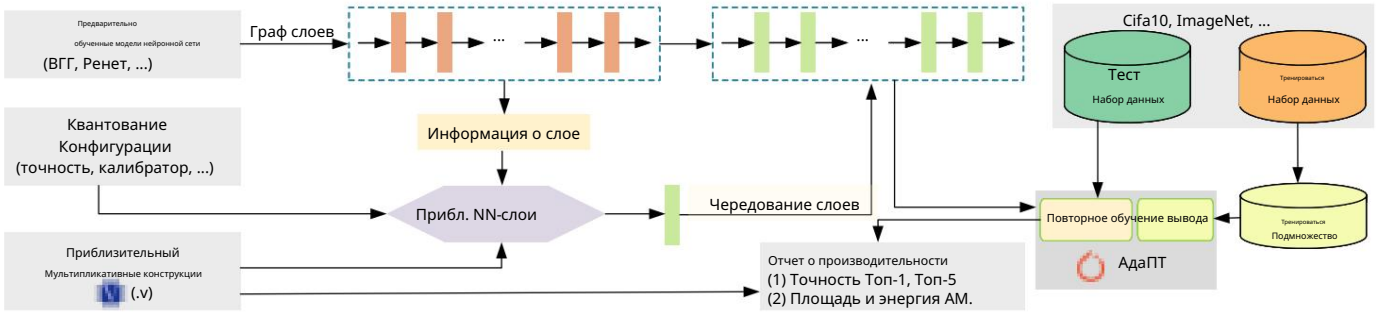


Рисунок 9. Порядок оценки на основе платформы AdaPT.

Несколько подписанных версий множителя MSAM были реализованы и тщательно по сравнению с точным множителем Бута по основанию 4, а также двумя логарифмическими множителями - а именно, Митчелл [14] и DR-ALM-7 [15], а также два нелогарифмических приближенных множители — C-SSM [18] и Approx-Booth [19]. Чтобы обеспечить справедливое сравнение, мы повторили конструкции 8-битного C-SSM (с m = 4) и Approx-Booth (с wG, wA = 6, 6). оценка этих знаковых множителей строго следовала ранее изложенной экспериментальной установке. Результаты этого комплексного сравнения представлены в Таблице 4.

При рассмотрении результатов набора данных CIFAR-10 MSAM приблизительно Умножитель показал наилучшие характеристики как с точки зрения точности, так и с точки зрения снижения энергопотребления. В частности, множитель MCAM-31 не только превосходил ДР-АЛМ-7 по показателям точности оценки сети, а также добились дополнительной экономии энергии на 9 % на каждый один множитель по сравнению с DR-ALM-7, что подчеркивает его эффективность. ДР-АЛМ-7, на с другой стороны, продемонстрировал превосходную производительность сети по сравнению с другими приблизительные множители, за исключением вариантов MSAM.

Таблица 4. Сравнение стоимости оборудования и точности 8-битных знаковых умножителей на CIFAR10 и наборы данных ImageNet.

Множитель	Задержка мощности в зоне NMED (10 2) (мк2) (мкВт) (нс)		Энергия (фДж)	СИФАР10		ImageNet
				VGG-19 ResNet-50	DenseNet-121 Топ-1 Топ-5	
Точное основание системы счисления-4	0	276,70 94,27	0,60 56,56 93,69% 0,62 32,75 (-42%)	93,57%	93,84%	57,37% 80,19%
Подписанный MSAM-21	0,34	173,88 52,82	93,66% 0,58 27,71 (-51%) 93,23% 0,55	92,81%	93,58%	54,96% 78,21%
Подписанный MSAM-32	0,40	169,60 47,78	23,55 (-58%) 93,60% 0,60 23,83 (-58%)	91,34%	93,30%	52,02% 76,00%
Подписанный MSAM-31	0,61	156,11 42,82	93,08% 0,49 16,71 70%) 91,24% 0,71	93,06%	93,64%	54,01% 77,60%
Подписанный MSAM-42	0,60	152,08 39,71	27,95 (-51%) 89,64% 0,73 28,64 (-49%)	91,71%	92,93%	50,37% 74,35%
Подписанный MSAM-41	1,20	138,47 34,10	93,29% 0,43 35,42 (-37%) 40,24% 0,55	91,26%	91,46%	43,31% 68,00%
Mitchell* [14]	1,12	132,30 39,37	40,25 (-29%) 12,29%	92,52%	91,75%	50,19% 74,20%
DR-ALM-7* [15]	0,78	127,13 39,23		92,24%	93,37%	53,92% 77,38%
Ц-ССМ (м = 4) [18]	3,62	276,32 82,37		61,17%	73,28%	2,22% 7,47%
Приблизительный стенд [19]	0,17	218,23 73,18		10,65%	9,99%	17,44% 36,33%

Три верхних приблизительных мультипликатора, продемонстрировавшие наибольшую адаптивность в набор данных ImageNet вошли MSAM-21, MSAM-31 и DR-ALM-7. Эти множители продемонстрировал исключительную производительность в различных сетевых архитектурах, подчеркнув их универсальность и надежность. Для приложений, в которых приоритетом является высокая точность сети,

Machine Translated by Google

MSAM-21 отличился как выдающийся производитель, добившись улучшения точности Top-1 на 1,04% и улучшения точности Top-5 на 0,83% по сравнению с DR-ALM-7. Эти улучшения значительны и могут оказать существенное влияние на общую производительность сети. Результаты, представленные в таблице 4, обеспечивают всестороннее сравнение стоимости оборудования и точности 8-битных знаковых умножителей в наборах данных CIFAR10 и ImageNet. Умножители MSAM и DR-ALM-7 неизменно демонстрировали высокие характеристики, при этом варианты MSAM демонстрировали особые надежды с точки зрения точности и энергоэффективности. Эти результаты имеют важное значение для разработки и оптимизации множителей для использования в приложениях глубокого обучения.

4.3.2. Оценка производительности ускорителя умножения трансформаторных матриц

Была проверена экономия ресурсов и энергии, достигнутая с помощью предложенного множителя при его развертывании в LLM на периферийных устройствах. Эта проверка была достигнута путем применения умножителя к ускорителю умножения матриц-трансформаторов, специально разработанному для обработки матриц размером 16 × 16. При оценке аппаратных накладных расходов, когда точные множители служили базой для сравнения, мы сосредоточились на оценке потребляемая мощность ускорителя связана с комбинационной логикой. Ссылка. [24] подчеркивает возможность выполнения вывода в LLM с количеством параметров до 175B без какого-либо снижения производительности с использованием LLM.int8(), который позволяет умножать более 99,9% значений в 8-битном формате. Ссылка. [25] демонстрирует эффективность приближенных множителей в контексте моделей преобразователей зрения. Как показано на рисунке 10, наш ускоритель матричного умножения включает в себя две 8-битные целочисленные матрицы: XI8 и WI8. Для ускорения процесса умножения мы заменили точное умножение приближенным методом.

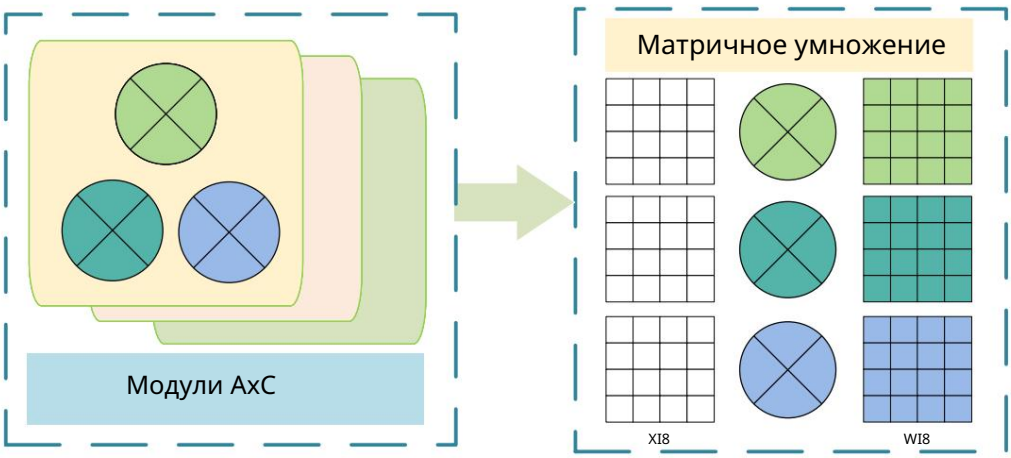


Рисунок 10. Ускоритель матричного умножения.

В таблице 5 представлены экспериментальные результаты модуля матричного умножения в ускорителе Transformer, подчеркивающие эффективное снижение аппаратных затрат, достигнутое с помощью предлагаемого умножителя. Результаты демонстрируют способность мультипликатора оптимизировать использование ресурсов и энергии при развертывании LLM на периферийных устройствах. По сравнению с базовым результатом, полученным с использованием точного множителя, MSAM демонстрируют значительное снижение площади, энергопотребления и задержки, при этом MSAMZ-41 демонстрирует наиболее выраженные улучшения. Использование приближенных вычислений продемонстрировало возможность повышения эффективности преобразователей за счет минимизации их вычислительной сложности и требований к памяти [25]. Учитывая, что у больших моделей более высокие вычислительные требования, это предполагает возможность включения боле Поэтому наши будущие исследования должны продолжать изучать и углубляться в более подходящую и рациональную интеграцию этих MSAM, стремясь снизить потребление энергии при сохранении приемлемой точности.

Machine Translated by Google

Таблица 5. Аппаратурные характеристики предлагаемой конструкции с трансформаторным ускорителем.

	Площадь (мкм2)	Гребень. Мощность (мВт)	Задержка (нс)
Базовый уровень	90 391,64	2.07	0,89
MCAM-41	67 618,91 (-25,19%) 47	1,82 (-12,08%) 1,46	0,79 (-11,24%)
MCAM3-41	618,35 (-47,32%) 51	(-29,47%) 1,55	0,75 (-15,73%)
MCAMK-41	618,67 (-42,89%)	(-25,12%)	0,76 (-14,61%)

4.4. Практический пример: ускоритель умножения матриц изображений

В нашем подходе мы проводим попиксельное умножение матриц между входными изображение и изображение-маска, где значение соответствующего пикселя выходного изображения определяется по полученному продукту. На рисунке 11 показано умножение изображений. результаты, где качество изображения положительно коррелирует с пиковым соотношением сигнал/шум. коэффициент (PSNR); по существу, значения PSNR для сгенерированных изображений включены в их соответствующие субтитры. Результаты, полученные с большинством наших MSAM, указывают на высокое качество, демонстрируя очень хороший PSNR по сравнению с точным множителем на рисунке 11c.
Из-за относительно равномерного распределения данных изображения, поскольку наша конструкция включает в себя ESM обладает высокой полезностью и надежностью в контексте обработки изображений.

(а) Исходное изображение (б) Изображение маски (с) Точный множитель (100 дБ) (г) MCAM-21 (57,65 дБ)

(д) MCAM-32 (51,71 дБ) (е) MCAM-31 (50,29 дБ) (г) MCAM-42 (47,48 дБ) (з) AxRM2 (47,28 дБ)

(i) MSAM-52 (44,46 дБ) (к) MSAM-41 (41,41 дБ) (к) ДР-АЛМ-7 (38,10 дБ) (л) Митчелл (37,98 дБ)

Рис. 11. Результаты умножения (а) входного изображения на (б) изображение маски с использованием выделенного множители.

Machine Translated by Google

4.5. Обсуждение предлагаемых MSAM 4.5.1. Преимущества и ограничения.

Предлагаемые MSAM обеспечивают значительную экономию энергии, одновременно ограничивая ошибки аппроксимирующих множителей за счет использования наиболее значимых стратегий побитовой аппроксимации . Эти стратегии обеспечивают несколько основных преимуществ по сравнению с другими работами. Во-первых, MSAM используют наиболее важные методы побитовой аппроксимации, облегчая выполнение умножения посредством сдвига. Это приводит к значительному уменьшению как площади, так и энергопотребления, что делает наш подход высокоэффективным для приложений нейронных сетей. Во-вторых, в отличие от логарифмических подходов, наш подход гарантирует, что приблизительный результат не будет односторонне смещенным. Эта функция обеспечивает компенсацию различных приближений, что приводит к повышению общей точности и надежности. Более того, ключевым преимуществом нашей стратегии аппроксимации является то, что она фокусируется только на одном операнде в операции умножения, полностью сохраняя при этом точность другого операнда. Этот особый подход эффективно обеспечивает точность, обеспечивая значительное превосходство над большинством других работ в этой области. Точный контроль над стратегиями аппроксимации и способность сохранять точность в одном операнде способствуют повышению производительности приложений нейронных сетей.

Одним из основных ограничений предлагаемых MSAM является относительно высокая вероятность ошибки, вызванная использованием заранее определенных компонентов для замены определенных битов. Это сделает наши множители неприменимыми в сценариях, где требуется высокая точность. Несмотря на это ограничение, предложенные MSAM обладают значительной практической ценностью, поскольку масштаб ошибок невелик, а учитывая, что ошибки могут быть как положительными, так и отрицательными, общая ошибка приемлема для пакетных операций.

4.5.2. Будущая работа

Умножение матриц служит фундаментальным строительным элементом в архитектуре трансформатора [25]. В будущем мы намерены продолжить исследование конструкции приближенных множителей с концентрированными ошибками и создать библиотеку приближенных множителей, адаптированную для LLM. Аппроксимации на программном уровне, такие как сети меньшей сложности и обрезка, способствуют энергоэффективному ускорению DNN [23]. Хотя как программные, так и аппаратные методы аппроксимации независимо друг от друга обеспечивают существенные энергетические преимущества для DNN, мы намерены объединить эти методы в синергетическую структуру в будущих усилиях по максимизации энергосбережения. Мы надеемся, что эта статья привлечет больше исследователей к проведению исследований в этом аспекте, позволяя применять MSAM в более сложных сценариях с ограниченными энергетическими бюджетами и строгими требованиями к точности.

5. Выводы

В этом исследовании мы предложили энергоэффективный наиболее значимый динамический приближенный умножитель с битовым управлением для ускорения нейронных сетей. Использование стратегии весовой аппроксимации , адаптированной для NN и LLM, в конечном итоге привело к созданию высокоэффективного приблизительного множителя, основанного на принципах сдвига MSO. В результате конструкция MSAM умело обеспечивает гармоничный баланс между точностью и стоимостью оборудования. При применении как к LLM, так и к DNN наш подход привел к существенному сокращению затрат на аппаратные ресурсы. По сравнению с традиционными точными множителями, используемыми в LLM и DNN, наша конструкция позволила добиться значительного снижения энергопотребления на 60 % без каких-либо существенных негативных последствий для точности задач. Кроме того, мы провели тематическое исследование по слиянию изображений, и результаты показывают, что наши множители можно использовать для создания высококачественных изображений. По сравнению с точным множителем наш подход показал себя очень хорошо при использовании PSNR в качестве показателя оценки.

Вклад авторов: Концептуализация, РН; Программное обеспечение, БГ; Валидация, БГ; Письмо — первоначальный вариант, РН; Написание — обзор и редактирование, КС; Надзор, CW. Все авторы прочитали и согласились с опубликованной версией рукописи.

Machine Translated by Google

Финансирование: Работа поддерживалась грантами Национального фонда естественных наук Китая (62101252 и 62134002).

Заявление о доступности данных: по обоснованному запросу соответствующий автор может предоставить данные, подтверждающие выводы этого исследования.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Райан, МАК; Мукта, МСХ; Фатема, К.; Фахад, Нью-Мексико; Сакиб, С.; Мим, ММЈ; Ахмад, Дж.; Али, МЭ; Азам, С. Обзор больших языковых моделей: архитектуры, приложения, таксономии, открытые проблемы и проблемы. IEEE Access 2024, 12, 26839–26874. [Перекрестная ссылка]

2. Раш, М.Дж.; Маккин, К.; Галло, МЛ; Чен, А.; Фасоли, А.; Одерматт, Ф.; Ли, Н.; Нандакумар, СР; Нарайанан, П.; Цай, Х.; и другие. Обучение с учетом аппаратного обеспечения для крупномасштабных и разнообразных рабочих нагрузок глубокого обучения с использованием ускорителей вычислений в памяти . Нат. Коммун. 2023, 14, 5282. [CrossRef] [ПабМед]

3. Тохиди, А.; Омиди, Р.; Мохаммади, К. О разработке итеративных приближенных множителей с плавающей запятой. IEEE Транс. Вычислить. 2023, 72, 1623–1635. [Перекрестная ссылка]

4. Чжан Х.; Ко, С.Б. Эффективные приближенные множители положительных значений для вычислений с глубоким обучением. IEEE J. Emerg. Сел. Вершина. Сист. цепей. 2023, 13, 201–211. [Перекрестная ссылка]

5. Саяди, Л.; Тимархи, С.; Шейх-Акбари, А. Два эффективных приближенных беззнаковых множителя путем разработки новой конфигурации для компрессоров с приблизительным соотношением 4:2. IEEE Транс. Сист. цепей. Я Регул. Пап. 2023, 70, 1649–1659. [Перекрестная ссылка]

6. Шефер, СЈ; Тахери, П.; Хорени, М.; Джоши, С. Аппаратное влияние квантования и сокращения весов в нейронных сетях с шипами. IEEE Транс. Сист. цепей. II Express Briefs 2023, 70, 1789–1793. [Перекрестная ссылка]

7. Пинос, М.; Мразек, В.; Ваверка, Ф.; Васичек, З.; Секанина Л. Методы ускорения автоматизированного проектирования приближенных моделей. Сверточные нейронные сети. IEEE J. Emerg. Сел. Вершина. Сист. цепей. 2023, 13, 212–224. [Перекрестная ссылка]

8. Ахмадинежад, М.; Моайери, М.Х. Энергоэффективные и качественные приближенные множители для нейронной сети и изображений Обработка заявок. IEEE Транс. Экстрен. Вершина. Вычислить. 2022, 10, 1105–1116. [Перекрестная ссылка]

9. Ли, З.; Чжэн, С.; Чжан, Дж.; Лу, Ю.; Гао, Дж.; Тао, Дж.; Ван, Л. Адаптируемый приблизительный дизайн множителя на основе распределения и полярности входных данных. IEEE Транс. Очень крупномасштабная интеграция. (СБИС) Сист. 2022, 30, 1813–1826 гг. [Перекрестная ссылка]

10. Амирафшар Н.; Баруги, А.С.; Шаххосейни, HS; ТахериНеджад, Н. Карри игнорирует приблизительные множители. IEEE Транс. Схемы Сист. Я Регул. Пап. 2023, 70, 4840–4853. [Перекрестная ссылка]

11. Лотрич, У.; Пилипович Р.; Булик, П. Гибридное основание счисления-4 и приближенный логарифмический множитель для энергоэффективной обработки изображений. Электроника 2021, 10, 1175. [CrossRef]

12. Фараоне Дж.; Кумм, М.; Хардик, М.; Зипф, П.; Лю, Х.; Боланд, Д.; Леонг, PH AddNet: Глубокие нейронные сети с использованием FPGA-оптимизированные множители. IEEE Транс. Очень крупномасштабная интеграция. (СБИС) Сист. 2019, 28, 115–128. [Перекрестная ссылка]

13. Гонг, Б.; Чен, К.; Хуанг, П.; Ву, Б.; Лю, В. Наиболее значимые однонаправленные динамические эффективные множители для моделей большого языка. В материалах Международного симпозиума IEEE по схемам и системам (ISCAS) 2024 г., Сингапур, 19–22 мая 2024 г.; стр. 1–5.

14. Митчелл Дж. Н. Компьютерное умножение и деление с использованием двоичных логарифмов. ИРЭ Транс. Электрон. Вычислить. 1962, EC-11, 512–517.

15. Инь, П.; Ван, К.; Варис, Х.; Лю, В.; Хан, Ю.; Ломбарди, Ф. Разработка и анализ приближенного энергоэффективного динамического диапазона. логарифмические множители для машинного обучения. IEEE Транс. Поддерживать. Вычислить. 2020, 6, 612–625. [Перекрестная ссылка]

16. Хашеми, С.; Бахар, Род-Айленд; Реда, С. DRAM: Несмещенный множитель динамического диапазона для приближенных приложений. В материалах Международной конференции IEEE/ACM по компьютерному проектированию (ICCAD) 2015 г., Остин, Техас, США, 2–6 ноября 2015 г.; стр. 418–425.

17. Варис, Х.; Ван, К.; Сюй, К.; Лю, В. AxRMs: Приближенные рекурсивные умножители с использованием высокопроизводительных строительных блоков. IEEE Транс. Экстрен. Вершина. Вычислить. 2021, 10, 1229–1235. [Перекрестная ссылка]

18. Строло, годовое общее собрание акционеров; Наполи, Э.; Де Каро, Д.; Петра, Н.; Саггезе, Г.; Ди Мео, Г. Приблизительные множители с использованием статической сегментации: Анализ ошибок и улучшения. IEEE Транс. Сист. цепей. Я Регул. Пап. 2022, 69, 2449–2462. [Перекрестная ссылка]

19. Парк, Г.; Кунг, Дж.; Ли, Ю. Упрощенные конструкции компрессора и кодера для недорогого приблизительного умножителя стенда Radix-4. IEEE Транс. Сист. цепей. II Express Briefs 2022, 70, 1154–1158. [Перекрестная ссылка]

20. Цзян, Х.; Сантьяго, ФДХ; Мо, Х.; Лю, Л.; Хан, Дж. Приближенные арифметические схемы: обзор, характеристика и недавние приложения. Учеб. IEEE 2020, 108, 2108–2135. [Перекрестная ссылка]

21. Мо, Х.; Ву, Ю.; Цзян, Х.; Ма, З.; Ломбарди, Ф.; Хан, Дж.; Лю, Л. Изучение особенностей ошибок приближенных множителей для нейронных сетей Сетевые приложения. IEEE Транс. Вычислить. 2024, 73, 842–856. [Перекрестная ссылка]

22. Данопулос, Д.; Зервакис, Г.; Сиозиос, К.; Судрис, Д.; Хенкель, Дж. AdaPT: Быстрая эмуляция приближенных ускорителей DNN в ПайТорч. IEEE Транс. Компьютерный дес. Интегр. Сист. цепей. 2023, 42, 2074–2078. [Перекрестная ссылка]

23. Саньял, С.; Неги, С.; Рагунатан, А.; Рой, К. Приближенные вычисления для рабочих нагрузок машинного обучения: взгляд на схемы и системы . В приближенных вычислениях; Лю В., Ломбарди Ф., ред.; Springer International Publishing: Чам, Швейцария, 2022 г.; стр. 365–395.

Machine Translated by Google

24. Деттмерс, Т.; Льюис, М.; Белкада, Ю.; Зеттлмойер, Л. GPT3.int8(): 8-битное умножение матриц для трансформаторов в масштабе. В Proceedings of the Advances in Neural Information Processing Systems, Новый Орлеан, Луизиана, США, 28 ноября – 9 декабря 2022 г.
25. Данопулос, Д.; Зервакис, Г.; Судрис, Д.; Хенкель, Дж. TransAxx: Эффективные трансформаторы с приближенными вычислениями. arXiv 2024, arXiv:2402.07545.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Однокаскадная электронная схема управления освещением на основе SiC Диоды Шоттки для питания светодиодов глубокого ультрафиолета Лампа для дезинфекции и стерилизации

Чун-Ань Чэн, Цзянь-Сюань Чанг, Хун-Лян Чэн, Энь-Чи Чанг*, Лун-Фу Лан, Шэн-Хун Хоу и Чэн-Куань Линь

Факультет электротехники, Университет И-Шоу, город Гаосюн 84001, Тайвань; кэшенг@isu.edu.tw (C.-AC); chchang@isu.edu.tw (C.-HC); hlcheng@isu.edu.tw (H.-LC); isu11001004m@cloud.isu.edu.tw (Л.-Флорида); isu11001001m@cloud.isu.edu.tw (C.-XX); isu11001014m@cloud.isu.edu.tw (C.-KL)
* Переписка: enchihchang@isu.edu.tw; Тел.: +886-7-6577711 (доб. 6642)

Аннотация: Недавно на рынках очистки помещений и стерилизации поверхностей появился новый тип источника освещения – светодиоды глубокого ультрафиолета (LED). В данной работе предложен и разработан новый тип электронного драйвера освещения для питания светодиодной стерилизационной лампы глубокого ультрафиолета . Основная схема объединяет понижающий преобразователь и обратноходовой преобразователь в одноступенчатый понижающе-обратноходовой преобразователь переменного тока в постоянный с коррекцией коэффициента мощности. Кроме того, предлагаемый электронный драйвер освещения использует широкозонный SiC-диод Шоттки в качестве выходного диода для снижения потерь на силовом диоде и повторно использует энергию, запасенную в индуктивности рассеяния трансформатора, для повышения эффективности схемы. Индуктор намагничивания внутри представленного преобразователя переменного тока в постоянный предназначен для работы в режиме прерывистой проводимости (DCM), что, естественно, обеспечивает коррекцию коэффициента мощности (PFC). Разработан и внедрен одноступенчатый прототип драйвера мощностью 3,6 Вт (90 В/40 мА) для обеспечения светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации. Результаты экспериментов показывают, что измеренный коэффициент мощности (PF) превышает 0,9, а измеренное общее гармоническое искажение (THD) входного тока составляет менее 18 % при входном напряжении сети 110 В. Кроме того, измеренный коэффициент пульсации выходного напряжения составляет менее 1%, а коэффициент пульсации выходного тока составляет менее 4%. Кроме того, предлагаемый одноступенчатый электронный драйвер освещения для питания светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом обеспечивает высокий КПД схемы (более 90%), малое количество компо

Ключевые слова: схема электронного драйвера освещения; светодиод глубокого ультрафиолета; лампа для дезинфекции и стерилизации; коррекция коэффициента мощности; полное гармоническое искажение

check for updates

Цитирование: Ченг, Калифорния; Чанг, Ч.-Х.; Ченг, Х.-Л.; Чанг, Э.-К.; Лан, Л.-Ф.; Хоу, С.-Х.; Лин, К.-К. Одноступенчатая электронная схема управления освещением, использующая SiC-диоды Шоттки для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации. Электроника 2024, 13, 3048. https://doi.org/10.3390/electronics13153048 .

Академические редакторы: Имэн Чжан, Лецзя Сунь и Юмин Чжан

Поступила: 24 июня 2024 г.
Пересмотрено: 27 июля 2024 г.
Принято: 29 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария.
Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Микроорганизмы, такие как бактерии, вирусы, кисты и плесень, используют прорастание или образование спор для размножения. Когда микроорганизмы подвергаются облучению или воздействию дезинфицирующих и стерилизующих ламп, свет проникает через стенки клеток и разрушает их ДНК, изменяя их ДНК так, что они становятся безвредными и неспособными к размножению или даже умирают, достигая, таким образом, эффекта дезинфекции и стерилизации. Лампы для стерилизации и дезинфекции используются в широком спектре применений в различных секторах, включая медицинские учреждения, пищевую промышленность и производство напитков, общественные места и транспорт, коммерческие и офисные здания, жилые здания, гостиничный бизнес, образовательные учреждения, розничную торговлю и продуктовые магазины. , водоочистка и сельское хозяйство. Эти разнообразные применения подчеркивают универсальность и эффективность дезинфицирующих и стерилизующих ламп в обеспечении здоровья и безопасности в различных средах [1–4]. Традиционным источником света для стерилизации и дезинфекции является ультрафиолетовая (УФ) ртутная лампа, которая имеет преимущества низкой цены и высокой световой мощности и обычно используется в больших помещениях, таких

Machine Translated by Google

а именно, появились карбид кремния (SiC) и нитрид галлия (GaN), которые характеризуются высокой энергоэффективностью и низким энергопотреблением. Кроме того, силовые устройства, изготовленные из широкозонных полупроводниковых материалов, могут работать при более высоких напряжениях, в генераторных ресторанах и школах. Однако недостатки ртутных УФ-ламп очевидны.
вырабатывают больше мощности и имеют меньшие потери энергии, при этом время прогрева значительно меньше , короткий срок службы ламп и то, что содержание ртути в лампах не соответствует требованиям
силовые устройства из кремниевых полупроводниковых материалов. Карбид кремния (SiC) соответствует экологическим требованиям. УФ-излучение невидимо для человеческого глаза и имеет длину волны
фунт полупроводникового материала, состоящего из кремния (Si) и углерода (C). SiC имеет толщину от 100 до 400 нанометров.
В зависимости от длины волны УФ можно разделить на три
сопротивление дрейфового слоя, чем Si, не требует модуляции проводимости, имеет энергетические типы: длинноволновое
УФ-А (длины волн 320–400 нм), средневолновое УФ-В (длины волн
зазор примерно в 3 раза шире, чем у Si, и имеет напряженность поля, разрушающего изоляцию, 280–320 нм) и коротковолновое
УФ-С (длины волн 100–280 нм). Коротковолновый УФ-С,
примерно в 10 раз выше, чем у Si. По сравнению с высоковольтными диодами Si-PN, которые также можно назвать глубоким
ультрафиолетом, обладает дезинфицирующими и бактерицидными свойствами.
высокоскоростные диоды Si-Fast Recovery, диоды Шоттки SiC, изготовленные из широкозонного полупроводника.
Использование коротковолнового ультрафиолета UV-C или глубокого ультрафиолета разрушает ДНК или
Торовые материалы обладают отличными скоростными и высоковольтными характеристиками. Поскольку обратная
молекулярная структура РНК микробных клеток предотвращает тем самым рост или регенерацию клеток,
Время восстановления SiC-диодов Шоттки чрезвычайно короткое, энергия теряется при реверсе и в конечном итоге достигается
эффект дезинфекции и стерилизации. Глубокие УФ-светодиоды
восстановление может быть значительно уменьшено, что подходит для применений с высокой частотой переключения,
характеризующихся защитой окружающей среды и энергосбережением, длительным сроком службы, отсутствием времени прогрева,
частота и способствует миниатюризации всей схемы, а также отсутствию ртути и миниатюризации источника света.
Традиционные УФ-ртутные лампы имеют
доказательство плотности мощности [21,22]. Дополнительные примечания о долгосрочной надежности и постепенно
заменяются светодиодами с глубоким УФ-излучением, которые стали новым источником света для
потенциальная деградация карбидокремниевых диодов Шоттки при термической и электрической очистке помещений и
стерилизации поверхностей, таких как дома, школы, больницы, офисы
напряжения приведены ниже. По данным литературы [21], карбидокремниевые диоды Шоттки используются в зданиях
супермаркетов или универмагов. Кроме того, глубокая ультрафиолетовая дезинфекция.
не имеют инъекции неосновных носителей, и время их восстановления не увеличивается при температуре и стерилизации.
Лампы могут легко убивать бактерии, но из-за низкой проникающей способности глубоких
по сравнению с обычными кремниевыми диодами. Кроме того, SiC-диоды Шоттки являются лампами для ультрафиолетовой
дезинфекции и стерилизации, на них могут воздействовать частицы пыли, если
Большинство несущих устройств и, следовательно, обратный ток меньше, а температура в ячейках, поврежденных глубокой
ультрафиолетовой дезинфекцией и стерилизационными лампами, не полностью
зависимый. Кроме того, по сравнению с кремниевыми диодами, SiC-диоды Шоттки обладают более высокой стойкостью к
разрушению, элементы все еще можно восстановить и оживить. Поэтому важно
Преимущество более быстрого восстановления, что снижает потери при переключении и шум, а также низкое вперед
обеспечивают соответствующее расстояние и достаточное время и интенсивность облучения глубокого
напряжения, что снижает потери проводимости. Кроме того, в литературе [22] упоминается использование ультрафиолетовой
лампы для дезинфекции и стерилизации дезинфицированных предметов с целью достижения
хороший стерилизующий эффект [5–8].

долговременная надежность SiC-диодов Шоттки в отношении отказа dV/dt и отказа dI/dt.

Кроме того, в кремниевых диодах, если dV/dt велик, наблюдается увеличение тока восстановления. На
(Irr), что может привести к повреждению устройства из-за перенапряжения тока. Напротив, для поставки светодиодных ламп для
стерилизации и дезинфекции глубоким ультрафиолетом. Первый этап
PFC с коррекцией коэффициента мощности (PFC), а второй каскад предлагает собой
закономерность вряд ли возникнет.
[8]. Двухкаскадный электронный драйвер освещения, который преобразует переменный ток в постоянный и подает необходимую мощность постоянного тока для стерилизации светодиодами глубокого УФ-излучения.
Для решения вышеуказанных проблем был разработан новый электронный драйвер освещения для питания глубокого
В данной модели предложена разработка LED-драйвера для светодиодных ламп для передней и стерилизации. Для
ряд, который состоит из двух каскадов, как показано на рисунке 1. Основная схема и задние ступени,
компоненты. Чтобы уменьшить количество силовых переключателей и компонентов схемы.
Кроме того, предлагается управление током. Добавленный драйвер светодиодных ламп для
В литературе [9–20] разработан ряд одноступенчатых драйверов светодиодов с одним переключателем.
широкозонный диод Шоттки из карбида кремния в качестве выходного диода для улучшения схемы питания светодиодных
эффективность.

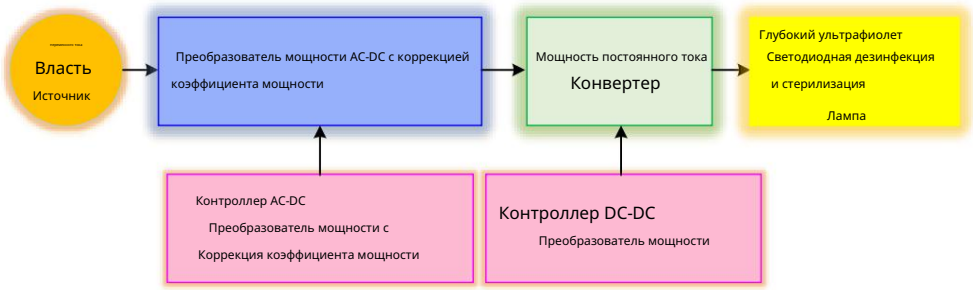


Рис. 1. Структурная схема существующего двухкаскадного электронного драйвера освещения для питания глубокого ультрафиолетового светодиодного лампа для дезинфекции и стерилизации [8].

Данная статья организована и представлена следующим образом. Раздел 2 описывает и анализирует, что в последние
режим работы одноступенчатой электронной схемы драйвера освещения, в которой используется сокращение выбросов
Диод Шоттки для питания светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом. Таким образом,
использование ископаемого топлива должно постепенно сокращаться и быстро внедряться в экологически чистую энергетику.
содержит рекомендации по проектированию некоторых параметров схемы. В разделе 3 показаны экспериментальные и
энергосберегающие приложения и, следовательно, высокая энергоэффективность и низкое энергопотребление.
Результаты прототипа схемы электронного драйвера освещения, использующей SiC-диод Шоттки для потребления, также

Поскольку экологические проблемы, вызванные глобальным потеплением и выбросами углекислого газа, стали
годы становится все более серьезным, люди занялись энергосбережением,
углерода SiC, и любовь к земле как общее основное направление развития.
постепенно принимаются в качестве цели в продуктах повседневной жизни. В Париже

В Соглашении Конференции ООН по изменению климата провозглашается, что

Machine Translated by Google

темпы глобального потепления должны удерживаться в пределах 2 С. Учитывая текущую тенденцию экономического развития, даже если потепление удержится в пределах 2 С в 2050 году, выбросы CO2 все равно увеличатся на 21%, а для удовлетворения различных потребностей человека и обеспечения различных потребностей человека потребуется производить до 50% больше электроэнергии. деятельность. Поэтому все более актуальной становится существенная модернизация и улучшение существующих источников энергии. Крупнейшим сырьем для полупроводников является производство кремниевых (Si) пластин первого поколения. Однако физические свойства этого материала достигли предела в существующих изделиях на основе Si и повысить мощность, снизить теплопотери и повысить скорость уже невозможно. В результате существует необходимость в разработке других материалов, которые смогут лучше использовать эффективность переноса электронов и низкое энергопотребление, и именно в этом контексте широкозонные полупроводники третьего поколения, которые в настоящее время горячо обсуждаются на рынке , а именно карбид кремния (SiC) и нитрид галлия (GaN), которые характеризуются высокой энергоэффективностью и низким энергопотреблением. Кроме того, силовые устройства из широкозонных полупроводниковых материалов могут работать при более высоких напряжениях, генерировать большую мощность и иметь меньшие потери энергии, при этом они существенно меньше силовых устройств из кремниевых полупроводниковых материалов. Карбид кремния (SiC) представляет собой сложный полупроводниковый материал, состоящий из кремния (Si) и углерода (C). SiC имеет меньшее сопротивление дрейфового слоя , чем Si, не требует модуляции проводимости, имеет энергетическую щель примерно в 3 раза шире , чем у Si, и имеет напряженность поля, разрушающего изоляцию, примерно в 10 раз выше, чем у Si. По сравнению с высоковольтными диодами Si-PN и быстродействующими диодами Si-Fast Recovery, SiC-диоды Шоттки, изготовленные из широкозонных полупроводниковых материалов, обладают превосходными быстродействующими и высоковольтными характеристиками. Поскольку время обратного восстановления SiC- диодов Шоттки чрезвычайно мало, потери энергии во время обратного восстановления могут быть значительно уменьшены, что подходит для приложений с высокой частотой переключения и способствует миниатюризации во Дополнительные примечания о долгосрочной надежности и потенциальной деградации карбидокремниевых диодов Шоттки под воздействием термических и электрических напряжений приведены ниже. По данным литературы [21], карбидокремниевые диоды Шоттки не имеют инжекции неосновных носителей заряда и время их восстановления не увеличивается с температурой по сравнению с обычными кремниевыми диодами. Кроме того, SiC-диоды Шоттки являются устройствами с основной несущей, поэтому обратный ток меньше и не зависит от температуры. Кроме того, по сравнению с кремниевыми диодами SiC-диоды Шоттки имеют преимущество более быстрого восстановления, что снижает потери переключения и шум, а также более низкое прямое напряжение, что снижает потери проводимости.

Кроме того, в литературе [22] упоминается долговременная надежность SiC-диодов Шоттки в отношении отказов dV/dt и dI/dt. Кроме того, в кремниевых диодах, если dI/dt велико, возникает картина повышенного тока восстановления (Irr), который может повредить устройство из-за концентрации тока. Напротив, поскольку ток восстановления в SiC- диодах Шоттки очень мал, можно предположить, что возникновение такой закономерности маловероятно.

Для решения вышеуказанных проблем в данной статье предложен и разработан новый электронный драйвер освещения для светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом, который является расширением и дальнейшим усовершенствованием литературы [20]. Основная схема объединяет понижающий преобразователь и обратноходовой преобразователь в одноступенчатый понижающе -обратноходовой преобразователь переменного тока в постоянный с одним переключателем и формированием входного тока. Кроме того, предлагаемый драйвер повторно использует энергию, накопленную в индуктивности рассеяния трансформатора, и использует диод Шоттки из карбида кремния с широкой запрещенной зоной в качестве выходного диода для повышения эффективности схемы.

Данная статья организована и представлена следующим образом. В разделе 2 описан и проанализирован режим работы однокаскадной электронной схемы драйвера освещения, использующей SiC -диод Шоттки для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации, а также приведены рекомендации по проектированию некоторых параметров схемы. В разделе 3 показаны экспериментальные результаты прототипа схемы электронного драйвера освещения, в которой используется SiC-диод Шоттки для питания светодиодной лампы для стерилизации и дезинфекции глубоким ультрафиолетом. Наконец, некоторые выводы приведены в разделе 4.

Machine Translated by Google

поставьте светодиодную лампу для стерилизации и дезинфекции глубоким ультрафиолетом. Наконец, некоторые выводы приведены в разделе 4.

2. Описания и анализ режимов работы предлагаемой одноступенчатой

Схема электронного драйвера освещения, использующая SiC-диоды Шоттки для питания Схема электронного драйвера освещения, использующая SiC-диоды Шоттки для питания глубокого Светодиодная лампа для дезинфекции и стерилизации глубоким ультрафиолетом Ультрафиолетовая светодиодная лампа для дезинфекции и стерилизации

Электронный драйвер освещения с PFC для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации. Электронный драйвер освещения с PFC для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации глубоким ультрафиолетом. В этой статье показано, разработанный в данной статье, показан на рисунке 2, для которого разработана лампа, на рисунке 2, который объединяет преобразует понижающий преобразователь и обратноходовой преобразователь в одноступенчатый преобразователь мощности переменного тока в постоянный ток, а обратноходовой преобразователь в одноступенчатый преобразователь мощности переменного тока в постоянный ток и включает в себя фильтрующий дроссель LF, фильтрующий конденсатор CF, полномостовой выпрямитель (включая D1, D2, D3) включает в себя фильтрующий дроссель LF, фильтрующий конденсатор CF, полномостовой выпрямитель (включая D1, D2, D3) Д1, Д2, Д3 и Д4), силовой ключ QB, два диода ДФ и ДБ, трансформатор TR с намагничиванием Д1, Д2, Д3 и Д4), силовой ключ QB, два диода ДФ и ДБ, трансформатор TR с намагничиванием CO2, дроссель LM и дроссель рассеяния Llk, два выходных конденсатора CO1 и CO2 и глубокий CO2, дроссель LM и дроссель рассеяния Llk, два выходных конденсатора CO1 и CO2 и глубокий ультрафиолетовый светодиодный лампа для дезинфекции и стерилизации. Способ. Разработанная индуктор намагничивания, LM в предлагаемой схеме расширения драйвера освещения для работы в режиме отключения - LM ультрафиолетовая светодиодная лампа для дезинфекции и стерилизации глубоким ультрафиолетом. Разработанная режим непрерывной проводимости (DCM). PFC можно быть реализован простого напряжения, режим трансформатора TR в предлагаемой схеме электронного драйвера освещения для режима прерывистой эффективность схемы может быть восстановлена, а эффективность схемы может быть улучшена за счет использования Диод Шоттки SiC вместо выходного диода.

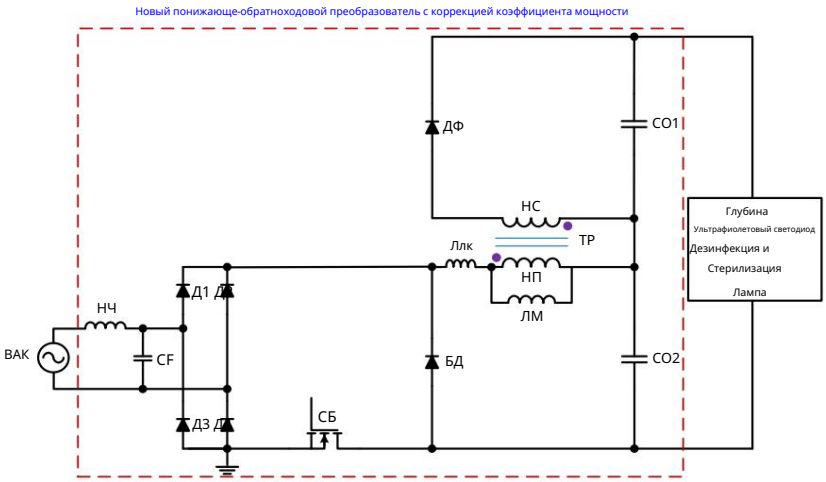


Рисунок 2. Предлагаемый одноступенчатый электронный драйвер освещения с использованием диодов Шоттки из SiC

использование светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом. поставка светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом.

На рисунке 3 представлена эквивалентная схема одноступенчатого электронного драйвера освещения. На рисунке 3 представлена эквивалентная схема одноступенчатого электронного драйвера освещения, использованием SiC диодов Шоттки для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и использованием SiC диодов Шоттки для питания светодиодной лампы глубокого ультрафиолета для дезинфекции стерилизации, разработанной в этой статье, с учетом коэффициента мощности. Теоретический и стерилизации, разработанной в этой статье, с учетом коэффициента мощности. Разработанная в данной Формулы предлагаемого одноступенчатого электронного драйвера освещения, использующего SiC-диоды Шоттки Формулы предлагаемого одноступенчатого электронного драйвера освещения, использующего SiC-диоды Шоттки Показаны диоды Шоттки для светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом. показаны на рисунке 4. Три предлагаемые схемы одноступенчатого электронного драйвера освещения для светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом. Во время Схема драйвера светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации. Во время описания драйвера режима работы в представленных описаниях и пояснениях режима работы в Схема драйвера показаны следующим образом:

(a) Эквивалентный источник напряжения входного сетевого напряжения после прохождения через (a) Эквивалентный источник напряжения входного сетевого напряжения после прохождения через двухполупериодную схему выпрямителя обозначается VREC. Поскольку частота переключения fs равна схема двухполупериодного выпрямителя обозначается VREC. Поскольку частота переключения fs равна намного превышающей частоту сети fAC, выпрямленное напряжение VREC можно рассматривать как намного превышает частоту сети fAC, выпрямленное напряжение VREC можно рассматривать как постоянное значение в течение одного цикла переключения при анализе режима схемы. величину в течение одного цикла переключения при анализе режима схемы.

(b) Учитывается индуктивность рассеяния Llk трансформатора TR. (b) Учитывается индуктивность рассеяния Llk трансформатора TR.

(c) Индуктор намагничивания LM трансформатора TR предназначен для работы в разрывном режиме (c) Индуктор намагничивания LM трансформатора TR предназначен для работы в режиме прерывистой проводимости (DCM). режим непрерывной проводимости (DCM).

(г) Пренебрежение падением напряжения проводимости и их эквивалентным сопротивлением всех диодов.

(e) Остальные компоненты схемы считаются идеальными.

Machine Translated by Google

(г) Пренебрежение падениями напряжения проводимости и их эквивалентными сопротивлениями всех диодов.

(e) Остальные компоненты схемы считаются идеальными.

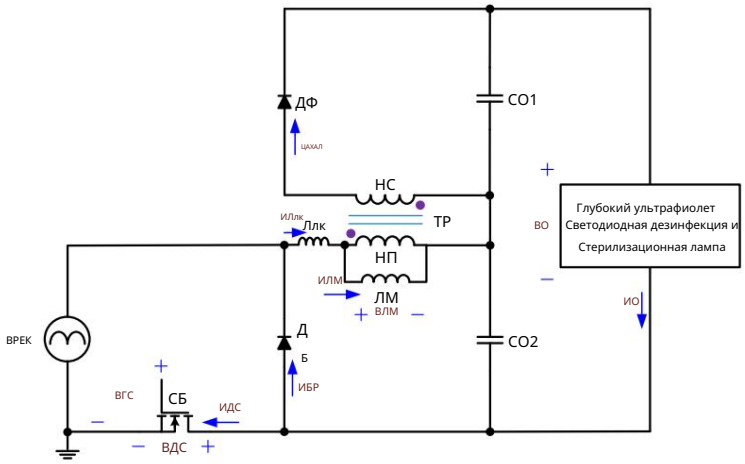


Рисунок 3. Эквивалентная схема предлагаемой схемы электронного драйвера освещения на основе SiC Шоттки

Рисунок 3. Эквивалентная схема предлагаемой схемы электронного драйвера освещения с использованием SiC-диодов Шоттки для питания светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом при анализе диодов для питания светодиода

режимы работы.

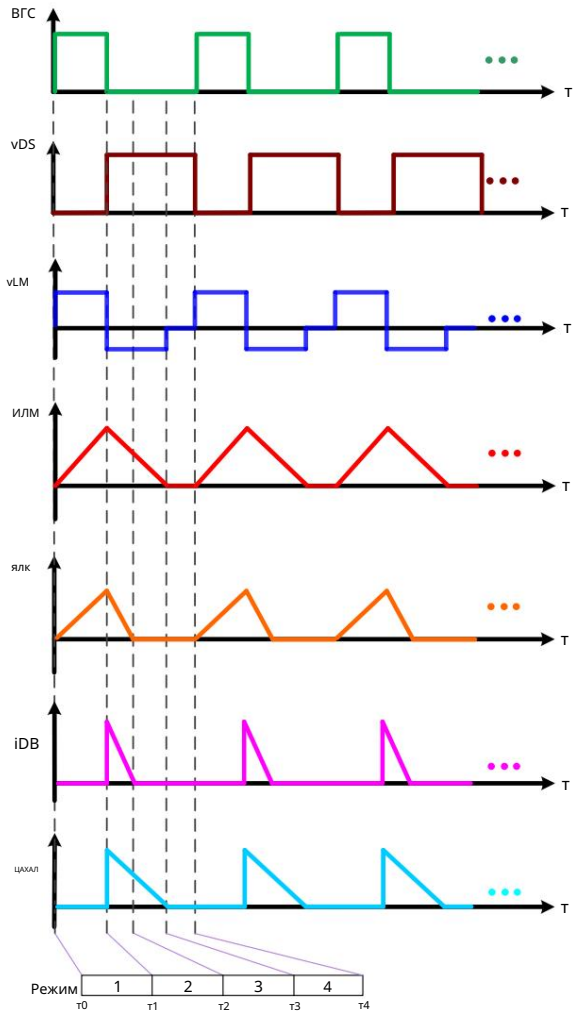


Рисунок 4. Теоретические формы сигналов предлагаемого одноканального электронного драйвера освещения, использующего SiC.

Рисунок 4. Теоретические формы сигналов предлагаемого одноступенчатого электронного драйвера освещения, использующего Schottky Diodes для питания светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом.

Режим работы 1 (t0 t < t1): На рисунке 5 показана эквивалентная схема предлагаемого одноступенчатая электронная схема драйвера освещения для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации в режиме 1. Выключатель питания SB включается, и

Machine Translated by Google

Режим работы 1 (t0 t < t1): На рисунке 5 показана эквивалентная схема предлагаемого одноступенчатая электронная схема драйвера освещения для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации в режиме 1. Выключатель питания SB включается, а к обоим концам диода DB подключается источник эквивалентного напряжения VREC, в результате чего диод DB в DB — неправильное состояние смещения. Источник эквивалентного напряжения VREC обеспечивает энергией выходной конденсатор CO2 и намагничивающую катушку LM трансформатора TR. а также утечка дросселя через ключ SB, что приводит к линейному увеличению тока намагничивающего тока индуктора намагничивания LM и полярность напряжения с обеих сторон индуктора намагничивания LM и индуктора рассеяния Llk положительна слева и отрицательна справа. ультрафиолетом. Выходные конденсаторы CO1 и CO2 обеспечивают энергией светодиодную лампу для и лампа для стерилизации. При выключении ключа SB ток намагничивающего индуктора LM достигает дроссель LM достигает максимального значения и режим завершается.

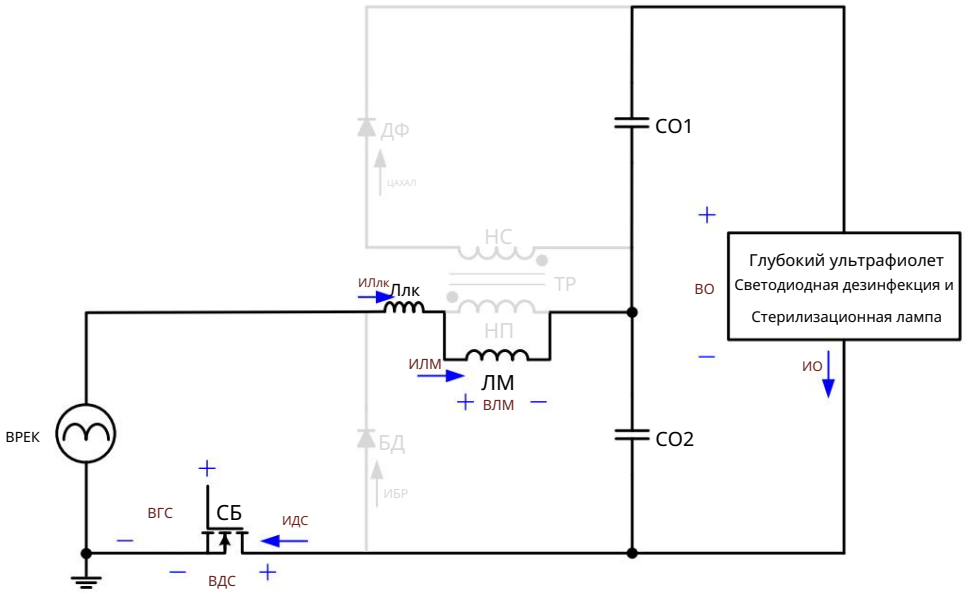


Рис. 5. Эквивалентная схема предлагаемой однокаскадной схемы электронного драйвера освещения с использованием SiC диодов Шоттки для дезинфекции светодиодной лампой глубокого ультрафиолета и стерилизации во время Режим 1.

Режим работы 2 (t1 t < t2): На рисунке 6 показана эквивалентная схема предлагаемого одноступенчатый электронный драйвер освещения для питания ламп глубокого ультрафиолета для обеззараживания и стерилизации светодиодов в режиме 2. При выключении силового выключателя SB напряжение полярность напряжения намагничивающего индуктора LM и дросселя рассеяния Llk изменяется на правый положительный и левый отрицательный в соответствии с законом Ленца, который переводит диод DB в состояние прямого смещения. Ток намагничивающего индуктора LM течет в положение, в положении, когда на первичной обмотке NP трансформатора TR нет точки, полярность напряжения двух концов обмотки NP право-положительная и лево-отрицательная. отрицательный. Следовательно, полярность напряжения, отраженного на оба конца вторичной обмотки, положительна слева и отрицательна справа, что заставляет диод DF работать в прямосмещенном режиме. Боковая обмотка NS положительна слева и отрицательна справа, что заставляет диод DF работать в проводящем состоянии. В это время дроссель намагничивания LM и дроссель рассеяния Llk С возрастом индуктор Llk находится в состоянии выделения энергии, в результате чего ток намагничивания дроссель LM убывает линейно. Так как диод DB находится в прямосмещенной проводимости намагничивающего состоянии, энергия намагничивающего индуктора LM и дросселя рассеяния Llk подается в конденсатор CO2 через диод DB, а ток намагничивающего индуктора LM показывает прямосмещенной проводимости диода DF энергия. Дюктор LM показывает линейное уменьшение. Из-за прямосмещенной проводимости диода DF питание намагничивающего дросселя LM через трансформатор TR и диод DF подается на энергия намагничивающего индуктора LM подается через трансформатор TR и выходные емкости CO1 и CO2 , диод DF к выходным емкостям CO1 и CO2 , а также к светодиоду глубокого ультрафиолета и стерилизационной лампе. Кроме того, выходные конденсаторы CO1 и CO2 продолжают обеспечивать лампа для дезинфекции и стерилизации. Кроме того, выходные конденсаторы CO1 и CO2 передают энергию светодиодной лампе для дезинфекции и стерилизации глубоким ультрафиолетом. Режим завершается когда индуктор рассеяния Llk заканчивает выделять энергию и ток утечки ILlk дроссель Llk линейно падает до нуля.

Machine Translated by Google

для обеспечения энергией светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом. Режим заканчивается, когда индуктор рассеяния Llk заканчивает выделять энергию, а ток Ilk индуктора рассеяния Llk линейно падает до нуля.

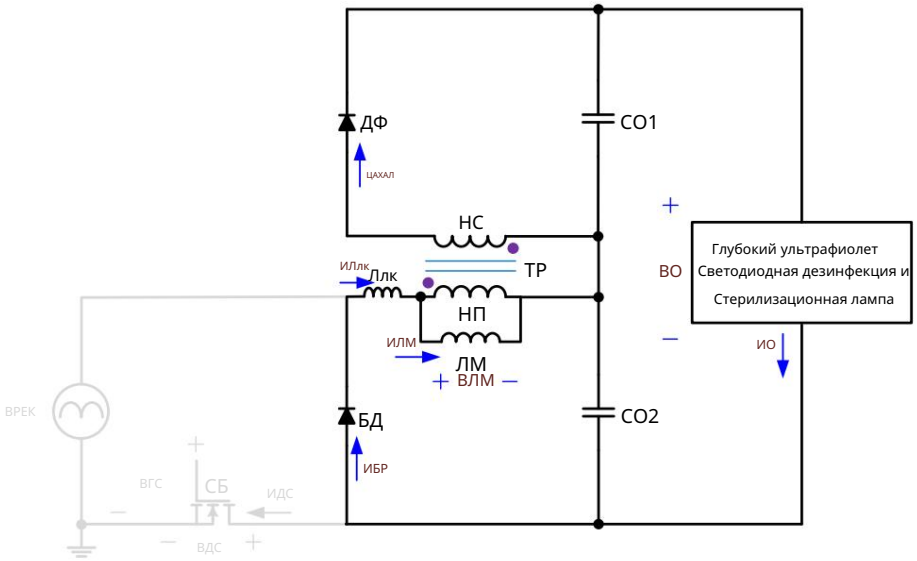


Рисунок 6. Эквивалентная схема предлагаемого однокаскадного электронного драйвера освещения с использованием SiC диодов Шоттки для питания светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом во время Режима 2.

Режим работы 3 (t2 < t3): на рисунке 7 показана эквивалентная схема предлагаемого однокаскадного электронного драйвера освещения для питания светодиодной лампы для дезинфекции и стерилизации в режиме 3. Энергия намагничивающего индуктора LM подается непрерывно через трансформатор TR и диод DF на выходные емкостные конденсаторы CO1 и CO2 и на светодиодную лампу для дезинфекции и стерилизации глубоким ультрафиолетом. Кроме того, выходные конденсаторы CO1 и CO2 продолжают подавать энергию для светодиодной лампы для глубокой ультрафиолетовой дезинфекции и стерилизации. Этот режим заканчивается, когда намагничивающий индуктор LM перестает выделять энергию, и ток ILM намагничивающей индуктивности линейно падает до нуля.

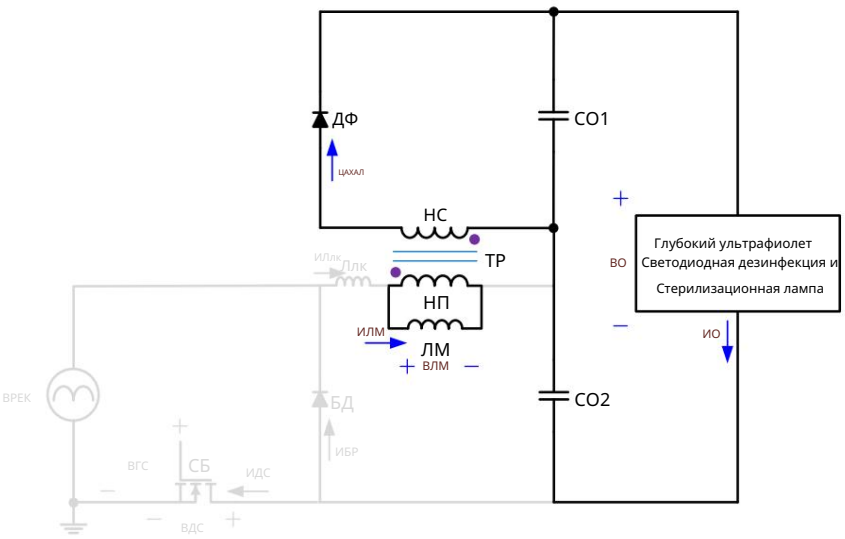


Рис. 7. Эквивалентная схема предлагаемого однокаскадного электронного драйвера освещения с использованием Диоды Шоттки SiC для подачи светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом во время Режима 3.

Режим работы 4 (t3 t < t4): на рисунке 8 показана эквивалентная схема предлагаемой однокаскадной схемы электронного драйвера освещения для питания светодиода глубокого ультрафиолета. лампа для дезинфекции и стерилизации в режиме 4. Выходные конденсаторы CO1 и CO2.

Machine Translated by Google

Режим работы 4 (t3 t < t4): На рисунке 8 показана эквивалентная схема предлагаемого одноступенчатая электронная схема драйвера освещения для питания светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом в режиме 4. Выходные конденсаторы CO1 и CO2 непрерывно обеспечивают энергию для светодиодной дезинфекции и стерилизации глубоким ультрафиолетом. Обеспечивает энергией светодиодную лампу для дезинфекции и стерилизации глубоким ультрафиолетом. лампа. Когда выключатель питания SB снова включается, этот режим заканчивается, и схема возвращается в исходное состояние. Когда выключатель питания SB снова включается, этот режим заканчивается, и схема возвращается к первому режиму работы. первый режим работы.

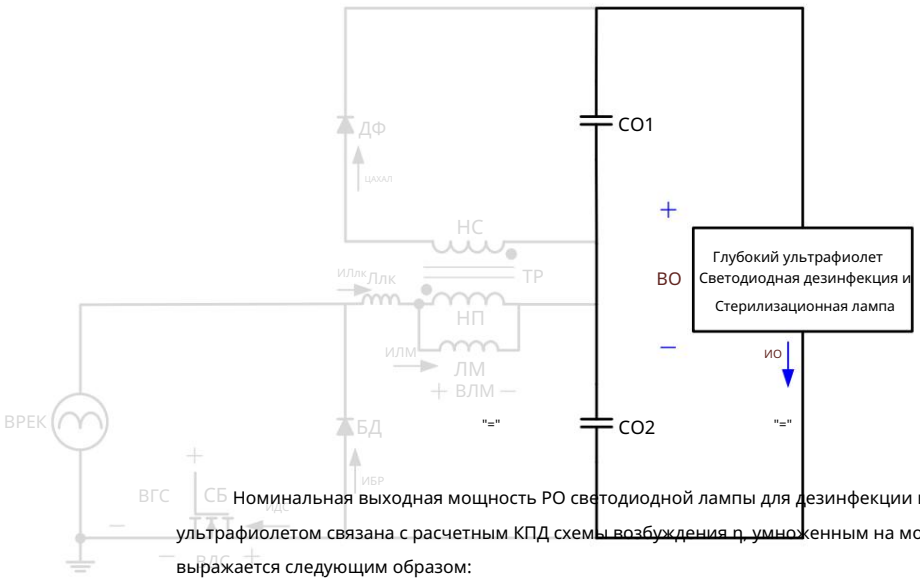


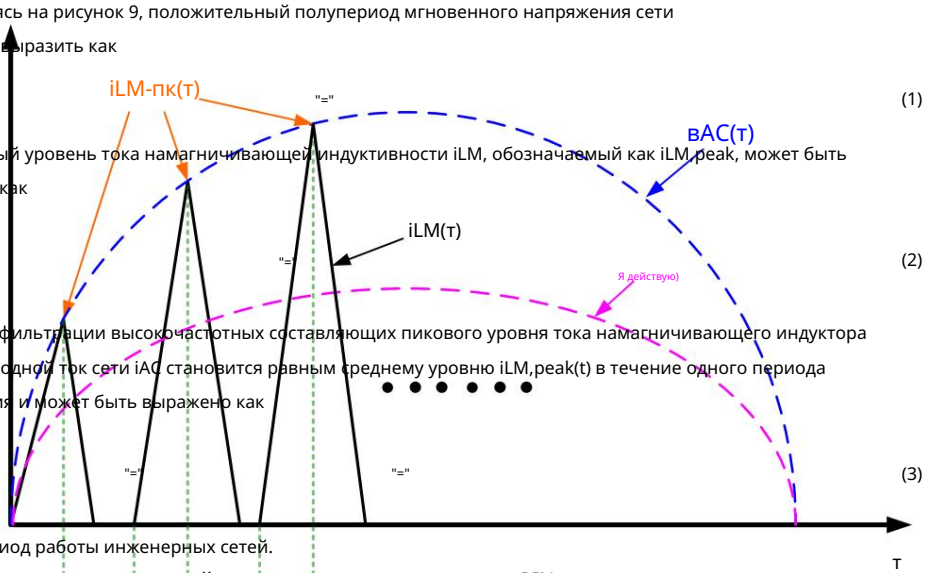
Рис. 8. Эквивалентная схема предлагаемой однокаскадной схемы электронного драйвера освещения с использованием карбидных диодов Шоттки для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации в режиме 4.

Рис. 8. Эквивалентная схема предлагаемой однокаскадной схемы электронного драйвера освещения с использованием SiC-диодов Шоттки для питания светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации в режиме 4.

2.1. Рекомендации по проектированию намагничивающего индуктора LM в трансформаторе TR

Чтобы показать рекомендации по проектированию намагничивающего индуктора LM в трансформаторе TR , на рисунке 9 показаны теоретические формы сигналов тока намагничивания LM. бывший TR, на рисунке 9 показаны теоретические формы тока намагничивающей индуктивности iLM(t), пикового уровня iLM-pk(t) и входного тока сети iAC(t) в положительном полупериоде полупериода напряжения сети vAC(t). напряжения сети электропитания vAC(t).

Ссылаясь на рисунок 9, положительный полупериод мгновенного напряжения сети vAC(t) можно выразить как Текущий



(1)

Пиковый уровень тока намагничивающей индуктивности iLM, обозначаемый как iLM,peak, может быть представлен как

(2)

Путем фильтрации высокочастотных составляющих пикового уровня тока намагничивающего индуктора iLM,peak(t) входной ток сети iAC становится равным среднему уровню iLM,peak(t) в течение одного периода переключения и может быть выражено как

(3)

где TAC – период работы инженерных сетей.

Среднее значение входной мощности линии электросети PIN получается путем умножения мгновенного значения напряжения линии электросети vAC(t) на мгновенное значение тока сети электросети iAC(t), а затем усреднения его за период в один цикл. , который рассчитывается следующим образом

Рис. 9. Наглядные формы тока намагничивающего индуктора iLM(t), пикового уровня iLM-pk(t) и входного тока сети электропитания iAC(t) в положительном полупериоде напряжения сети электропитания vAC(t).

Рис.9. Наглядная форма тока намагничивающей индуктивности iLM(t), пикового уровня iLM-pk(t) сетевого тока iAC(t) в положительном полупериоде сетевого напряжения vAC

2.2. Руководство по проектированию выходных конденсаторов CO1 и CO2

Выбор выходных конденсаторов CO1 и CO2 зависит от выходного напряжения, допустимого перенапряжения, выходной мощности и желаемого напряжения. Пульсации напряжения VO составляют половину размаха выходного напряжения и частоты сети и являются функцией емкости CO и пиковой емкости конденсатора.

Machine Translated by Google

Ссылаясь на рисунок 9, положительный полупериод мгновенного напряжения сети vAC(t) можно выразить как

$$|vAC(t)| = \sqrt{2}vAC_{rms}|\sin(2\pi fACt)| \quad (1)$$

Пиковый уровень тока намагничивающей индуктивности iLM, обозначаемый как iLM,peak, может быть представлен выражением

$$iLM,peak(t) = \frac{\sqrt{2}vAC\text{-}rms|\sin(2\pi fACt)|DutyTS}{2LM} \quad (2)$$

Путем фильтрации высокочастотных составляющих пикового уровня тока намагничивания iLM,peak(t) входной ток сети iAC равен среднему уровню iLM,peak(t) в течение одного периода переключения и может быть выражается как

$$iAC(t) = \frac{1}{TAC}\int_0^{T} rmsDuty2TS\, iLM,peak(t)dt = \sin(2\pi fACt)\frac{2\ В\ переменного\ тока\text{-}}{4LM} \quad (3)$$

где TAC – период работы инженерных сетей.

Среднее значение входной мощности линии электропередачи PIN получается путем умножения мгновенного значения напряжения линии электросети vAC(t) на мгновенное значение тока сети электросети iAC(t), а затем усреднения его за период в один цикл. , который рассчитывается следующим образом

$$ПИН = \frac{1}{TAC}\int_0^{T} vAC(t)iAC(t)dt = \frac{vAC\text{-}rms\ 2Duty2TS}{4LM} \quad (4)$$

Номинальная выходная мощность PO светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации связана с расчетным КПД схемы возбуждения η , умноженным на входную мощность PIN, которая выражается следующим образом:

$$ПО = \eta ПИН = \frac{\eta vAC\text{-}rms\ 2Duty2TS}{4LM} \quad (5)$$

Переставив (5), получим расчетную формулу индуктивности намагничивающего индуктора LM определяется

$$LM = \frac{\eta vAC\text{-}rms\ 2Duty2TS}{4PO} \quad (6)$$

2.2. Руководство по проектированию выходных конденсаторов CO1 и CO2

Выбор выходных конденсаторов CO1 и CO2 зависит от выходного напряжения постоянного тока, допустимого перенапряжения, выходной мощности и желаемой пульсации напряжения. Пульсации напряжения VO составляют половину размаха выходного напряжения при удвоенной частоте сети и являются функцией емкости CO и пикового тока конденсатора (который равен выходному току IO) и может быть выражено в следующем уравнении [23].

$$VO = IO\sqrt{\left(\frac{1}{2\pi \times 2\ fAC \times CO}\right)^2 + RECP2} \quad (7)$$

Пренебрегая эквивалентным последовательным сопротивлением, которое обозначается как RESR, на выходе конденсатора, емкость CO выходного конденсатора можно выразить как

$$CO \frac{IO}{4\pi\ fAC \times VO} \text{"="} \frac{}{PO\ 4\pi\ fAC \times VO \times VO} \quad (8)$$

Кроме того, пульсации напряжения VO обычно выбирают в пределах от 1% до 5% выходного напряжения.

Machine Translated by Google

3. Экспериментальные результаты предлагаемой схемы электронного драйвера освещения на основе SiC.

Schottky Didoes за поставку оборудования для глубокой ультрафиолетовой светодиодной дезинфекции и
Стерилизационная лампа

На рисунке 10 показана фотография процесса дезинфекции и стерилизации с помощью светодиодов глубокого ультрафиолета.
модуль, используемый в этой статье, со спецификациями, которые включают длину волны 275 нм и
Угол заточки 120 градусов. Более того, производитель подключил постоянный ток.
управляющая ИС последовательно и диод Шоттки параллельно для защиты от обратной полярности на
алюминиевая подложка в каждом модуле светодиодной дезинфекции и стерилизации глубоким ультрафиолетом.
Спецификация используемой светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом
в этой статье состоит из десяти светодиодных модулей глубокой ультрафиолетовой дезинфекции и стерилизации.
соединены последовательно с номинальным напряжением 90 вольт, номинальным током 40 миллиампер, и
номинальная мощность 3,6 Вт. В таблице 1 показано сравнение глубокого ультрафиолета.
Светодиодный модуль дезинфекции и стерилизации в [24] и модуль, использованный в данной статье. Как
Как видно из Таблицы 1, модуль дезинфекции и стерилизации с помощью светодиодов глубокого ультрафиолета.
используемый в этой статье, имеет большую пиковую длину волны, немного узкий угол обзора и
немного большую выходную оптическую мощность и полуширину спектра, чем у модуля [24].
В таблице 2 приведены электрические параметры прототипа одноступенчатого электронного освещения.
Схема драйвера для светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом, разработанная в
Эта бумага. Среднеквадратичное входное напряжение переменного тока составляло 110 В, а характеристики глубокого сверхвысокого напряжения
В качестве выходных параметров использовались светодиодные лампы для дезинфекции и стерилизации, в том числе:
выходная мощность 3,6 Вт, выходное напряжение 90 вольт и выходной ток 40 миллиампер.
Кроме того, вопросы безопасности и обеспечения соответствия нормативным стандартам для УФ-излучения
воздействия и электробезопасности включают в себя: Не смотрите прямо на источник УФ-излучения, когда
использование лампы для глубокой ультрафиолетовой дезинфекции и стерилизации, поскольку глубокое УФ-излучение может вызвать
эритема кожи, раздражение конъюнктивы и усталость. Кроме того, не злоупотребляйте глубоким УФ-излучением.
лампы для дезинфекции и стерилизации при длительном использовании или контакте с УФ-лампами или при прямом
Воздействие УФ-излучения может вызвать повреждение кожи и глаз.

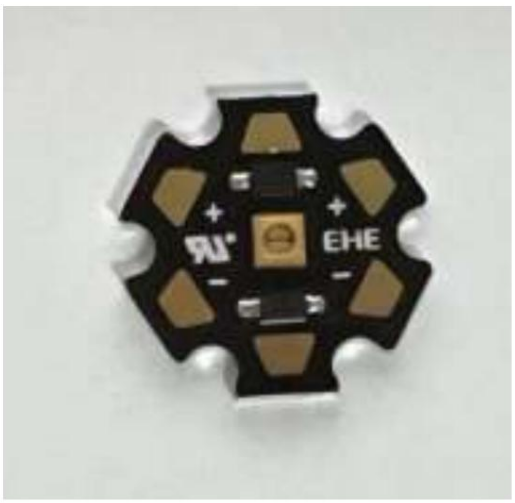


Рисунок 10. Фотография модуля дезинфекции и стерилизации с помощью светодиодов глубокого ультрафиолета, используемого в этом документе. Рисунок 10. Фотография модуля светодиодной дезинфекции и стерилизации глубоким ультрафиолетом, используемого в этом документе.
бумага со спецификациями, включающими длину волны 275 нм и угол при вершине 120 градусов. бумага со спецификациями, включающими длину волны 275 нм и угол при вершине 120 градусов.

3.1. Расчет намагничивающего индуктора LM в трансформаторе TR

3.1. Расчет намагничивающего индуктора LM в трансформаторе TR

Ссылаясь на (6), с η 0,8, Duty 0,25, PO 3,6 Вт, частотой переключения fS
Ссылаясь на (6), с η 0,8, Duty 0,25, PO 3,6 Вт, частотой переключения fS
50 кГц (что является обратной величиной периода переключения TS) и среднеквадратичным напряжением переменного тока
50 кГц (что обратно пропорционально периоду переключения TS) и среднеквадратичное напряжение переменного тока 110 В.
110 В, индуктивность намагничивающего индуктора LM рассчитывается по формуле
индуктивность намагничивающего индуктора LM рассчитывается по формуле

$$LM = \frac{\eta vAC\text{-}rms\ 2Duty2TS}{4PO} = \frac{0,8 \times 1102 \times 0,252 \times \frac{1}{50\,000}}{4 \times 3,6} = 834\ мкГн \quad (9)$$

Чтобы управлять током намагничивающей индуктивности в режиме прерывистой проводимости, чтобы
позволить схеме драйвера иметь функцию формирования входного тока и улучшить коэффициент мощности,
намагничивающая индуктивность LM выбрана равной 800 мкГн для прототипной схемы драйвера. .

Machine Translated by Google

Таблица 1. Сравнение светодиодного модуля глубокой ультрафиолетовой дезинфекции и стерилизации в [24] и тот, который используется в этой статье.

Элемент	Светодиодная дезинфекция глубоким ультрафиолетом и модуль стерилизации в [24]	Светодиодная дезинфекция глубоким ультрафиолетом и модуль стерилизации, используемый в Эта бумага
Типичное значение пиковой длины волны	255 нм	275 нм
Типичное значение угла обзора выходной оптической мощности	3,5 мВт	4 мВт
	125 градусов	120 градусов
Полуширина спектра	11 нм	12 нм

Таблица 2. Технические характеристики предлагаемой схемы электронного драйвера освещения на основе SiC-диодов Шоттки на поставку светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом.

Параметр	Ценить
Входное переменное напряжение В переменного тока	110 В
Номинальная выходная мощность ПО	3,6 Вт
Номинальное выходное напряжение VO	90 В
Номинальный выходной ток IO	40 мА

Чтобы управлять током намагничивающей индуктивности в условиях прерывистой проводимости режим, чтобы позволить схеме драйвера иметь функцию формирования входного тока и для улучшения коэффициента мощности дроссель намагничивания LM выбран равным 800 мкГн. прототип схемы драйвера.

3.2. Расчет выходных конденсаторов CO1 и CO2

Подставив параметры схемы в уравнение (8), при мощности PO 3,6 Вт, fAC 60 Гц, BO 90 В, VO 2,7 В (что составляет 3% от выходного напряжения VO), емкость CO можно получить:

$$CO \quad \frac{ПО}{4\pi\, fAC \times VO \times \quad VO} \text{ "=" } \frac{3.6}{4\pi \times 60 \times 90 \times 2{,}7} = 196{,}49\ \text{мкФ} \qquad (10)$$

При реализации схемы-прототипа значения емкостей выходных конденсаторов CO1 и CO2 были выбраны равными 220 мкФ и выдерживаемое напряжение 250 В.

В таблице 3 показаны характеристики компонентов одноступенчатого электронного освещения. прототип схемы драйвера светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом развито в этой статье. Кроме того, диоды основной несущей с технологией Шоттки на Широкозонный материал из карбида кремния имеет более высокие характеристики и используется в качестве выходного сигнала. диоды ДБ и ДФ. Путем варьирования количества последовательно соединенных обеззараживателей глубоким ультрафиолетом и используемые модули стерилизации, масштабируемая конструкция предлагаемого электронного драйвера освещения. могут быть созданы схемы для работы с приложениями с более высокой мощностью и переменной нагрузкой. условия. В результате выходную мощность PO и выходное напряжение VO можно изменять при этом. выходной ток IO остается постоянным и может использоваться в различных практических приложениях. сценарии. Ключевые параметры предлагаемой схемы привода, такие как намагничивание Индуктивность и выходной конденсатор можно затем перепроектировать.

На рис . 11 показана измеренная форма тока намагничивающего индуктора iLM. На рис . 12 показана измеренная развернутая форма тока намагничивающего индуктора. ИЛМ. Из рисунка 12 видно, что ток намагничивающего индуктора iLM действует в прерывистый режим проводимости. На рисунке 13 показаны измеренные формы выходного сигнала. напряжение VO и выходной ток IO, а также средние значения выходного напряжения и тока составили 91,35 вольт и 38,57 миллиампер соответственно. На рисунке 14 показаны формы пульсаций. выходного напряжения VO и выходного тока IO.

Machine Translated by Google

На рис. 11 показана измеренная форма тока намагничивающего индуктора iLM.

На рисунке 12 показана измеренная развернутая форма тока намагничивающего индуктора. Таблица 3.
Ключевые компоненты, используемые в предлагаемой схеме электронного драйвера освещения на основе SiC Шоттки.
светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации.

ИДМ. Из рисунка 12 видно, что ток намагничивающего индуктора iLM работает в диодах для питания	
прерывистый режим проводимости. На рисунке 13 показаны измеренные формы выходного сигнала. Компонент	
Значение	
напряжение VO и выходной ток IO, а средние значения выходного напряжения и тока составили 91,35 вольт и	
фильтра Cf	1 мГн
38,57 миллиампер соответственно. На рисунке 14 показаны формы пульсаций дросселя фильтра Lf Конденсатор	220 нФ
напряжения VO и **выходного** тока IO.	
Диоды D1 , D2 , D3 , D4 В	МУР460
прототипа схемы	ИРФ730

таблице 4 показаны коэффициент пульсации выходного напряжения и коэффициент пульсации выходного тока
электронного драйвера освещения силового переключателя SB для глубокой ультрафиолетовой светодиодной дезинфекции
и намагничивающего индуктора LM 800 мкГн.
стерилизационная лампа с входным напряжением переменного тока 110 В (среднеквадратичное значение). Как видно из таблицы 4,
делительная индуктивность утечки Llk 12 мкГн
пульсации
размах выходного напряжения 143,8 милливольт при среднем значении 91,35 вольт. Деление 220 мкФ/200 В, 2
напряжения 0,16%; деление размаха выходного тока на 1,1658. Выходной конденсатор CO1 , CO2 220 мкФ/200 В.
миллиампер на среднее значение 38,57 миллиампер дает коэффициент пульсации тока 3,02%.

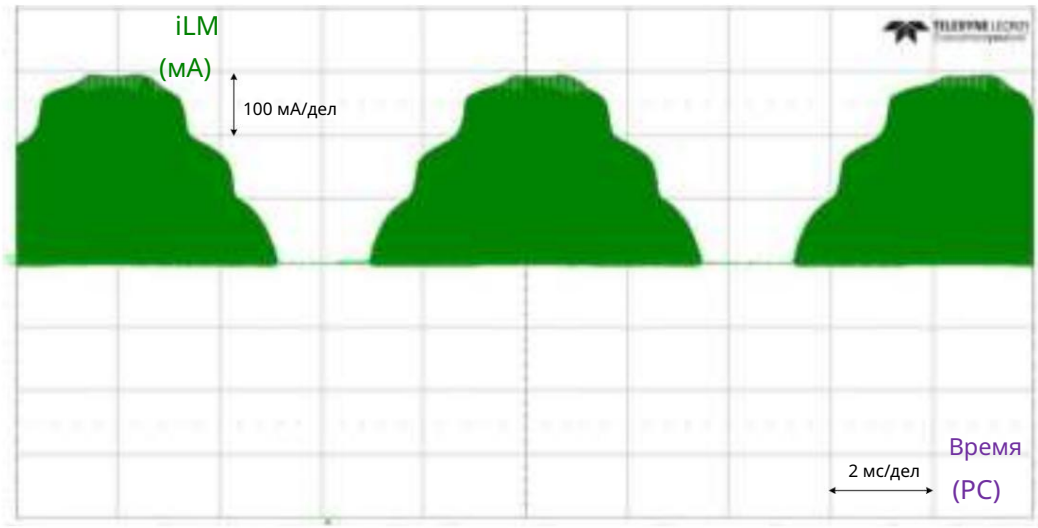


Рисунок 11. Рисунок 11. Измеренный ток индуктивности намагничивания 2 мс/дел. (100 мА/дел), масштаб времени: 2 мс/дел.

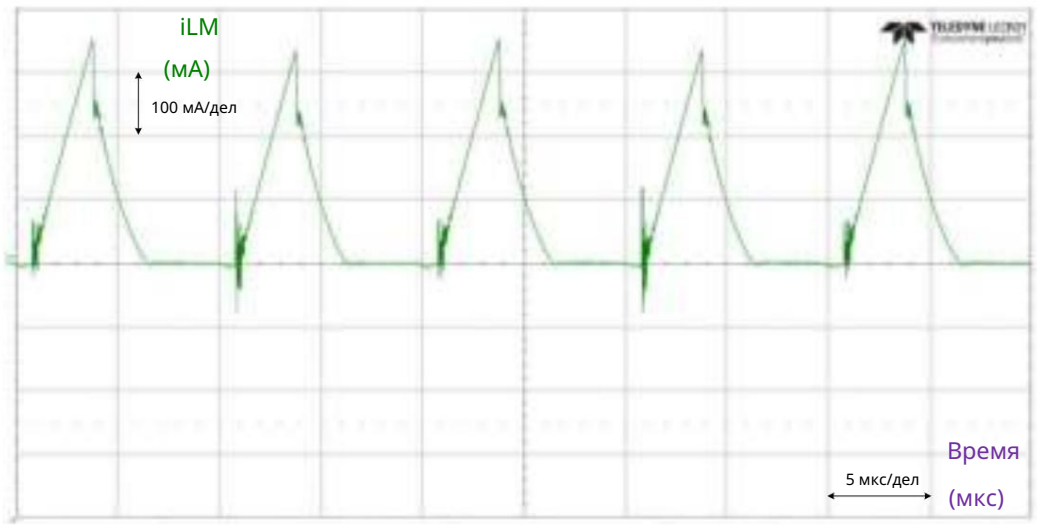


Рис. 12. Измеренная развернутая форма тока намагничивающего индуктора iLM (100 мА/дел); время Рис. 12. Измеренная развернутая
форма тока намагничивающего индуктора iLM (100 мА/дел),
масштаб: 5 мкс/дел.
масштаб времени: 5 мкс/дел.

Machine Translated by Google

Рис. 12. Измеренная развернутая форма тока намагничивающего индуктора iLM (100 мА/дел); время мА/д: 5 мкс/

Рис. 12. Измеренная развернутая форма тока намагничивающего индуктора iLM (шкала 100

дел. Шкала: 5 мкс/дел.

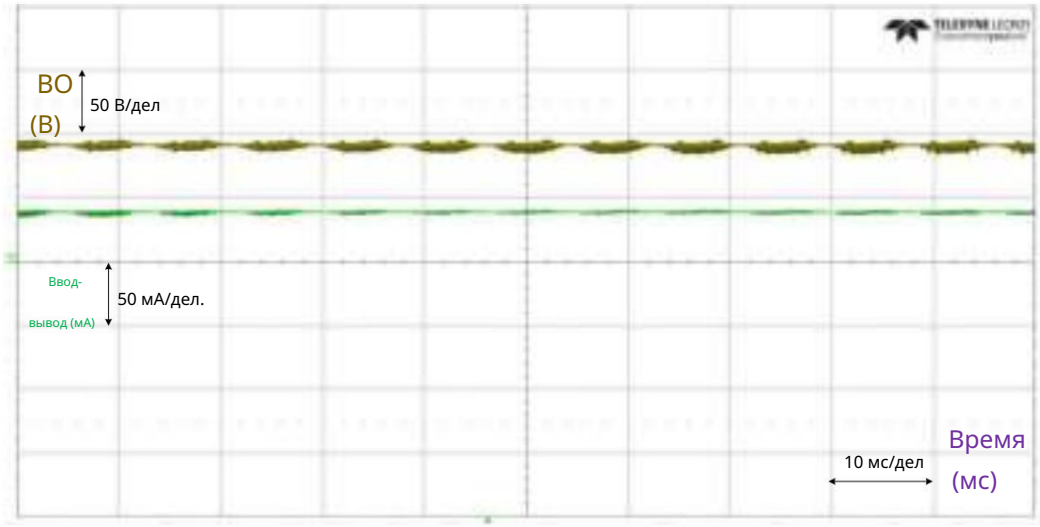


Рисунок 13. Измеренные формы выходного напряжения Vo (50 В/дел) и выходного тока Io (50

13. Измеренные формы выходного напряжения Vo (50 В/дел) и выходного тока Io (50) Рисунок

Измеренные формы сигналов выходное напряжение Vo (50 В/дел) и выходной ток мА/дел);

(50 мА/дел); масштаб времени: 10 мс/

масштаб времени: 10 мс/

дел. мА/дел); масштаб времени: 10 мс/дел.

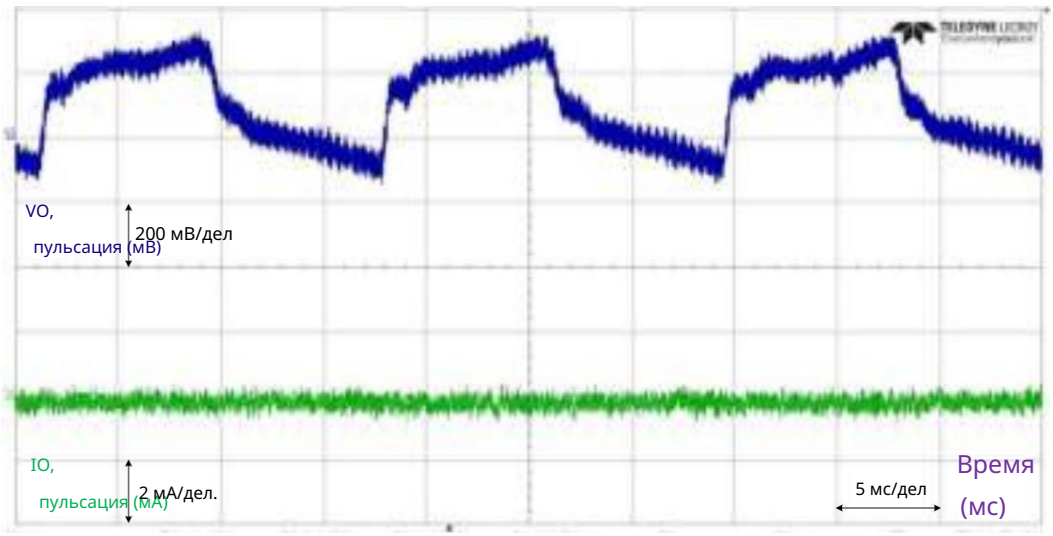


Рис. 14. Измеренные формы пульсаций выходного напряжения Vo,ripple (200 мВ/дел) и пульсаций выходного тока Io,ripple (2 мА/дел); масштаб времени: 5 мс/дел.

В таблице 4 показаны коэффициент пульсации выходного напряжения и коэффициент пульсации выходного тока прототипа схемы электронного драйвера освещения для светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации при среднеквадратичном входном напряжении переменного тока 110 В. Как видно из таблицы 4, деление размаха выходного напряжения 143,8 милливольта на среднее значение 91,35 вольт дает коэффициент пульсации напряжения 0,16%; деление размаха выходного тока 1,1658 миллиампер на среднее значение 38,57 миллиампер дает коэффициент пульсации тока 3,02%.

На рисунке 15 показаны измеренные формы входного переменного напряжения vAC и входного тока iAC. Из рисунка видно, что входной переменный ток соответствует входному напряжению, а фазы двух сигналов одинаковы; Таким образом, прототип предлагаемой однокаскадной схемы драйвера питания светодиодной дезинфицирующе-стерилизующей лампы глубокого ультрафиолета имеет эффект коррекции коэффициента мощности. С помощью анализатора мощности были измерены коэффициент мощности и общий коэффициент гармонических искажений входного тока, которые составили 0,9236 и 17,401% соответственно . Кроме того, измеренная входная мощность схемы возбуждения прототипа составляет 3,901 Вт, выходная мощность - 3,583 Вт, а КПД схемы прототипа - 91,85%.

Machine Translated by Google

Параметры	Ценности
Пиковое значение выходного напряжения	143,8 мВ
Среднее значение выходного напряжения	91,35 В
Коэффициент пульсации выходного напряжения	0,16%
Пиковое значение выходного тока	1,1658 мА
Среднее значение выходного тока	38,57 мА
Коэффициент пульсации выходного тока	3,02%

Таблица 4. Измерение коэффициента пульсации выходного напряжения и коэффициента пульсации выходного тока электронных устройств.

схема прототипа драйвера освещения с использованием диодов Шоттки из карбида кремния для глубокой ультрафиолетовой дезинфекции светодиодов стерилизационной лампы.

На рисунке 15 показаны измеренные формы входного переменного напряжения vAC , входного тока и iAC. Из рисунка видно, что входной переменный ток соответствует входному напряжению.
и фазы двух сигналов одинаковы; Таким образом, прототип про- пикового значения выходного напряжения,
драйвера для питания светодиодов глубокого ультрафиолета для дезинфекции и среднего значения выходного
представленного одноступенчатой схемой
имеет эффект коррекции коэффициента мощности. Коэффициент мощности и общий коэффициент пульсаций
напряжения стерилизующей лампы,
коэффициент гармонических искажений входного тока составил 0,9236 и 17,401%,
выходного напряжения. Пиковое
использованием анализатора мощности. Кроме того, измеренная входная мощность про- Среднее значение
составляет 3,901 Вт, а выходная мощность - 3,583 Вт, а эффективность коэффициента пульсации схемы-прототипа
выходного тока для схемы возбуждения
выходного тока составляет 91,85%.

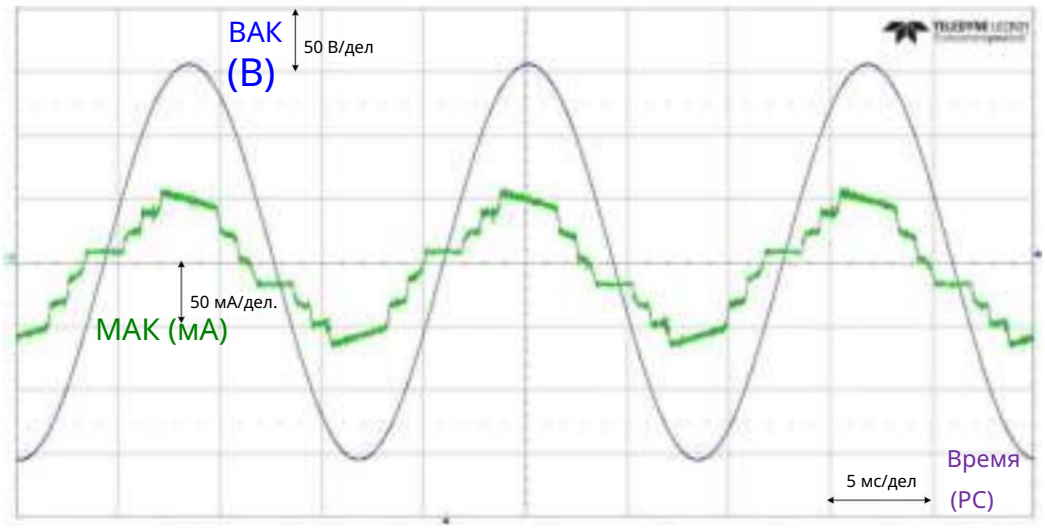


Рисунок 15. Измеренное входное напряжение сети переменного тока (50 В/дел) и ток iAC (50 мА/дел), временная шкала: Рисунок 15. Измеренное входное напряжение сети переменного тока (50 В/дел) и ток iAC (50 мА/дел), время
5 мс/дел.
масштаб: 5 мс/дел.

В Таблице 5 показано сравнение существующей драйвера светодиодов AC-DC [19] и в Таблице 5
инверсный повышающе-понижающий преобразователь с демпфером без потерь и обеспечивает мощность 18 Вт.
питания (60 В/0,3 А), есть (60 В/0,3 А) и предлагаемый вариант включает преобразователь
обратноходовой преобразователь и обеспечивает мощность 3,6 Вт (90 В/0,04 А). Как видно из
В таблице 5 светодиодные драйверы питаются входом переменного напряжения 110 В с использованием источника питания
выключатель. Предложенная схема драйвера AC-DC светодиодов имеет только конденсатор
для лучшего сравнения с другими [19]. Кроме того, [19] имеет THD, а эффективность схемы
предлагаемый драйвер светодиода AC-DC лучше, чем существующий драйвер.
Эффективность предлагаемого драйвера светодиодов AC-DC выше, чем существующего драйвера.

Таблица 5. Сравнение существующего AC-DC светодиодного драйвера в [19] и предлагаемого.

Элемент	Существующий драйвер светодиодов AC-DC Ссылка [19]	Предлагаемый драйвер светодиода AC-DC
Топология схемы	Интеграция обратного понижающе-повышающего режима преобразователь с демпфером без потерь	Интеграция понижающего преобразователя и обратноходовой преобразователь
Входное переменное напряжение	110 В	110 В
Выходная мощность	18 Вт (60 В/0,3 А) 1	3,6 Вт (90 В/0,04 А)
Количество необходимых переключателей		1
Количество необходимых конденсаторов	4	3
Количество необходимого магнитного элемента	3	2
Количество необходимых диодов	7	6
Измеренный коэффициент мощности	0,9737	0,9236
Измеренный ток THD	18,422%	17,401%
Измеренная эффективность схемы	85,01%	91,85%

Machine Translated by Google

4. Выводы

В данной статье предлагается одноступенчатый прототип схемы драйвера электронного освещения, в которой используются диоды Шоттки из SiC для светодиодной лампы для дезинфекции и стерилизации глубоким ультрафиолетом, которая объединяет понижающий преобразователь и обратноходовой преобразователь для формирования одноступенчатого преобразователя переменного тока в постоянный с одним силовым переключателем. Архитектура схемы силового преобразователя с функцией коэффициентной коррекции. Путем разработки индуктора в преобразователе переменного тока в постоянный для работы в режиме прерывистой проводимости эффект коррекции коэффициента мощности может быть достигнут естественным путем. Кроме того, в предлагаемой однокаскадной электронной схеме драйвера освещения используется диод Шоттки с основной несущей, изготовленный из карбидокремниевого широкозонного полупроводникового материала, который восстанавливает энергию из индуктивности рассеяния трансформатора, тем самым улучшая эффективность преобразования схемы драйвера. В этой статье был разработан и протестирован прототип схемы драйвера электронного освещения со светодиодами глубокого ультрафиолета мощностью 3,6 Вт для лампы для дезинфекции и стерилизации . При входном напряжении переменного тока 110 В (среднеквадратичное значение) и номинальной выходной мощности результаты были следующими: коэффициент пульсации выходного напряжения менее 1 %, коэффициент пульсации выходного тока менее 4 %, коэффициент мощности более 0,9, общий коэффициент гармонических искажений входного тока менее более 18%, а эффективность схемы более 90%. Однокаскадный прототип схемы драйвера электронного освещения светодиодов глубокого ультрафиолета, предложенный в этой статье, может упростить количество силовых переключателей, необходимых в схем

Вклад авторов: C.-AC разработал топологию схемы электронного драйвера освещения и разработал методологию; C.-HC, H.-LC и E.-CC организовали ресурсы программного обеспечения и настроили моделирование, а также выполнили моделирование схем; Л.-ФЛ, С.-ХХ и К.-КЛ реализовали прототип схемы электронного драйвера освещения и провели измерения схемы, а также проанализировали результаты экспериментов под руководством К.-АС; C.-AC подготовил первый вариант статьи и отредактировал рукопись для подачи. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Это исследование финансировалось Национальным советом по науке и технологиям (NSTC) Тайваня по номерам грантов MOST 110-2221-E-214-014, NSTC 111-2221-E-214-011 и NSTC 112-2221 -E. -214-005.

Заявление о доступности данных: данные, представленные в этом исследовании, доступны в статье.

Благодарность: Авторы хотели бы выразить искреннюю благодарность Национальному совету по науке и технологиям (NSTC) Тайваня за номера грантов MOST 110-2221-E-214-014, NSTC 111-2221-E-214-011 и NSTC. 112-2221-E-214-005.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Кебби, Ю.; Мухаммед, А.И.; Сант'Ана, AS; Прадо-Сильва, LD; Лю, Д.; Дин, Т. Последние достижения в применении технологии УФ-светодиодов для инактивации микробов: прогресс и механизм. Компр. Преподобный Food Sci. Еда Саф. 2020, 19, 3501–3527. [Перекрестная ссылка] [ПабМед]

2. Шаталов М.; Джайн, Р.; Саксена, Т.; Добринский А.; Шур М. Разработка светодиодов глубокого УФ-излучения и текущие проблемы в технологии материалов и устройств. В полупроводниках и полуметаллах; Elsevier: Амстердам, Нидерланды, 2017 г.; Том 96, стр. 45–83.

3. Кучма Т.; Попович, В.; Грин А. Обзор технологии ультрафиолетовых (УФ) светодиодов для применения в производстве продуктов питания. В Ультрафиолетовая светодиодная технология для пищевой промышленности; Academic Press: Кембридж, Массачусетс, США, 2019; стр. 1–23.

4. Акгюн, депутат; Юнлютюрк, С. Влияние ультрафиолетовых светодиодов (светодиодов) на микробную и ферментативную инактивацию яблочного сока. Межд. J. Пищевая микробиол. 2017, 260, 65–74. [Перекрестная ссылка] [ПабМед]

5. Кая, З.; Йылдыз, С.; Унлютюрк, С. Влияние УФ-облучения и термической обработки на стабильность срока годности лимонно-дынного сока смесь: многомерный статистический подход. Иннов. Пищевая наука. Экстрен. Технол. 2015, 29, 230–239. [Перекрестная ссылка]

6. Гуань З.; Лю, П.; Чжоу, Т.; Чжоу, Л.; Чжан, Д.; Се, К.; Ю, К.; Привет.; Ван, С.; Ван, Х.; и другие. Исследование регулирования светового поля при дезинфекции УФ-светодиодами при транспортировке в холодовой цепи. Прил. наук. 2022, 12, 1285. [CrossRef]

7. Лян, Джей-Джей; Ляо, СС; Чанг, CS; Ли, Калифорния; Чен, С.Ю.; Хуанг, С.Б.; Да, ЮФ; Сингх, К.Дж.; Куо, ХК; Лин, Ю.Л.; и другие. Эффективность прототипов устройств дальнего ультрафиолета (УФ) с разными длинами волн при дезинфекции SARS-CoV-2. Прил. наук. 2021, 11, 10661. [CrossRef]

8. Хуарес-Леон, ФА; Сориано-Санчес, АГ; Родригес-Лисеа, Массачусетс; Перес-Пиналь, Ф.Дж. Разработка и внедрение бактерицидного средства UVC-LED лампа. IEEE Access 2020, 8, 196951–196962. [Перекрестная ссылка] [ПабМед]

Machine Translated by Google

9. Алонсо, Ж.М.; Коста, безумец; Ордис, К. Интегрированный понижающе-обратно-ходовой преобразователь как автономный источник питания с высоким коэффициентом мощности. IEEE Пер. Индийский Электрон. 2008, 55, 1090–1100. [Перекрестная ссылка]

10. Гасио, Д.; Алонсо, Дж. М.; Каллеха, АЈ; Гарсия, Дж.; Рико-Секадес, М. Одноступенчатый источник питания 2e с высоким коэффициентом мощности и универсальным входом для HB-светодиодов на основе встроенного понижающе-обратноходового преобразователя. IEEE Транс. Индийский Электрон. 2010, 58, 589–599. [Перекрестная ссылка]

11. Абдельмесих, Г.З.; Алонсо, Дж. М.; Коста, MAD Анализ потерь для повышения эффективности интегрированного драйвера светодиодов с понижающе-обратным ходом. IEEE Транс. Индийское прил. 2018, 54, 6543–6553. [Перекрестная ссылка]

12. Абдельмесих, Г.З.; Алонсо, Дж. М.; Цай, В.Т. Анализ и экспериментирование нового автономного светодиодного драйвера с высоким коэффициентом мощности на основе на встроенном обратноходовом преобразователе с чередованием. IEEE Транс. Индийское прил. 2019, 55, 4359–4369. [Перекрестная ссылка]

13. Луз, ПК; Болан, ЧП; Мабони, Т.; Косетин, г-н; да Силва, МФ; до Прадо, Р.Н. Встроенный изолированный понижающе-повышающий-обратноходовой преобразователь для питания светодиодных ламп уличного освещения с пониженной емкостью. В материалах 11-й Международной конференции IEEE/IAS по отраслевым приложениям 2014 г., Жуис-де-Фора, Бразилия, 7–10 декабря 2014 г.; стр. 1–6.

14. Дивья, К.М.; Паракал, Р. Встроенный повышающе-понижающий обратноходовой преобразователь с высоким коэффициентом мощности, управляющий несколькими выходами. В материалах Международной онлайн-конференции по зеленой инженерии и технологиям 2015 г. (IC-GET), Коимбатур, Индия, 27 ноября 2015 г.; стр. 1–5.

15. Ченг, Калифорния; Чанг, Тайвань Новый интегрированный драйвер с PFC для питания светодиодов. В материалах 41-й ежегодной конференции Общества промышленной электроники IEEE (IECON 2015), Иокогама, Япония, 9–12 ноября 2015 г.; стр. 3514–3518.

16. Пурали, Б.; Адиб, Э. Анализ встроенного обратноходового преобразователя SEPIC как одноступенчатого однопереключательного устройства коррекции коэффициента мощности. Светодиодный драйвер. IEEE Транс. Индийский Электрон. 2016, 63, 3562–3570. [Перекрестная ссылка]

17. Ченг, Калифорния; Чанг, Ч.-Х.; Ченг, Х.-Л.; Чанг, М.-Т. Новый светодиодный драйвер с коррекцией коэффициента мощности, подходящий для уличного освещения. В материалах Международной конференции по энергетической электронике будущего 2017 (IFEEC 2017) — ECCE Asia, Гаосюн, Тайвань, 3–7 июня 2017 г.; стр. 1–4.

18. Ченг, Калифорния; Чанг, Швейцария; Ченг, HL; Чанг, EC; Лай, CC Новый драйвер светодиодной трубки с коррекцией коэффициента мощности для внутреннего освещения. В материалах 7-й Глобальной конференции IEEE по бытовой электронике (GCCE) 2018 г., Нара, Япония, 9–12 октября 2018 г.; стр. 700–701.

19. Ченг, Калифорния; Чен, ЮЛ; Чен, Ю.Дж.; Лай, Юв. Новый экономичный источник питания переменного и постоянного тока для внутреннего светодиодного освещения. В материалах Азиатско-Тихоокеанской конференции по энергетике и энергетике IEEE PES 2019 (APPEEC), Макао, Китай, 1–4 декабря 2019 г.; стр. 1–3.

20. Ченг, Калифорния; Лан, LF; Хоу, Ш; Лин, К.К. Одноступенчатый электронный осветительный драйвер для светодиодной лампы глубокого ультрафиолета для дезинфекции и стерилизации. В материалах семинара IEEE 2023 г. по силовым устройствам и приложениям с широкой полосой пропускания в Азии (WiPDA Asia), Синьчжу, Тайвань, 27–29 августа 2023 г.; стр. 1–3.

21. Эффективное использование силовых устройств — понимание силовых устройств из карбида кремния и примеры применения с использованием их преимуществ. ROHM Semiconductor, 2017. Доступно онлайн: https://pages.rohm.com.tw/rs/247-PYD-578/images/TechWeb_E_DL_SiC.pdf. (по состоянию на 31 июля 2024 г.).

22. Рекомендации по применению силовых устройств и модулей SiC, ROHM Semiconductor, 2020 г. Доступно в Интернете: https://fscdn.rohm.com/en/продукты/сборник данных/applinote/discrete/sic/common/sic_appli-e.pdf (по состоянию на 31 июля 2024 г.).

23. ПРИМЕЧАНИЕ ПО ПРИМЕНЕНИЮ AN966, STMicroelectronics, «L6561, РАСШИРЕННЫЙ КОРРЕКТОР ПЕРЕХОДНОГО РЕЖИМА МОЩНОСТИ», стр. 1–23, март 2003 г. Доступно в Интернете: https://www.st.com/resource/en/application_note/cd00004002-l6561 . -улучшенный -переходный режим-корректор коэффициента мощности-stmicroelectronics.pdf (по состоянию на 31 июля 2024 г.).

24. СЕУЛЬВИОСИС. Серия UV CA3535 (CUD5GF1B). Доступно онлайн: https://media.digikey.com/pdf/Data%20Sheets/Seoul%20.Полупроводник/CUD5GF1B_R0.1_Prelliminary_2-3-20.pdf (по состоянию на 31 июля 2024 г.).

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Эффективное уменьшение размерности речевого сигнала с использованием Комплексные методы

Сонгюн Ко [1] и Минхо Пак 2,*

[1] Департамент конвергенции ИИ и ИТ, Университет Сунгсил, Сеул 06978, Республика Корея;
[2] kosungkyun@naver.com Школа электронной инженерии, Университет Сунгсил, Сеул 06978, Республика Корея
* Переписка: mhp@ssu.ac.kr

Аннотация: В данном исследовании мы предлагаем алгоритм CVMFCC-DR (Complex-Valued Mel-Frequency Cepstral Coefficients Dimensionality Reduction) в качестве эффективного метода снижения размерности речевых сигналов. Используя комплексный метод MFCC, который учитывает как действительные, так и мнимые компоненты, наш алгоритм позволяет уменьшить размерность без потери информации и одновременно снизить вычислительные затраты. Эффективность предложенного алгоритма подтверждена экспериментами, которые демонстрируют его эффективность при построении модели распознавания речи с использованием комплексной нейронной сети. Кроме того, представлен комплексный метод интерпретации softmax для комплексных чисел. Результаты экспериментов показывают, что этот подход обеспечивает более высокую производительность по сравнению с традиционными методами на основе MFCC, тем самым подчеркивая его потенциал в области распознавания речи.

Ключевые слова: мел-частотные кепстральные коэффициенты; комплексная коррекция ошибок нейронной сети; обратное распространение ошибки; комплексный softmax

check for updates

Цитирование: Ко, С.; Парк, М. Эффективное уменьшение размерности речевого сигнала с использованием комплексных методов. Электроника 2024, 13, 3046. https://doi.org/10.3390/электроника13153046

Академические редакторы: Синь Гэн, Падма Айенгар и Эльке Пульвермюллер

Поступило: 3 июня 2024 г.
Пересмотрено: 31 июля 2024 г.
Принято: 31 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария.
Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Сегодня обработка и распознавание речевых сигналов играют решающую роль в различных областях применения. Эти технологии используются для различных целей, таких как преобразование речевых команд в текст или идентификация владельца голоса. Развитие технологий распознавания речи может улучшить взаимодействие человека и машины, тем самым повышая производительность и эффективность в различных отраслях. Однако речевые сигналы обладают весьма разнообразными формами и многомерными характеристиками, что делает их эффективную обработку и распознавание постоянной проблемой. Разнообразие и высокая размерность речевых сигналов затрудняют уменьшение размерности и распознавание образов, что приводит к увеличению вычислительных затрат. Поэтому эффективное снижение размерности речевых сигналов и повышение производительности моделей распознавания речи на основе нейронных сетей является серьезной проблемой.

Технологии обработки и распознавания речи необходимы в различных приложениях, таких как голосовые интерфейсы, виртуальные помощники и автоматический перевод. Достижение более точных и эффективных результатов обработки и распознавания речевых сигналов может значительно улучшить качество обслуживания пользователей сопутствующих услуг. В связи с этим ведутся постоянные исследования в области обработки и распознавания речевых сигналов. Речевые сигналы обладают многомерными характеристиками, что затрудняет эффективное снижение разме Более того, точное распознавание разнообразных речевых образов требует сложных моделей нейронных сетей, что, в свою очередь, увеличивает вычислительные затраты [1]. Поэтому эффективное снижение размерности речевых сигналов без потери информации и одновременное повышение точности и эффективности моделей распознавания речи на основе нейронных сетей является важнейшей задачей.

Несмотря на богатую литературу по комплексным сетям, примеров их применения в обработке звука или речи было мало. Учитывая, что слуховая система челове

Machine Translated by Google

обрабатывает звук в частотной области, также имеет смысл изучить комплексные нейронные сети для обработки звука, а не полагаться исключительно на нейронные сети с действительными значениями во временной области [2].

Недавние исследования достигли значительного прогресса в распознавании речи путем объединения традиционных методов, таких как мел-частотные кепстральные коэффициенты (MFCC), с нейронными сетями [3]. Хотя метод MFCC широко используется для извлечения речевых характеристик, у него есть недостаток: он не может выполнить уменьшение размерности без потерь, поскольку использует только действительные значения. Другими словами, уменьшение количества MFCC приводит к потере информации, что снижает точность распознавания речи. Эти методы по-прежнему демонстрируют ограниченную эффективность, и необходимы более совершенные подходы, выходящие за рамки традиционных методов обучения.

Использование комплексного представления MFCC позволяет учитывать как действительные , так и мнимые части, обеспечивая уменьшение размерности без потерь. Кроме того, использование комплексных нейронных сетей может отражать изменения фазы посредством сложных операций умножения [4], лучше представляя характеристики речевых сигналов. Ожидается, что этот подход обеспечит более эффективное снижение размерности и распознавание речи на основе нейронных сетей по сравнению с существующими методами.

В этом исследовании мы предлагаем алгоритм CVMFCC-DR для эффективного уменьшения размерности речевого сигнала без потерь. Объединив CVMFCC-DR с комплексными нейронными сетями, этот метод может продемонстрировать превосходную производительность распознавания речи по сравнению с существующими методами.

Основные результаты данного исследования заключаются в следующем:

• Предложение эффективного метода уменьшения размерности (CVMFCC-DR), использующего комплексное представление MFCC. • Повышение производительности распознавания речи за счет объединения комплексных MFCC и комплексных нейронных сетей. • Предложение нового метода интерпретации комплексной функции softmax с сложные входы.

Эта статья начинается с представления существующей техники MFCC и комплексных нейронных сетей. Затем он дает подробное объяснение предлагаемого алгоритма CVMFCC-DR и метода обучения комплексных нейронных сетей [5]. В дальнейшем эффективность предложенного метода подтверждается экспериментально и анализируются результаты . Наконец, в документе приводится краткое изложение результатов исследования и предлагаются направления будущих исследований.

2. Сопутствующие работы

В области распознавания речи используются различные подходы и методы на основе нейронных сетей. Используются различные типы нейронных сетей, такие как CNN, RNN, LSTM и DNN, каждая из которых имеет свои преимущества и недостатки [3]. Кроме того, исследуются также гибридные модели, сочетающие традиционные методы со структурами нейронных сетей . Гибридные модели могут оптимизировать производительность за счет объединения сильных сторон каждой структуры, но они также имеют недостаток, заключающийся в повышенной сложности реализации и интерпретации [6].

На фоне этих разработок традиционные методы извлечения признаков, такие как метод MFCC, по-прежнему широко используются в качестве входных данных для нейронных сетей [3]. Кроме того, изучение новых нейронных архитектур, таких как комплекснозначные нейронные сети, может открыть новые перспективы и потенциальные прорывы в возможностях распознавания речи. Недавно были проведены исследования, показывающие, что FPGA можно использовать для значительного повышения скорости вычислений комплексных чисел [7]. Результаты этих исследований эмпирически показывают, что аппаратная реализация операций с комплексными числами может привести к практическому повышению производительности. Кроме того, недавнее исследование реализовало комплексные нейронные сети на FPGA для классификации, продемонстрировав превосходную точность по сравнению с нейронными сетями с действительными значениями и добившись значительного улучшения скорости по сравнению с системами на базе ЦП [8]. Это дает основание ожидать, что в будущем процессоры, которые в настоящее время поддерживают только операции с действительными числами, однажды смогут поддерживать и операции с ком

Machine Translated by Google

Основываясь на этих исследованиях, использование представлений комплексных чисел позволяет кодировать пары действительных чисел как отдельные комплексные числа, эффективно уменьшая входную размерность вдвое. Такое уменьшение размерности уменьшает количество параметров модели и повышает эффективность вычислений. Следовательно, экспериментальные данные показывают, что комплексные нейронные сети могут достичь равной или превосходящей производительности с меньшим количеством параметров по сравнению с обычными нейронными сетями с действительными значениями.

Поскольку исследования продолжают расширять границы, синергетическое сочетание проверенных методов и инновационных подходов может проложить путь к следующему поколению надежных и высокопроизводительных систем распознавания речи.

3. Базовые знания

В области распознавания и классификации речи проводятся различные исследования, и в этом разделе мы представляем MFCC (мел-частотные кепстральные коэффициенты) [3]. Метод MFCC — это один из методов выделения признаков, обычно используемых в области обработки речевых сигналов и распознавания речи. Метод MFCC представляет частотные характеристики речевых сигналов и в основном используется для эффективного представления и распознавания речевых сигналов. Основной принцип метода MFCC показан на рисунке 1.

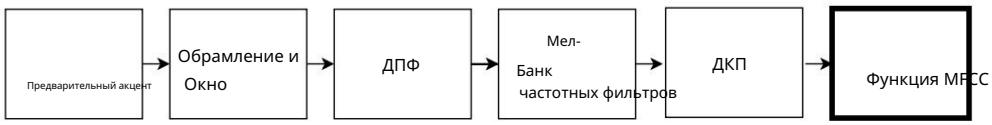


Рисунок 1. Обзор процесса применения метода MFCC.

Предварительное выделение: Предварительное выделение – это, по сути, фильтр верхних частот. Из-за строения человеческого тела высокочастотные компоненты имеют тенденцию уменьшаться в реальном издаваемом звуке во время произнесения речи. Таким образом, предыскажение — это процесс применения фильтра верхних частот для усиления сначала ослабленных высокочастотных составляющих.

Формирование и формирование окон: речевые сигналы для анализа делятся на короткие сегменты, называемые кадрами. Это связано с тем, что речевые характеристики остаются относительно стабильными в течение коротких периодов времени. Эти кадры слегка перекрываются, чтобы более эффективно фиксировать изменения ре Каждый кадр обрабатывается с использованием оконной технологии, что расширяет возможности анализа. Распространенные типы окон включают окна Хэмминга и Хэннинга. Этот метод кадрирования позволяет более точно анализировать сложные речевые данные, что имеет решающее значение для разработки технологий распознавания и синтеза речи. Хотя конкретные детали реализации могут различаться, эти фундаментальные принципы широко используются в системах обработки речи.

Дискретное преобразование Фурье: для расчета спектра мощности используется ДПФ (дискретное преобразование Фурье) для расчета спектра мощности для каждого кадра, полученного ранее. Этот процесс позволяет распознавать частотно-зависимое распределение энергии речевого сигнала в каждом кадре.

Набор фильтров Mel-Frequency: Полосовой фильтр Mel представляет собой набор фильтров, предназначенных для извлечения речевых сигналов нелинейным образом, имитируя слуховую систему человека.

Обычно он состоит из 40 треугольных фильтров и анализирует частотные характеристики речевых сигналов.

Дискретное косинусное преобразование: DCT (Дискретное косинусное преобразование) выражает конечную последовательность точек данных как сумму косинусных функций, колеблющихся на разных частотах. В процессе MFCC DCT применяется к банку фильтров mel для выбора коэффициентов ускорения или для отделения отношений в логарифмических спектральных величинах от банка фильтров.

Распознавание речи с использованием MFCC может эффективно использоваться в различных средах [9], и этот метод может достигать высокой точности в сочетании с моделями глубокого обучения. Таким образом, MFCC служит решающим фактором для моделей глубокого обучения для распознавания речи [10].

Machine Translated by Google

4. Предлагаемый метод

В этом разделе мы предлагаем метод эффективного снижения размерности речевого сигнала с использованием комплексных MFCC и комплексных нейронных сетей.
Обзор предлагаемого подхода представлен на рисунке 2. Этот рисунок иллюстрирует предлагаемый подход к классификации голоса с использованием комплексных нейронных сетей. Процесс начинается с голосового ввода, за которым следует извлечение признаков с использованием мел-частотных кепстральных коэффициентов (MFCC). Эти функции затем подвергаются уменьшению размерности и преобразуются в комплексные представления, в результате чего получаются комплексные MF
Обработанные функции впоследствии вводятся в комплексную нейронную сеть.
Производительность сети повышается за счет выбора соответствующих функций активации и алгоритмов обучения. В конечном итоге выходные данные нейронной сети используются для целей классификации.

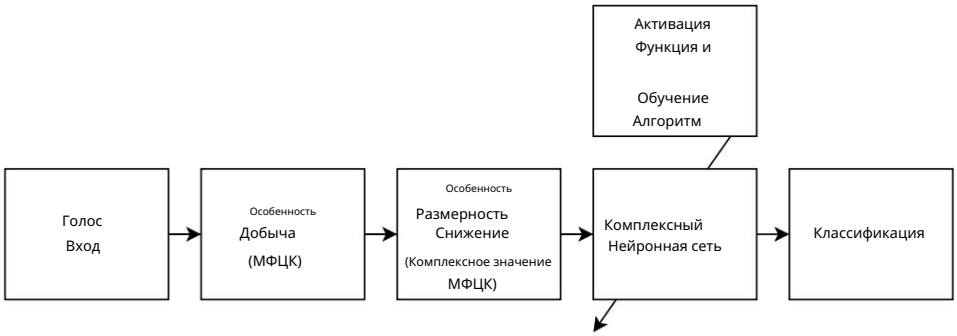


Рисунок 2. Обзор предлагаемого подхода.

Комплекснозначные MFCC: количество MFCC, используемых при распознавании речи, может варьироваться в зависимости от конкретных приложений и моделей. Однако обычно используются значения MFCC в диапазоне от 13 до 40. Известно, что этот диапазон эффективно работает в различных системах распознавания речи. Как правило, большее количество MFCC позволяет модели улавливать больше частотных и спектральных деталей, но это также увеличивает вычислительные затраты модели. Таким образом, выбранное количество MFCC определяется с учетом баланса между эффективностью вычислений и производительностью модели.

При представлении физических задач с использованием комплексных чисел существует два основных метода. Один из них — простой метод замены, а другой — более элегантный подход, известный как математическая эквивалентность, например преобразование Фурье. При замене берутся два действительных физических параметра и помещаются один в действительную часть комплексного числа, а другой в мнимую часть. Это позволяет манипулировать двумя значениями как единым объектом, т. е. одним комплексным числом [11]. Метод, предложенный для уменьшения размерности в этой статье, включает преобразование значений n MFCC в комплексные числа. Первый MFCC используется как действительная часть, а второй MFCC используется как мнимая часть комплексного числа.
Впоследствии каждый MFCC поочередно является действительной и мнимой частями комплексного числа. Этот шаблон продолжается до n-го числа MFCC, в результате чего получается n/2 комплексных чисел. Например, когда у нас есть n значений MFCC, они могут быть следующими:

$$[x1, x2, x3, x4, \ldots, xn \quad 1, xn]$$

В этом случае первый MFCC x1 используется как действительная часть первого комплексного числа. Второй MFCC x2 используется как мнимая часть первого комплексного числа. Третий MFCC x3 используется как действительная часть второго комплексного числа. Четвертый MFCC x4 используется как мнимая часть второго комплексного числа. Этот шаблон продолжает чередовать действительную и мнимую части до n-го MFCC xn, который становится мнимой частью (n/2)-го комплексного числа. Следовательно, при n значениях MFCC генерируется n/2 комплексных чисел. Первое комплексное число — x1 + jx2, второе комплексное число

Machine Translated by Google

число равно x3 + jx4 и так далее до последнего комплексного числа, которое равно xn 1 + jxn. Мы проиллюстрировали это объяснение на рисунке 3.

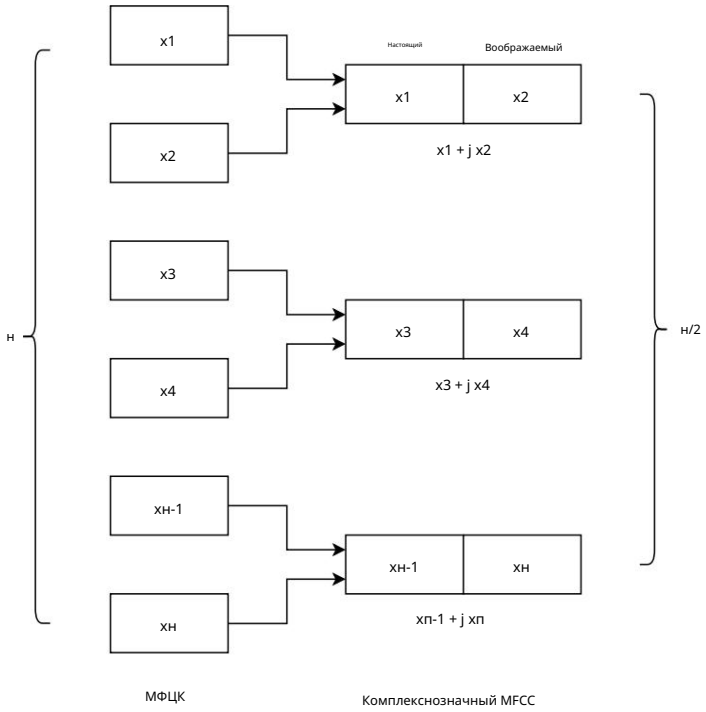


Рисунок 3. Процесс генерации комплексных MFCC.

Нейронная сеть с комплексными значениями. Нейронные сети с комплексными значениями — это модели нейронных сетей, которые используют комплексные числа в качестве входных значений или весов, и у них есть преимущество в обработке более обширной информации по сравнению с нейронными сетями с действительными значениями. Кроме того , в различных областях были проведены различные исследования и эффективно используются методы комплексного уменьшения размерности или классификации и обнаружения с использованием комплексных чисел. Для этого выполняется умножение комплексных входов и комплексных весов, что подразумевает вращение в комплексной плоскости и, таким образом, вносит изменения в фазу [5]. Сумма умножения входа X и веса W может быть представлена в виде уравнений (1) и (2).

$$Z = W0 + W1X1 + W2X2 + \ldots + WnXn \quad (1)$$

$$Z = \quad n\, Wn \cdot Xn \quad (2)$$

Функция активации и алгоритм обучения. В комплексных нейронных сетях используется фазозависимая функция активации, которая приводит к нормализации информации о величине к единичному кругу, равному 1, и только к извлечению фазовой информации, как показано в уравнении (3).).

$$P(3) = \frac{3}{|3|} = e^{\,j\cdot \arg(Z)} \quad (3)$$

Алгоритм обучения комплексных нейронных сетей следующий. Когда есть желаемая цель t и выходное значение z, как показано на рисунке 4, веса корректируются в соответствии с уравнением (4) для постепенного перемещения в направлении желаемой цели [12].

Machine Translated by Google

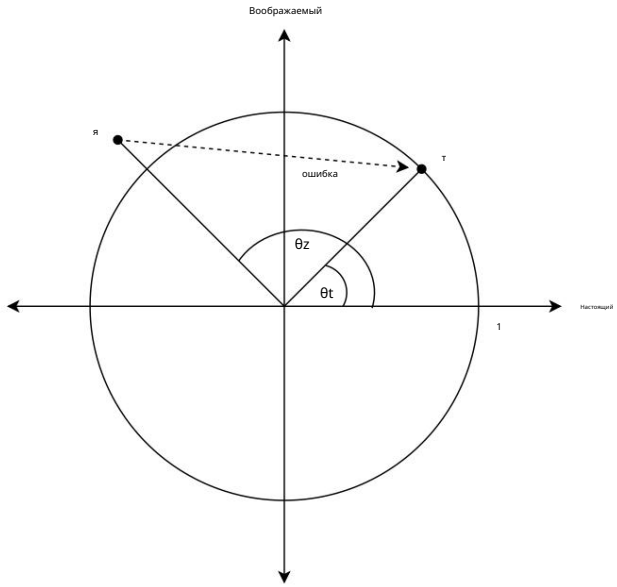


Рисунок 4. Алгоритм обучения CVNN.

$$t = n\,(Wn + Wn)\cdot Xn \tag{4}$$

Ошибка относительно цели обучения t равна e = t - z. , как показано в уравнении (5).

$$e = n\,(Wn + Wn)\cdot Xn \quad n\,(Wn)\cdot Xn = n\,(Wn)\cdot Xn \tag{5}$$

Чтобы упростить и ускорить вычисление ошибки относительно изменения веса, предполагая, что каждый узел вносит равный вклад в ошибку, символ суммирования σ удаляется из уравнения, и обе части делятся на количество вносящих вклад узлов N в комплексе. -значная нейронная сеть для удобства вычислений. Полученное уравнение представляет собой уравнение (6).

$$Wn \cdot Xn = \frac{e}{H} \tag{6}$$

Чтобы получить Wn и изменить веса, умножьте обе части уравнения (6) на Xn при умножении $Икс_н^{1}$. $Икс_н^{1}$ на комплексно-сопряженное число Xn, где n является и числителем, и знаменателем ; таким образом, $X_н^{1} = \frac{\overline{Xн}}{|Xn|^2}$, где |Xn| = 1 для Xn внутри единицы круг. Следовательно, $X_н^{1} = \overline{Xn}$, и уравнение (6) снова получается как уравнения (7) и (8).

$$Wn \cdot Xn \cdot X_н^{1} = \frac{e}{H} \cdot Икс_н^{1} = \frac{e}{H} \cdot \overline{Xн} \tag{7}$$

$$Wn = \frac{e}{H} \cdot \overline{Xн} \tag{8}$$

Таким образом, полученное изменение весов можно наблюдать в уравнении (8). Это изменение представляет собой корректировку комплексных весов для каждого узла, вносящего вклад в ошибку относительно цели обучения, и равномерно применяется по всей сети.

Благодаря этому процессу комплексные нейронные сети подстраиваются ближе к цели, используя фазозависимые функции активации и эффективные алгоритмы обучения, одновременно обрабатывая богатую информацию посредством сложных операций умножения между входными данными и весами [13].

Обратное распространение ошибки в сложной области: при использовании скрытых слоев ошибка из выходного слоя распространяется обратно на скрытые слои, чтобы вычислить, насколько велика ошибка каждого узла.

Machine Translated by Google

в скрытых слоях способствовали этой ошибке. Исходя из этого, веса между скрытым и выходным слоями необходимо соответствующим образом настроить. Этот процесс называется обратным распространением ошибки, посредством которого нейронная сеть может корректировать веса, чтобы минимизировать ошибку в выходных значениях во время обучения. Структура многослойной комплексной нейронной сети показана на рисунке 5.

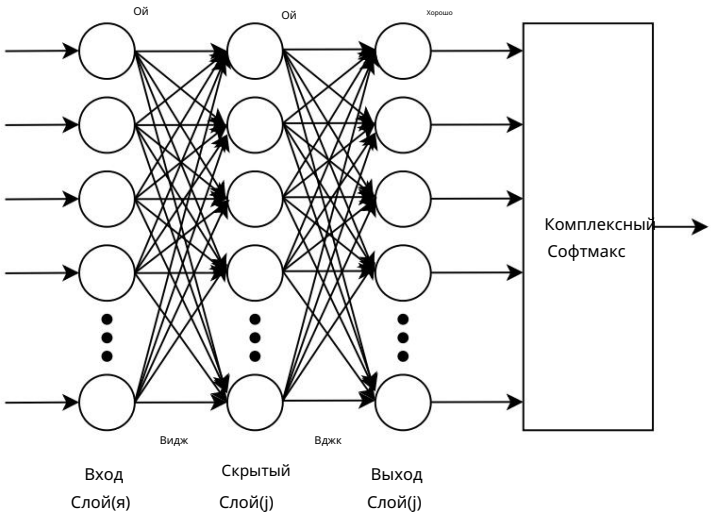


Рисунок 5. Многоуровневая CVNN.

В этой статье метод исправления ошибок используется для выходного слоя k для обновления весов, а метод обратного распространения ошибки используется для скрытого слоя j, чтобы изучить его влияние на выходные данные. Другими словами, выходной слой k использует метод исправления ошибок, чтобы избежать попадания в локальные минимумы, в то время как метод обратного распространения ошибки применяется к скрытому слою j, чтобы обеспечить лучшее распределение ошибок от выходного слоя к скрытому слою. Можно ожидать , что эта комбинация обратного распространения ошибок в скрытом слое и коррекции ошибок в выходном слое улучшит проблему застревания всей нейронной сети в локальных минимумах, одновременно обеспечивая более быструю сходи

Метод обратного распространения ошибки для комплексных нейронных сетей заключается в следующем.

При использовании градиентного спуска для обновления веса в комплексных нейронных сетях исчисление Виртингера используется для дифференциации сложных функций. Поскольку сложные функции состоят из вещественной и мнимой частей, исчисление Виртингера проводит дифференцирование отдельно для действительной и мнимой частей. Для комплексного числа z = x + jy , определенного с действительной и мнимой частями x и y соответственно, и комплексной функции f(z) = u(x, y) + jv(x, y), производные Виртингера определяются как следует [14].

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - j\frac{\partial}{\partial y}\right) = \frac{1}{2}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + j\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\right) \quad (9)$$

$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + j\frac{\partial}{\partial y}\right) = \frac{1}{2}\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} + j\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)\right) \quad (10)$$

$$= + z \frac{\partial}{\partial \bar{z}} \frac{df}{} \quad (11)$$

В этом контексте $\bar{z}$ обозначает комплексно-сопряженное число z. В методе производных Виртингера комплексные функции можно рассматривать как расширения действительных функций, разделив их на действительную и мнимую части для дифференцирования. Ниже описан процесс обновления комплексных переменных с использованием размера шага α/2 и функции потерь L.

Machine Translated by Google

$$xn+1 = xn - \frac{\alpha}{2} \frac{Л}{x} \tag{12}$$

$$yn+1 = yn - \frac{\alpha}{2} \frac{L}{y} \tag{13}$$

Подставляя zn+1 = xn+1 + j yn+1 в уравнения (12) и (13) , а затем подставляя Производная Виртингера в конце процесса вывода становится следующей.

$$zn+1 = xn - \frac{\alpha}{2} \frac{L + j\,yn}{x} \frac{\alpha}{2} \frac{Л}{y}$$

$$= xn - \frac{\alpha}{2} \frac{L + jyn}{x} j \frac{\alpha}{2} \frac{L}{y}$$

$$= xn + jyn \frac{\alpha}{2} \frac{L}{y} \frac{L + j}{x}$$

$$1 = zn \alpha \frac{L}{2} \frac{L + j}{y} x$$

$$= zn \alpha \frac{L}{z} \tag{14}$$

Частная производная L по комплексно-сопряженному элементу как сопряженная $\frac{L}{z}$ ссылается на производная Виртингера. Это демонстрирует, как можно использовать сопряженную производную Виртингера для упрощения формулы обновления для сложных переменных. С точки зрения использования описанного выше метода, переписывание уравнения (5) как ошибки выходного слоя дает уравнение (15), где tk представляет собой целевое значение, а ok представляет фактическое выходное значение выходного слоя k.

$$ek = tk - ok \tag{15}$$

А функция потерь определяется как уравнение (16).

$$L = e\,k^2 = |tk - ok|\,|tk - ok| \tag{16}$$

Если веса велики, ошибка с большей вероятностью распространится обратно на скрытые слои. Ошибка скрытого слоя ej может быть выражена следующим уравнением.

$$ej = WT_{джк} \cdot эк \tag{17}$$

Используя результат уравнения (14), градиент функции ошибок относительно веса Wij между входным слоем i и скрытым слоем j можно выразить следующим образом. Здесь oi — выходное значение входного слоя i, а oj — выходное значение скрытого слоя j.

$$\frac{Л}{Видж} = \frac{Л}{ой} \cdot \frac{oj}{Wij} \tag{18}$$

$$\frac{Л}{Wij} ej \cdot P\ Wij \cdot oi = \tag{19}$$

Machine Translated by Google

$$\frac{Л}{} = ej \cdot P \quad Wij' \cdot ой \tag{20}$$

В уравнениях (19) и (20) P представляет фазозависимую функцию активации, используемую в выходных данных скрытого слоя, а P. Обновление весов скрытого является производной зависимой функции активации. слоя можно выразить следующим образом.

$$новый\ Wij = старый\ Wij \quad lr \cdot \frac{Л}{Wij} \tag{21}$$

При работе с комплексными переменными для обновления необходимо вместо Wij использовать $\frac{Л}{Wij}$. Здесь lr играет роль в $\frac{L}{}$ регулировании интенсивности изменения, чтобы предотвратить перерегулирование, и он называется скоростью обучения. Кроме того, производная функции активации P является комплексной функцией, и для нахождения производной комплексной функции используется производная Виртингера [15].

Softmax с комплексным значением: функция softmax, которая используется в качестве функции активации в выходном слое, принимает входные значения и нормализует каждое значение в диапазоне от 0 до 1. Она гарантирует, что сумма нормализованных значений всегда равна 1. Эта функция позволяет интерпретировать результаты как вероятности. Это можно выразить уравнением (22).

$$Итак,\ f\ tmax(Z)i = e\ Zi \quad \frac{e^{Зи}}{\sum_{i=1}^{н}} \tag{22}$$

Однако функция softmax может интерпретировать результаты как вероятности только для входных данных действительных чисел, а для входных данных комплексных чисел необходим другой метод интерпретации. В этой статье мы предлагаем комплекснозначный метод интерпретации softmax для входных комплексных чисел. Результат функции softmax для входных комплексных чисел дает комплексные выходные значения, и поскольку сумма всех выходных значений равна 1+0j, мы предлагаем алгоритм, который вычисляет расстояние между выходными значениями и 1 + 0j в комплексной плоскости и выбирает индекс ближайшего значения. Это можно выразить уравнением (24).

$$расстояниеi = |So\ f\ tmax(Z)i \quad \omega| \tag{23}$$

$$ближайший_индекс = argmin(расстояние) \tag{24}$$

Здесь определено ω = 1 + 0j . Следовательно, мы решили классифицировать комплексное число соответствующий ближайшему_индексу как класс 1.

5. Анализ производительности. В

наших экспериментах по распознаванию речевых сигналов было проведено детальное сравнение результатов вещественных и комплексных нейронных сетей.
Анализ производительности проводился наряду с предварительной обработкой комплекснозначных кепстральных коэффициентов мел-частоты (MFCC) для уменьшения размерности. Кроме того, наблюдалось влияние комплексных MFCC на результаты распознавания речи. В частности, при распознавании заданных речевых сигналов с помощью нейронных сетей оценивалось влияние значений MFCC и их эффект комплексификации на производительность модели и сравнивалось с целью определить, можно ли повысить производительность при уменьшении количества параметров.

Экспериментальный метод: на рисунке 6 изображена архитектура многослойной комплекснозначной нейронной сети, используемой для обучения и тестирования с набором данных, состоящим из 8000 коротких речевых образцов, каждый из которых содержит восемь разных слов и длится менее одной секунды. Речевые данные, использованные в этом исследовании, являются частью набора данных речевых команд, собранного Google и распространяемого по лицензии CC BY.

Machine Translated by Google

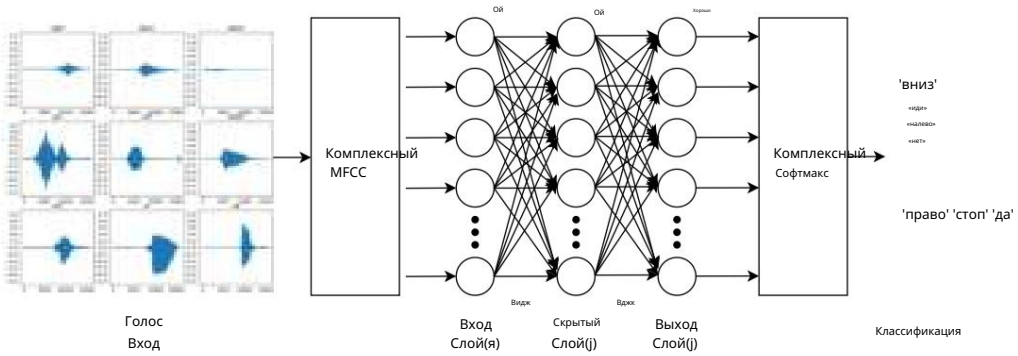


Рисунок 6. Экспериментальная установка многослойной CVNN.

Нейронные сети с комплексными значениями структурно похожи на нейронные сети с действительными значениями, но разница заключается в использовании комплексных чисел, которые вводят фазовые изменения, когда сумма произведений между входными данными и весами вращается в комплексной плоскости. Эти фазовые изменения играют важную роль в комплексных нейронных Комплексная нейронная сеть состоит из входного слоя с комплекснозначными MFCC уменьшенной размерности, полученными из речевых данных посредством процесса комплексификации, показанного на рисунке 3, скрытого слоя с 500 узлами и выходного слоя с 8 узлами для классификация. Все узлы и веса используют комплексные числа. Окончательные результаты классифицировались на выходном слое. Набор данных включал 7000 образцов для обучения и 1000 образцов для тестирования. В качестве данных использовались аудиоклипы из восьми слов, и эксперименты проводились путем деления входных данных в зависимости от того, применялось ли комплексное преобразование MFCC или нет. В случае использования вещественных MFCC использовалась вещественная нейронная сеть, тогда как комплексная нейронная сеть использовалась с комплексными MFCC. Нейронные сети как с действительными, так и с комплексными значениями обучались в одинаковых условиях: lr = 0,01, размер пакета = 50 и количество эпох = 50. Для входных данных использовались результаты трех различных этапов предварительной обработки уменьшения размерности. сравнивались с действительной нейронной сетью. Эти этапы предварительной обработки включали размеры MFCC 20, 10 и 5. Длина входного сигнала для нейронной сети может быть определена, как показано в уравнении (25). В частности, входной речевой сигнал был разделен на 32 кадра, и для каждого кадра генерировались MFCC.

$$\text{Длина данных входного сигнала} = \text{MFCC} \times \text{кадры} \quad (25)$$

Например, в случае комплексного значения MFCC 20 сначала вычисляется вещественное значение MFCC 40. Используя уравнение (25), длина входного сигнала для действительного MFCC 40 становится 40 × 32 = 1280. Если мы затем преобразуем эти 40 MFCC в комплексные числа, как показано на рисунке 3, уменьшение размерности будет достигнуто до половины длины. , в результате чего получилось 16 кадров. Следовательно, длина входного сигнала становится 40 × 16 = 640, что эквивалентно длине входного сигнала действительного MFCC 20. Аналогичное уменьшение размерности может быть выполнено для комплексного MFCC 10 и комплексного MFCC 5, используя тот же метод . . Общее количество параметров для всей нейронной сети, полученной в результате этого процесса, представлено в таблице 1. В этой таблице I(n) представляет количество нейронов во входном слое, H(n) представляет количество нейронов в скрытом слое. слой, а O(n) представляет количество нейронов в выходном слое.

Таблица 1. Общее количество параметров в нейронной сети.

МФЦК	Реальный MFCC	Комплекснозначный MFCC
20	I(20 × 32 = 640) × H(500) + H(500) × O(8) = 324000	I(40 × 16 = 640) × H(500) + H(500) × O(8) = 324000
10	I(10 × 32 = 320) × H(500) + H(500) × O(8) = 164 000	I(20 × 16 = 320) × H(500) + H(500) × O(8) = 164 000
5	I(5 × 32 = 160) × H(500) + H(500) × O(8) = 84 000	I(10 × 16 = 160) × H(500) + H(500) × O(8) = 84 000

Machine Translated by Google

MFCC 20 с комплексным значением и MFCC 20 с действительным значением. Как показано на рисунке 7, MFCC 20 с комплексным значением достиг максимальной точности обучения 99,40 % и точности тестирования 84,90 %, несмотря на ту же длину входных данных, что и вещественное значение. MFCC 20, который достиг максимальной точности обучения 98,13% и точности тестирования 83,10%. Результаты , полученные с использованием комплексных MFCC, превзошли результаты, полученные с использованием действительных MFCC, с точки зрения точности обучения и тестирования.

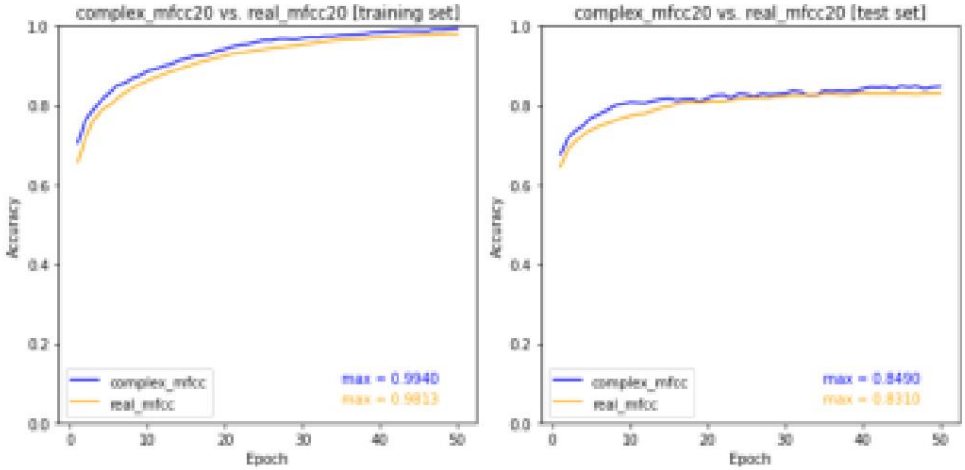


Рисунок 7. Сравнение комплексных и действительных результатов MFCC 20.

MFCC 10 с комплексным значением и MFCC 10 с действительным значением: Как показано на рисунке 8, MFCC 10 с комплексным значением достиг максимальной точности обучения 96,90% и точности тестирования 83,80%, несмотря на ту же длину входных данных, что и вещественное значение. MFCC 10, который достиг максимальной точности обучения 94,24% и точности тестирования 81,20%. Результаты , полученные с использованием комплексных MFCC, превзошли результаты, полученные с использованием действительных MFCC, с точки зрения точности обучения и тестирования.

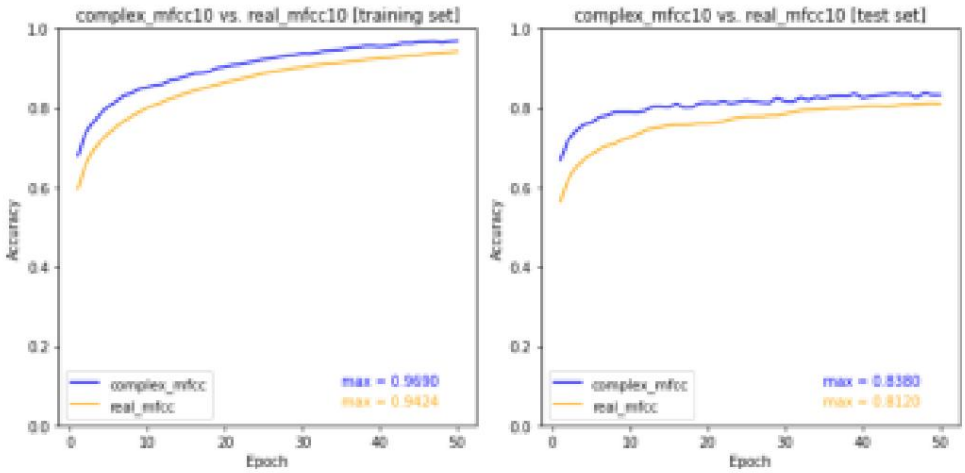


Рисунок 8. Сравнение комплексных и действительных результатов MFCC 10.

Комплексный MFCC5 и действительный MFCC5: Как показано на рисунке 9, комплексный MFCC 5 достиг максимальной точности обучения 92,33% и точности тестирования 81,40%, несмотря на ту же длину входных данных, что и вещественный MFCC 5. , который достиг максимальной точности обучения 86,34% и точности тестирования 74,90%. Результаты , полученные с использованием комплексных MFCC, превзошли результаты, полученные с использованием действительных MFCC, с точки зрения точности обучения и тестирования.

Machine Translated by Google

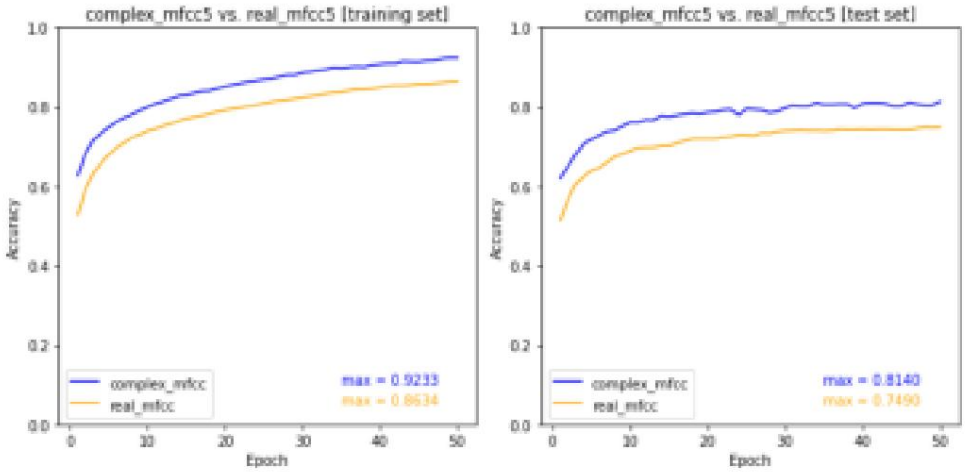


Рисунок 9. Сравнение комплексных и действительных результатов MFCC 5.

Общие результаты анализа: При рассмотрении экспериментальных результатов было замечено, что использование MFCC с комплексными значениями дает более высокую производительность по сравнению с использованием MFCC с действительными значениями, даже при одинаковой длине входных данных. Таким образом, было подтверждено, что использование комплексных MFCC и комплексных нейронных сетей приводит к лучшим результатам. Изменение точности в зависимости от количества функций MFCC показано на рисунке 10. Хотя здесь это не показано, те же результаты были получены с использованием оценки F1.

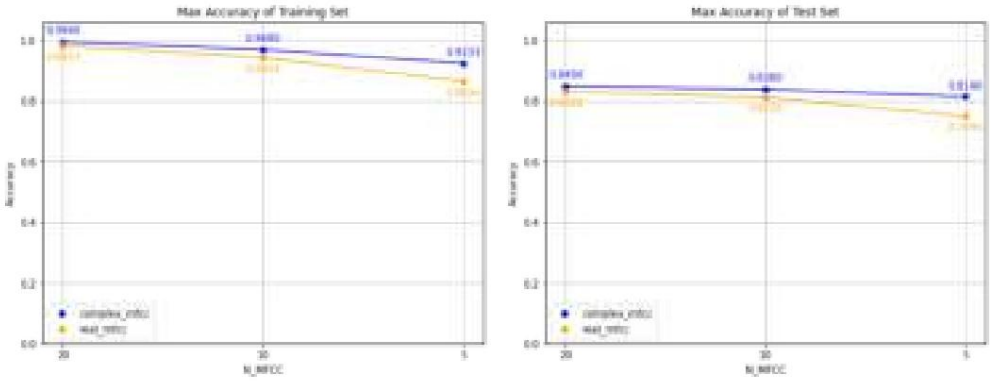


Рисунок 10. Изменение точности в зависимости от количества MFCC.

Если имеется достаточное количество MFCC, с увеличением эпох оказывается, что производительность при использовании действительных чисел начинает приближаться к производительности при использовании комплексных чисел. Однако на рисунке 9 показан другой результат. Комплексный MFCC 5 имеет 5 размерно уменьшенных комплексных MFCC, полученных из 10 MFCC посредством процесса комплексификации, показанного на рисунке 3. MFCC 5 с действительным знаком только с 5 MFCC не может догнать комплексный MFCC 5 даже при увеличении эпох. . Другими словами, когда количество MFCC невелико, даже при увеличении эпох результаты MFCC с действительным знаком не соответствуют производительности MFCC с комплексным знаком. Как показано на рисунке 10, несмотря на одинаковое количество параметров, результаты MFCC 5 с комплексными значениями превосходят результаты MFCC 5 с действительными значениями. Таким образом, было замечено, что при использовании MFCC с комплексными значениями изменение точности по мере уменьшения количества функций MFCC происходит более плавно по сравнению с использованием вещественных MFCC.

6. Выводы

В этой статье мы представили комплексные нейронные сети для решения проблемы распознавания речи. В наших экспериментах использовались комплексные MFCC в качестве входных данных дл

Machine Translated by Google

комплексная нейронная сеть. Результаты продемонстрировали превосходную производительность при использовании MFCC с комплексными значениями с более постепенным изменением точности по мере уменьшения количества функций MFCC по сравнению с использованием MFCC с действительными значениями. Это говорит о том, что комплексные MFCC могут помочь поддерживать производительность при меньшем количестве функций. Эти результаты дают ценную информацию о потенциале комплексных нейронных сетей и демонстрируют эффективность сочетания комплексных MFCC с комплексными нейронными сетями в распознавании речи.

Несмотря на эти многообещающие результаты, мы признаем определенные ограничения нашего подхода . Это исследование было проведено на ограниченном наборе данных, который может не полностью отражать разнообразие реальных речевых моделей. Более того, архитектура комплексной нейронной сети, описанная в этом исследовании, относительно проста и не может охватить все тонкости речевых сигналов.

Заглядывая в будущее, мы стремимся развивать эту работу для дальнейшего повышения точности и производительности моделей обработки и распознавания речевых сигналов. Наши будущие исследования будут сосредоточены на разработке более сложных моделей нейронных сетей с комплексными значениями и проведении экспериментов на различных наборах данных для повышения производительности. Кроме того, мы планируем изучить применение комплекснозначных нейронных сетей в других областях, помимо распознавания речи, потенциально расширяя их использование в различных областях обработки сигналов [16].

Вклад авторов: Концептуализация, SK; администрация проекта, депутат; надзор, депутат; написание — первоначальный вариант рукописи, С.К. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Эта работа проводилась при совместной поддержке Национального исследовательского фонда Кореи (NRF) посредством гранта, предоставленного правительством Кореи (MSIT) (грант № NRF-2023R1A2C1005461), а также MSIT (Министерством науки и информационных технологий), Корея. , в рамках Программы поддержки бизнеса по обучению талантов Convergence Security (IITP-2024-RS-2024-00426853), контролируемой IITP (Институтом планирования и оценки информационных и коммуникационных технологий).

Заявление о доступности данных: оригинальные материалы, представленные в исследовании, включены в статью. Дополнительные запросы можно направить соответствующим авторам.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Тебельскис, Дж. Распознавание речи с использованием нейронных сетей. Кандидат наук. Диссертация, Школа компьютерных наук, Университет Карнеги-Меллон, Питтсбург, Пенсильвания, США, 1995.
2. Саррофф, АМ Комплексные нейронные сети для аудио. Кандидат наук. Диссертация, Дартмутский колледж, Ганновер, Нью-Хэмпшир, США, 2018.
3. Абдул, З.К.; Аль-Талабани, А.К. Частотный кепстральный коэффициент Мела и его применение: обзор. IEEE Access 2022, 10, 122136–122158. [Перекрестная ссылка]
4. Баррачина, Дж.А.; Рен, К.; Вьейяр, Г.; Мориссо, К.; Оварлес, Дж. П. Теория и реализация комплексных нейронных сетей Сети. arXiv 2023, arXiv:2302.08286.
5. Айзенберг И. Комплекснозначные нейронные сети с многозначными нейронами; Шпрингер: Берлин/Гейдельберг, Германия, 2016.
6. Грейвс, А.; Джейтли, Н.; Мохамед, А.Р. Гибридное распознавание речи с глубоким двунаправленным LSTM. В материалах семинара IEEE 2013 г. по автоматическому распознаванию и пониманию речи, Оломоуц, Чешская Республика, 8–12 декабря 2013 г.; стр. 273–278. [Перекрестная ссылка]
7. Пас, П.; Гарридо, М. Эффективная реализация комплексных умножителей на ПЛИС с использованием срезов DSP. J. Сигнальный процесс. Сист. 2023, 95, 543–550. [Перекрестная ссылка]
8. Ахмад, М.; Чжан, Л.; Чоудхури, МЕН Реализация FPGA комплексной нейронной сети для полярно представленных Классификация изображений. Датчики 2024, 23, 2627. [CrossRef] [ПабМед]
9. Анггрени, Д.; Санджая, штат Вашингтон; Нурасийдик, МЮ; Мунаваро, М. Реализация распознавания речи с использованием мел- частотных коэффициентов кепстра (MFCC) и метода опорных векторов (SVM) на основе Python для управления манипулятором робота. Конференция ИОП. Сер. Матер. наук. англ. 2018, 288, 012042. [CrossRef]
10. Дханжал, А.С.; Сингх, В. Комплексное исследование автоматического распознавания речи с использованием нейронных сетей; Springer Nature: Берлин/Гейдельберг, Германия, 2023 г.
11. Смит, С.В. Руководство для ученых и инженеров по цифровой обработке сигналов, 2-е изд.; Калифорнийское техническое издательство: Сан-Диего, Калифорния, США, 1999 г.; стр. 557–558.
12. Бэсси Дж.; Ли, Х.; Цянь, Л. Обзор комплексных нейронных сетей. arXiv 2021, arXiv:2101.12249v1.
13. МОЙНейралНет. Комплексные нейронные сети — эксперименты. 2016. Доступно онлайн: http://makeourownneuralnetwork. blogspot.com/2016/05/complex-valued-neural-networks.html (по состоянию на 19 февраля 2021 г.).

Machine Translated by Google

14. Фишер Р. Исчисление Виртингера. В области предварительного кодирования и формирования сигналов для цифровой передачи; John Wiley & Sons, Inc.: Хобокен, Нью-Джерси, США, 2002 г.; стр. 405–413.
15. ПиТорч. Пакет автоматической дифференциации — torch.autograd, без даты Доступно в Интернете: https://pytorch.org/docs/stable/notes/autograd.html (по состоянию на 3 ноября 2023 г.).
16. Ли, К.; Хасэгава, Х.; Гао, С. Нейронные сети с комплексными значениями: комплексный обзор. IEEE/CAA J. Autom. Грех. 2022, 9, 1433–1454. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Рекомендации по работе: Сравнительный анализ совместной фильтрации Методы для объявлений

Роберт Квечинский 1,2, Томаш Гурецкий 2,*, Агата Филиповская 3 и Вячеслав Дубров 1

1 OLX Group, Królowej Jadwigi 43, 61-872 Познань, Польша; r.kwiecinskipl@gmail.com (PK); slavadubrov@gmail.com (ВД)
2 Факультет математики и информатики, Университет Адама Мицкевича, Uniwersytetu Pozna´nskiego 4, 61-614 Pozna´n, Польша
3 Кафедра информационных систем, Познаньский университет экономики и бизнеса, Ал. Niepodległo´sci 10, 61-875 Познань, Польша; agata.filipowska@ue.poznan.pl
* Переписка: tomasz.gorecki@amu.edu.pl

Аннотация: Объявления создают многочисленные проблемы для методов рекомендаций, включая временную видимость рекламы, анонимность большинства пользователей и тот факт, что обычно только один пользователь может потреблять рекламируемый товар. В этой работе мы решаем эти проблемы, выбирая модели и процедуры оценки, которые считаются точными, разнообразными и эффективными (с точки зрения потребления памяти и времени во время обучения и прогнозирования). Целью данной статьи является сравнение различных методов рекомендаций для объявлений о вакансиях на примере OLX Jobs для повышения коэффициента конверсии рекламы и удовлетворенности пользователей. В наших исследованиях мы реализуем масштабируемые методы и представляем различные подходы к рекомендациям: метод альтернативных наименьших квадратов (ALS), LightFM, Prod2Vec, RP3Beta и разреженные линейные методы (SLIM). Мы провели A/B-тестирование, отправив миллионы сообщений с рекомендациями провести онлайн-оценку выбранных методов. Кроме того, мы опубликовали набор данных, созданный для нашего исследования. Насколько нам известно, это первый набор данных такого рода. Он содержит 65 502 201 событие, выполненное в OLX Jobs 3 295 942 пользователями, которые взаимодействовали (отображали, отвечали или добавляли в закладки) 185 395 объявлений о вакансиях за две недели в 2020 году. Мы демонстрируем, что RP3Beta, SLIM и ALS при тестировании работают значительно лучше, чем Prod2Vec и LightFM. в лабораторных условиях. Онлайн-тесты A/B также показывают, что отправка сообщений с рекомендациями, сгенерированными моделями ALS и RP3Beta, увеличивает количество пользователей, обращающихся к рекламодателям. Кроме того, RP3Beta оказал на этот показатель на 20% более существенн

Ключевые слова: рекомендации по трудоустройству; объявления; совместная фильтрация; тестирование масштабируемости; А/Б-тесты; рекомендательные системы; матричная факторизация; рекомендательные системы на основе графов; Набор данных о взаимодействии с работой OLX ; показатели разнообразия

check for updates

Цитирование: Квечинский, Р.; Гурецкий, Т.; Филиповская А.; Дубров В. Рекомендации по работе: бенчмаркинг методов совместной фильтрации для Объявления. Электроника 2024, 13, 3049. https://doi.org/10.3390/электроника13153049

Академические редакторы: Марко Ломбарди, Доменико Сантаньелло, Зилу Лян и Мухаммад Хан

Поступила: 30 мая 2024 г.
Пересмотрено: 24 июля 2024 г.
Принято: 27 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Интернет-объявления — это веб-сайты, на которых рекламодатели размещают объявления о продаже или аренде услуг или продуктов [1]. В этой статье рассматривается онлайн-реклама на примере OLX Jobs.

Объявления стремятся удовлетворить потребности как рекламодателей, так и тех, кто отвечает на онлайн-рекламу. Чтобы соответствовать этим группам, более эффективные методы рекомендаций становятся все более важными. По сравнению с традиционной электронной коммерцией (например, Amazon (https://www.amazon.com/, по состоянию на 26 июня 2024 г.) или eBay (https://www.ebay.com/, по состоянию на 26 июня 2024 г.)), объявлениям приходится сталкиваться с различными проблемами, которые мы решаем в нашей процедуре оценки. Одним из них является временная видимость рекламы, которую мы решили решить путем частого переобучения модели. Следовательно, эффективность модели, о которой мы сообщаем, имеет важное значение для оценки. Другая проблема заключается в том, что рекламодателю обычно нужен только один пользователь для завершения транзакции. Этот аспект был решен путем включения анализа разнообразия в нашу оценку.

Machine Translated by Google

В наших исследованиях мы фокусируемся на подходах, возникающих на основе совместной фильтрации. Модели совместной фильтрации доказали свою точность, даже несмотря на то, что они не используют дополнительную информацию о пользователях или объектах, которую иногда трудно получить или обработать на практике. Целью данной статьи является оценка различных методов рекомендаций для использования в секторе объявлений. Высокая разреженность нашего набора данных, а также значительное количество рекламы и пользователей усложняют задачу исследования.

Ниже мы перечисляем наш основной вклад.

- Мы всесторонне оцениваем разнообразные, масштабируемые и совместные подходы к фильтрации в лабораторных условиях на предмет точности, разнообразия и эффективности. Мы оцениваем точность для разных групп пользователей в зависимости от количества элементов, с которыми они взаимодействовали, а также статистический анализ различий. Предлагаемая методология оценки решает основные проблемы, связанные с объявлениями.
- Мы представляем результаты двух онлайн-A/B-тестов, демонстрирующих влияние моделей ALS и RP3Beta на пользователей OLX. В каждом тесте участвовало более 1 миллиона пользователей. Сообщаемое влияние может быть использовано для оценки важности рекомендательных систем в сфере объявлений и позволит другим компаниям принимать более обоснованные решения о разработке таких моделей.
- Мы публикуем и представляем набор данных о взаимодействии на работе. Насколько нам известно, это самый крупный общедоступный набор данных о взаимодействии на рабочем месте.

Данная статья состоит из семи разделов. Раздел 2 представляет собой обзор литературы по основным разработкам. В разделе 3 описан набор данных, разработанный и опубликованный на Kaggle. План лабораторного эксперимента, а также реализация методов описаны в разделе 4. Раздел 5 посвящен результатам лабораторной оценки наших реализаций, а также статистической значимости результатов. В разделе 6 представлены результаты онлайн-внедрений, выполненных в ходе A/B-тестирования с пользователями, а также обсуждение результатов. Выводы представлены в разделе 7.

2. Сопутствующая работа

В этом разделе мы начнем с описания уникальных характеристик объявлений и областей вакансий. Затем мы предоставим обзор рекомендательных систем. Далее мы обсудим факторы, которые влияют на наш выбор моделей, и предоставим подробное описание каждой модели. Наконец, мы указываем пробелы в исследованиях.

2.1. Сайты объявлений. Сайты

объявлений (объявления) представляют собой онлайн-версии разделов газетной рекламы с объявлениями, касающимися транспортных средств, недвижимости, трудоустройства, домашних животных и т. д. Объявления представляют собой особый тип сайтов электронной коммерции, поскольку они не поддерживают завершение транзакций между сторонами — вместо этого они используются для общения о продукте или услуге и для назначения встреч, пока транзакция выполняется в автономном режиме; это отличает их от традиционных сайтов электронной коммерции, таких как eBay или Amazon [2].

Наши исследования сосредоточены на компании OLX, которая присутствует более чем в 30 странах мира. Одним из типов объявлений, публикуемых на объявлениях, являются предложения о работе. В ходе нашего исследования мы выявили несколько характеристик, которые отличают объявления о вакансиях на рекламных объявлениях, например, от объявлений о продаже автомобилей или недвижимости:

- Требования к рекламе указывают тип пользователя, который может реагировать на рекламу, а не все пользователи обладают необходимыми компетенциями.
- Более интенсивное вовлечение пользователя в процесс поиска, поскольку поиск работы является первоочередной необходимостью по сравнению, например, с покупкой автомобиля.
- Местоположение работы существенно влияет на пользователя: чтобы купить мобильный телефон или автомобиль, пользователь может путешествовать или воспользоваться службой доставки посылок для доставки товара. Тем не менее, работа требует переезда, если она находится далеко от дома пользователя.
- С технической точки зрения описание вакансии имеет множество качественных аспектов, и в зависимости от должности различия могут быть существенными даже между объявлениями одной и той же компании.

Machine Translated by Google

Помимо этих особенностей, в ходе нашего исследования мы также выявили следующие проблемы, связанные с методами рекомендаций по объявлениям, которые повлияли на нашу реализацию этих методов:

- Количество пользователей и предложений: с помощью OLX Jobs мы имели дело с миллионами пользователей и десятками тысяч предложений, которые постоянно менялись. •
Пользователю не нужно создавать профиль для взаимодействия с рекламой. Это означает, что информация из профиля не может быть использована для рекомендаций, а рекомендации определяются поведением пользователя во время просмотра.
- Объявление обычно имеет ограниченное количество посетителей (взаимодействующих с объявлением), поскольку объявление обычно касается уникального предложения. После устранения потребности реклама отключается. • Нет информации о том, произошла ли транзакция; на объявлениях конверсия связана с получением ответа на опубликованное объявление, поскольку любая транзакция происходит офлайн.

Эти проблемы создают трудности не только для реализации, но и для проверки точности предлагаемых методов. Мы учли эти проблемы при подготовке набора данных, а также при реализации и оценке методов, описанных в этой статье.

2.2. Рекомендательные системы

Было показано, что проблема избыточного выбора [3] может снизить потребление продукции [4] и снизить доход сайта [5]. Поставщики могут запрашивать у своих пользователей дополнительную информацию об их ожиданиях в отношении товаров, чтобы уменьшить количество представленных вариантов. Даже если пользователь готов потратить время на предоставление этой информации, количество возможностей обычно слишком велико, чтобы представить его пользователю. Чтобы решить эту проблему, многие компании, такие как Netflix [6] и Amazon [7], создали системы персонализированных рекомендаций, роль которых заключается в предложении пользователям соответствующих товаров [8].

Рекомендации используются во многих областях, включая музыку, туризм, моду и еду. Многочисленные исследования были проведены в области электронной коммерции [9], объявлений [10] и рекомендаций по работе [11–13]. Ли и др. [14] значительно улучшили свои рекомендации в LinkedIn, используя глубокое трансферное обучение для создания моделей понимания работы, специфичных для конкретной предметной области. Чжу и др. [15] предложили модель междоменных рекомендаций, использующую гетерогенный граф и алгоритм обнаружения сообществ для рекомендации курсов и вакансий. Лачич и др. [16] оценили различные архитектуры автокодировщиков для рекомендаций по заданиям на основе сеансов и представили результаты на популярных наборах данных рекомендаций по заданиям.

Обычно мы различаем два подхода к предоставлению рекомендаций: рекомендации на основе контента и совместную фильтрацию [8].

В рекомендациях на основе контента [17] мы используем информацию о пользователях (например, хобби, образование, навыки) и товарах (например, цена, местоположение, категория). Эти системы обычно используют историю взаимодействия (рейтинги, посещения, покупки и ответы) одного пользователя, но не учитывают рейтинги всех пользователей одновременно.

Сбор/обработка информации о пользователях и объектах не всегда осуществим. В подходе совместной фильтрации [18,19] мы используем только информацию о взаимодействии пользователя между элементами. Даже не имея информации о функциях пользователя и характеристиках товаров, эти системы могут дать очень точные рекомендации. Мы выбрали этот подход из-за отсутствия пользовательских функций и сложности превзойти совместную фильтрацию за счет использования функций элементов в производственной среде (подробнее об этом мы поговорим в разделе 6.3).

Одним из наиболее существенных недостатков совместной фильтрации является невозможность предоставления рекомендаций относительно пользователей и элементов без взаимодействия, известная как проблема холодного запуска [20]. Одним из возможных решений этой проблемы является частое переобучение модели. Учитывая множество взаимодействий с пользователем, эта стратегия требует сосредоточения внимания на масштабируемости модели [21].

Еще одним важным фактором, определяющим выбор модели, является тип взаимодействия пользователя и предмета [20]. Мы называем обратную связь явной, когда пользователь явно указывает свое предпочтение предмету (например, оценивает его по шкале от 1 до 5). Другой тип — неявная обратная связь, когда мы можем только предположить предпочтения пользователя по отношению к элементу на ос

Machine Translated by Google

о поведении пользователя (например, покупка товара или просмотр фильма). Неявная обратная связь менее точна, более обширна и обычно выражает только благоприятные предпочтения [22].

Лабораторные оценки рекомендательных систем обычно основаны на показателях точности [23]. В последнее время большое внимание уделяется другим аспектам, таким как разнообразие, неопределенность, новизна и охват [24,25]. В случае с объявлениями элемент обычно может быть использован только одним пользователем, поэтому нам следует избегать рекомендации одного и того же элемента значительному числу пользователей. Следовательно, необходимо сосредоточиться на охвате каталога [26] . Одной из наиболее сложных задач при оценке рекомендательных систем является поиск офлайн-метрики, наиболее коррелирующей с онлайн-целью [27].

В этой статье мы сравниваем модели совместной фильтрации для значительного набора данных неявной обратной связи. Мы сравниваем точность, разнообразие и эффективность выбранных рекомендательных методов в лабораторных условиях, а затем доказываем качество методов в онлайн-тестах с пользователями.

2.3. Методы рекомендаций

Ниже мы перечисляем наиболее важные аспекты, влияющие на выбор метода нашего исследования.

- Популярность в литературе. Мы выбрали методы, эффективность которых была тщательно оценена. Все выбранные методы, кроме одного, цитировались как минимум сотни раз.
- Популярность в индустрии. Оценка этого аспекта была сложной задачей, поскольку компании редко делятся результатами проведенных экспериментов. Некоторые показатели популярности включают доступность реализации и количество людей, которые ее используют (например, это отражается в количестве звезд и вилок на GitHub). Мы также рассмотрели опыт специалистов по обработке данных, работающих в нашей компании, которые в прошлом применяли большинство выбранных методов для некоторых других случаев использования.
- Масштабируемость. Мы рассматривали только решения, способные обрабатывать миллионы пользователей и объектов. Кроме того, мы уменьшили проблему холодного запуска за счет частого переобучения модели . Следовательно, нам пришлось выбирать методы для обучения на нашем наборе данных в течение нескольких часов.
- Разнообразие подходов. Мы хотели протестировать методы, представляющие разные «семейства», чтобы оценить их применимость к классифицированным объявлениям. Мы планировали выявить наиболее перспективное семейство подходов для продолжения исследований в этом направлении.

Наконец, для тестирования и оценки были выбраны ALS, LightFM, Prod2Vec, RP3Beta и SLIM . В следующих подразделах описываются причины выбора каждого конкретного метода и его определение. Общим ограничением этих методов является их неспособность рекомендовать элементы, которые не посещал ни один пользователь, и генерировать рекомендации для пользователей , которые не посещали ни один элемент (проблема холодного запуска, [21]). С нашим набором данных, без особенностей пользователя и элементов, невозможно решить эту проблему каким-либо персонализированным методом.

Отобранные методы уже сравнивались в разных областях. Анелли и др. [28] сравнили RP3Beta, SLIM, ALS и BPRMF (что эквивалентно LightFM при использовании функции потерь BPR) на трех наборах данных: Amazon Digital Music, Epinions и MovieLens 1M, с точки зрения точности, разнообразия и эффективности. Количество взаимодействий в этих наборах данных составляет соответственно 145 523, 300 475 и 571 531. В нашем исследовании использовался набор данных, содержащий десятки миллионов взаимодействий, аналогичный тем, которые используются в производственных системах OLX.

Кроме того, мы предоставляем результаты онлайн-оценки.

Дакрема и др. [29] сравнили точность SLIM, RP3Beta и ALS на нескольких наборах данных (CiteULike-a, MovieLens 1M, Pinterest, Amazon Music, Amazon Movies, MovieLens 20M, Netflix, Yelp и Gowalla). Они показали, что каждый из этих алгоритмов превосходил два других по крайней мере на одном наборе данных. Некоторые из их выводов были дополнительно подтверждены Anelli et al. в наборах данных MovieLens 1M и Pinterest. [30]. Чжу и др. [31] сравнили BPRMF, SLIM и Prod2Vec на трех наборах данных (Amazon Books, Yelp и Gowalla), показав превосходство SLIM. Донг и др. [32] проанализировали точность RP3Beta, SLIM и ALS на девяти наборах данных, при этом SLIM показал лучшие результаты на восьми наборах данных и ALS.

Machine Translated by Google

на одной. В отличие от нашего исследования, в этих статьях выбранные методы не сравнивались с точки зрения эффективности.
разнообразие или эффективность, а также не проводили онлайн-оценку.

В последнее время появилось несколько методов рекомендаций, основанных на нейронных сетях.
предложенный. Было показано, что адекватно определенные и настроенные простые модели часто
превосходят их [28,31–33]. Кроме того, из-за вычислительной сложности модели, основанные на
нейронных сетях реже тестируются на наборах данных, содержащих десятки миллионов
взаимодействия [28]. Следовательно, мы оставляем на будущее исследование, чтобы предложить более продвинутые и
масштабируемые методы, которые обеспечивают превосходные результаты в нашем наборе данных.

2.3.1. Модели матричной факторизации

Методы матричной факторизации успешно используются в рекомендательных системах
после конкурса Netflix Prize [34]. Основная идея заключается в аппроксимации разреженных
Матрица взаимодействия пользователя и элемента как произведение двух меньших плотных матриц. Мы оцениваем два
подходы матричной факторизации: LightFM и ALS.

Модель LightFM [35] была предложена для решения проблемы холодного запуска методов матричной
факторизации. В нашем случае это сводится к классической матричной факторизации без дополнительных
информация о пользователях и предметах. Несмотря на это, мы решили использовать этот подход из-за
его эффективная официальная реализация при поддержке нескольких функций потерь: логистика, BPR [36],
WARP [37] и -OS WARP [38].

Пусть U — набор пользователей, а I — набор элементов. Для данного пользователя € U и предмета € I мы
define = 1, если пользователь взаимодействует с элементом, , и = 0 в противном случае. Модель ЛайтФМ
прогнозирует оценку следующим образом:

$$\hat{} = \cdot + +\quad ,$$

где и обозначают трехмерные скрытые представления пользователя и элемента, а и — предвзятости
пользователя и элемента. Все эти параметры изучаются во время
тренировка.

Функцию логистических потерь следует использовать, когда как положительная, так и отрицательная обратная связь
типы доступны, поэтому он не подходит для нашего набора данных, что мы подтвердили
экспериментально во время настройки гиперпараметра.

Функции потерь BPR, WARP и -OS WARP представляют собой парные подходы к ранжированию,
которые более подходят для нашей задачи по выработке лучших рекомендаций. Они пытаются изменить
параметры модели, чтобы максимизировать разницу между оценками элементов D с
с каким пользователем взаимодействует и элементов D, с которыми пользователь не взаимодействует.

В частности, в функции потерь BPR [36] мы пытаемся максимизировать выражение:
следующее:

$$\ln(\hat{}\quad \hat{}) \quad \Theta\,||\Theta||2\ ,$$

$$(,,)\ S$$

где

$$S\ \text{знак равно}\ \{(,,)|\quad U,\quad D\ \text{и}\quad D\},$$

$$\text{" ="}\quad \frac{1}{1+}\ ,$$

Θ — параметр регуляризации, специфичный для модели, Θ — вектор параметров, а || · || это
Евклидова норма.

Мы представляем WARP [37] как случай WARP -OS [38]. Функция потери -OS WARP
сумма потерь для каждого пользователя:

$$(\hat{}\ ,Д),$$

$$€U$$

где ˆ — вектор оценочных оценок ˆ для каждого элемента I.

Machine Translated by Google

Для данного , мы упорядочиваем взаимодействующие элементы D в порядке убывания пользователя расчетные баллы. А именно, мы упорядочиваем элементы как 1, 2, . . . , | Д | , где индексы выбрано таким образом, что имеем следующее:

$$\hat{}_1 \quad \hat{}_2 \quad \ldots \quad \hat{}_{|Д|}.$$

Тогда у нас есть следующее:

$$(\hat{} , Д) = \frac{1}{} \sum_{=1}^{| Д |} ( ) Ф(\text{ранг} (\hat{})),$$

где ранг (ˆ) = I\D (1 + ( ) ˆ ˆ), является индикаторной функцией, Ф( ) = $\sum_{=1} \frac{1}{}$, и = $\sum_{=1}^{| Д |}$ нормализует веса, индуцированные .

На практике мы выбираем положительный элемент, касающийся весовой функции, и случайным образом выбирать невидимые предметы до тех пор, пока после выборки не будет 1 + . Затем мы выполняем шаг градиента для минимизации Ф($\frac{|Я\Д|}{}$) (1 + ˆ).

Наконец, потеря -OS WARP достигается для ( ) = 1 и ( ) = 0, если . ВАРП потеря достигается, когда мы устанавливаем ( ) = 1 для N. Мы можем пропустить заказ положительных элементов из D, чтобы упростить вычисления.

В нашем исследовании учитывались все функции потерь (логистика, BPR, WARP и -OS WARP). в качестве гиперпараметров модели LightFM и оптимизируются во время обучения.

Модель ALS (попеременные наименьшие квадраты), также известная как WRMF – взвешенная регуляризованная матричная факторизация [39], является примером метода матричной факторизации, разработанного для наборов данных неявной обратной связи. Мы выбрали этот метод из-за его доказанной эффективности. масштабируемость и популярность в исследованиях и промышленности. Кроме того, модель БАС имеет уже реализован и используется в OLX.

Мы определяем достоверность наблюдения как = 1 + , где – гиперпараметр . Мы также определяем предпочтение как = (> 0). Затем в ALS мы оптимизируем следующее выражение:

$$\text{мин} \quad (-)^2 + (\quad | | | |2 + \quad | | | |2),$$

где — трехмерное внедрение пользователя, — трехмерное внедрение элемента, и является параметром регуляризации. Название модели относится к оптимизированному методу. для этой функции. Процедура попеременных наименьших квадратов предполагает, что пользователь или элемент скрыты. факторы фиксированы и аналитически определяют остальные. Процедура эффективно повторяется несколько раз [39,40].

В подходах матричной факторизации встраивание для каждого пользователя изучается отдельно, что может привести к увеличению использования времени и памяти, если число пользователей велико. Кроме того, рекомендации могут быть необъяснимыми, поскольку два пользователя с одни и те же взаимодействия могут получить разные рекомендации. Другие рассмотренные модели (SLIM, RP3Beta, Prod2Vec) предоставляют одинаковый набор рекомендаций для пользователей с те же взаимодействия.

2.3.2. Модели на основе соседства

Модель SLIM (разреженные линейные методы для рекомендательных систем Top-N [41]) на основе регрессии ближайших соседей на основе элементов [8]. Мы выбрали этот метод из-за его преимущества по сравнению с моделями ближайших соседей на основе элементов и моделями матричной факторизации [41] и хорошие результаты по сравнению с ненейронными и нейронными подходами [29].

В этой модели мы аппроксимируем матрицу пользователь-элемент, изучая матрицу элемент-элемент. разреженная матрица сходства. Явно это изучается путем минимизации следующего выражения:

$$\frac{1}{2} | | \quad | |2 + \frac{}{| | 2} \quad | |2 + | | \quad | | |1,$$

Machine Translated by Google

где 0 и = 0 для всех , , 1 , |I|, | | · | | – матричная норма Фробениуса:

$$||\quad||=\sqrt{\sum_{=1}^{|Я|}\sum_{=1}^{|Я|}\quad^{2}},$$

| | · | |1 — поэлементная ℓ1-норма:

$$||\quad||1=\sum_{=1}^{|Я|}\sum_{=1}^{|Я|}\left|\quad\right|$$

и – параметры регуляризации.

Мы видим, что каждый столбец можно изучать независимо, что улучшает масштабируемость этой модели [41].

Ограничением этого метода является то, что он не генерирует внедрения пользователей и элементов, что может быть полезно для других задач (например, кластеризации пользователей или элементов). Мы можем рассматривать столбцы матрицы как вложения элементов, но они представляют собой разреженные |I|-мерные векторы, поэтому их, вероятно, нельзя напрямую использовать в других задачах.

2.3.3. Подходы на основе графов

RP3Beta [42] — это подход совместной фильтрации на основе графов. Аналогично Dacrema et al. В работе [29] мы представляем эту модель как расширение модели P3alpha [43]. Мы рассматриваем двудольный и неориентированный граф = (V, E), где каждая вершина V представляет либо пользователя, либо элемент, а каждое ребро { , } E представляет взаимодействие между данным пользователем и данным элементом. Пусть – матрица смежности и матриц Затем оценки P3alpha сохраняются в матрице

$$((\quad^{1})\quad)^{3},$$

где обозначает поэлементную степень матрицы (степень Адамара). Точнее, оценка ˆ сохраняется на пересечении строки, представляющей пользователя, и столбца, представляющего элемент. Мы видим, что эта оценка представляет собой сумму оценок, присвоенных путям длиной 3, соединяющим пользователя и элемент в графе.

В RP3Beta оценка вычисляется с помощью P3alpha и делится на популярность элементов, возведенную в степень, чтобы смягчить смещение популярности.

Мы выбрали эту модель из-за ее простоты, масштабируемости и проверенной производительности [29,30], особенно с учетом высокой разреженности нашего набора данных, что значительно снижает использование памяти при обучении модели. Это позволяет нам напрямую вычислять и сохранять матрицу сходства между предметами в памяти вместо аппроксимации случайных блужданий, предложенной в оригинальной статье.

RP3Beta, как и SLIM, не создает встраивания пользователей и элементов. Еще одним связанным ограничением является невозможность рекомендовать элементы, расстояние которых на двудольном графе пользователь-элемент больше трех от данного пользователя. В частности, элемент может быть рекомендован пользователю только в том случае, если хотя бы один пользователь ′ посетил элемент и хотя бы один из элементов, посещенных пользователем . Кроме того, модель RP3Beta не имеет параметров, оптимизированных во время обучения; следовательно, он может быть не в состоянии использовать сложные отношения в данных.

2.3.4. Прод2Век

Модель Prod2Vec [44] (или Item2vec [45]) основана на модели Word2Vec [46], широко используемой при обработке естественного языка. Модель фокусируется на изучении векторных представлений элементов, используя последовательность элементов как «предложение», а элементы внутри этой последовательности как «слова». Этот метод был выбран потому, что он масштабируем и сильно отличается от сравниваемых методов. Кроме того, этот подход основан на современной системе рекомендаций гибридных глубоких нейронных сетей, используемой в других категори

Machine Translated by Google

Prod2Vec использует модель пропуска граммы, максимизируя целевую функцию по множеству S последовательностей элементов, определенных следующим образом:

$$= \sum_{\in S} \sum_{\in} \text{журнал } P(+ \mid),$$

$$0$$

где длина контекста для последовательностей элементов, а элементы из одной и той же последовательности заказано произвольно. Вероятность P(+ |) наблюдения соседнего элемента + при условии текущий элемент определяется с помощью функции soft-max:

$$P(+ \mid \quad " =" \frac{\text{опыт}(\quad ' \quad)}{\sum_{=1}^{|\text{Я}|} \text{опыт}(\quad ' \quad)},$$

где и — входные и ' выходные векторные представления элемента и |I| это , количество уникальных слов в словаре. Prod2Vec моделирует контекст последовательности элементов, где элементы со схожим контекстом (т. е. с похожими соседними взаимодействиями) имеют схожие векторные представления.

Прежде чем генерировать рекомендации, мы нормализуем все представления входных векторов, что повышает эффективность. Для формирования рекомендаций мы представляем пользователя как среднее значение входных векторных представлений взаимодействующих элементов. Затем ищем предметы чьи входные представления имеют наибольшее косинусное сходство с пользовательским представлением (-ближайшие соседи в скрытом пространстве).

Потенциальным ограничением этого метода является строгая зависимость между пользователем и элементом. представления. Некоторые взаимодействия потенциально должны оказывать большее влияние на представления пользователей, чем другие (что возможно в моделях матричной факторизации).

2.4. Пробел в исследованиях

Целью данной статьи является сравнение различных методов рекомендаций в области классификаций должностей, устраняя конкретный пробел в исследованиях. Насколько нам известно, рассматриваемые методы рекомендаций никогда не тестировались и не оценивались на крупных масштабируйтесь, работая онлайн в своей сфере деятельности. В данной работе мы сравниваем эти методы в лабораторных условиях с использованием опубликованного нами набора данных и онлайн-тестов A/B, проведенных в производственная среда.

3. Набор данных о взаимодействии с работой OLX

Это исследование предоставляет коммерческий набор данных, содержащий 65 502 201 событие, выполненное на OLX Jobs 3 295 942 пользователей, которые взаимодействовали с 185 395 объявлениями о вакансиях за две недели в 2020 году. Мы опубликовали набор данных на Kaggle (https://www.kaggle.com/olxdatascience/olx-jobs-interactions , доступ 26 июня 2024 г.).

Каждая строка набора данных представляет событие, которое данный пользователь выполняет относительно данного объявления в данный момент времени (табл. 1). Были включены следующие события, связанные с поведением пользователей:

- клик: пользователь посетил страницу с описанием товара;
- закладка: пользователь добавил объект в закладки;
- chat_click: пользователь открыл чат, чтобы связаться с владельцем предмета;
- contact_phone_click_1: пользователь раскрыл номер телефона, прикрепленный к элементу;
- contact_phone_click_2: пользователь нажал кнопку, чтобы позвонить владельцу элемента;
- contact_phone_click_3: пользователь нажал, чтобы отправить SMS-сообщение владельцу элемента;
- contact_partner_click: пользователь нажал, чтобы получить доступ к внешней странице владельца элемента;
- contact_chat: пользователь отправил владельцу элемента сообщение.

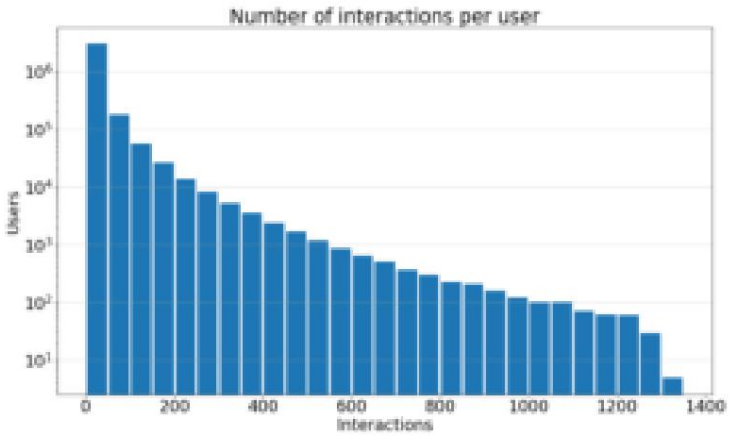
Machine Translated by Google

Таблица 1. Примеры строк из набора данных.

Пользователь	Элемент	Событие	Временная метка
1745587	168661		1582216025
843008	62838		1582485868
14285	30469	нажмите	1582247367
1142944	80122		1581805847
2835659	23728	нажмите закладку нажмите чат_клик	1582397836

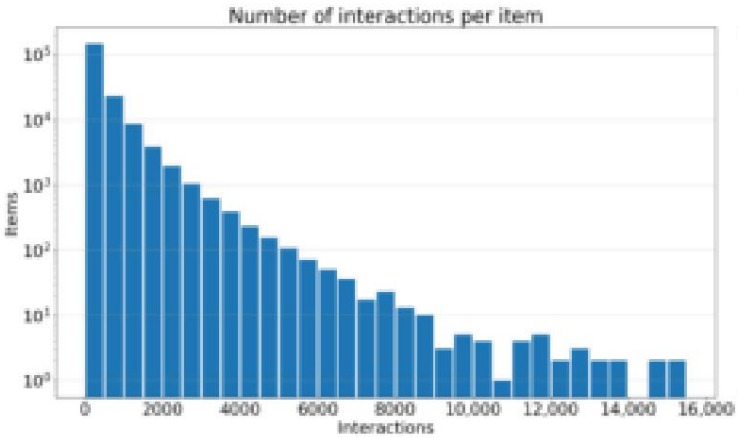
Источник: собственное исследование.

Разреженность данных (% отсутствующих записей в матрице пользователь-элемент) равна 99,9923%, что очень мало, но типично для классификаторов (для сравнения: разреженность данных на Набор данных MovieLens 20M составляет 99,47%, что делает его почти в 70 раз плотнее). Распределение взаимодействия на пользователя, взаимодействия на элемент и события представлены на рисунках 1 и 2. и таблица 2 соответственно. Мы использовали логарифмическую шкалу, поскольку распределения сильно выражены. перекошенный вправо.



Quantile	Value
0.01	1
0.10	1
0.25	2
0.50	6
0.75	18
0.90	46
0.99	221

Рисунок 1. Количество взаимодействий на пользователя. Источник: собственное исследование.



Quantile	Value
0.01	1
0.10	1
0.25	3
0.50	137
0.75	428
0.90	950
0.99	2944

Рисунок 2. Количество взаимодействий на элемент. Источник: собственное исследование.

По данным Де Рюйта и Бхулаи [47], наиболее популярными наборами данных, используемыми для оценки систем рекомендаций по вакансиям, являются CareerBuilder12 (https://www.kaggle.com/конкурсы/рекомендации по работе, по состоянию на 16 июня 2024 г.), RecSys16 [12] и Rec-Sys17 [13]. В таблице 3 представлена статистика рассматриваемых наборов данных. Количество уникальных Взаимодействия — это количество уникальных пар пользователь-элемент, в которых пользователь взаимодействовал с элемент. Несмотря на то, что исследователи используют наборы данных RecSys16 и RecSys17, они публикуются во время соревнований по науке о данных в строго ограниченных условиях и не являются дольше в наличии. Следовательно, мы не смогли подсчитать количество уникальных взаимодействий для этих

Machine Translated by Google

наборы данных. Мы заметили, что набор данных о взаимодействии с работой OLX имеет наибольшее количество взаимодействия и типы между рассматриваемыми наборами данных.

Таблица 2. Распределение событий.

Действие	Частота (%)
	89,794
нажмите contact_phone_click_1	2,628
закладку	2.511
chat_click	2.136
contact_chat	1,448
contact_partner_click	0,701
contact_phone_click_2	0,679
contact_phone_click_3 Источник:	0,103

собственное исследование.

Таблица 3. Сравнение наборов данных о вакансиях. Буква обозначает стандартное отклонение.

	OLX Вакансии Взаимодействия	CareerBuilder12	RecSys16	RecSys17
Взаимодействия	65,50 M	1,60 M	8,83 M	8,27 M
Уникальные взаимодействия	47,17 M	1,60 M	-	-
Пользователи	3,30 M	0,3 2M	1,37 M	1,50 M
Среднее количество уникальных взаимодействий на пользователя	14.31 (= 29,23)	4,99 (= 11,42)	<6,46	<5,51
Предметы	0,19 M	0,37 M	1,36 M	1,31 M
Среднее количество уникальных взаимодействия на элемент	254,42 (= 426,02)	4.38 (= 8,19)	<6,50	<6,31
Плотность матрицы взаимодействия (%)	0,0077%	0,0014%	<0,0005%	<0,0004%
Виды взаимодействия	8	1	4	5
Временные метки				
Возможности пользователя				
Особенности товара				

Источник: собственное исследование.

Следует отметить, что сохранение конфиденциальности рекламы и пользователей было приоритетом. при подготовке этого набора данных. Меры, принятые для защиты конфиденциальности, включали следующее:

- Исходные идентификаторы пользователей и элементов были заменены уникальными случайными целыми числами;
- К каждой временной метке было добавлено какое-то нераскрытое постоянное целое число;
- Некоторые части взаимодействий были отфильтрованы;
- Добавлены некоторые дополнительные искусственные взаимодействия.

4. Экспериментальная установка

4.1. Выполнение

Мы реализовали упомянутые выше методы на основе предыдущих разработок. Подробности о том, чем наши реализации отличаются от ранее протестированных подходов, см. представлены в таблице 4. Реализация методов для обеспечения воспроизводимости результатов доступно на GitHub (https://github.com/rob-kwiec/olx-jobs-recommendations, доступ 29 мая 2024 г.).

Machine Translated by Google

Таблица 4. Методы, повторно использованные для исследования.

Метод	Семейство методов	Источник	Отличие от Оригинальный метод
Факторизация матрицы LightFM		Реализация на основе: [48]. Сопроводительный документ: [35].	Никто
БАС	Матричная факторизация	Реализация на основе: [49]. Подтверждающие документы: [39,40].	Никто
СТРОЙНЫЙ	На основе района	Источники, вдохновившие нас на реализацию: [50,51]. Сопроводительный документ: [41].	Никто
RP3Beta	На основе графа	Источник, вдохновивший на нашу реализацию: [52]. Подтверждающие документы: [29,53].	Выполняется напрямую вычисления на разреженных матрицах вместо случайных приближение прогулок.
Прод2Век	Word2Vec	Реализация на основе: [54]. Подтверждающие документы: [44,45].	Последовательности взаимодействий, упорядоченные по временной метке. Представление пользователя как среднего из представлений взаимодействующих элементов. Мы также экспериментировали с CBOW (непрерывным набором слов).

Источник: собственное исследование.

4.2. Цели лабораторных исследований и онлайн-тестирования Нашей

конечной бизнес-целью было увеличить количество пользователей, подающих заявки на работу через OLX Jobs. Чтобы добиться этого, мы пытаемся решить проблему ранжирования; точнее, мы рекомендуем десять элементов с наибольшей вероятностью получения взаимодействия с пользователем. Вопрос поиска офлайн-метрики, наиболее коррелирующей с уровнем заявок в сфере работы, рассматривался Mogenet et al. [27]. Поскольку наша цель аналогична, мы последовали их результатам, оптимизировав точность @ 10 (точность для первых десяти рекомендуемых элементов [23]). Однако мы также сообщаем и обсуждаем другие показатели точности, разнообразия и эффективности.

Мы ориентированы на пользователей, которые активно ищут работу, взаимодействуя с объявлениями о работе. Поэтому мы не рассматриваем пользователей без взаимодействия (пользователей с холодным запуском).

4.3. Разделение данных

Набор данных, описанный в разделе 3, был разделен на наборы данных для обучения-проверки и тестирования, как того требует методология исследования. В тестовый набор входит 20% новейших взаимодействий. Это означает, что из 14 дней в тестовый набор вошли примерно 3 дня (см. рисунок 3). По мнению Meng et al., такой способ разделения набора данных на обучающий и тестовый наборы является наиболее реалистичным . [55], которые изучили несколько стратегий разделения для оценки рекомендательных систем.

В обоих наборах данных мы также ограничили количество взаимодействий уникальными взаимодействиями между пользователем и элементом. Если пользователь взаимодействовал с элементом более одного раза, учитывалось только первое взаимодействие, и с элементом была связана временная метка этого взаимодействия. Мы также не различали типы взаимодействий. Кроме того, были применены еще две модификации: в тестовый набор были включены только пользователи, которые фигурировали в наборе обучающих и проверочных данных (поскольку мы не могли предоставлять персонализированные рекомендации другим пользователям без дополнительных знаний о них), и мы отфильтровывали все пользовательские данные. пары элементов , которые присутствовали в обучающем наборе (чтобы не рекомендовать элементы, которые уже были просмотрены). В результате у нас получилось около 38 М строк в обучающем и проверочном наборе и около 6 М строк в тестовом наборе. Дополнительная статистика этих двух наборов данных представлена в Таблице 5.

Machine Translated by Google

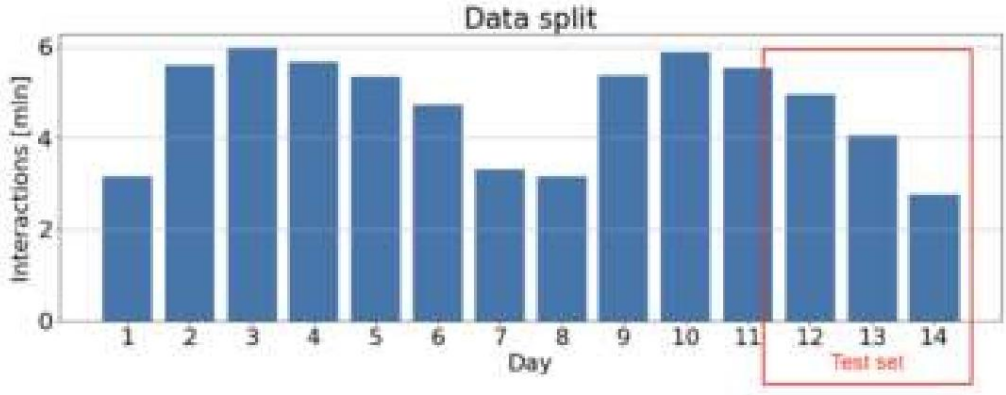


Рисунок 3. Подготовка набора данных (разделение на обучение-проверку и тестирование, в зависимости от количества дни). Источник: собственное исследование.

Таблица 5. Сравнение обучающих/проверочных и тестовых наборов. Буква обозначает стандарт. отклонение.

	Взаимодействие с сотрудниками OLX Взаимодействия	Набор для обучения и проверки	Тестовый набор
Взаимодействия	65,50 M	52,40 M 6,11 M	
Уникальные взаимодействия	47,17 M	38,25 млн 6,11 млн	
Пользователи	3,30 M	2,83 M 0,62 M	
Среднее количество уникальных взаимодействий на пользователя	14.31 (= 29,23)	13.50 (= 26,11)	9,87 (= 14,38)
Предметы	0,19 M	0,18 M 0,13 M	
Среднее количество уникальных взаимодействия на элемент	254,42 (= 426,02)	214.05 (= 365,93)	47,51 (= 83,84)
Плотность матрицы взаимодействия (%)	0,0077%	0,0076%	0,0077%

Источник: собственное исследование.

4.4. Тюнинг модели

В целях экономии средств и времени при выполнении оптимизации параметров мы сделали ограничение до 20% пользователей и 20% предметов обучения и валидации набор. Затем мы разделяем ограниченный набор данных на набор данных для обучения и проверки с помощью правила, описанные в разделе выше. Для каждой модели было выбрано 100 итераций. путем байесовской оптимизации с использованием гауссовских процессов, реализованных в scikit-learn (https://scikit-optimize.github.io/stable/modules/generated/skopt.gp_minimize.html, доступ 29 мая 2024 г.). Файл конфигурации определяет места для гиперпараметров и доступен в нашем репозитории кода (https://github.com/rob-kwiec/olx-jobs-recommendations/blob/основной/src/tuning/config.py, доступ 29 мая 2024 г.). Например, для RP3Beta мы ограничились поиском и в интервале [0, 2].

Рекомендации в процессе оптимизации подготовлены для 30 000 пользователей, а как упоминалось выше, использованный показатель оптимизации составлял точность @ 10. На рисунке 4 показано показатель точности в зависимости от параметров. Видно, что БАС не был высоко оптимизирован внутри процесса, но для других методов эта оптимизация имела решающее значение для улучшение точности. Гиперпараметры для всех моделей представлены в таблице 6.

Machine Translated by Google

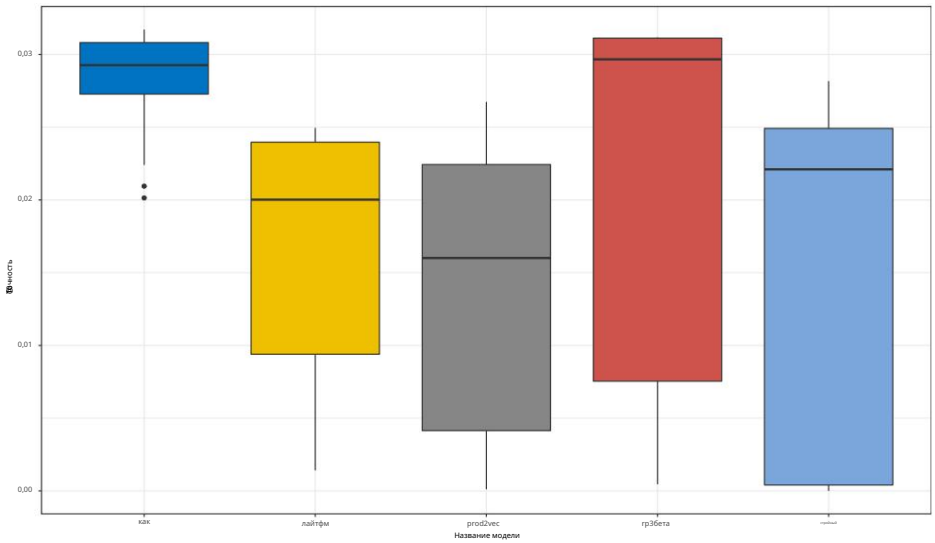


Рисунок 4. Точность для каждой модели в зависимости от параметров. Источник: собственное исследование.

Таблица 6. Гиперпараметры модели.

Модель	Гиперпараметры модели
как	{'факторы': 357, 'регуляризация': 0,001, 'итерации': 20, 'event_weights_multiplier': 63}
лайтфм	{'no_comComponents': 512, 'k': 3, 'n': 20, 'learning_schedule': «ададельта», «потеря»: «деформация», «max_sampled»: 61, «эпохи»: 11}
prod2vec	{'vector_size': 168, 'альфа': 0,028728, 'окно': 20, 'min_count': 16, «образец»: 0,002690026, «min_alpha»: 0,0, «sg»: 1, «hs»: 1, «отрицательный »: 200, «ns_exdependent»: -0,16447846705441527, «cbow_mean»: 0, 'эпохи': 22}
RP3Beta	{'альфа': 0,61447198, 'бета': 0,1443548}
тонкий	{'альфа': 0,00181289, 'l1_ratio': 0,0, 'итерации': 3}

Источник: собственное исследование.

4.5. Оценка модели

Мы оцениваем наши модели с точки зрения точности, разнообразия и эффективности.

Точность метода рекомендаций для наборов данных с неявной обратной связью обычно составляет оценивается путем изучения того, какие позиции из списка лучших рекомендаций пользователь взаимодействовали с тестовым набором [8]. Многие метрики оценки определяют метрики для каждого отдельно пользователю и представить среднее значение как окончательную оценку метода [23]. рекомендательную систему можно рассматривать как классификатор, в котором взаимодействие пользователя в набор тестов представляет собой положительные примеры, а верхние рекомендации — это элементы, классифицированные как положительный по модели. Следовательно, методы, используемые для оценки классификаторов, могут быть непосредственно применяется к рекомендательным системам. Например, значение точности@ для данного пользователя равно количество рекомендуемых элементов, с которыми пользователь взаимодействовал в тестовом наборе, разделенное на количество товаров, рекомендованных этому пользователю. Другими показателями, обсуждаемыми в этой работе, были описано Таммом и др. [23].

В сфере объявлений, в первую очередь в сфере вакансий, обычно требуется только один пользователь для завершить транзакцию относительно данного объявления. В этом случае разнообразие рекомендаций даже более критичен, чем в других областях (таких как музыка, фильмы или другая электронная коммерция, кроме объявлений). Для оценки этого аспекта мы используем три показателя: тестовое покрытие, Шеннон энтропия [56] и индекс Джини [56]. Покрытие тестами относится к доле тестовых заданий. рекомендуется хотя бы одному пользователю. Используя энтропию Шеннона и индекс Джини, мы игнорируем элементы вне тестового набора. Больший охват тестами и значения энтропии Шеннона указывают на более высокий уровень разнообразия, тогда как более высокие значения индекса Джини указывают на меньшее разнообразие.

Machine Translated by Google

Заключительный этап лабораторной оценки фокусируется на эффективности. Для этой цели мы использовали экземпляр AWS SageMaker ml.m5.4xlarge, оснащенный 64 ГБ ОЗУ и 16 виртуальными ЦП, с процессором Intel(R) Xeon(R) Platinum 8259CL @ 2,50 ГГц, работающим на Amazon Linux AMI 2018.03. Все методы оценивались по трем функциям: предварительная обработка данных, Примерка модели и рекомендации по товару. Время выполнения рассчитывалось для общее время, необходимое для расчета модели (в часах). Пик памяти был рассчитан в ГБ, а максимальное использование памяти учитывалось при выполнении данной функции. Эта оценка была выполнена с использованием библиотеки трассировки (https://docs.python.org/3/). библиотека/tracemalloc.html, доступ 29 мая 2024 г.).

5. Лабораторная оценка

В этом разделе описывается лабораторная оценка методов, обсуждаемых в разработанном наборе данных. Мы использовали наиболее эффективные гиперпараметры, найденные в предыдущем разделе. Каждая модель была обучена на всем наборе обучающих и проверочных данных, а рекомендации были подготовлены для всех 619 389 пользователей из тестового набора. Чтобы лучше представить ценности каждую офлайн-метрику мы сравнивали описанные методы с простыми, неперсонализированными базовые показатели: наиболее популярный продуктовый подход и случайные рекомендации.

5.1. Точность

Все метрики в этом разделе учитывают первые десять рекомендаций, например точность@10, отзыв@10. Более высокие значения показателя указывают на более высокое качество рекомендаций. Ценности не следует напрямую сравнивать с результатами, полученными на других наборах данных, поскольку метрики сильно зависят от распределения набора данных (например, высокая разреженность) и стратегия разделения обучения/теста [55].

5.1.1. Оценка точности

В Таблице 7 представлены результаты оценки. Большинство наблюдаемых различий статистически значимы из-за большого количества оцениваемых пользователей. Мы подробно рассказываем о статистическая значимость результатов приведена в разделе 5.1.2.

Таблица 7. Точность: результаты оценки. Все представленные показатели были описаны Tamm et al. [23]. Для Для метрики mAP мы использовали вариант c = min(,).

Метрический	RP3Beta	SLIM	ALS	Prod2Vec	LightFM	Самый популярный	Случайный
точность	0,0484	0,0472	0,0434	0,0368	0,0736	0,0359	0,0012	0,00006
отзыв	0,0783	0,0657	0,0580	0,0721	0,0657	0,0564	0,0012	0,00005
ndcg	0,0759	0,0567	0,0365	0,0329	0,0282	0,0545	0,0016	0,00007
mAP	0,0393	0,1314	0,1230	0,1065	0,5368	0,0264	0,0006	0,00002
MRR	0,1365	0,5328	0,5 289	0,3066	0,2878	0,1034	0,0038	0,00019
LAUC	0,5391	0,2537				0,5281	0,5006	0,49999
HR	0,3131					0,2547	0,0112	0,00059

Источник: собственное исследование.

Наши подходы значительно превосходят случайные и наиболее популярные рекомендации по товарам . Низкая эффективность самого популярного подхода обусловлена природой объявления, в первую очередь о сфере деятельности, где пользователи больше всего интересуются конкретной местоположение и категория. Создание более сложных, персонализированных систем рекомендаций. разумно в таком случае.

Показано, что RP3Beta превосходит другие подходы по точности. метрики. Мы полагаем, что это связано с высокой разреженностью нашего набора данных. RP3Beta рассчитывает рекомендации детерминированно (путем использования путей длиной три в пользовательском элементе). двудольный граф). Все остальные подходы используют методы машинного обучения для поиска пользователя. представления (ALS, LightFM), представления элементов (ALS, LightFM, Prod2Vec) или прямые Сходство между элементами (SLIM), которое может иметь место в тех случаях, когда пользователи или элементы имеют низкий уровень количество взаимодействий.

Machine Translated by Google

5.1.2. Статистическое сравнение для Precision@10

Мы представляем подробное статистическое сравнение нашего основного показателя точности @ 10, чтобы выявить различия между методами. Для начала мы проверяем нулевую гипотезу о том, что все методы работают одинаково, а наблюдаемые различия являются просто случайными (омнибусный тест). Тест Фридмана [57,58] с расширением Имана и Давенпорта [59] является, вероятно, самым популярным комплексным тестом, и обычно он является хорошим выбором при сравнении более чем пяти различных алгоритмов [60,61]. Пусть – ранг th методов на th выборках и

$$"=" \frac{1}{} \quad .$$
$$=1$$

В нашем случае количество выборок — это количество пользователей в тестовом наборе = 619 389. Тест сравнивает средние ранги методов и основан на следующей статистике:

$$"=" \frac{(- 1) (- \quad ^2}{1) - \quad ^2},$$

где

$$2 \quad "=" \frac{12}{(+ 1)} \quad \quad ^2 \quad 3 (+ 1)$$
$$=1$$

– это статистика Фридмана, имеющая распределение с – 1 и (– 1) (– 1) степенями свободы. Значение этого теста равно 0. Полученное значение указывает на то, что мы можем смело отвергнуть нулевую гипотезу о том, что все алгоритмы работают одинаково. Следовательно, мы можем приступить к апостериорным тестам, чтобы обнаружить значительные попарные различия между всеми методами. Демшар [62] предлагает использовать тест Неменьи [63], который сравнивает все алгоритмы попарно. Для уровня значимости тест определяет критическую разницу (CD). Если разница между средними рейтингами двух алгоритмов превышает

$$\text{CD} = \quad \sqrt{\frac{}{(+ 1) 6}},$$

нулевая гипотеза о том, что алгоритмы имеют одинаковую производительность, основан отклоняется (по статистике стьюдентизированного диапазона, деленной на 2). Демшар [62] предложил график проверки различий (график CD визуально). Алгоритмы, не соединенные линией, можно считать разными по сюжету.

В нашем случае при значимости = 0,05 любые два алгоритма с разницей в средний ранг выше 0,0114 будет считаться неравным (рис. 5).

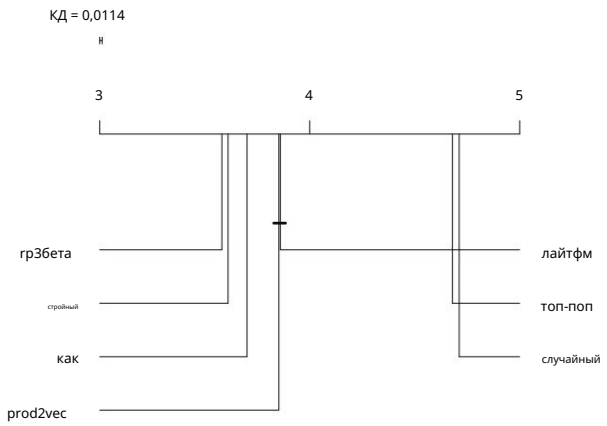


Рисунок 5. График критической разницы (для точности). Источник: собственное исследование.

Machine Translated by Google

5.1.3. Точность для пользователей с разным количеством взаимодействий

Единственная информация, которой мы располагаем о наших пользователях, — это их взаимодействие. На рисунке 6 мы исследуем, как количество элементов, с которыми взаимодействовал пользователь, влияет на нашу основную метрику — точность@10. Мы делим наших пользователей на десять групп одинакового размера в зависимости от количества взаимодействий.

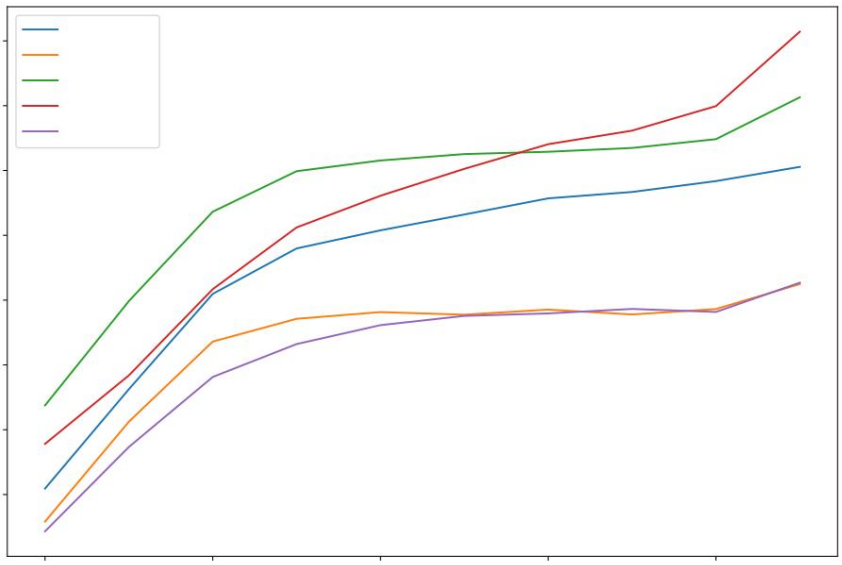


Рисунок 6. Точность@10 для каждой модели в зависимости от количества элементов, с которыми взаимодействовал пользователь. Лучше всего рассматривать его в цвете. Источник: собственное исследование.

Мы видим, что порядок моделей, отсортированных по точности @ 10, не зависит от группы, за исключением того, что SLIM работал лучше, чем RP3Beta для пользователей, совершивших не менее 22 взаимодействий. Разница статистически значима для каждой группы пользователей (обсуждается в разделе 5.1.4). Модель RP3Beta можно рассматривать как частный случай подхода совместной фильтрации на основе элементов, который использует детерминированную меру сходства между элементами [64]. С другой стороны, SLIM использует методы машинного обучения для выявления сходств.
Мы предполагаем, что такие оценки выученного сходства менее предвзяты, но имеют более высокую дисперсию, что приводит к повышению производительности для пользователей с большим количеством взаимодействий, поскольку усреднение выполняется по более значительному числу оценок.

5.1.4. Статистическая оценка точности для пользователей с разным количеством взаимодействий. В этом подразделе мы обсуждаем разницу между моделями RP3Beta и SLIM.
в зависимости от количества взаимодействий пользователя, как показано на рисунке 6.

Для статистического сравнения двух методов на нескольких наборах данных Демшар [62] рекомендует знаково-ранговый критерий Уилкоксона [65]. Знаковый ранговый критерий Уилкоксона — это непараметрическая альтернатива парному тесту, который ранжирует различия в производительности двух алгоритмов для каждого набора данных, игнорируя знаки. Он сравнивает ранги положительных и отрицательных различий.

В таблице 8 показано, что разница между моделями RP3Beta и SLIM статистически значимы для каждой группы пользователей.

Machine Translated by Google

Таблица 8. Результаты знаково-рангового теста Уилкоксона между моделями RP3Beta и SLIM в зависимости от количество элементов, с которыми взаимодействовал пользователь.

Бин	p-значение
[1,0, 3,0)	0
[3,0, 5,0)	0
[5,0, 8,0)	0
[8,0, 11,0)	0
[11,0, 16,0)	0
[16,0, 22,0)	1,25 × 10 6
[22,0, 31,0)	9,27 × 10 7
[31,0, 45,0)	0
[45,0, 74,0)	0
[74,0, 852,0)	0

Источник: собственное исследование.

5.2. Разнообразие

В Таблице 9 мы видим, что Prod2Vec и LightFM обеспечивают наибольшее разнообразие. после исключения случайных рекомендаций (которые должны иметь наибольшее разнообразие). Разнообразие наиболее точных моделей RP3Beta, SLIM и ALS схоже, за исключением случай более низкого тестового покрытия БАС. Связь между тестовым покрытием и точностью @ 10 показано на рисунке 7.

Таблица 9. Разнообразие: результаты оценки.

Метрика	RP3Beta	SLIM	ALS	Prod2Vec	LightFM	Самые популярные	Случайные
тестовое покрытие	0,5725	0,5171	0,3038	0,7400	9,6728	0,7031	0,9778
Шеннон	9,5271	9,6270	10,4031	0,9029	0,9120	0,0002 10,1385	11,7267
Джини	0,9083	2,3296	0,8397	0,9999	0,7956		0,1159

Источник: собственное исследование.

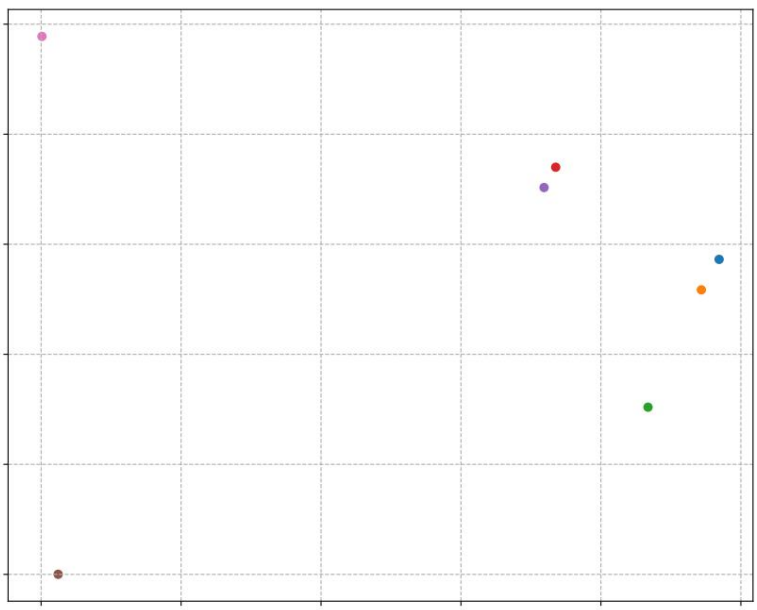


Рисунок 7. Тестовое покрытие в сравнении с точностью@10 для рассматриваемых моделей. Источник: собственное исследование.

Machine Translated by Google

5.3. Перекрытие методов

В наших исследованиях мы фокусируемся на методах, которые стоит протестировать на реальных пользователях. Поэтому, мы изучили совпадение рекомендаций, предлагаемых реализованными методами относительно пар пользователь-элемент (коэффициент перекрытия [66]). Мы ожидаем более значительных различия в рекомендациях при сравнении методов с низким перекрытием.

Результаты этого эксперимента представлены в Таблице 10. RP3Beta и SLIM предлагают аналогичные рекомендации. Хотя они рассчитываются по-разному, они используют разреженный элемент матрица сходства. Группа сравнения с БАС имеет среднее сходство с другими методы. Наиболее разными с точки зрения перекрытия являются также самые разнообразные модели, а именно Prod2Vec и LightFM. Низкое перекрытие ALS и LightFM подчеркивает важность выбора функции потерь в подходах матричной факторизации.

Таблица 10. Перекрытие моделей.

Модель	РП3Бета	СТРОЙНЫЙ	БАС	Прод2Век	ЛайтФМ
RP3Beta	100%	73%	53%	37%	38%
SLIM	73%	100%	50%	35%	35%
ALS	53%	50%	100%	38%	37%
Prod2Vec	37%	35%	38%	100%	28%
LightFM	38%	35%	37%	28%	100%

Источник: собственное исследование.

5.4. Оценка эффективности

Результаты оценки эффективности представлены на рисунке 8. RP3Beta, ALS и LightFM превосходит SLIM и Prod2Vec по времени выполнения, тогда как Prod2Vec требует меньше всего памяти. Чтобы часто переобучать модель, мы установили максимально приемлемое время выполнения составило 3 часа, что соответствует ALS, LightFM и RP3Beta. От стоимости В перспективе допустимое использование памяти было установлено на уровне 110 ГБ, что соответствует всем моделям. выполнено. Мы можем заметить, что наши модели можно обучать даже на одном ноутбуке.

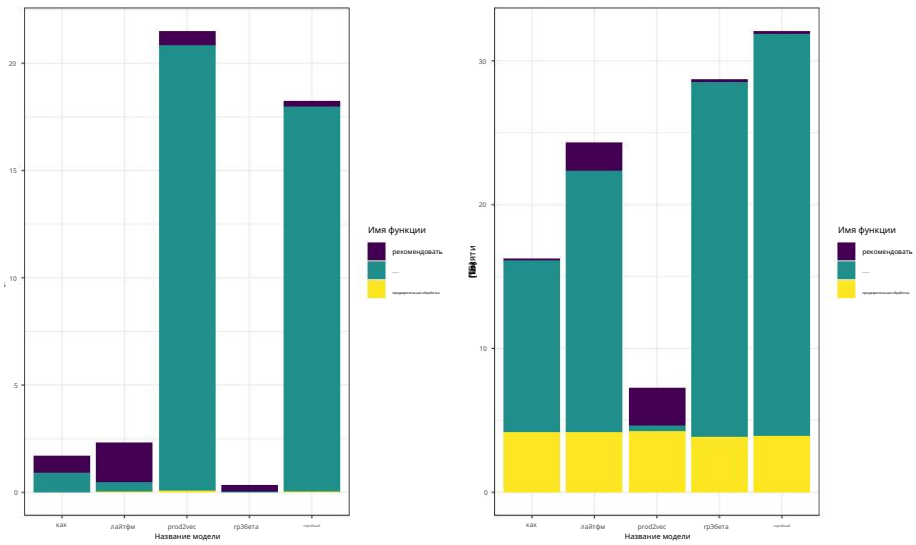


Рисунок 8. Время выполнения и использование памяти. Лучше всего рассматривать его в цвете. Источник: собственное исследование.

5.5. Краткое содержание

Лабораторные исследования позволили нам выбрать модели для онлайн-тестирования с помощью OLX. пользователи. Во-первых, мы выбрали модель RP3Beta для онлайн-тестирования, потому что это лучший подход. относительно точности и времени выполнения. Относительно высокое потребление памяти по-прежнему достаточно низкий для тестирования модели в производстве. Исследование влияния этого относительно небольшое разнообразие метода может стать интересным направлением будущих исследований.

Machine Translated by Google

SLIM — вторая лучшая модель с точки зрения точности @ 10. Тем не менее, мы не выбрали этот метод для онлайн-оценки из-за его низкой эффективности и высокого совпадения с уже выбранной моделью RP3Beta. Повышение эффективности реализации SLIM может изменить это решение.

Среди других моделей наибольшая точность@10 достигается у ALS. Кроме того, ALS уже был внедрен и использовался для рекомендаций по вакансиям в OLX; поэтому мы решили включить его в онлайн-сравнение.

Несмотря на то, что Prod2Vec и LightFM выдают самые разнообразные рекомендации, мы решили не тестировать их онлайн из-за значительно худшей точности@10 и эффективности у Prod2Vec.

6. Онлайн-оценка

Чтобы оценить эффективность выбранных методов, мы провели два онлайн-A/B -теста с пользователями OLX Jobs. Мы сосредоточились на пользователях, которые недавно просмотрели объявления о работе на нашей платформе. Мы отправили им электронные письма с десятью рекомендованными объявлениями о работе, а когда пользователь установил приложение OLX, ему также было отправлено push-уведомление с одним рекомендуемым объявлением о работе. Затем мы наблюдали, подавал ли пользователь заявку на какую-либо работу в течение следующих 48 часов (конвертированный пользователь).

6.1. Влияние модели БАС

Первый эксперимент был направлен на ответ на вопрос: увеличивает ли внедрение рекомендательного метода количество пользователей, претендующих на работу? В этом тесте мы разделили пользователей на две группы: контрольную группу без рекомендаций и группу с рекомендациями, сформированными с помощью модели ALS. Лишь 10% пользователей попали в контрольную группу, поскольку модель ALS уже была внедрена и использовалась для рекомендаций по работе в OLX еще до теста. Мы предпочли собрать минимальную выборку, проведя более расширенное тестирование, чтобы наблюдать долгосрочное влияние на наших пользователей. Тест проводился в течение 25 дней в марте 2021 г. Результаты представлены в таблице 11. Статистическую значимость преимущества БАС над контрольной группой исследовали с помощью критерия хи-квадрат, значение - оказалось равным 0. Этот эксперимент подтвердил нашу гипотезу о том, что рекомендации влияют на активность: процент конвертированных пользователей после получения сообщения значительно увеличился на (16,83% - 15,98%)/15,98% 5,3%.

Наблюдаемый рост означает, что сотни дополнительных пользователей ежедневно подают заявки на работу через OLX Jobs.

Таблица 11. Сравнение активности с рекомендациями и без них.

Вариант	Пользователи	% совершивших конверсию пользователей
контроль	129 308 1	15,98%
БАС	170 262	16,83%

Источник: собственное исследование.

6.2. Сравнение моделей ALS и RP3Beta

Во втором эксперименте мы проверяли, превзошла ли лучшая модель лабораторных испытаний БАС. Мы разделили наших пользователей на три группы, получая рекомендации от разных рекомендательных систем: ALS, RP3Beta и смешанного варианта. Мы создали половину рекомендаций с помощью ALS, а другую половину — с помощью RP3Beta. Второй эксперимент проводился в течение 28 дней в марте и апреле 2021 года. Результаты представлены в таблице 12.

Здесь рассчитывается процент конвертированных пользователей относительно всех пользователей; однако изменение системы рекомендаций не коснулось пользователей, не открывавших сообщение. Мы решили их отфильтровать, поскольку их поведение добавляло к нашим данным ненужный шум. В таблице 13 представлено влияние на пользователей, открывших сообщение. Преимущество RP3Beta над ALS статистически значимо (значение p 10–6). Разница между вариантами RP3Beta и ALS+RP3Beta не является статистически значимой (значение p составляет 0,19).

Machine Translated by Google

Таблица 12. Сравнение различных методов рекомендаций.

Вариант	Пользователи	% совершивших конверсию пользователей
БАС	343 892	15,25%
РП3Бета	345 273	15,40%
БАС+RP3Бета	343 896	15,30%

Источник: собственное исследование.

Таблица 13. Сравнение различных методов рекомендаций для пользователей, открывших сообщение.

Вариант	Пользователи	% совершивших конверсию пользователей
БАС	44 775	19,66%
РП3Бета	46 097	20,94%
БАС+RP3Бета	45 469	20,59%

Источник: собственное исследование.

Из пользователей, открывших сообщение, мы наблюдали 1,28 п.п. (процентных пунктов) больше конвертированных пользователей из RP3Beta, чем из ALS. Поскольку уровень открытия сообщений составил 13,2%, это эквивалентно увеличению на 1,28 × 0,132 0,17 п.п. всех целевых пользователей. На основе По результатам первого эксперимента, БАС дает увеличение на 0,85 п.п. по сравнению с контрольной группой. Это означает что замена ALS на RP3Beta увеличивает наше влияние на всех целевых пользователей примерно 0,17/0,85 = 20%.

6.3. Обсуждение

Мы доказали, что отправка рекомендаций о вакансиях из модели ALS увеличивает количество пользователей, откликнувшихся на объявления о работе, увеличилось более чем на 5%. Более того, RP3Beta значительно превосходит БАС.

Кроме того, мы провели еще один онлайн-тест A/B, сравнив модель RP3Beta с вариант, в котором мы заменили половину рекомендаций RP3Beta на рекомендации созданный на основе гибридной рекомендательной системы глубокого обучения, используемой в OLX (https://tech.olx.com/item2vec-neural-item-embeddings-to-enhance-recommendations-1fd948a6f293, по состоянию на 29 мая 2024 г.). Мы решили не описывать этот тест в этой статье, так как данная работа основное внимание уделяется подходам совместной фильтрации. Мы также хотели обеспечить воспроизводимость наших результатов в опубликованном наборе данных. К сожалению, мы не смогли опубликовать все функции. используется нашей внутренней моделью. Мы не заметили никаких улучшений по сравнению с RP3Beta, хотя вторая модель использует дополнительные знания о объявлениях о работе. Следовательно, мы подтвердил, что правильно определенная и настроенная простая модель может превзойти более передовые глубокие решения. Это согласуется с результатами Dacrema et al. [64] и Анелли и др. [30], которые сравнили несколько ненейронных подходов, включая RP3Beta, с более продвинутые нейронные модели.

Эффективность модели RP3Beta позволяет OLX Jobs обучать ее с нуля раз в день с очень низкими затратами (на процессор). Современная совместная фильтрация Модель в OLX Jobs в настоящее время используется для генерации электронных писем и отправки рекомендаций для наши пользователи. Недавно мы адаптировали эту модель для предоставления рекомендаций на месте в режиме реального времени. учитывая недавнее взаимодействие пользователя с элементами [67]. Мы также предложили обучение P3, чтобы ранговая модель (P3LTR) [68], обобщение RP3Beta, где мы узнаем о важности отношений пользователь-элемент на основе типов прошлых взаимодействий пользователя и элемента и временных меток.

В этой работе мы восполнили пробел в исследованиях, связанный с отсутствием сравнений классические подходы совместной фильтрации в сфере работы, предоставляя комплексные лабораторные исследования на реальном наборе данных и онлайн-оценка с участием миллионов пользователей. Кроме того, мы сделали общедоступным исходный код для автономной оценки и набор данных. доступный. Мы призываем других исследователей разработать новые методы рекомендаций для помочь пользователям быстрее найти идеальную работу.

Методы, рассматриваемые в этой работе, требуют только триплетов (пользователь, элемент и временная метка). в качестве входных данных для обучения и оценки; следовательно, их можно применять в нескольких областях,

Machine Translated by Google

даже для наборов данных, состоящих из десятков миллионов взаимодействий. Мы ожидаем хороших результатов в области объявлений, основываясь на наших первоначальных результатах, полученных в других категориях OLX.

7. Резюме

В этой статье рассматривается тема рекомендаций о вакансиях в онлайн-объявлениях на примере OLX Jobs. Мы провели обширную оценку подходов, представляющих различные семейства рекомендательных методов.

В ходе лабораторных исследований мы продемонстрировали, что RP3Beta, SLIM и ALS работают значительно лучше, чем LightFM и Prod2Vec. Мы также обнаружили, что Prod2Vec предлагает самое необычайное разнообразие, которое может понадобиться при работе над новизной рекомендаций. Более того, мы проверили, насколько схожи рекомендации, генерируемые разными моделями.

Мы также провели оценку, отправив миллионы сообщений онлайн-пользователям. A/B-тесты проводились с использованием ALS и RP3Beta. Результаты этих тестов показали , что отправка рекомендаций о вакансиях, сгенерированных этими моделями, увеличивает количество пользователей, связывающихся с рекламодателями из OLX Jobs. Благодаря доступности RP3Beta в производственной среде мы повысили эффективность на 20 % (с учетом реакции пользователей).

При оценке методов с точки зрения объявлений модель, рекомендовавшая наиболее популярные (еще не встречавшиеся) предметы, оказалась непригодной для рекомендаций по вакансиям. Также для сравнения была добавлена рекомендательная модель, основанная на случайных рекомендациях, чтобы продемонстрировать преимущество исследуемых методов.

Еще одним важным результатом этого исследования является набор данных, который позволит проводить дальнейшие исследования по рекомендациям для классифицированных потребностей.

Одним из недостатков нашей работы является отсутствие сравнения с глубокими нейронными сетями (DNN) или нейронными сетями на графах (GNN). Для таких методов сложно удовлетворить наши временные ограничения без более дорогой инфраструктуры (например, с графическим процессором). Ослабление наших ограничений или сравнение наиболее эффективных DNN и GNN было бы интересно для будущей работы. Еще одним недостатком является то, что сравнение проводится только на одном наборе данных. Размер набора данных оказывает существенное влияние на выбор метода. Было бы интересно протестировать одни и те же методы на наборах данных разного размера, либо на подмножествах исходного набора данных, либо на других наборах данных из предметной области (например, с разных рынков, на которых работает OLX). Кроме того, было бы интересно изучить корреляцию между разнообразием и онлайн-результатами, чтобы разработать офлайн-метрику, которая лучше соответствовала бы бизнес-целям.

Вклад авторов: Концептуализация, РК; методология, РК и ТГ; программное обеспечение, РК и ВД; валидация, РК, ВД, ТГ и АФ; формальный анализ, РК и ТГ; расследование, РК и ТГ; ресурсы, РК; курирование данных, РК и ВД; письменное – подготовка оригинального проекта, РК, АФ и ТГ; написание – рецензирование и редактирование, РК, А.Ф. и ТГ; визуализация, РК и ТГ; наблюдение, ТГ и ФП; администрация проекта, РК Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: Набор данных: https://www.kaggle.com/datasets/olxdatascience/olx-jobs-interactions , доступ 29 мая 2024 г.; исходный код: https://github.com/rob-kwiec/olx-jobs-рекомендации , доступ 29 мая 2024 г.

Конфликты интересов: авторы Роберт Квецинский, Агата Филиповска и Вячеслав Дубров работали в компании OLX Group. Оставшийся автор заявляет, что исследование проводилось при отсутствии каких-либо коммерческих или финансовых отношений, которые могли бы быть истолкованы как потенциальный конфликт интересов.

Рекомендации

1. Дэвис, М. Что такое онлайн-объявления? Журнал Big Easy, 4 января 2021 г. 2.
Джонс, С. Интернет-объявления; Проект Pew Internet & American Life: Вашингтон, округ Колумбия, США, 2009 г.
3. Гурвиль, Дж.; Соман, Д. Чрезмерный выбор и тип ассортимента: когда и почему разнообразие имеет неприятные последствия. Отметка. наук. 2005, 24, 382–395.
[Перекрестная ссылка]

Machine Translated by Google

4. Чернев А. Когда больше значит меньше, а меньше значит больше: роль наличия и ассортимента идеальных точек в потребительском выборе. Дж. Потребление. Рез. 2003, 30, 170–183. [Перекрестная ссылка]

5. Боутрайт, П.; Нуньес, Дж. Сокращение ассортимента: подход, основанный на атрибутах. Дж. Марк. 2001, 65, 50–63. [Перекрестная ссылка]

6. Гомес-Урибе, К.; Хант, Н. Рекомендательная система Netflix. АКМ Транс. Менеджер. Инф. Сист. 2015, 6, 1–19. [Перекрестная ссылка]

7. Смит, Б.; Линден, Г. Два десятилетия рекомендательных систем на Amazon.com. IEEE Интернет-вычисления. 2017, 21, 12–18. [Перекрестная ссылка]

8. Аггарвал, СС Рекомендательные системы: Учебник, 1-е изд.; Springer: Берлин/Гейдельберг, Германия, 2016. 9. Ли, С.; Караханна, Э. Системы онлайн-рекомендаций в контексте электронной коммерции B2C: обзор и будущие направления. Дж. доц. Инф. Сист. 2015, 16, 72–107. [Перекрестная ссылка]

10. Твардовский Б. Моделирование контекстной информации в рекомендательных системах с учетом сеанса с помощью нейронных сетей. В материалах 10-й конференции ACM по рекомендательным системам, Бостон, Массачусетс, США, 15–19 сентября 2016 г.; стр. 273–276. [Перекрестная ссылка]

11. Алотаиби С. Обзор систем рекомендации вакансий. Межд. Дж. Физ. наук. 2012, 7, 5127–5142. [Перекрестная ссылка]

12. Абель, Ф.; Бенчур, А.; Кольсдорф, Д.; Ларсон, М.; Паловичс, Р. RecSys Challenge'16: Материалы конкурса рекомендательных систем; Ассоциация вычислительной техники: Нью-Йорк, штат Нью-Йорк, США, 2016 г.

13. Абель, Ф.; Делджу, Ю.; Элахи, М.; Колсдорф, Д. RecSys Challenge'17: Материалы конкурса рекомендательных систем 2017; Ассоциация вычислительной техники: Нью-Йорк, штат Нью-Йорк, США, 2017 г.

14. Ли, С.; Ши, Б.; Ян, Дж.; Ян, Дж.; Ван, С.; Чен, Ф.; Он, В. Глубокое понимание работы в LinkedIn. В материалах 43-й Международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, виртуально, 25–30 июля 2020 г.; стр. 2145–2148. [Перекрестная ссылка]

15. Чжу, Г.; Чен, Ю.; Ван, С. Персонализированные рекомендации для курсовой работы с поддержкой графического сообщества и междоменные данные Интеграция. Устойчивое развитие 2022, 14, 7439. [CrossRef]

16. Лачич Э.; Райтер-Хаас, М.; Ковальд, Д.; Редди Даредди, М.; Чо, Дж.; Лекс, Э. Использование автоэнкодеров для рекомендаций заданий на основе сеанса. свидания. Модель пользователя. Адаптация пользователя. Взаимодействуйте. 2020, 30, 617–658. [Перекрестная ссылка]

17. Паццани, МЮ; Биллсус, Д. Системы рекомендаций на основе контента. В книге «Адаптивный Интернет: методы и стратегии Интернета». Персонализация; Springer: Берлин/Гейдельберг, Германия, 2007 г.; стр. 325–341. [Перекрестная ссылка]

18. Чен, Р.; Хуа, Ц.; Чанг, Ю.С.; Ван, Б.; Чжан, Л.; Конг, Х. Обзор рекомендательных систем на основе совместной фильтрации: от традиционных методов к гибридным методам на основе социальных сетей. Доступ IEEE 2018, 6, 64301–64320. [Перекрестная ссылка]

19. Ван, Х.; Он, Х.; Ван, М.; Фэн, Ф.; Чуа, Т. С. Совместная фильтрация нейронных графов. В материалах 42-й Международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, Париж, Франция, 21–25 июля 2019 г.; стр. 165–174. [Перекрестная ссылка]

20. Шафер, Дж.Б.; Франковски, Д.; Херлокер, Дж.; Сен, С. Рекомендательные системы совместной фильтрации. В Адаптивной сети; Springer: Берлин/Гейдельберг, Германия, 2007 г.; стр. 291–324. [Перекрестная ссылка]

21. Такач, Г.; Пиласи, И.; Немет, Б.; Тикк, Д. Масштабируемые подходы к совместной фильтрации для больших рекомендательных систем. Дж. Мах. Учиться. Рез. 2009, 10, 623–656. [Перекрестная ссылка]

22. Джавахир, Г.; Сомсзор, М.; Косткова П. Сравнение неявной и явной обратной связи от онлайн- сервиса музыкальных рекомендаций. В материалах 1-го международного семинара по неоднородности и слиянию информации в рекомендательных системах, Барселона, Испания, 26 сентября 2010 г.; стр. 47–51. [Перекрестная ссылка]

23. Тамм Ю.М.; Дамдинов Р.; Васильев А. Метрики качества в рекомендательных системах: единообразно ли мы рассчитываем метрики? В материалах пятнадцатой конференции ACM по рекомендательным системам, Амстердам, Нидерланды, 27 сентября – 1 октября 2021 г.; стр. 708–713. [Перекрестная ссылка]

24. Каминскас, М.; Бридж, Д. Разнообразие, случайность, новизна и охват: обзор и эмпирический анализ целей сверхточности в рекомендательных системах. АКМ Транс. Взаимодействуйте. Интел. Сист. 2016, 7, 1–42. [Перекрестная ссылка]

25. Кунавер, М.; Позрл, Т. Разнообразие в рекомендательных системах, Обзор. Информационная система. 2017, 123, 154–162. [Перекрестная ссылка]

26. Ге, М.; Дельгадо-Баттенфельд, К.; Яннах, Д. За пределами точности: оценка рекомендательных систем по охвату и случайности. В материалах четвертой конференции ACM по рекомендательным системам, Барселона, Испания, 26–30 сентября 2010 г.; стр. 257–260. [Перекрестная ссылка]

27. Могенет А.; Фам, ТАН; Казама, М.; Конг, Дж. Прогнозирование онлайн-производительности систем рекомендации вакансий с помощью автономной оценки. В материалах 13-й конференции ACM по рекомендательным системам, Копенгаген, Дания, 16–20 сентября 2019 г.; стр. 477–480. [Перекрестная ссылка]

28. Анелли, Фольксваген; Беллогин, А.; Ди Ноя, Т.; Яннах, Д.; Помо, К. Алгоритмы рекомендаций Top-N: поиск новейших достижений. В материалах 30-й конференции ACM по моделированию, адаптации и персонализации пользователей, Барселона, Испания, 4–7 июля 2022 г.; стр. 121–131. [Перекрестная ссылка]

29. Дакрема, МФ; Больо, С.; Кремонези, П.; Яннак, Д. Тревожный анализ воспроизводимости и прогресса в рекомендациях Системные исследования. АКМ Транс. Инф. Сист. 2021, 39, 1–49. [Перекрестная ссылка]

30. Анелли, Фольксваген; Беллогин, А.; Ди Ноя, Т.; Помо, К. Новый взгляд на сравнение нейронной совместной фильтрации и матричной факторизации. В материалах пятнадцатой конференции ACM по рекомендательным системам, Амстердам, Нидерланды, 27 сентября – 1 октября 2021 г.; стр. 521–529. [Перекрестная ссылка]

Machine Translated by Google

31. Чжу Дж.; Дай, К.; Су, Л.; Ма, Р.; Лю, Дж.; Кай, Г.; Сяо, Х.; Чжан, Р. БАРС: На пути к открытому сравнительному анализу рекомендательных систем. В материалах 45-й Международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, Мадрид, Испания, 11–15 июля 2022 г.; стр. 2912–2923. [Перекрестная ссылка]
32. Донг, Ю.; Ли, Дж.; Шнабель, Т. Когда новее не значит лучше: действительно ли глубокое обучение приносит пользу рекомендациям неявной обратной связи? В материалах 46-й Международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, Тайбэй, Тайвань, 23–27 июля 2023 г.; стр. 942–952. [Перекрестная ссылка]
33. Анелли, Фольксваген; Малитеста, Д.; Помо, К.; Беллогин А.; Ди Шашио, Э.; Ди Нойа, Т. Оспаривание мифа о совместной фильтрации графов: обоснованный анализ, основанный на воспроизводимости. В материалах 17-й конференции ACM по рекомендательным системам, Сингапур, 18–22 сентября 2023 г.; стр. 350–361. [Перекрестная ссылка]
34. Корень Ю.; Белл, Р.; Волинский, К. Методы матричной факторизации для рекомендательных систем. Компьютер 2009, 42, 30–37. [Перекрестная ссылка]
35. Кула, М. Встраивание метаданных для рекомендаций по холодному запуску пользователей и элементов. arXiv 2015, arXiv:1507.08439.
36. Рендл, С.; Фройденталер, К.; Гантнер, З.; Шмидт-Тиме, Л. БПР: Байесовский персонализированный рейтинг на основе неявной обратной связи. arXiv 2012, arXiv:1205.2618
37. Уэстон, Дж.; Бенджио, С.; Усунье, Н. Всаби: Масштабирование аннотаций изображений до большого словарного запаса. В материалах двадцать второй Международной совместной конференции по искусственному интеллекту, Барселона, Испания, 16–22 июля 2011 г.; стр. 2764–2770.
38. Уэстон, Дж.; Да, Х.; Вайс, Р.Дж. Учимся ранжировать рекомендации с потерей статистики K-порядка. В материалах 7-й конференции ACM по рекомендательным системам, Гонконг, Китай, 12–16 октября 2013 г.; стр. 245–248. [Перекрестная ссылка]
39. Ху, Ю.; Корен, Ю.; Волинский, К. Совместная фильтрация наборов данных неявной обратной связи. В материалах восьмой конференции IEEE 2008 г. Международная конференция по интеллектуальному анализу данных, Пиза, Италия, 15–19 декабря 2008 г.; стр. 263–272. [Перекрестная ссылка]
40. Такач, Г.; Пиласи, И.; Тикк, Д. Применение метода сопряженных градиентов для совместной фильтрации с неявной обратной связью. В материалах пятой конференции ACM по рекомендательным системам, Чикаго, Иллинойс, США, 23–27 октября 2011 г.; стр. 297–300. [Перекрестная ссылка]
41. Нин, Х.; Карипис, Г. СЛИМ: Разреженные линейные методы для рекомендательных систем Top-N. В материалах 11-й Международной конференции IEEE по интеллектуальному анализу данных 2011 г., Ванкувер, Британская Колумбия, Канада, 11–14 декабря 2011 г.; стр. 497–506. [Перекрестная ссылка]
42. Кристоффель Ф.; Паудель, Б.; Ньюэлл, К.; Бернштейн, А. Блокбастеры и желтофиоль: точные, разнообразные и масштабируемые рекомендации с помощью случайных блужданий. В материалах 9-й конференции ACM по рекомендательным системам, Вена, Австрия, 16–20 сентября 2015 г.; стр. 163–170. [Перекрестная ссылка]
43. Купер, К.; Ли, Ш.; Радзик Т.; Сиантос, Ю. Случайные блуждания в рекомендательных системах: точные вычисления и моделирование. В материалах 23-й Международной конференции по Всемирной паутине, Нью-Йорк, штат Нью-Йорк, США, 7–11 апреля 2014 г.; стр. 811–816. [Перекрестная ссылка]
44. Грбович, М.; Радосавлевич, В.; Джурич, Н.; Бхамидипати, Н.; Савла, Дж.; Бхагван, В.; Шарп, Д. Электронная коммерция в вашем почтовом ящике: рекомендации по продуктам в масштабе. В материалах 21-й Международной конференции ACM SIGKDD по обнаружению знаний и интеллектуальному анализу данных, Сидней, Новый Южный Уэльс, Австралия, 10–13 августа 2015 г.; стр. 1809–1818. [Перекрестная ссылка]
45. Баркан, О.; Кенигштейн, Н. ITEM2VEC: Встраивание нейронных элементов для совместной фильтрации. В материалах 26-го Международного семинара IEEE по машинному обучению для обработки сигналов (MLSP) 2016 г., Виетри-суль-Маре, Италия, 13–16 сентября 2016 г.; стр. 1–6. [Перекрестная ссылка]
46. Миколов Т.; Суцкевер И.; Чен, К.; Коррадо, Г.; Дин, Дж. Распределенные представления слов и фраз и их композиционность. В материалах 26-й Международной конференции по нейронным системам обработки информации — Том 2, Ред-Хук, Нью-Йорк, США, 5–10 декабря 2013 г.; стр. 3111–3119.
47. Де Рюйт, К.; Бхулаи, С. Системы рекомендации вакансий: обзор. arXiv 2021, arXiv:2111.13576.
48. Кула, М. LightFM — репозиторий Github. 2023 г. Доступно онлайн: https://github.com/lyst/lightfm. (по состоянию на 29 мая 2024 г.).
49. Фредериксон, Б. ALS — репозиторий Github. 2022 г. Доступно онлайн: https://github.com/benfred/implicit/blob/main/implicit/. als.py (по состоянию на 29 мая 2024 г.).
50. Леви, М.; Гризель, О. СЛИМ — репозиторий Github. 2014. Доступно онлайн: https://github.com/Mendeley/mrec/blob/master/. mrec/item_similarity/slim.py (по состоянию на 29 мая 2024 г.).
51. Квадрана, М. СЛИМ — репозиторий Github. 2021 г. Доступно онлайн: https://github.com/MaurizioFD/RecSys2019_DeepLearning_. Evaluation/blob/master/SLIM_ElasticNet/ SLIMElasticNetRecommender.py (по состоянию на 29 мая 2024 г.).
52. Бернардис, К. RP3Beta — репозиторий Github. 2021 г. Доступно онлайн: https://github.com/MaurizioFD/RecSys2019_DeepLearning_. Evaluation/blob/master/GraphBased/ RP3betaRecommender.py (по состоянию на 29 мая 2024 г.).
53. Паудель, Б.; Кристоффель, Ф.; Ньюэлл, К.; Бернштейн А. Обновляемые, точные, разнообразные и масштабируемые рекомендации для интерактивных технологий. Приложения. АКМ Транс. Взаимодействуйте. Интел. Сист. 2016, 7, 1–34. [Перекрестная ссылка]
54. Рехůрек, Р.; Сойка, П. Программная среда для тематического моделирования в крупных корпорациях. В материалах семинара LREC 2010 «Новые вызовы для структур НЛП», Валлетта, Мальта, 22 мая 2010 г.; стр. 45–50, ISBN 2-9517408-6-7. Доступно онлайн: http://is.muni.cz/publication/884893/en. (по состоянию на 29 мая 2012 г.).
55. Мэн, З.; Маккриди, Р.; Макдональд, К.; Оунис, И. Исследование стратегий разделения данных для оценки моделей рекомендаций. В материалах 14-й конференции ACM по рекомендательным системам, виртуально, 22–26 сентября 2020 г.; стр. 681–686. [Перекрестная ссылка]
56. Шани, Г.; Гунавардана А. Оценка рекомендательных систем. В Справочнике по рекомендательным системам; Спрингер: Бостон, Массачусетс, США, 2011 г.; стр. 257–297. [Перекрестная ссылка]

Machine Translated by Google

57. Фридман М. Использование рангов, чтобы избежать предположения о нормальности, подразумеваемого при дисперсионном анализе. Варенье. Стат. доц. 1937, 32, 675–701. [Перекрестная ссылка]

58. Фридман М. Сравнение альтернативных критериев значимости проблемы ранжирования. Анна. Математика. Стат. 1940, 11, 86–92. [Перекрестная ссылка]

59. Иман, Р.Л.; Давенпорт, Дж. М. Приближения критической области статистики Фридмана. Коммун. Статистические методы 1980, 9, 571–595. [Перекрестная ссылка]

60. Гарсия, С.; Эррера, Ф. Расширение «Статистического сравнения классификаторов по нескольким наборам данных» для всех парных сравнений. Дж. Мах. Учиться. Рез. 2008, 9, 2677–2694.

61. Гарсия, С.; Фернандес А.; Луенго, Дж.; Эррера, Ф. Расширенные непараметрические тесты для множественных сравнений при планировании экспериментов в области вычислительного интеллекта и интеллектуального анализа данных: экспериментальный анализ мощности. Инф. наук. 2010, 180, 2044–2064. [Перекрестная ссылка]

62. Демшар Дж. Статистическое сравнение классификаторов по нескольким наборам данных. Дж. Мах. Учиться. Рез. 2006, 7, 1–30.

63. Неменьи П. Множественные сравнения без распределения. Кандидат наук. Диссертация, Принстонский университет, Принстон, Нью-Джерси, США, 1963 г.

64. Дакрема, МФ; Кремонези, П.; Яннах, Д. Действительно ли мы добиваемся большого прогресса? Тревожный анализ последних подходов к нейронным рекомендациям. В материалах 13-й конференции ACM по рекомендательным системам, Копенгаген, Дания, 16–20 сентября 2019 г.; стр. 101–109. [Перекрестная ссылка]

65. Уилкоксон Ф. Индивидуальные сравнения методами ранжирования. Биом. Бык. 1945, 1, 80–83. [Перекрестная ссылка]

66. Виджаймина, М.; Кавита, К. Обзор мер сходства при анализе текста. Мах. Учиться. Прил. Межд. Дж. 2016, 3, 19–28. [Перекрестная ссылка]

67. Квечинский Р.; Мельничак, Г.; Гурецкий, Т. Сравнение рекомендаций для работы в режиме реального времени и пакетной обработки. IEEE Access 2023, 11, 20553–20559. [Перекрестная ссылка]

68. Квечинский Р.; Гурецкий, Т.; Филиповска, А. Изучение важности ребер в рекомендациях на основе двудольных графов. В материалах 17-й конференции по информатике и интеллектуальным системам (FedCSIS) 2022 г., София, Болгария, 4–7 сентября 2022 г.; стр. 227–233. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Эффективное уменьшение размерности речевого сигнала с использованием Комплексные методы

Сонгюн Ко [1] и Минхо Пак 2,*

[1] Департамент конвергенции ИИ и ИТ, Университет Сунгсил, Сеул 06978, Республика Корея;
[2] kosungkyun@naver.com Школа электронной инженерии, Университет Сунгсил, Сеул 06978, Республика Корея
* Переписка: mhp@ssu.ac.kr

Аннотация: В данном исследовании мы предлагаем алгоритм CVMFCC-DR (Complex-Valued Mel-Frequency Cepstral Coefficients Dimensionality Reduction) в качестве эффективного метода снижения размерности речевых сигналов. Используя комплексный метод MFCC, который учитывает как действительные, так и мнимые компоненты, наш алгоритм позволяет уменьшить размерность без потери информации и одновременно снизить вычислительные затраты. Эффективность предложенного алгоритма подтверждена экспериментами, которые демонстрируют его эффективность при построении модели распознавания речи с использованием комплексной нейронной сети. Кроме того, представлен комплексный метод интерпретации softmax для комплексных чисел. Результаты экспериментов показывают, что этот подход обеспечивает более высокую производительность по сравнению с традиционными методами на основе MFCC, тем самым подчеркивая его потенциал в области распознавания речи.

Ключевые слова: мел-частотные кепстральные коэффициенты; комплексная коррекция ошибок нейронной сети; обратное распространение ошибки; комплексный softmax

check for updates

Цитирование: Ко, С.; Парк, М. Эффективное уменьшение размерности речевого сигнала с использованием комплексных методов. Электроника 2024, 13, 3046. https://doi.org/10.3390/электроника13153046

Академические редакторы: Синь Гэн, Падма Айенгар и Эльке Пульвермюллер

Поступило: 3 июня 2024 г.
Пересмотрено: 31 июля 2024 г.
Принято: 31 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария.
Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Сегодня обработка и распознавание речевых сигналов играют решающую роль в различных областях применения. Эти технологии используются для различных целей, таких как преобразование речевых команд в текст или идентификация владельца голоса. Развитие технологий распознавания речи может улучшить взаимодействие человека и машины, тем самым повышая производительность и эффективность в различных отраслях. Однако речевые сигналы обладают весьма разнообразными формами и многомерными характеристиками, что делает их эффективную обработку и распознавание постоянной проблемой. Разнообразие и высокая размерность речевых сигналов затрудняют уменьшение размерности и распознавание образов, что приводит к увеличению вычислительных затрат. Поэтому эффективное снижение размерности речевых сигналов и повышение производительности моделей распознавания речи на основе нейронных сетей является серьезной проблемой.

Технологии обработки и распознавания речи необходимы в различных приложениях, таких как голосовые интерфейсы, виртуальные помощники и автоматический перевод. Достижение более точных и эффективных результатов обработки и распознавания речевых сигналов может значительно улучшить качество обслуживания пользователей сопутствующих услуг. В связи с этим ведутся постоянные исследования в области обработки и распознавания речевых сигналов. Речевые сигналы обладают многомерными характеристиками, что затрудняет эффективное снижение разме Более того, точное распознавание разнообразных речевых образов требует сложных моделей нейронных сетей, что, в свою очередь, увеличивает вычислительные затраты [1]. Поэтому эффективное снижение размерности речевых сигналов без потери информации и одновременное повышение точности и эффективности моделей распознавания речи на основе нейронных сетей является важнейшей задачей.

Несмотря на богатую литературу по комплексным сетям, примеров их применения в обработке звука или речи было мало. Учитывая, что слуховая система челове

Machine Translated by Google

обрабатывает звук в частотной области, также имеет смысл изучить комплексные нейронные сети для обработки звука, а не полагаться исключительно на нейронные сети с действительными значениями во временной области [2].

Недавние исследования достигли значительного прогресса в распознавании речи путем объединения традиционных методов, таких как мел-частотные кепстральные коэффициенты (MFCC), с нейронными сетями [3]. Хотя метод MFCC широко используется для извлечения речевых характеристик, у него есть недостаток: он не может выполнить уменьшение размерности без потерь, поскольку использует только действительные значения. Другими словами, уменьшение количества MFCC приводит к потере информации, что снижает точность распознавания речи. Эти методы по-прежнему демонстрируют ограниченную эффективность, и необходимы более совершенные подходы, выходящие за рамки традиционных методов обучения.

Использование комплексного представления MFCC позволяет учитывать как действительные , так и мнимые части, обеспечивая уменьшение размерности без потерь. Кроме того, использование комплексных нейронных сетей может отражать изменения фазы посредством сложных операций умножения [4], лучше представляя характеристики речевых сигналов. Ожидается, что этот подход обеспечит более эффективное снижение размерности и распознавание речи на основе нейронных сетей по сравнению с существующими методами.

В этом исследовании мы предлагаем алгоритм CVMFCC-DR для эффективного уменьшения размерности речевого сигнала без потерь. Объединив CVMFCC-DR с комплексными нейронными сетями, этот метод может продемонстрировать превосходную производительность распознавания речи по сравнению с существующими методами.

Основные результаты данного исследования заключаются в следующем:

- Предложение эффективного метода уменьшения размерности (CVMFCC-DR), использующего комплексное представление MFCC.
- Повышение производительности распознавания речи за счет объединения комплексных MFCC и комплексных нейронных сетей.
- Предложение нового метода интерпретации комплексной функции softmax с сложные входы.

Эта статья начинается с представления существующей техники MFCC и комплексных нейронных сетей. Затем он дает подробное объяснение предлагаемого алгоритма CVMFCC-DR и метода обучения комплексных нейронных сетей [5]. В дальнейшем эффективность предложенного метода подтверждается экспериментально и анализируются результаты . Наконец, в документе приводится краткое изложение результатов исследования и предлагаются направления будущих исследований.

2. Сопутствующие работы

В области распознавания речи используются различные подходы и методы на основе нейронных сетей. Используются различные типы нейронных сетей, такие как CNN, RNN, LSTM и DNN, каждая из которых имеет свои преимущества и недостатки [3]. Кроме того, исследуются также гибридные модели, сочетающие традиционные методы со структурами нейронных сетей . Гибридные модели могут оптимизировать производительность за счет объединения сильных сторон каждой структуры, но они также имеют недостаток, заключающийся в повышенной сложности реализации и интерпретации [6].

На фоне этих разработок традиционные методы извлечения признаков, такие как метод MFCC, по-прежнему широко используются в качестве входных данных для нейронных сетей [3]. Кроме того, изучение новых нейронных архитектур, таких как комплекснозначные нейронные сети, может открыть новые перспективы и потенциальные прорывы в возможностях распознавания речи. Недавно были проведены исследования, показывающие, что FPGA можно использовать для значительного повышения скорости вычислений комплексных чисел [7]. Результаты этих исследований эмпирически показывают, что аппаратная реализация операций с комплексными числами может привести к практическому повышению производительности. Кроме того, недавнее исследование реализовало комплексные нейронные сети на FPGA для классификации, продемонстрировав превосходную точность по сравнению с нейронными сетями с действительными значениями и добившись значительного улучшения скорости по сравнению с системами на базе ЦП [8]. Это дает основание ожидать, что в будущем процессоры, которые в настоящее время поддерживают только операции с действительными числами, однажды смогут поддерживать и операции с ком

Machine Translated by Google

Основываясь на этих исследованиях, использование представлений комплексных чисел позволяет кодировать пары действительных чисел как отдельные комплексные числа, эффективно уменьшая входную размерность вдвое. Такое уменьшение размерности уменьшает количество параметров модели и повышает эффективность вычислений. Следовательно, экспериментальные данные показывают, что комплексные нейронные сети могут достичь равной или превосходящей производительности с меньшим количеством параметров по сравнению с обычными нейронными сетями с действительными значениями.

Поскольку исследования продолжают расширять границы, синергетическое сочетание проверенных методов и инновационных подходов может проложить путь к следующему поколению надежных и высокопроизводительных систем распознавания речи.

3. Базовые знания

В области распознавания и классификации речи проводятся различные исследования, и в этом разделе мы представляем MFCC (мел-частотные кепстральные коэффициенты) [3]. Метод MFCC — это один из методов выделения признаков, обычно используемых в области обработки речевых сигналов и распознавания речи. Метод MFCC представляет частотные характеристики речевых сигналов и в основном используется для эффективного представления и распознавания речевых сигналов. Основной принцип метода MFCC показан на рисунке 1.

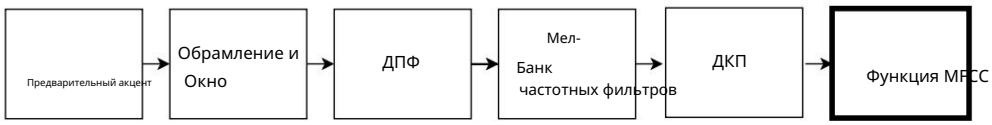


Рисунок 1. Обзор процесса применения метода MFCC.

Предварительное выделение: Предварительное выделение – это, по сути, фильтр верхних частот. Из-за строения человеческого тела высокочастотные компоненты имеют тенденцию уменьшаться в реальном издаваемом звуке во время произнесения речи. Таким образом, предыскажение — это процесс применения фильтра верхних частот для усиления сначала ослабленных высокочастотных составляющих.

Формирование и формирование окон: речевые сигналы для анализа делятся на короткие сегменты, называемые кадрами. Это связано с тем, что речевые характеристики остаются относительно стабильными в течение коротких периодов времени. Эти кадры слегка перекрываются, чтобы более эффективно фиксировать изменения ре Каждый кадр обрабатывается с использованием оконной технологии, что расширяет возможности анализа. Распространенные типы окон включают окна Хэмминга и Хэннинга. Этот метод кадрирования позволяет более точно анализировать сложные речевые данные, что имеет решающее значение для разработки технологий распознавания и синтеза речи. Хотя конкретные детали реализации могут различаться, эти фундаментальные принципы широко используются в системах обработки речи.

Дискретное преобразование Фурье: для расчета спектра мощности используется ДПФ (дискретное преобразование Фурье) для расчета спектра мощности для каждого кадра, полученного ранее. Этот процесс позволяет распознавать частотно-зависимое распределение энергии речевого сигнала в каждом кадре.

Набор фильтров Mel-Frequency: Полосовой фильтр Mel представляет собой набор фильтров, предназначенных для извлечения речевых сигналов нелинейным образом, имитируя слуховую систему человека. Обычно он состоит из 40 треугольных фильтров и анализирует частотные характеристики речевых сигналов.

Дискретное косинусное преобразование: DCT (Дискретное косинусное преобразование) выражает конечную последовательность точек данных как сумму косинусных функций, колеблющихся на разных частотах. В процессе MFCC DCT применяется к банку фильтров mel для выбора коэффициентов ускорения или для отделения отношений в логарифмических спектральных величинах от банка фильтров.

Распознавание речи с использованием MFCC может эффективно использоваться в различных средах [9], и этот метод может достигать высокой точности в сочетании с моделями глубокого обучения. Таким образом, MFCC служит решающим фактором для моделей глубокого обучения для распознавания речи [10].

Machine Translated by Google

4. Предлагаемый метод

В этом разделе мы предлагаем метод эффективного снижения размерности речевого сигнала с использованием комплексных MFCC и комплексных нейронных сетей.
Обзор предлагаемого подхода представлен на рисунке 2. Этот рисунок иллюстрирует предлагаемый подход к классификации голоса с использованием комплексных нейронных сетей. Процесс начинается с голосового ввода, за которым следует извлечение признаков с использованием мел-частотных кепстральных коэффициентов (MFCC). Эти функции затем подвергаются уменьшению размерности и преобразуются в комплексные представления, в результате чего получаются комплексные MF
Обработанные функции впоследствии вводятся в комплексную нейронную сеть.
Производительность сети повышается за счет выбора соответствующих функций активации и алгоритмов обучения. В конечном итоге выходные данные нейронной сети используются для целей классификации.

Рисунок 2. Обзор предлагаемого подхода.

Комплекснозначные MFCC: количество MFCC, используемых при распознавании речи, может варьироваться в зависимости от конкретных приложений и моделей. Однако обычно используются значения MFCC в диапазоне от 13 до 40. Известно, что этот диапазон эффективно работает в различных системах распознавания речи. Как правило, большее количество MFCC позволяет модели улавливать больше частотных и спектральных деталей, но это также увеличивает вычислительные затраты модели. Таким образом, выбранное количество MFCC определяется с учетом баланса между эффективностью вычислений и производительностью модели.

При представлении физических задач с использованием комплексных чисел существует два основных метода. Один из них — простой метод замены, а другой — более элегантный подход, известный как математическая эквивалентность, например преобразование Фурье. При замене берутся два действительных физических параметра и помещаются один в действительную часть комплексного числа, а другой в мнимую часть. Это позволяет манипулировать двумя значениями как единым объектом, т. е. одним комплексным числом [11]. Метод, предложенный для уменьшения размерности в этой статье, включает преобразование значений n MFCC в комплексные числа. Первый MFCC используется как действительная часть, а второй MFCC используется как мнимая часть комплексного числа.
Впоследствии каждый MFCC поочередно является действительной и мнимой частями комплексного числа. Этот шаблон продолжается до n-го числа MFCC, в результате чего получается n/2 комплексных чисел. Например, когда у нас есть n значений MFCC, они могут быть следующими:

$$[x1, x2, x3, x4, \ldots, xn \quad 1, xn]$$

В этом случае первый MFCC x1 используется как действительная часть первого комплексного числа. Второй MFCC x2 используется как мнимая часть первого комплексного числа. Третий MFCC x3 используется как действительная часть второго комплексного числа. Четвертый MFCC x4 используется как мнимая часть второго комплексного числа. Этот шаблон продолжает чередовать действительную и мнимую части до n-го MFCC xn, который становится мнимой частью (n/2)-го комплексного числа. Следовательно, при n значениях MFCC генерируется n/2 комплексных чисел. Первое комплексное число — x1 + jx2, второе комплексное число

Machine Translated by Google

число равно x3 + jx4 и так далее до последнего комплексного числа, которое равно xn 1 + jxn. Мы проиллюстрировали это объяснение на рисунке 3.

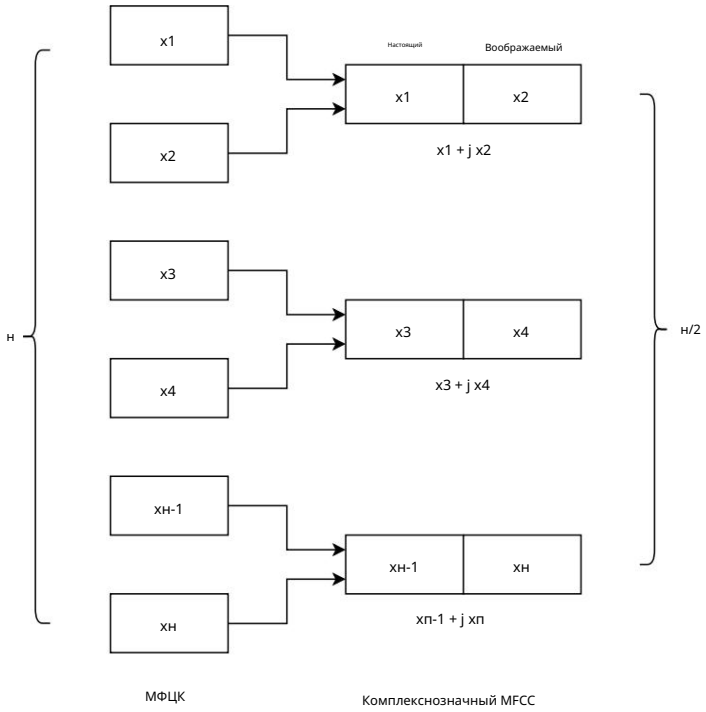


Рисунок 3. Процесс генерации комплексных MFCC.

Нейронная сеть с комплексными значениями. Нейронные сети с комплексными значениями — это модели нейронных сетей, которые используют комплексные числа в качестве входных значений или весов, и у них есть преимущество в обработке более обширной информации по сравнению с нейронными сетями с действительными значениями. Кроме того , в различных областях были проведены различные исследования и эффективно используются методы комплексного уменьшения размерности или классификации и обнаружения с использованием комплексных чисел. Для этого выполняется умножение комплексных входов и комплексных весов, что подразумевает вращение в комплексной плоскости и, таким образом, вносит изменения в фазу [5]. Сумма умножения входа X и веса W может быть представлена в виде уравнений (1) и (2).

$$Z = W0 + W1X1 + W2X2 + \ldots + WnXn \quad (1)$$

$$Z = \quad n\, Wn \cdot Xn \quad (2)$$

Функция активации и алгоритм обучения. В комплексных нейронных сетях используется фазозависимая функция активации, которая приводит к нормализации информации о величине к единичному кругу, равному 1, и только к извлечению фазовой информации, как показано в уравнении (3).).

$$P(3) = \frac{3}{|3|} = e^{\,j\cdot\arg(Z)} \quad (3)$$

Алгоритм обучения комплексных нейронных сетей следующий. Когда есть желаемая цель t и выходное значение z, как показано на рисунке 4, веса корректируются в соответствии с уравнением (4) для постепенного перемещения в направлении желаемой цели [12].

Machine Translated by Google

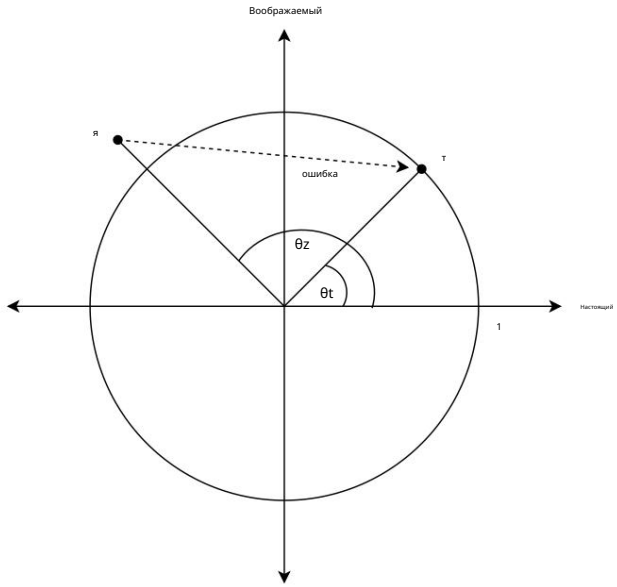


Рисунок 4. Алгоритм обучения CVNN.

$$t = n\,(Wn + Wn)\cdot Xn \tag{4}$$

Ошибка относительно цели обучения t равна e = t - z. , как показано в уравнении (5).

$$e = n\,(Wn + Wn)\cdot Xn \quad n\,(Wn)\cdot Xn = n\,(Wn)\cdot Xn \tag{5}$$

Чтобы упростить и ускорить вычисление ошибки относительно изменения веса, предполагая, что каждый узел вносит равный вклад в ошибку, символ суммирования σ удаляется из уравнения, и обе части делятся на количество вносящих вклад узлов N в комплексе. -значная нейронная сеть для удобства вычислений. Полученное уравнение представляет собой уравнение (6).

$$Wn \cdot Xn = \frac{e}{H} \tag{6}$$

Чтобы получить Wn и изменить веса, умножьте обе части уравнения (6) на Xn при умножении $Икс_н^{1}$. $Икс_н^{1}$ на комплексно-сопряженное число Xn, где n является и числителем, и знаменателем ; таким образом, $X_н^{1} = \frac{\overline{Xн}}{|Xn|^2}$, где |Xn| = 1 для Xn внутри единицы круг. Следовательно, $X_н^{1} = \overline{Xn}$, и уравнение (6) снова получается как уравнения (7) и (8).

$$Wn \cdot Xn \cdot X_н^{1} = \frac{e}{H} \cdot Икс_н^{1} = \frac{e}{H} \cdot \overline{Xн} \tag{7}$$

$$Wn = \frac{e}{H} \cdot \overline{Xн} \tag{8}$$

Таким образом, полученное изменение весов можно наблюдать в уравнении (8). Это изменение представляет собой корректировку комплексных весов для каждого узла, вносящего вклад в ошибку относительно цели обучения, и равномерно применяется по всей сети.

Благодаря этому процессу комплексные нейронные сети подстраиваются ближе к цели, используя фазозависимые функции активации и эффективные алгоритмы обучения, одновременно обрабатывая богатую информацию посредством сложных операций умножения между входными данными и весами [13].

Обратное распространение ошибки в сложной области: при использовании скрытых слоев ошибка из выходного слоя распространяется обратно на скрытые слои, чтобы вычислить, насколько велика ошибка каждого узла.

Machine Translated by Google

в скрытых слоях способствовали этой ошибке. Исходя из этого, веса между скрытым и выходным слоями необходимо соответствующим образом настроить. Этот процесс называется обратным распространением ошибки, посредством которого нейронная сеть может корректировать веса, чтобы минимизировать ошибку в выходных значениях во время обучения. Структура многослойной комплексной нейронной сети показана на рисунке 5.

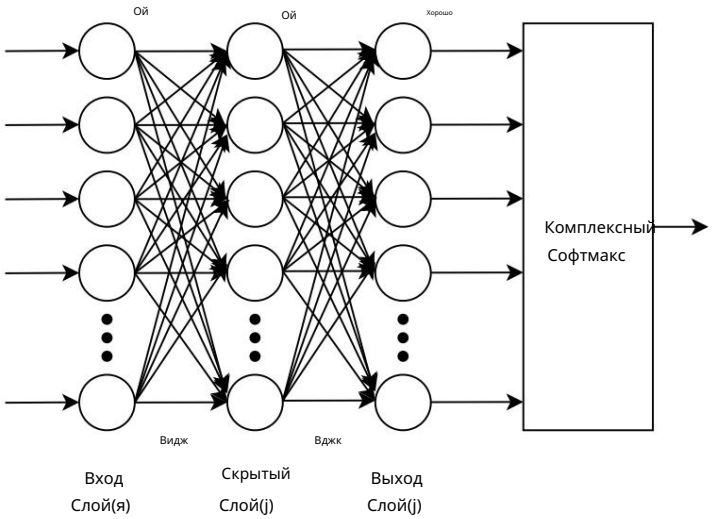


Рисунок 5. Многоуровневая CVNN.

В этой статье метод исправления ошибок используется для выходного слоя k для обновления весов, а метод обратного распространения ошибки используется для скрытого слоя j, чтобы изучить его влияние на выходные данные. Другими словами, выходной слой k использует метод исправления ошибок, чтобы избежать попадания в локальные минимумы, в то время как метод обратного распространения ошибки применяется к скрытому слою j, чтобы обеспечить лучшее распределение ошибок от выходного слоя к скрытому слою. Можно ожидать , что эта комбинация обратного распространения ошибок в скрытом слое и коррекции ошибок в выходном слое улучшит проблему застревания всей нейронной сети в локальных минимумах, одновременно обеспечивая более быструю сходи
Метод обратного распространения ошибки для комплексных нейронных сетей заключается в следующем.

При использовании градиентного спуска для обновления веса в комплексных нейронных сетях исчисление Виртингера используется для дифференциации сложных функций. Поскольку сложные функции состоят из вещественной и мнимой частей, исчисление Виртингера проводит дифференцирование отдельно для действительной и мнимой частей. Для комплексного числа z = x + jy , определенного с действительной и мнимой частями x и y соответственно, и комплексной функции f(z) = u(x, y) + jv(x, y), производные Виртингера определяются как следует [14].

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - j\frac{\partial}{\partial y}\right) = \frac{1}{2}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + j\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\right) \quad (9)$$

$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + j\frac{\partial}{\partial y}\right) = \frac{1}{2}\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} + j\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)\right) \quad (10)$$

$$= + z \frac{\partial}{\partial \bar{z}} \frac{df}{} \quad (11)$$

В этом контексте $\bar{z}$ обозначает комплексно-сопряженное число z. В методе производных Виртингера комплексные функции можно рассматривать как расширения действительных функций, разделив их на действительную и мнимую части для дифференцирования. Ниже описан процесс обновления комплексных переменных с использованием размера шага α/2 и функции потерь L.

Machine Translated by Google

$$x_{n+1} = x_n - \frac{\alpha}{2} \frac{\partial}{\partial x} \tag{12}$$

$$y_{n+1} = y_n - \frac{\alpha}{2} \frac{\partial L}{\partial y} \tag{13}$$

Подставляя $z_{n+1} = x_{n+1} + j\, y_{n+1}$ в уравнения (12) и (13) , а затем подставляя Производная Виртингера в конце процесса вывода становится следующей.

$$z_{n+1} = x_n - \frac{\alpha}{2} \frac{\partial}{\partial x} L + j\, y_n \quad x \quad \frac{\alpha}{2} \frac{\partial}{\partial y}$$

$$= x_n - \frac{\alpha}{2} \frac{\partial}{\partial x} L + jy_n \quad j \quad \frac{\alpha}{2} \frac{\partial L}{\partial y}$$

$$= x_n + jy_n \quad \frac{\alpha}{2} \frac{\partial L}{\partial y} \quad x \frac{\partial L + j}{\partial}$$

$$1 = z_n \quad \alpha \frac{\partial}{\partial 2} \quad \frac{\partial L}{\partial y} \quad x \frac{\partial L + j}{\partial}$$

$$= z_n \quad \alpha \frac{\partial L}{\partial z} \tag{14}$$

Частная производная L по комплексно-сопряженному элементу как сопряженная $\frac{\partial L}{\partial z}$ ссылается на производная Виртингера. Это демонстрирует, как можно использовать сопряженную производную Виртингера для упрощения формулы обновления для сложных переменных. С точки зрения использования описанного выше метода, переписывание уравнения (5) как ошибки выходного слоя дает уравнение (15), где tk представляет собой целевое значение, а ok представляет фактическое выходное значение выходного слоя k.

$$e_k = t_k - o_k \tag{15}$$

А функция потерь определяется как уравнение (16).

$$L = e_k^2 = |t_k - o_k| |t_k - o_k| \tag{16}$$

Если веса велики, ошибка с большей вероятностью распространится обратно на скрытые слои. Ошибка скрытого слоя ej может быть выражена следующим уравнением.

$$e_j = WT_{jk} \cdot e_k \tag{17}$$

Используя результат уравнения (14), градиент функции ошибок относительно веса Wij между входным слоем i и скрытым слоем j можно выразить следующим образом. Здесь oi — выходное значение входного слоя i, а oj — выходное значение скрытого слоя j.

$$\frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial o_j} \cdot \frac{\partial o_j}{\partial W_{ij}} \tag{18}$$

$$\frac{\partial E}{\partial W_{ij}} = e_j \cdot \frac{\partial}{\partial W_{ij}} P \quad W_{ij} \cdot o_i = \tag{19}$$

Machine Translated by Google

$$\frac{\partial L}{\partial W_{ij}} = e_j \cdot P' \cdot \bar{o}_i \quad (20)$$

В уравнениях (19) и (20) P представляет фазозависимую функцию активации, используемую в выходных данных скрытого слоя, а P′ является производной зависимой функции активации. Обновление весов скрытого слоя можно выразить следующим образом.

$$\text{новый } W_{ij} = \text{старый } W_{ij} - lr \cdot \frac{\partial L}{\partial W_{ij}} \quad (21)$$

При работе с комплексными переменными для обновления необходимо вместо W_{ij} использовать $\frac{\partial L}{\partial W_{ij}}$. Здесь lr играет роль в регулировании интенсивности изменения, чтобы предотвратить перерегулирование, и он называется скоростью обучения. Кроме того, производная функции активации P является комплексной функцией, и для нахождения производной комплексной функции используется производная Виртингера [15].

Softmax с комплексным значением: функция softmax, которая используется в качестве функции активации в выходном слое, принимает входные значения и нормализует каждое значение в диапазоне от 0 до 1. Она гарантирует, что сумма нормализованных значений всегда равна 1. Эта функция позволяет интерпретировать результаты как вероятности. Это можно выразить уравнением (22).

$$\text{Итак, } f_{tmax}(Z)_i = e^{Z_i} \frac{e^{Z_i}}{\sum_{i=1}^{H}} \quad (22)$$

Однако функция softmax может интерпретировать результаты как вероятности только для входных данных действительных чисел, а для входных данных комплексных чисел необходим другой метод интерпретации. В этой статье мы предлагаем комплекснозначный метод интерпретации softmax для входных комплексных чисел. Результат функции softmax для входных комплексных чисел дает комплексные выходные значения, и поскольку сумма всех выходных значений равна 1+0j, мы предлагаем алгоритм, который вычисляет расстояние между выходными значениями и 1 + 0j в комплексной плоскости и выбирает индекс ближайшего значения. Это можно выразить уравнением (24).

$$\text{расстояние}_i = |S_{of\,tmax}(Z)_i - \omega| \quad (23)$$

$$\text{ближайший_индекс} = \text{argmin}(\text{расстояние}) \quad (24)$$

Здесь определено ω = 1 + 0j . Следовательно, мы решили классифицировать комплексное число соответствующий ближайшему_индексу как класс 1.

5. Анализ производительности. В

наших экспериментах по распознаванию речевых сигналов было проведено детальное сравнение результатов вещественных и комплексных нейронных сетей.
Анализ производительности проводился наряду с предварительной обработкой комплекснозначных кепстральных коэффициентов мел-частоты (MFCC) для уменьшения размерности. Кроме того, наблюдалось влияние комплексных MFCC на результаты распознавания речи. В частности, при распознавании заданных речевых сигналов с помощью нейронных сетей оценивалось влияние значений MFCC и их эффект комплексификации на производительность модели и сравнивалось с целью определить, можно ли повысить производительность при уменьшении количества параметров.

Экспериментальный метод: на рисунке 6 изображена архитектура многослойной комплекснозначной нейронной сети, используемой для обучения и тестирования с набором данных, состоящим из 8000 коротких речевых образцов, каждый из которых содержит восемь разных слов и длится менее одной секунды. Речевые данные, использованные в этом исследовании, являются частью набора данных речевых команд, собранного Google и распространяемого по лицензии CC BY.

Machine Translated by Google

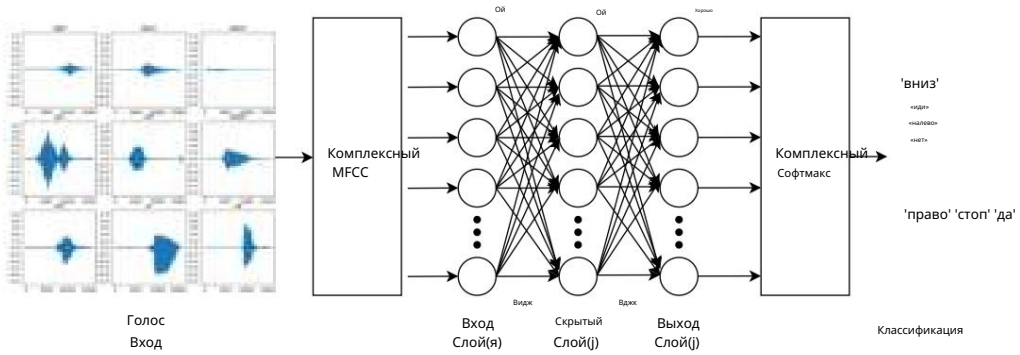


Рисунок 6. Экспериментальная установка многослойной CVNN.

Нейронные сети с комплексными значениями структурно похожи на нейронные сети с действительными значениями, но разница заключается в использовании комплексных чисел, которые вводят фазовые изменения, когда сумма произведений между входными данными и весами вращается в комплексной плоскости. Эти фазовые изменения играют важную роль в комплексных нейронных Комплексная нейронная сеть состоит из входного слоя с комплекснозначными MFCC уменьшенной размерности, полученными из речевых данных посредством процесса комплексификации, показанного на рисунке 3, скрытого слоя с 500 узлами и выходного слоя с 8 узлами для классификация. Все узлы и веса используют комплексные числа. Окончательные результаты классифицировались на выходном слое. Набор данных включал 7000 образцов для обучения и 1000 образцов для тестирования. В качестве данных использовались аудиоклипы из восьми слов, и эксперименты проводились путем деления входных данных в зависимости от того, применялось ли комплексное преобразование MFCC или нет. В случае использования вещественных MFCC использовалась вещественная нейронная сеть, тогда как комплексная нейронная сеть использовалась с комплексными MFCC. Нейронные сети как с действительными, так и с комплексными значениями обучались в одинаковых условиях: lr = 0,01, размер пакета = 50 и количество эпох = 50. Для входных данных использовались результаты трех различных этапов предварительной обработки уменьшения размерности. сравнивались с действительной нейронной сетью. Эти этапы предварительной обработки включали размеры MFCC 20, 10 и 5. Длина входного сигнала для нейронной сети может быть определена, как показано в уравнении (25). В частности, входной речевой сигнал был разделен на 32 кадра, и для каждого кадра генерировались MFCC.

$$\text{Длина данных входного сигнала} = \text{MFCC} \times \text{кадры} \tag{25}$$

Например, в случае комплексного значения MFCC 20 сначала вычисляется вещественное значение MFCC 40. Используя уравнение (25), длина входного сигнала для действительного MFCC 40 становится 40 × 32 = 1280. Если мы затем преобразуем эти 40 MFCC в комплексные числа, как показано на рисунке 3, уменьшение размерности будет достигнуто до половины длины. , в результате чего получилось 16 кадров. Следовательно, длина входного сигнала становится 40 × 16 = 640, что эквивалентно длине входного сигнала действительного MFCC 20. Аналогичное уменьшение размерности может быть выполнено для комплексного MFCC 10 и комплексного MFCC 5, используя тот же метод . . Общее количество параметров для всей нейронной сети, полученной в результате этого процесса, представлено в таблице 1. В этой таблице I(n) представляет количество нейронов во входном слое, H(n) представляет количество нейронов в скрытом слое. слой, а O(n) представляет количество нейронов в выходном слое.

Таблица 1. Общее количество параметров в нейронной сети.

МФЦК	Реальный MFCC	Комплекснозначный MFCC
20	I(20 × 32 = 640) × H(500) + H(500) × O(8) = 324000	I(40 × 16 = 640) × H(500) + H(500) × O(8) = 324000
10	I(10 × 32 = 320) × H(500) + H(500) × O(8) = 164 000	I(20 × 16 = 320) × H(500) + H(500) × O(8) = 164 000
5	I(5 × 32 = 160) × H(500) + H(500) × O(8) = 84 000	I(10 × 16 = 160) × H(500) + H(500) × O(8) = 84 000

Machine Translated by Google

MFCC 20 с комплексным значением и MFCC 20 с действительным значением. Как показано на рисунке 7, MFCC 20 с комплексным значением достиг максимальной точности обучения 99,40 % и точности тестирования 84,90 %, несмотря на ту же длину входных данных, что и вещественное значение. MFCC 20, который достиг максимальной точности обучения 98,13% и точности тестирования 83,10%. Результаты , полученные с использованием комплексных MFCC, превзошли результаты, полученные с использованием действительных MFCC, с точки зрения точности обучения и тестирования.

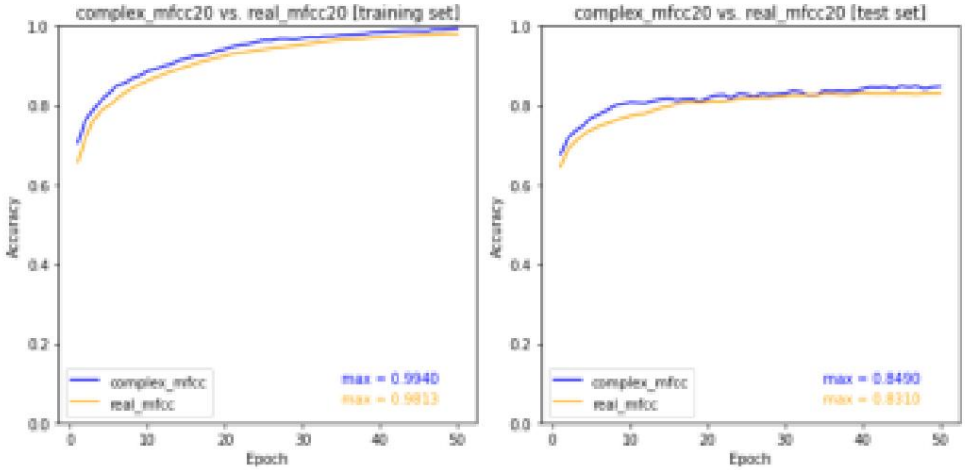


Рисунок 7. Сравнение комплексных и действительных результатов MFCC 20.

MFCC 10 с комплексным значением и MFCC 10 с действительным значением: Как показано на рисунке 8, MFCC 10 с комплексным значением достиг максимальной точности обучения 96,90% и точности тестирования 83,80%, несмотря на ту же длину входных данных, что и вещественное значение. MFCC 10, который достиг максимальной точности обучения 94,24% и точности тестирования 81,20%. Результаты , полученные с использованием комплексных MFCC, превзошли результаты, полученные с использованием действительных MFCC, с точки зрения точности обучения и тестирования.

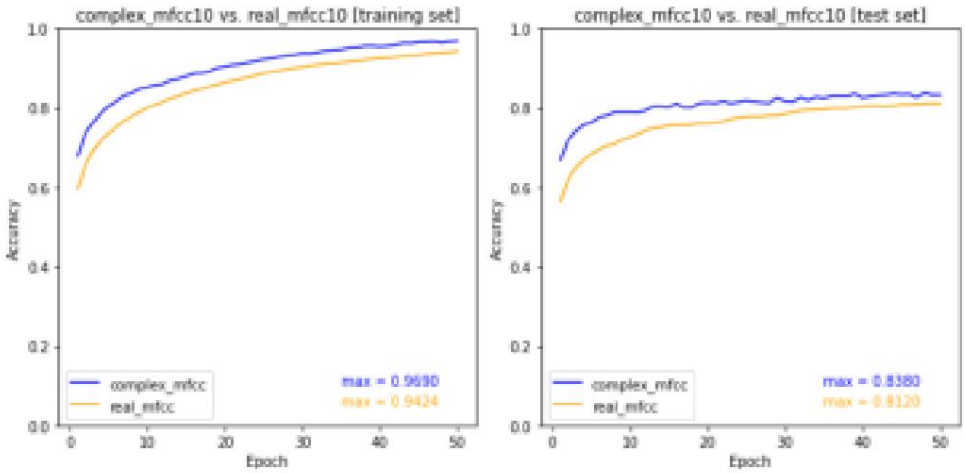


Рисунок 8. Сравнение комплексных и действительных результатов MFCC 10.

Комплексный MFCC5 и действительный MFCC5: Как показано на рисунке 9, комплексный MFCC 5 достиг максимальной точности обучения 92,33% и точности тестирования 81,40%, несмотря на ту же длину входных данных, что и вещественный MFCC 5. , который достиг максимальной точности обучения 86,34% и точности тестирования 74,90%. Результаты , полученные с использованием комплексных MFCC, превзошли результаты, полученные с использованием действительных MFCC, с точки зрения точности обучения и тестирования.

Machine Translated by Google

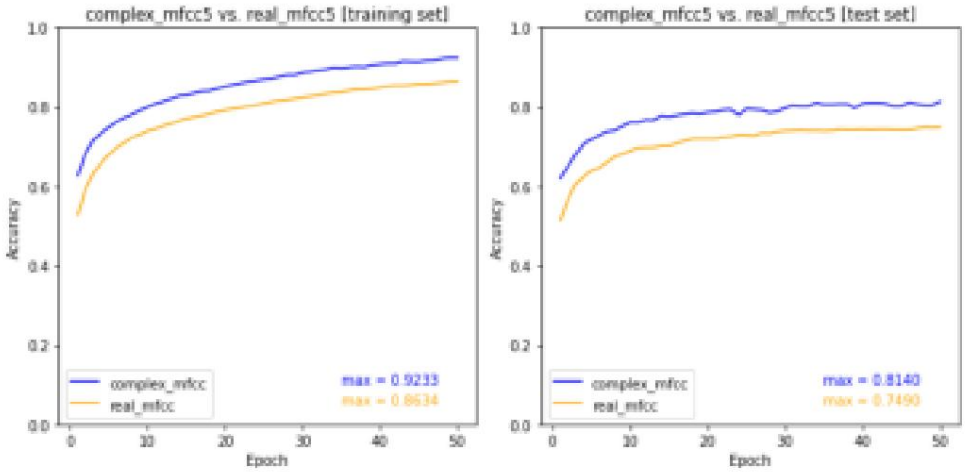


Рисунок 9. Сравнение комплексных и действительных результатов MFCC 5.

Общие результаты анализа: При рассмотрении экспериментальных результатов было замечено, что использование MFCC с комплексными значениями дает более высокую производительность по сравнению с использованием MFCC с действительными значениями, даже при одинаковой длине входных данных. Таким образом, было подтверждено, что использование комплексных MFCC и комплексных нейронных сетей приводит к лучшим результатам. Изменение точности в зависимости от количества функций MFCC показано на рисунке 10. Хотя здесь это не показано, те же результаты были получены с использованием оценки F1.

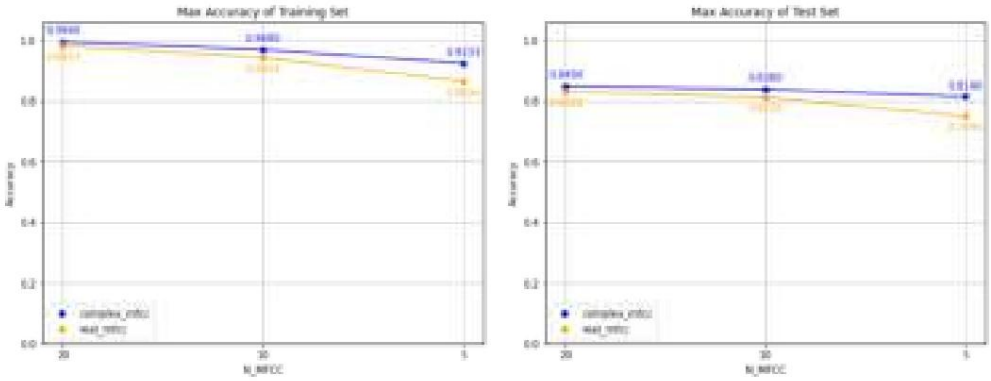


Рисунок 10. Изменение точности в зависимости от количества MFCC.

Если имеется достаточное количество MFCC, с увеличением эпох оказывается, что производительность при использовании действительных чисел начинает приближаться к производительности при использовании комплексных чисел. Однако на рисунке 9 показан другой результат. Комплексный MFCC 5 имеет 5 размерно уменьшенных комплексных MFCC, полученных из 10 MFCC посредством процесса комплексификации, показанного на рисунке 3. MFCC 5 с действительным знаком только с 5 MFCC не может догнать комплексный MFCC 5 даже при увеличении эпох. . Другими словами, когда количество MFCC невелико, даже при увеличении эпох результаты MFCC с действительным знаком не соответствуют производительности MFCC с комплексным знаком. Как показано на рисунке 10, несмотря на одинаковое количество параметров, результаты MFCC 5 с комплексными значениями превосходят результаты MFCC 5 с действительными значениями. Таким образом, было замечено, что при использовании MFCC с комплексными значениями изменение точности по мере уменьшения количества функций MFCC происходит более плавно по сравнению с использованием вещественных MFCC.

6. Выводы

В этой статье мы представили комплексные нейронные сети для решения проблемы распознавания речи. В наших экспериментах использовались комплексные MFCC в качестве входных данных дл

Machine Translated by Google

комплексная нейронная сеть. Результаты продемонстрировали превосходную производительность при использовании MFCC с комплексными значениями с более постепенным изменением точности по мере уменьшения количества функций MFCC по сравнению с использованием MFCC с действительными значениями. Это говорит о том, что комплексные MFCC могут помочь поддерживать производительность при меньшем количестве функций. Эти результаты дают ценную информацию о потенциале комплексных нейронных сетей и демонстрируют эффективность сочетания комплексных MFCC с комплексными нейронными сетями в распознавании речи.

Несмотря на эти многообещающие результаты, мы признаем определенные ограничения нашего подхода . Это исследование было проведено на ограниченном наборе данных, который может не полностью отражать разнообразие реальных речевых моделей. Более того, архитектура комплексной нейронной сети, описанная в этом исследовании, относительно проста и не может охватить все тонкости речевых сигналов.

Заглядывая в будущее, мы стремимся развивать эту работу для дальнейшего повышения точности и производительности моделей обработки и распознавания речевых сигналов. Наши будущие исследования будут сосредоточены на разработке более сложных моделей нейронных сетей с комплексными значениями и проведении экспериментов на различных наборах данных для повышения производительности. Кроме того, мы планируем изучить применение комплекснозначных нейронных сетей в других областях, помимо распознавания речи, потенциально расширяя их использование в различных областях обработки сигналов [16].

Вклад авторов: Концептуализация, SK; администрация проекта, депутат; надзор, депутат; написание — первоначальный вариант рукописи, С.К. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Эта работа проводилась при совместной поддержке Национального исследовательского фонда Кореи (NRF) посредством гранта, предоставленного правительством Кореи (MSIT) (грант № NRF-2023R1A2C1005461), а также MSIT (Министерством науки и информационных технологий), Корея. , в рамках Программы поддержки бизнеса по обучению талантов Convergence Security (IITP-2024-RS-2024-00426853), контролируемой IITP (Институтом планирования и оценки информационных и коммуникационных технологий).

Заявление о доступности данных: оригинальные материалы, представленные в исследовании, включены в статью. Дополнительные запросы можно направить соответствующим авторам.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Тебельскис, Дж. Распознавание речи с использованием нейронных сетей. Кандидат наук. Диссертация, Школа компьютерных наук, Университет Карнеги-Меллон, Питтсбург, Пенсильвания, США, 1995.
2. Саррофф, АМ Комплексные нейронные сети для аудио. Кандидат наук. Диссертация, Дартмутский колледж, Ганновер, Нью-Хэмпшир, США, 2018.
3. Абдул, З.К.; Аль-Талабани, А.К. Частотный кепстральный коэффициент Мела и его применение: обзор. IEEE Access 2022, 10, 122136–122158. [Перекрестная ссылка]
4. Баррачина, Дж.А.; Рен, К.; Вьейяр, Г.; Мориссо, К.; Оварлес, Дж. П. Теория и реализация комплексных нейронных сетей Сети. arXiv 2023, arXiv:2302.08286.
5. Айзенберг И. Комплекснозначные нейронные сети с многозначными нейронами; Шпрингер: Берлин/Гейдельберг, Германия, 2016.
6. Грейвс, А.; Джейтли, Н.; Мохамед, А.Р. Гибридное распознавание речи с глубоким двунаправленным LSTM. В материалах семинара IEEE 2013 г. по автоматическому распознаванию и пониманию речи, Оломоуц, Чешская Республика, 8–12 декабря 2013 г.; стр. 273–278. [Перекрестная ссылка]
7. Пас, П.; Гарридо, М. Эффективная реализация комплексных умножителей на ПЛИС с использованием срезов DSP. J. Сигнальный процесс. Сист. 2023, 95, 543–550. [Перекрестная ссылка]
8. Ахмад, М.; Чжан, Л.; Чоудхури, МЕН Реализация FPGA комплексной нейронной сети для полярно представленных Классификация изображений. Датчики 2024, 23, 2627. [CrossRef] [ПабМед]
9. Анггрени, Д.; Санджая, штат Вашингтон; Нурасийдик, МЮ; Мунаваро, М. Реализация распознавания речи с использованием мел- частотных коэффициентов кепстра (MFCC) и метода опорных векторов (SVM) на основе Python для управления манипулятором робота. Конференция ИОП. Сер. Матер. наук. англ. 2018, 288, 012042. [CrossRef]
10. Дханжал, А.С.; Сингх, В. Комплексное исследование автоматического распознавания речи с использованием нейронных сетей; Springer Nature: Берлин/Гейдельберг, Германия, 2023 г.
11. Смит, С.В. Руководство для ученых и инженеров по цифровой обработке сигналов, 2-е изд.; Калифорнийское техническое издательство: Сан-Диего, Калифорния, США, 1999 г.; стр. 557–558.
12. Бэсси Дж.; Ли, Х.; Цянь, Л. Обзор комплексных нейронных сетей. arXiv 2021, arXiv:2101.12249v1.
13. МОЙНейралНет. Комплексные нейронные сети — эксперименты. 2016. Доступно онлайн: http://makeourownneuralnetwork. blogspot.com/2016/05/complex-valued-neural-networks.html (по состоянию на 19 февраля 2021 г.).

Machine Translated by Google

14. Фишер Р. Исчисление Виртингера. В области предварительного кодирования и формирования сигналов для цифровой передачи; John Wiley & Sons, Inc.: Хобокен, Нью-Джерси, США, 2002 г.; стр. 405–413.
15. ПиТорч. Пакет автоматической дифференциации — torch.autograd, без даты Доступно в Интернете: https://pytorch.org/docs/stable/notes/autograd.html (по состоянию на 3 ноября 2023 г.).
16. Ли, К.; Хасэгава, Х.; Гао, С. Нейронные сети с комплексными значениями: комплексный обзор. IEEE/CAA J. Autom. Грех. 2022, 9, 1433–1454. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Рекомендации по работе: Сравнительный анализ совместной фильтрации Методы для объявлений

Роберт Квечинский 1,2, Томаш Гурецкий 2,*, Агата Филиповская 3 и Вячеслав Дубров 1

1 OLX Group, Królowej Jadwigi 43, 61-872 Познань, Польша; r.kwiecinskipl@gmail.com (PK); slavadubrov@gmail.com (ВД)
2 Факультет математики и информатики, Университет Адама Мицкевича, Uniwersytetu Pozna´nskiego 4, 61-614 Pozna´n, Польша
3 Кафедра информационных систем, Познаньский университет экономики и бизнеса, Ал. Niepodległo´sci 10, 61-875 Познань, Польша; agata.filipowska@ue.poznan.pl
* Переписка: tomasz.gorecki@amu.edu.pl

Аннотация: Объявления создают многочисленные проблемы для методов рекомендаций, включая временную видимость рекламы, анонимность большинства пользователей и тот факт, что обычно только один пользователь может потреблять рекламируемый товар. В этой работе мы решаем эти проблемы, выбирая модели и процедуры оценки, которые считаются точными, разнообразными и эффективными (с точки зрения потребления памяти и времени во время обучения и прогнозирования). Целью данной статьи является сравнение различных методов рекомендаций для объявлений о вакансиях на примере OLX Jobs для повышения коэффициента конверсии рекламы и удовлетворенности пользователей. В наших исследованиях мы реализуем масштабируемые методы и представляем различные подходы к рекомендациям: метод альтернативных наименьших квадратов (ALS), LightFM, Prod2Vec, RP3Beta и разреженные линейные методы (SLIM). Мы провели A/B-тестирование, отправив миллионы сообщений с рекомендациями провести онлайн-оценку выбранных методов. Кроме того, мы опубликовали набор данных, созданный для нашего исследования. Насколько нам известно, это первый набор данных такого рода. Он содержит 65 502 201 событие, выполненное в OLX Jobs 3 295 942 пользователями, которые взаимодействовали (отображали, отвечали или добавляли в закладки) 185 395 объявлений о вакансиях за две недели в 2020 году. Мы демонстрируем, что RP3Beta, SLIM и ALS при тестировании работают значительно лучше, чем Prod2Vec и LightFM. в лабораторных условиях. Онлайн-тесты A/B также показывают, что отправка сообщений с рекомендациями, сгенерированными моделями ALS и RP3Beta, увеличивает количество пользователей, обращающихся к рекламодателям. Кроме того, RP3Beta оказал на этот показатель на 20% более существенн

Ключевые слова: рекомендации по трудоустройству; объявления; совместная фильтрация; тестирование масштабируемости; А/Б-тесты; рекомендательные системы; матричная факторизация; рекомендательные системы на основе графов; Набор данных о взаимодействии с работой OLX ; показатели разнообразия

check for updates

Цитирование: Квечинский, Р.; Гурецкий, Т.; Филиповская А.; Дубров В. Рекомендации по работе: бенчмаркинг методов совместной фильтрации для Объявления. Электроника 2024, 13, 3049. https://doi.org/10.3390/электроника13153049

Академические редакторы: Марко Ломбарди, Доменико Сантаньелло, Зилу Лян и Мухаммад Хан

Поступила: 30 мая 2024 г.
Пересмотрено: 24 июля 2024 г.
Принято: 27 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Интернет-объявления — это веб-сайты, на которых рекламодатели размещают объявления о продаже или аренде услуг или продуктов [1]. В этой статье рассматривается онлайн-реклама на примере OLX Jobs.

Объявления стремятся удовлетворить потребности как рекламодателей, так и тех, кто отвечает на онлайн-рекламу. Чтобы соответствовать этим группам, более эффективные методы рекомендаций становятся все более важными. По сравнению с традиционной электронной коммерцией (например, Amazon (https://www.amazon.com/, по состоянию на 26 июня 2024 г.) или eBay (https://www.ebay.com/, по состоянию на 26 июня 2024 г.)), объявлениям приходится сталкиваться с различными проблемами, которые мы решаем в нашей процедуре оценки. Одним из них является временная видимость рекламы, которую мы решили решить путем частого переобучения модели. Следовательно, эффективность модели, о которой мы сообщаем, имеет важное значение для оценки. Другая проблема заключается в том, что рекламодателю обычно нужен только один пользователь для завершения транзакции. Этот аспект был решен путем включения анализа разнообразия в нашу оценку.

Machine Translated by Google

В наших исследованиях мы фокусируемся на подходах, возникающих на основе совместной фильтрации. Модели совместной фильтрации доказали свою точность, даже несмотря на то, что они не используют дополнительную информацию о пользователях или объектах, которую иногда трудно получить или обработать на практике. Целью данной статьи является оценка различных методов рекомендаций для использования в секторе объявлений. Высокая разреженность нашего набора данных, а также значительное количество рекламы и пользователей усложняют задачу исследования.

Ниже мы перечисляем наш основной вклад.

- Мы всесторонне оцениваем разнообразные, масштабируемые и совместные подходы к фильтрации в лабораторных условиях на предмет точности, разнообразия и эффективности. Мы оцениваем точность для разных групп пользователей в зависимости от количества элементов, с которыми они взаимодействовали, а также статистический анализ различий. Предлагаемая методология оценки решает основные проблемы, связанные с объявлениями.
- Мы представляем результаты двух онлайн-A/B-тестов, демонстрирующих влияние моделей ALS и RP3Beta на пользователей OLX. В каждом тесте участвовало более 1 миллиона пользователей. Сообщаемое влияние может быть использовано для оценки важности рекомендательных систем в сфере объявлений и позволит другим компаниям принимать более обоснованные решения о разработке таких моделей.
- Мы публикуем и представляем набор данных о взаимодействии на работе. Насколько нам известно, это самый крупный общедоступный набор данных о взаимодействии на рабочем месте.

Данная статья состоит из семи разделов. Раздел 2 представляет собой обзор литературы по основным разработкам. В разделе 3 описан набор данных, разработанный и опубликованный на Kaggle. План лабораторного эксперимента, а также реализация методов описаны в разделе 4. Раздел 5 посвящен результатам лабораторной оценки наших реализаций, а также статистической значимости результатов. В разделе 6 представлены результаты онлайн-внедрений, выполненных в ходе A/B-тестирования с пользователями, а также обсуждение результатов. Выводы представлены в разделе 7.

2. Сопутствующая работа

В этом разделе мы начнем с описания уникальных характеристик объявлений и областей вакансий. Затем мы предоставим обзор рекомендательных систем. Далее мы обсудим факторы, которые влияют на наш выбор моделей, и предоставим подробное описание каждой модели. Наконец, мы указываем пробелы в исследованиях.

2.1. Сайты объявлений. Сайты

объявлений (объявления) представляют собой онлайн-версии разделов газетной рекламы с объявлениями, касающимися транспортных средств, недвижимости, трудоустройства, домашних животных и т. д. Объявления представляют собой особый тип сайтов электронной коммерции, поскольку они не поддерживают завершение транзакций между сторонами — вместо этого они используются для общения о продукте или услуге и для назначения встреч, пока транзакция выполняется в автономном режиме; это отличает их от традиционных сайтов электронной коммерции, таких как eBay или Amazon [2].

Наши исследования сосредоточены на компании OLX, которая присутствует более чем в 30 странах мира. Одним из типов объявлений, публикуемых на объявлениях, являются предложения о работе. В ходе нашего исследования мы выявили несколько характеристик, которые отличают объявления о вакансиях на рекламных объявлениях, например, от объявлений о продаже автомобилей или недвижимости:

- Требования к рекламе указывают тип пользователя, который может реагировать на рекламу, а не все пользователи обладают необходимыми компетенциями.
- Более интенсивное вовлечение пользователя в процесс поиска, поскольку поиск работы является первоочередной необходимостью по сравнению, например, с покупкой автомобиля.
- Местоположение работы существенно влияет на пользователя: чтобы купить мобильный телефон или автомобиль, пользователь может путешествовать или воспользоваться службой доставки посылок для доставки товара. Тем не менее, работа требует переезда, если она находится далеко от дома пользователя.
- С технической точки зрения описание вакансии имеет множество качественных аспектов, и в зависимости от должности различия могут быть существенными даже между объявлениями одной и той же компании.

Machine Translated by Google

Помимо этих особенностей, в ходе нашего исследования мы также выявили следующие проблемы, связанные с методами рекомендаций по объявлениям, которые повлияли на нашу реализацию этих методов:

- Количество пользователей и предложений: с помощью OLX Jobs мы имели дело с миллионами пользователей и десятками тысяч предложений, которые постоянно менялись. •
Пользователю не нужно создавать профиль для взаимодействия с рекламой. Это означает, что информация из профиля не может быть использована для рекомендаций, а рекомендации определяются поведением пользователя во время просмотра.
- Объявление обычно имеет ограниченное количество посетителей (взаимодействующих с объявлением), поскольку объявление обычно касается уникального предложения. После устранения потребности реклама отключается. • Нет информации о том, произошла ли транзакция; на объявлениях конверсия связана с получением ответа на опубликованное объявление, поскольку любая транзакция происходит офлайн.

Эти проблемы создают трудности не только для реализации, но и для проверки точности предлагаемых методов. Мы учли эти проблемы при подготовке набора данных, а также при реализации и оценке методов, описанных в этой статье.

2.2. Рекомендательные системы

Было показано, что проблема избыточного выбора [3] может снизить потребление продукции [4] и снизить доход сайта [5]. Поставщики могут запрашивать у своих пользователей дополнительную информацию об их ожиданиях в отношении товаров, чтобы уменьшить количество представленных вариантов. Даже если пользователь готов потратить время на предоставление этой информации, количество возможностей обычно слишком велико, чтобы представить его пользователю. Чтобы решить эту проблему, многие компании, такие как Netflix [6] и Amazon [7], создали системы персонализированных рекомендаций, роль которых заключается в предложении пользователям соответствующих товаров [8].

Рекомендации используются во многих областях, включая музыку, туризм, моду и еду. Многочисленные исследования были проведены в области электронной коммерции [9], объявлений [10] и рекомендаций по работе [11–13]. Ли и др. [14] значительно улучшили свои рекомендации в LinkedIn, используя глубокое трансферное обучение для создания моделей понимания работы, специфичных для конкретной предметной области. Чжу и др. [15] предложили модель междоменных рекомендаций, использующую гетерогенный граф и алгоритм обнаружения сообществ для рекомендации курсов и вакансий. Лачич и др. [16] оценили различные архитектуры автокодировщиков для рекомендаций по заданиям на основе сеансов и представили результаты на популярных наборах данных рекомендаций по заданиям.

Обычно мы различаем два подхода к предоставлению рекомендаций: рекомендации на основе контента и совместную фильтрацию [8].

В рекомендациях на основе контента [17] мы используем информацию о пользователях (например, хобби, образование, навыки) и товарах (например, цена, местоположение, категория). Эти системы обычно используют историю взаимодействия (рейтинги, посещения, покупки и ответы) одного пользователя, но не учитывают рейтинги всех пользователей одновременно.

Сбор/обработка информации о пользователях и объектах не всегда осуществим. В подходе совместной фильтрации [18,19] мы используем только информацию о взаимодействии пользователя между элементами. Даже не имея информации о функциях пользователя и характеристиках товаров, эти системы могут дать очень точные рекомендации. Мы выбрали этот подход из-за отсутствия пользовательских функций и сложности превзойти совместную фильтрацию за счет использования функций элементов в производственной среде (подробнее об этом мы поговорим в разделе 6.3).

Одним из наиболее существенных недостатков совместной фильтрации является невозможность предоставления рекомендаций относительно пользователей и элементов без взаимодействия, известная как проблема холодного запуска [20]. Одним из возможных решений этой проблемы является частое переобучение модели. Учитывая множество взаимодействий с пользователем, эта стратегия требует сосредоточения внимания на масштабируемости модели [21].

Еще одним важным фактором, определяющим выбор модели, является тип взаимодействия пользователя и предмета [20]. Мы называем обратную связь явной, когда пользователь явно указывает свое предпочтение предмету (например, оценивает его по шкале от 1 до 5). Другой тип — неявная обратная связь, когда мы можем только предположить предпочтения пользователя по отношению к элементу на ос

Machine Translated by Google

о поведении пользователя (например, покупка товара или просмотр фильма). Неявная обратная связь менее точна, более обширна и обычно выражает только благоприятные предпочтения [22].

Лабораторные оценки рекомендательных систем обычно основаны на показателях точности [23]. В последнее время большое внимание уделяется другим аспектам, таким как разнообразие, неопределенность, новизна и охват [24,25]. В случае с объявлениями элемент обычно может быть использован только одним пользователем, поэтому нам следует избегать рекомендации одного и того же элемента значительному числу пользователей. Следовательно, необходимо сосредоточиться на охвате каталога [26] . Одной из наиболее сложных задач при оценке рекомендательных систем является поиск офлайн-метрики, наиболее коррелирующей с онлайн-целью [27].

В этой статье мы сравниваем модели совместной фильтрации для значительного набора данных неявной обратной связи. Мы сравниваем точность, разнообразие и эффективность выбранных рекомендательных методов в лабораторных условиях, а затем доказываем качество методов в онлайн-тестах с пользователями.

2.3. Методы рекомендаций

Ниже мы перечисляем наиболее важные аспекты, влияющие на выбор метода нашего исследования.

- Популярность в литературе. Мы выбрали методы, эффективность которых была тщательно оценена. Все выбранные методы, кроме одного, цитировались как минимум сотни раз.
- Популярность в индустрии. Оценка этого аспекта была сложной задачей, поскольку компании редко делятся результатами проведенных экспериментов. Некоторые показатели популярности включают доступность реализации и количество людей, которые ее используют (например, это отражается в количестве звезд и вилок на GitHub). Мы также рассмотрели опыт специалистов по обработке данных, работающих в нашей компании, которые в прошлом применяли большинство выбранных методов для некоторых других случаев использования.
- Масштабируемость. Мы рассматривали только решения, способные обрабатывать миллионы пользователей и объектов. Кроме того, мы уменьшили проблему холодного запуска за счет частого переобучения модели . Следовательно, нам пришлось выбирать методы для обучения на нашем наборе данных в течение нескольких часов.
- Разнообразие подходов. Мы хотели протестировать методы, представляющие разные «семейства», чтобы оценить их применимость к классифицированным объявлениям. Мы планировали выявить наиболее перспективное семейство подходов для продолжения исследований в этом направлении.

Наконец, для тестирования и оценки были выбраны ALS, LightFM, Prod2Vec, RP3Beta и SLIM . В следующих подразделах описываются причины выбора каждого конкретного метода и его определение. Общим ограничением этих методов является их неспособность рекомендовать элементы, которые не посещал ни один пользователь, и генерировать рекомендации для пользователей , которые не посещали ни один элемент (проблема холодного запуска, [21]). С нашим набором данных, без особенностей пользователя и элементов, невозможно решить эту проблему каким-либо персонализированным методом.

Отобранные методы уже сравнивались в разных областях. Анелли и др. [28] сравнили RP3Beta, SLIM, ALS и BPRMF (что эквивалентно LightFM при использовании функции потерь BPR) на трех наборах данных: Amazon Digital Music, Epinions и MovieLens 1M, с точки зрения точности, разнообразия и эффективности. Количество взаимодействий в этих наборах данных составляет соответственно 145 523, 300 475 и 571 531. В нашем исследовании использовался набор данных, содержащий десятки миллионов взаимодействий, аналогичный тем, которые используются в производственных системах OLX.

Кроме того, мы предоставляем результаты онлайн-оценки.

Дакрема и др. [29] сравнили точность SLIM, RP3Beta и ALS на нескольких наборах данных (CiteULike-a, MovieLens 1M, Pinterest, Amazon Music, Amazon Movies, MovieLens 20M, Netflix, Yelp и Gowalla). Они показали, что каждый из этих алгоритмов превосходил два других по крайней мере на одном наборе данных. Некоторые из их выводов были дополнительно подтверждены Anelli et al. в наборах данных MovieLens 1M и Pinterest. [30]. Чжу и др. [31] сравнили BPRMF, SLIM и Prod2Vec на трех наборах данных (Amazon Books, Yelp и Gowalla), показав превосходство SLIM. Донг и др. [32] проанализировали точность RP3Beta, SLIM и ALS на девяти наборах данных, при этом SLIM показал лучшие результаты на восьми наборах данных и ALS.

Machine Translated by Google

на одной. В отличие от нашего исследования, в этих статьях выбранные методы не сравнивались с точки зрения эффективности.
разнообразие или эффективность, а также не проводили онлайн-оценку.

В последнее время появилось несколько методов рекомендаций, основанных на нейронных сетях.
предложенный. Было показано, что адекватно определенные и настроенные простые модели часто
превосходят их [28,31–33]. Кроме того, из-за вычислительной сложности модели, основанные на
нейронных сетях реже тестируются на наборах данных, содержащих десятки миллионов
взаимодействия [28]. Следовательно, мы оставляем на будущее исследование, чтобы предложить более продвинутые и
масштабируемые методы, которые обеспечивают превосходные результаты в нашем наборе данных.

2.3.1. Модели матричной факторизации

Методы матричной факторизации успешно используются в рекомендательных системах
после конкурса Netflix Prize [34]. Основная идея заключается в аппроксимации разреженных
Матрица взаимодействия пользователя и элемента как произведение двух меньших плотных матриц. Мы оцениваем два
подходы матричной факторизации: LightFM и ALS.

Модель LightFM [35] была предложена для решения проблемы холодного запуска методов матричной
факторизации. В нашем случае это сводится к классической матричной факторизации без дополнительных
информация о пользователях и предметах. Несмотря на это, мы решили использовать этот подход из-за
его эффективная официальная реализация при поддержке нескольких функций потерь: логистика, BPR [36],
WARP [37] и -OS WARP [38].

Пусть U — набор пользователей, а I — набор элементов. Для данного пользователя € U и предмета € I мы
define = 1, если пользователь взаимодействует с элементом, , и = 0 в противном случае. Модель ЛайтФМ
прогнозирует оценку следующим образом:

$$\hat{} = \cdot + +\quad ,$$

где и обозначают трехмерные скрытые представления пользователя и элемента, а и — предвзятости
пользователя и элемента. Все эти параметры изучаются во время
тренировка.

Функцию логистических потерь следует использовать, когда как положительная, так и отрицательная обратная связь
типы доступны, поэтому он не подходит для нашего набора данных, что мы подтвердили
экспериментально во время настройки гиперпараметра.

Функции потерь BPR, WARP и -OS WARP представляют собой парные подходы к ранжированию,
которые более подходят для нашей задачи по выработке лучших рекомендаций. Они пытаются изменить
параметры модели, чтобы максимизировать разницу между оценками элементов D с
с каким пользователем взаимодействует и элементов D, с которыми пользователь не взаимодействует.

В частности, в функции потерь BPR [36] мы пытаемся максимизировать выражение:
следующее:

$$\ln(\hat{}\quad \hat{}) \quad \Theta\,||\Theta||2\ ,$$

$$(,,)\ S$$

где

$$S\ \text{знак равно}\ \{(,,)|\quad U,\quad D\ \text{и}\quad D\},$$

$$\text{" ="}\quad \frac{1}{1+}\ ,$$

Θ — параметр регуляризации, специфичный для модели, Θ — вектор параметров, а || · || это
Евклидова норма.

Мы представляем WARP [37] как случай WARP -OS [38]. Функция потери -OS WARP
сумма потерь для каждого пользователя:

$$(\hat{}\ ,Д),$$

$$€U$$

где ˆ — вектор оценочных оценок ˆ для каждого элемента I.

Machine Translated by Google

Для данного , мы упорядочиваем взаимодействующие элементы D в порядке убывания пользователя расчетные баллы. А именно, мы упорядочиваем элементы как 1, 2, . . . , | Д | , где индексы выбрано таким образом, что имеем следующее:

$$\hat{}_1 \quad \hat{}_2 \quad \ldots \quad \hat{}_{|Д|}.$$

Тогда у нас есть следующее:

$$(\hat{} , Д) = \frac{1}{} \sum_{=1}^{| Д |} ( ) Ф(\text{ранг} (\hat{})),$$

где ранг (^) = I\D (1 + ( ) ^), является индикаторной функцией, Ф( ) = $\sum_{=1} \frac{1}{}$, и = $\sum_{=1}^{| Д |}$ нормализует веса, индуцированные .

На практике мы выбираем положительный элемент, касающийся весовой функции, и случайным образом выбирать невидимые предметы до тех пор, пока после выборки не будет 1 + . Затем мы выполняем шаг градиента для минимизации Ф($\frac{|Я\Д|}{}$) (1 + ^).

Наконец, потеря -OS WARP достигается для ( ) = 1 и ( ) = 0, если . ВАРП потеря достигается, когда мы устанавливаем ( ) = 1 для N. Мы можем пропустить заказ положительных элементов из D, чтобы упростить вычисления.

В нашем исследовании учитывались все функции потерь (логистика, BPR, WARP и -OS WARP). в качестве гиперпараметров модели LightFM и оптимизируются во время обучения.

Модель ALS (попеременные наименьшие квадраты), также известная как WRMF – взвешенная регуляризованная матричная факторизация [39], является примером метода матричной факторизации, разработанного для наборов данных неявной обратной связи. Мы выбрали этот метод из-за его доказанной эффективности. масштабируемость и популярность в исследованиях и промышленности. Кроме того, модель БАС имеет уже реализован и используется в OLX.

Мы определяем достоверность наблюдения как = 1 + , где – гиперпараметр . Мы также определяем предпочтение как = (> 0). Затем в ALS мы оптимизируем следующее выражение:

$$\text{мин}_{,} \quad (-)^2 + (\quad | | | 2 + \quad | | | 2),$$

где — трехмерное внедрение пользователя, — трехмерное внедрение элемента, и является параметром регуляризации. Название модели относится к оптимизированному методу. для этой функции. Процедура попеременных наименьших квадратов предполагает, что пользователь или элемент скрыты. факторы фиксированы и аналитически определяют остальные. Процедура эффективно повторяется несколько раз [39,40].

В подходах матричной факторизации встраивание для каждого пользователя изучается отдельно, что может привести к увеличению использования времени и памяти, если число пользователей велико. Кроме того, рекомендации могут быть необъяснимыми, поскольку два пользователя с одни и те же взаимодействия могут получить разные рекомендации. Другие рассмотренные модели (SLIM, RP3Beta, Prod2Vec) предоставляют одинаковый набор рекомендаций для пользователей с те же взаимодействия.

2.3.2. Модели на основе соседства

Модель SLIM (разреженные линейные методы для рекомендательных систем Top-N [41]) на основе регрессии ближайших соседей на основе элементов [8]. Мы выбрали этот метод из-за его преимущества по сравнению с моделями ближайших соседей на основе элементов и моделями матричной факторизации [41] и хорошие результаты по сравнению с ненейронными и нейронными подходами [29].

В этой модели мы аппроксимируем матрицу пользователь-элемент, изучая матрицу элемент-элемент. разреженная матрица сходства. Явно это изучается путем минимизации следующего выражения:

$$\frac{1}{2} | | \quad | |2 + \frac{}{| | 2} \quad | |2 + | | \quad | | |1,$$

Machine Translated by Google

где 0 и = 0 для всех , , 1 , |Ɪ|, | | · | | – матричная норма Фробениуса:

$$|| \quad || = \sqrt{\sum_{=1}^{|Я|}\sum_{=1}^{|Я|} 2},$$

| | · | |1 — поэлементная ℓ1-норма:

$$|| \quad ||1 = \sum_{=1}^{|Я|}\sum_{=1}^{|Я|} | \quad |$$

и – параметры регуляризации.

Мы видим, что каждый столбец можно изучать независимо, что улучшает масштабируемость этой модели [41].

Ограничением этого метода является то, что он не генерирует внедрения пользователей и элементов, что может быть полезно для других задач (например, кластеризации пользователей или элементов). Мы можем рассматривать столбцы матрицы как вложения элементов, но они представляют собой разреженные |Ɪ|-мерные векторы, поэтому их, вероятно, нельзя напрямую использовать в других задачах.

2.3.3. Подходы на основе графов

RP3Beta [42] — это подход совместной фильтрации на основе графов. Аналогично Dacrema et al. В работе [29] мы представляем эту модель как расширение модели P3alpha [43]. Мы рассматриваем двудольный и неориентированный граф = (V, E), где каждая вершина V представляет либо пользователя, либо элемент, а каждое ребро { , } E представляет взаимодействие между данным пользователем и данным элементом. Пусть – матрица смежности и матриц Затем оценки P3alpha сохраняются в матрице

$$((\quad ^{1} \quad) \quad)^{3},$$

где обозначает поэлементную степень матрицы (степень Адамара). Точнее, оценка ˆ сохраняется на пересечении строки, представляющей пользователя, и столбца, представляющего элемент. Мы видим, что эта оценка представляет собой сумму оценок, присвоенных путям длиной 3, соединяющим пользователя и элемент на графе.

В RP3Beta оценка вычисляется с помощью P3alpha и делится на популярность элементов, возведенную в степень, чтобы смягчить смещение популярности.

Мы выбрали эту модель из-за ее простоты, масштабируемости и проверенной производительности [29,30], особенно с учетом высокой разреженности нашего набора данных, что значительно снижает использование памяти при обучении модели. Это позволяет нам напрямую вычислять и сохранять матрицу сходства между предметами в памяти вместо аппроксимации случайных блужданий, предложенной в оригинальной статье.

RP3Beta, как и SLIM, не создает встраивания пользователей и элементов. Еще одним связанным ограничением является невозможность рекомендовать элементы, расстояние которых на двудольном графе пользователь-элемент больше трех от данного пользователя. В частности, элемент может быть рекомендован пользователю только в том случае, если хотя бы один пользователь ′ посетил элемент и хотя бы один из элементов, посещенных пользователем . Кроме того, модель RP3Beta не имеет параметров, оптимизированных во время обучения; следовательно, он может быть не в состоянии использовать сложные отношения в данных.

2.3.4. Прод2Век

Модель Prod2Vec [44] (или Item2vec [45]) основана на модели Word2Vec [46], широко используемой при обработке естественного языка. Модель фокусируется на изучении векторных представлений элементов, используя последовательность элементов как «предложение», а элементы внутри этой последовательности как «слова». Этот метод был выбран потому, что он масштабируем и сильно отличается от сравниваемых методов. Кроме того, этот подход основан на современной системе рекомендаций гибридных глубоких нейронных сетей, используемой в других категори

Machine Translated by Google

Prod2Vec использует модель пропуска граммы, максимизируя целевую функцию по множеству S последовательностей элементов, определенных следующим образом:

$$= \sum_{\in S} \sum_{\in} \text{журнал } P(+ \mid),$$

где длина контекста для последовательностей элементов, а элементы из одной и той же последовательности заказано произвольно. Вероятность P(+ |) наблюдения соседнего элемента + при условии текущий элемент определяется с помощью функции soft-max:

$$P(+ \mid \quad " =" \frac{\text{опыт}(\quad ' \quad)}{\sum_{=1}^{|Я|} \text{опыт}(\quad ' \quad)},$$

где и — входные ' и выходные векторные представления элемента и |I| это , количество уникальных слов в словаре. Prod2Vec моделирует контекст последовательности элементов, где элементы со схожим контекстом (т. е. с похожими соседними взаимодействиями) имеют схожие векторные представления.

Прежде чем генерировать рекомендации, мы нормализуем все представления входных векторов, что повышает эффективность. Для формирования рекомендаций мы представляем пользователя как среднее значение входных векторных представлений взаимодействующих элементов. Затем ищем предметы чьи входные представления имеют наибольшее косинусное сходство с пользовательским представлением (-ближайшие соседи в скрытом пространстве).

Потенциальным ограничением этого метода является строгая зависимость между пользователем и элементом. представления. Некоторые взаимодействия потенциально должны оказывать большее влияние на представления пользователей, чем другие (что возможно в моделях матричной факторизации).

2.4. Пробел в исследованиях

Целью данной статьи является сравнение различных методов рекомендаций в области классификаций должностей, устраняя конкретный пробел в исследованиях. Насколько нам известно, рассматриваемые методы рекомендаций никогда не тестировались и не оценивались на крупных масштабируйтесь, работая онлайн в своей сфере деятельности. В данной работе мы сравниваем эти методы в лабораторных условиях с использованием опубликованного нами набора данных и онлайн-тестов A/B, проведенных в производственная среда.

3. Набор данных о взаимодействии с работой OLX

Это исследование предоставляет коммерческий набор данных, содержащий 65 502 201 событие, выполненное на OLX Jobs 3 295 942 пользователей, которые взаимодействовали с 185 395 объявлениями о вакансиях за две недели в 2020 году. Мы опубликовали набор данных на Kaggle (https://www.kaggle.com/olxdatascience/olx-jobs-interactions , доступ 26 июня 2024 г.).

Каждая строка набора данных представляет событие, которое данный пользователь выполняет относительно данного объявления в данный момент времени (табл. 1). Были включены следующие события, связанные с поведением пользователей:

- клик: пользователь посетил страницу с описанием товара;
- закладка: пользователь добавил объект в закладки;
- chat_click: пользователь открыл чат, чтобы связаться с владельцем предмета;
- contact_phone_click_1: пользователь раскрыл номер телефона, прикрепленный к элементу;
- contact_phone_click_2: пользователь нажал кнопку, чтобы позвонить владельцу элемента;
- contact_phone_click_3: пользователь нажал, чтобы отправить SMS-сообщение владельцу элемента;
- contact_partner_click: пользователь нажал, чтобы получить доступ к внешней странице владельца элемента;
- contact_chat: пользователь отправил владельцу элемента сообщение.

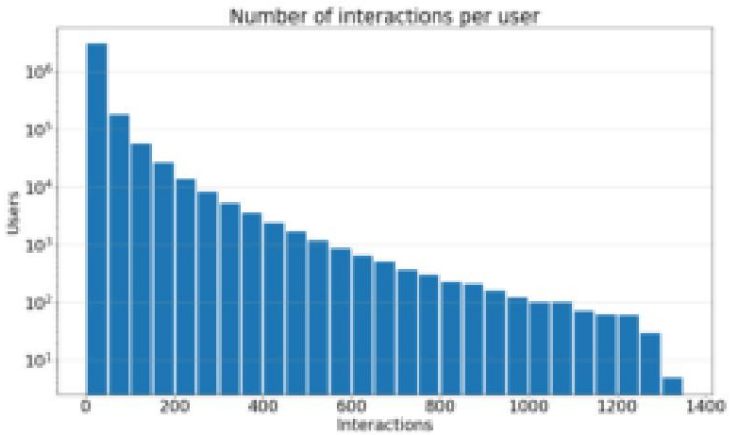
Machine Translated by Google

Таблица 1. Примеры строк из набора данных.

Пользователь	Элемент	Событие	Временная метка
1745587	168661		1582216025
843008	62838		1582485868
14285	30469	нажмите	1582247367
1142944	80122		1581805847
2835659	23728	нажмите закладку нажмите чат_клик	1582397836

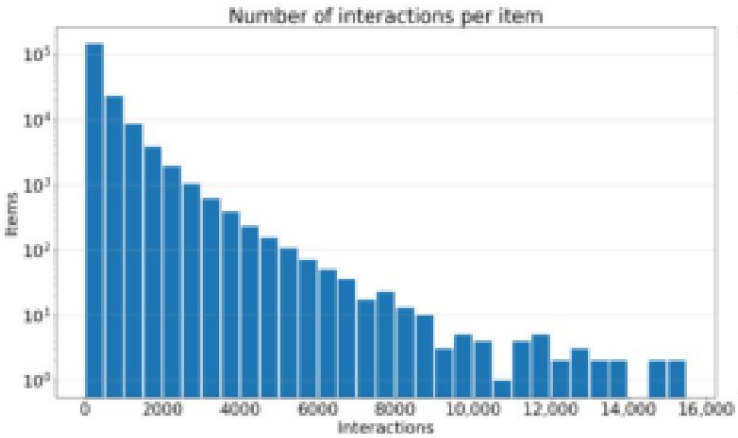
Источник: собственное исследование.

Разреженность данных (% отсутствующих записей в матрице пользователь-элемент) равна 99,9923%, что очень мало, но типично для классификаторов (для сравнения: разреженность данных на Набор данных MovieLens 20M составляет 99,47%, что делает его почти в 70 раз плотнее). Распределение взаимодействия на пользователя, взаимодействия на элемент и события представлены на рисунках 1 и 2. и таблица 2 соответственно. Мы использовали логарифмическую шкалу, поскольку распределения сильно выражены. перекошенный вправо.



Quantile	Value
0.01	1
0.10	1
0.25	2
0.50	6
0.75	18
0.90	46
0.99	221

Рисунок 1. Количество взаимодействий на пользователя. Источник: собственное исследование.



Quantile	Value
0.01	1
0.10	1
0.25	3
0.50	137
0.75	428
0.90	950
0.99	2944

Рисунок 2. Количество взаимодействий на элемент. Источник: собственное исследование.

По данным Де Рюйта и Бхулаи [47], наиболее популярными наборами данных, используемыми для оценки систем рекомендаций по вакансиям, являются CareerBuilder12 (https://www.kaggle.com/конкурсы/рекомендации по работе, по состоянию на 16 июня 2024 г.), RecSys16 [12] и Rec-Sys17 [13]. В таблице 3 представлена статистика рассматриваемых наборов данных. Количество уникальных Взаимодействия — это количество уникальных пар пользователь-элемент, в которых пользователь взаимодействовал с элемент. Несмотря на то, что исследователи используют наборы данных RecSys16 и RecSys17, они публикуются во время соревнований по науке о данных в строго ограниченных условиях и не являются дольше в наличии. Следовательно, мы не смогли подсчитать количество уникальных взаимодействий для этих

Machine Translated by Google

наборы данных. Мы заметили, что набор данных о взаимодействии с работой OLX имеет наибольшее количество взаимодействия и типы между рассматриваемыми наборами данных.

Таблица 2. Распределение событий.

Действие	Частота (%)
	89,794
нажмите contact_phone_click_1	2,628
закладку	2.511
chat_click	2.136
contact_chat	1,448
contact_partner_click	0,701
contact_phone_click_2	0,679
contact_phone_click_3 Источник:	0,103

собственное исследование.

Таблица 3. Сравнение наборов данных о вакансиях. Буква обозначает стандартное отклонение.

	OLX Вакансии Взаимодействия	CareerBuilder12	RecSys16	RecSys17
Взаимодействия	65,50 M	1,60 M	8,83 M	8,27 M
Уникальные взаимодействия	47,17 M	1,60 M	-	-
Пользователи	3,30 M	0,3 2M	1,37 M	1,50 M
Среднее количество уникальных взаимодействий на пользователя	14.31 (= 29,23)	4,99 (= 11,42)	<6,46	<5,51
Предметы	0,19 M	0,37 M	1,36 M	1,31 M
Среднее количество уникальных взаимодействия на элемент	254,42 (= 426,02)	4.38 (= 8,19)	<6,50	<6,31
Плотность матрицы взаимодействия (%)	0,0077%	0,0014%	<0,0005%	<0,0004%
Виды взаимодействия	8	1	4	5
Временные метки				
Возможности пользователя				
Особенности товара				

Источник: собственное исследование.

Следует отметить, что сохранение конфиденциальности рекламы и пользователей было приоритетом. при подготовке этого набора данных. Меры, принятые для защиты конфиденциальности, включали следующее:

- Исходные идентификаторы пользователей и элементов были заменены уникальными случайными целыми числами;
- К каждой временной метке было добавлено какое-то нераскрытое постоянное целое число;
- Некоторые части взаимодействий были отфильтрованы;
- Добавлены некоторые дополнительные искусственные взаимодействия.

4. Экспериментальная установка

4.1. Выполнение

Мы реализовали упомянутые выше методы на основе предыдущих разработок. Подробности о том, чем наши реализации отличаются от ранее протестированных подходов, см. представлены в таблице 4. Реализация методов для обеспечения воспроизводимости результатов доступно на GitHub (https://github.com/rob-kwiec/olx-jobs-recommendations, доступ 29 мая 2024 г.).

Machine Translated by Google

Таблица 4. Методы, повторно использованные для исследования.

Метод	Семейство методов	Источник	Отличие от Оригинальный метод
Факторизация матрицы LightFM		Реализация на основе: [48]. Сопроводительный документ: [35].	Никто
БАС	Матричная факторизация	Реализация на основе: [49]. Подтверждающие документы: [39,40].	Никто
СТРОЙНЫЙ	На основе района	Источники, вдохновившие нас на реализацию: [50,51]. Сопроводительный документ: [41].	Никто
RP3Beta	На основе графа	Источник, вдохновивший на нашу реализацию: [52]. Подтверждающие документы: [29,53].	Выполняется напрямую вычисления на разреженных матрицах вместо случайных приближение прогулок.
Прод2Век	Word2Vec	Реализация на основе: [54]. Подтверждающие документы: [44,45].	Последовательности взаимодействий, упорядоченные по временной метке. Представление пользователя как среднего из представлений взаимодействующих элементов. Мы также экспериментировали с CBOW (непрерывным набором слов).

Источник: собственное исследование.

4.2. Цели лабораторных исследований и онлайн-тестирования Нашей

конечной бизнес-целью было увеличить количество пользователей, подающих заявки на работу через OLX Jobs. Чтобы добиться этого, мы пытаемся решить проблему ранжирования; точнее, мы рекомендуем десять элементов с наибольшей вероятностью получения взаимодействия с пользователем. Вопрос поиска офлайн-метрики, наиболее коррелирующей с уровнем заявок в сфере работы, рассматривался Mogenet et al. [27]. Поскольку наша цель аналогична, мы последовали их результатам, оптимизировав точность @ 10 (точность для первых десяти рекомендуемых элементов [23]). Однако мы также сообщаем и обсуждаем другие показатели точности, разнообразия и эффективности.

Мы ориентированы на пользователей, которые активно ищут работу, взаимодействуя с объявлениями о работе. Поэтому мы не рассматриваем пользователей без взаимодействия (пользователей с холодным запуском).

4.3. Разделение данных

Набор данных, описанный в разделе 3, был разделен на наборы данных для обучения-проверки и тестирования, как того требует методология исследования. В тестовый набор входит 20% новейших взаимодействий. Это означает, что из 14 дней в тестовый набор вошли примерно 3 дня (см. рисунок 3). По мнению Meng et al., такой способ разделения набора данных на обучающий и тестовый наборы является наиболее реалистичным . [55], которые изучили несколько стратегий разделения для оценки рекомендательных систем.

В обоих наборах данных мы также ограничили количество взаимодействий уникальными взаимодействиями между пользователем и элементом. Если пользователь взаимодействовал с элементом более одного раза, учитывалось только первое взаимодействие, и с элементом была связана временная метка этого взаимодействия. Мы также не различали типы взаимодействий. Кроме того, были применены еще две модификации: в тестовый набор были включены только пользователи, которые фигурировали в наборе обучающих и проверочных данных (поскольку мы не могли предоставлять персонализированные рекомендации другим пользователям без дополнительных знаний о них), и мы отфильтровывали все пользовательские данные. пары элементов , которые присутствовали в обучающем наборе (чтобы не рекомендовать элементы, которые уже были просмотрены). В результате у нас получилось около 38 М строк в обучающем и проверочном наборе и около 6 М строк в тестовом наборе. Дополнительная статистика этих двух наборов данных представлена в Таблице 5.

Machine Translated by Google

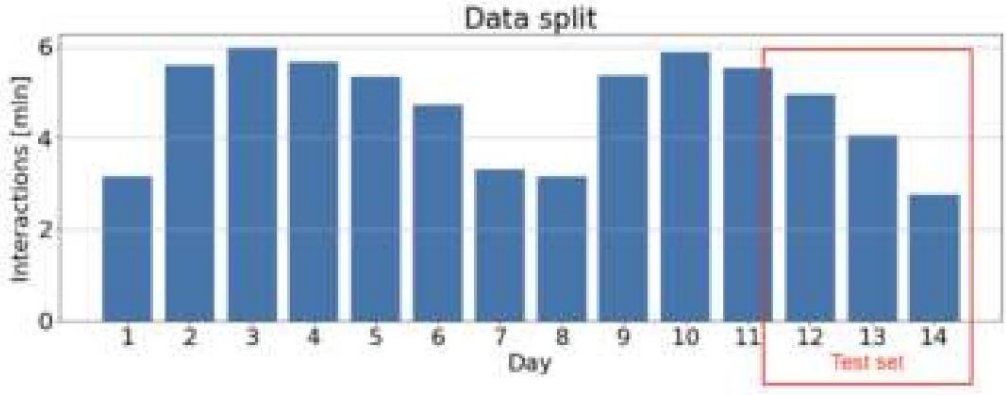


Рисунок 3. Подготовка набора данных (разделение на обучение-проверку и тестирование, в зависимости от количества дни). Источник: собственное исследование.

Таблица 5. Сравнение обучающих/проверочных и тестовых наборов. Буква обозначает стандарт. отклонение.

	Взаимодействие с сотрудниками OLX Взаимодействия	Набор для обучения и проверки	Тестовый набор
Взаимодействия	65,50 M	52,40 M 6,11 M	
Уникальные взаимодействия	47,17 M	38,25 млн 6,11 млн	
Пользователи	3,30 M	2,83 M 0,62 M	
Среднее количество уникальных взаимодействий на пользователя	14.31 (= 29,23)	13.50 (= 26,11)	9,87 (= 14,38)
Предметы	0,19 M	0,18 M 0,13 M	
Среднее количество уникальных взаимодействия на элемент	254,42 (= 426,02)	214.05 (= 365,93)	47,51 (= 83,84)
Плотность матрицы взаимодействия (%)	0,0077%	0,0076%	0,0077%

Источник: собственное исследование.

4.4. Тюнинг модели

В целях экономии средств и времени при выполнении оптимизации параметров мы сделали ограничение до 20% пользователей и 20% предметов обучения и валидации набор. Затем мы разделяем ограниченный набор данных на набор данных для обучения и проверки с помощью правила, описанные в разделе выше. Для каждой модели было выбрано 100 итераций. путем байесовской оптимизации с использованием гауссовских процессов, реализованных в scikit-learn (https://scikit-optimize.github.io/stable/modules/generated/skopt.gp_minimize.html, доступ 29 мая 2024 г.). Файл конфигурации определяет места для гиперпараметров и доступен в нашем репозитории кода (https://github.com/rob-kwiec/olx-jobs-recommendations/blob/основной/src/tuning/config.py, доступ 29 мая 2024 г.). Например, для RP3Beta мы ограничились поиском и в интервале [0, 2].

Рекомендации в процессе оптимизации подготовлены для 30 000 пользователей, а как упоминалось выше, использованный показатель оптимизации составлял точность @ 10. На рисунке 4 показано показатель точности в зависимости от параметров. Видно, что БАС не был высоко оптимизирован внутри процесса, но для других методов эта оптимизация имела решающее значение для улучшение точности. Гиперпараметры для всех моделей представлены в таблице 6.

Machine Translated by Google

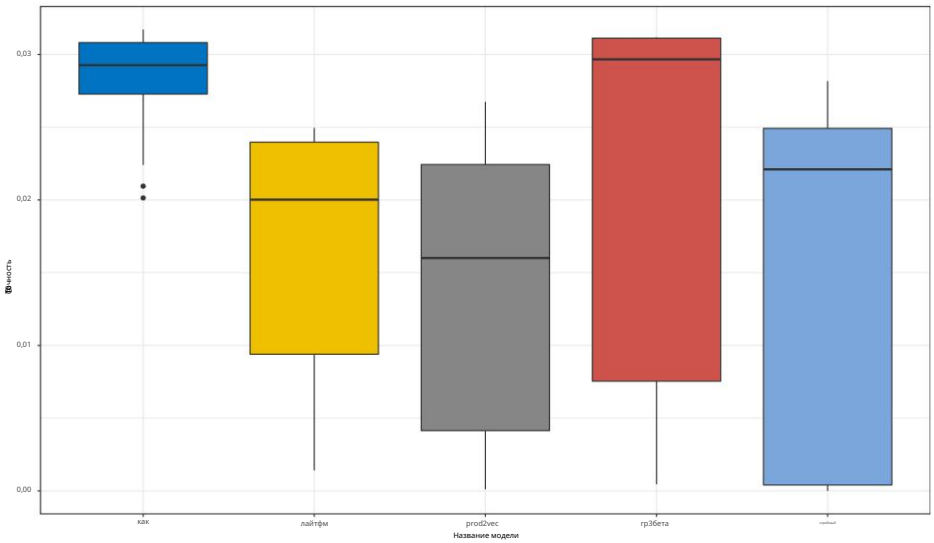


Рисунок 4. Точность для каждой модели в зависимости от параметров. Источник: собственное исследование.

Таблица 6. Гиперпараметры модели.

Модель	Гиперпараметры модели
как	{'факторы': 357, 'регуляризация': 0,001, 'итерации': 20, 'event_weights_multiplier': 63}
лайтфм	{'no_comComponents': 512, 'k': 3, 'n': 20, 'learning_schedule': «ададельта», «потеря»: «деформация», «max_sampled»: 61, «эпохи»: 11}
prod2vec	{'vector_size': 168, 'альфа': 0,028728, 'окно': 20, 'min_count': 16, «образец»: 0,002690026, «min_alpha»: 0,0, «sg»: 1, «hs»: 1, «отрицательный »: 200, «ns_exdependent»: -0,16447846705441527, «cbow_mean»: 0, 'эпохи': 22}
RP3Beta	{'альфа': 0,61447198, 'бета': 0,1443548}
тонкий	{'альфа': 0,00181289, 'l1_ratio': 0,0, 'итерации': 3}

Источник: собственное исследование.

4.5. Оценка модели

Мы оцениваем наши модели с точки зрения точности, разнообразия и эффективности.

Точность метода рекомендаций для наборов данных с неявной обратной связью обычно составляет оценивается путем изучения того, какие позиции из списка лучших рекомендаций пользователь взаимодействовали с тестовым набором [8]. Многие метрики оценки определяют метрики для каждого отдельно пользователю и представить среднее значение как окончательную оценку метода [23]. рекомендательную систему можно рассматривать как классификатор, в котором взаимодействие пользователя в набор тестов представляет собой положительные примеры, а верхние рекомендации — это элементы, классифицированные как положительный по модели. Следовательно, методы, используемые для оценки классификаторов, могут быть непосредственно применяется к рекомендательным системам. Например, значение точности@ для данного пользователя равно количество рекомендуемых элементов, с которыми пользователь взаимодействовал в тестовом наборе, разделенное на количество товаров, рекомендованных этому пользователю. Другими показателями, обсуждаемыми в этой работе, были описано Таммом и др. [23].

В сфере объявлений, в первую очередь в сфере вакансий, обычно требуется только один пользователь для завершить транзакцию относительно данного объявления. В этом случае разнообразие рекомендаций даже более критичен, чем в других областях (таких как музыка, фильмы или другая электронная коммерция, кроме объявлений). Для оценки этого аспекта мы используем три показателя: тестовое покрытие, Шеннон энтропия [56] и индекс Джини [56]. Покрытие тестами относится к доле тестовых заданий. рекомендуется хотя бы одному пользователю. Используя энтропию Шеннона и индекс Джини, мы игнорируем элементы вне тестового набора. Больший охват тестами и значения энтропии Шеннона указывают на более высокий уровень разнообразия, тогда как более высокие значения индекса Джини указывают на меньшее разнообразие.

Machine Translated by Google

Заключительный этап лабораторной оценки фокусируется на эффективности. Для этой цели мы использовали экземпляр AWS SageMaker ml.m5.4xlarge, оснащенный 64 ГБ ОЗУ и 16 виртуальными ЦП, с процессором Intel(R) Xeon(R) Platinum 8259CL @ 2,50 ГГц, работающим на Amazon Linux AMI 2018.03. Все методы оценивались по трем функциям: предварительная обработка данных, Примерка модели и рекомендации по товару. Время выполнения рассчитывалось для общее время, необходимое для расчета модели (в часах). Пик памяти был рассчитан в ГБ, а максимальное использование памяти учитывалось при выполнении данной функции. Эта оценка была выполнена с использованием библиотеки трассировки (https://docs.python.org/3/). библиотека/tracemalloc.html, доступ 29 мая 2024 г.).

5. Лабораторная оценка

В этом разделе описывается лабораторная оценка методов, обсуждаемых в разработанном наборе данных. Мы использовали наиболее эффективные гиперпараметры, найденные в предыдущем разделе. Каждая модель была обучена на всем наборе обучающих и проверочных данных, а рекомендации были подготовлены для всех 619 389 пользователей из тестового набора. Чтобы лучше представить ценности каждую офлайн-метрику мы сравнивали описанные методы с простыми, неперсонализированными базовые показатели: наиболее популярный продуктовый подход и случайные рекомендации.

5.1. Точность

Все метрики в этом разделе учитывают первые десять рекомендаций, например точность@10, отзыв@10. Более высокие значения показателя указывают на более высокое качество рекомендаций. Ценности не следует напрямую сравнивать с результатами, полученными на других наборах данных, поскольку метрики сильно зависят от распределения набора данных (например, высокая разреженность) и стратегия разделения обучения/теста [55].

5.1.1. Оценка точности

В Таблице 7 представлены результаты оценки. Большинство наблюдаемых различий статистически значимы из-за большого количества оцениваемых пользователей. Мы подробно рассказываем о статистическая значимость результатов приведена в разделе 5.1.2.

Таблица 7. Точность: результаты оценки. Все представленные показатели были описаны Tamm et al. [23]. Для Для метрики mAP мы использовали вариант c = min(,).

Метрический	RP3Beta	SLIM	ALS	Prod2Vec	LightFM	Самый популярный	Случайный
точность	0,0484	0,0472	0,0434	0,0368	0,0736	0,0359	0,0012	0,00006
отзыв	0,0783	0,0657	0,0580	0,0721	0,0657	0,0564	0,0012	0,00005
ndcg	0,0759	0,0567	0,0365	0,0329	0,0282	0,0545	0,0016	0,00007
mAP	0,0393	0,1314	0,1230	0,1065	0,5368	0,0264	0,0006	0,00002
MRR	0,1365	0,5328	0,5 289	0,3066	0,2878	0,1034	0,0038	0,00019
LAUC	0,5391	0,2537				0,5281	0,5006	0,49999
HR	0,3131					0,2547	0,0112	0,00059

Источник: собственное исследование.

Наши подходы значительно превосходят случайные и наиболее популярные рекомендации по товарам . Низкая эффективность самого популярного подхода обусловлена природой объявления, в первую очередь о сфере деятельности, где пользователи больше всего интересуются конкретной местоположение и категория. Создание более сложных, персонализированных систем рекомендаций. разумно в таком случае.

Показано, что RP3Beta превосходит другие подходы по точности. метрики. Мы полагаем, что это связано с высокой разреженностью нашего набора данных. RP3Beta рассчитывает рекомендации детерминированно (путем использования путей длиной три в пользовательском элементе). двудольный граф). Все остальные подходы используют методы машинного обучения для поиска пользователя. представления (ALS, LightFM), представления элементов (ALS, LightFM, Prod2Vec) или прямые Сходство между элементами (SLIM), которое может иметь место в тех случаях, когда пользователи или элементы имеют низкий уровень количество взаимодействий.

Machine Translated by Google

5.1.2. Статистическое сравнение для Precision@10

Мы представляем подробное статистическое сравнение нашего основного показателя точности @ 10, чтобы выявить различия между методами. Для начала мы проверяем нулевую гипотезу о том, что все методы работают одинаково, а наблюдаемые различия являются просто случайными (омнибусный тест). Тест Фридмана [57,58] с расширением Имана и Давенпорта [59] является, вероятно, самым популярным комплексным тестом, и обычно он является хорошим выбором при сравнении более чем пяти различных алгоритмов [60,61]. Пусть – ранг th методов на th выборках и

$$"=" \frac{1}{} \quad .$$
$$=1$$

В нашем случае количество выборок — это количество пользователей в тестовом наборе = 619 389. Тест сравнивает средние ранги методов и основан на следующей статистике:

$$"=" \frac{(- 1) (- \quad ^2}{1) - \quad ^2},$$

где

$$2 \quad "=" \frac{12}{(+ 1)} \quad \quad ^2 \quad 3 (+ 1)$$
$$=1$$

– это статистика Фридмана, имеющая распределение с – 1 и (– 1) (– 1) степенями свободы. Значение этого теста равно 0. Полученное значение указывает на то, что мы можем смело отвергнуть нулевую гипотезу о том, что все алгоритмы работают одинаково. Следовательно, мы можем приступить к апостериорным тестам, чтобы обнаружить значительные попарные различия между всеми методами. Демшар [62] предлагает использовать тест Неменьи [63], который сравнивает все алгоритмы попарно. Для уровня значимости тест определяет критическую разницу (CD). Если разница между средними рейтингами двух алгоритмов превышает

$$\text{CD} = \quad \frac{\quad}{\quad}{(+ 1) 6},$$

нулевая гипотеза о том, что алгоритмы имеют одинаковую производительность, основан отклоняется (по статистике стьюдентизированного диапазона, деленной на 2). Демшар [62] предложил график проверки различий (график CD визуально). Алгоритмы, не соединенные линией, можно считать разными по сюжету.

В нашем случае при значимости = 0,05 любые два алгоритма с разницей в средний ранг выше 0,0114 будет считаться неравным (рис. 5).

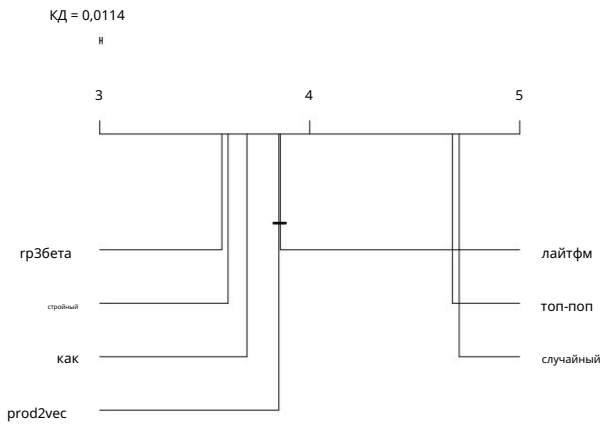


Рисунок 5. График критической разницы (для точности). Источник: собственное исследование.

Machine Translated by Google

5.1.3. Точность для пользователей с разным количеством взаимодействий

Единственная информация, которой мы располагаем о наших пользователях, — это их взаимодействие. На рисунке 6 мы исследуем, как количество элементов, с которыми взаимодействовал пользователь, влияет на нашу основную метрику — точность@10. Мы делим наших пользователей на десять групп одинакового размера в зависимости от количества взаимодействий.

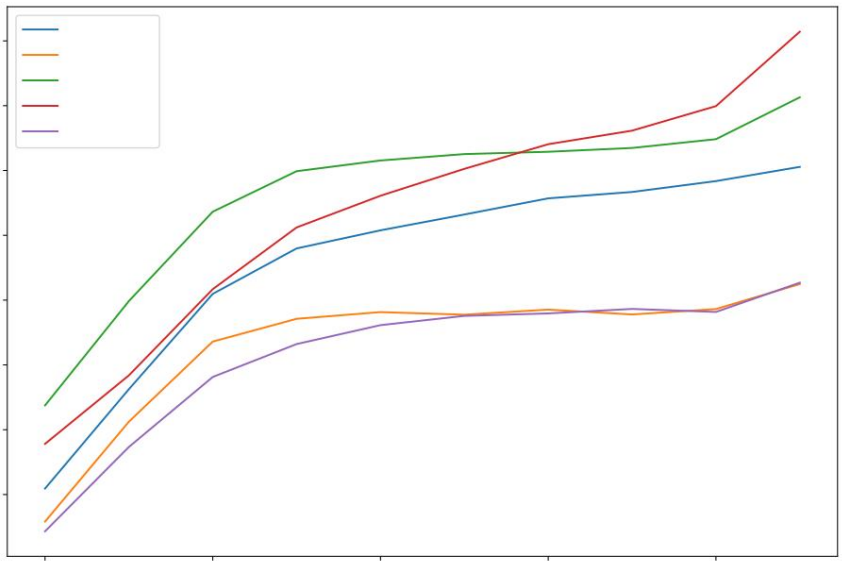


Рисунок 6. Точность@10 для каждой модели в зависимости от количества элементов, с которыми взаимодействовал пользователь. Лучше всего рассматривать его в цвете. Источник: собственное исследование.

Мы видим, что порядок моделей, отсортированных по точности @ 10, не зависит от группы, за исключением того, что SLIM работал лучше, чем RP3Beta для пользователей, совершивших не менее 22 взаимодействий. Разница статистически значима для каждой группы пользователей (обсуждается в разделе 5.1.4). Модель RP3Beta можно рассматривать как частный случай подхода совместной фильтрации на основе элементов, который использует детерминированную меру сходства между элементами [64]. С другой стороны, SLIM использует методы машинного обучения для выявления сходств.
Мы предполагаем, что такие оценки выученного сходства менее предвзяты, но имеют более высокую дисперсию, что приводит к повышению производительности для пользователей с большим количеством взаимодействий, поскольку усреднение выполняется по более значительному числу оценок.

5.1.4. Статистическая оценка точности для пользователей с разным количеством взаимодействий. В этом подразделе мы обсуждаем разницу между моделями RP3Beta и SLIM.
в зависимости от количества взаимодействий пользователя, как показано на рисунке 6.

Для статистического сравнения двух методов на нескольких наборах данных Демшар [62] рекомендует знаково-ранговый критерий Уилкоксона [65]. Знаковый ранговый критерий Уилкоксона — это непараметрическая альтернатива парному тесту, который ранжирует различия в производительности двух алгоритмов для каждого набора данных, игнорируя знаки. Он сравнивает ранги положительных и отрицательных различий.

В таблице 8 показано, что разница между моделями RP3Beta и SLIM статистически значимы для каждой группы пользователей.

Machine Translated by Google

Таблица 8. Результаты знаково-рангового теста Уилкоксона между моделями RP3Beta и SLIM в зависимости от количество элементов, с которыми взаимодействовал пользователь.

Бин	p-значение
[1,0, 3,0)	0
[3,0, 5,0)	0
[5,0, 8,0)	0
[8,0, 11,0)	0
[11,0, 16,0)	0
[16,0, 22,0)	1,25 × 10 6
[22,0, 31,0)	9,27 × 10 7
[31,0, 45,0)	0
[45,0, 74,0)	0
[74,0, 852,0)	0

Источник: собственное исследование.

5.2. Разнообразие

В Таблице 9 мы видим, что Prod2Vec и LightFM обеспечивают наибольшее разнообразие. после исключения случайных рекомендаций (которые должны иметь наибольшее разнообразие). Разнообразие наиболее точных моделей RP3Beta, SLIM и ALS схоже, за исключением случай более низкого тестового покрытия БАС. Связь между тестовым покрытием и точностью @ 10 показано на рисунке 7.

Таблица 9. Разнообразие: результаты оценки.

Метрика	RP3Beta	SLIM	ALS	Prod2Vec	LightFM	Самые популярные	Случайные
тестовое покрытие	0,5725	0,5171	0,3038	0,7400	9,6728	0,7031	0,9778
Шеннон	9,5271	9,6270	10,4031	0,9029	0,9120	0,0002 10,1385	11,7267
Джини	0,9083	2,3296	0,8397	0,9999	0,7956		0,1159

Источник: собственное исследование.

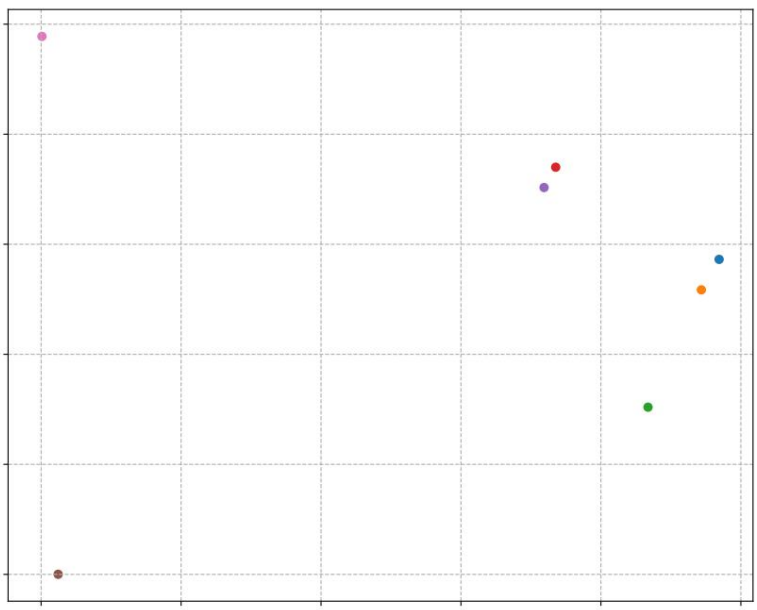


Рисунок 7. Тестовое покрытие в сравнении с точностью@10 для рассматриваемых моделей. Источник: собственное исследование.

Machine Translated by Google

5.3. Перекрытие методов

В наших исследованиях мы фокусируемся на методах, которые стоит протестировать на реальных пользователях. Поэтому, мы изучили совпадение рекомендаций, предлагаемых реализованными методами относительно пар пользователь-элемент (коэффициент перекрытия [66]). Мы ожидаем более значительных различия в рекомендациях при сравнении методов с низким перекрытием.

Результаты этого эксперимента представлены в Таблице 10. RP3Beta и SLIM предлагают аналогичные рекомендации. Хотя они рассчитываются по-разному, они используют разреженный элемент матрица сходства. Группа сравнения с БАС имеет среднее сходство с другими методы. Наиболее разными с точки зрения перекрытия являются также самые разнообразные модели, а именно Prod2Vec и LightFM. Низкое перекрытие ALS и LightFM подчеркивает важность выбора функции потерь в подходах матричной факторизации.

Таблица 10. Перекрытие моделей.

Модель	РП3Бета	СТРОЙНЫЙ	БАС	Прод2Век	ЛайтФМ
RP3Beta	100%	73%	53%	37%	38%
SLIM	73%	100%	50%	35%	35%
ALS	53%	50%	100%	38%	37%
Prod2Vec	37%	35%	38%	100%	28%
LightFM	38%	35%	37%	28%	100%

Источник: собственное исследование.

5.4. Оценка эффективности

Результаты оценки эффективности представлены на рисунке 8. RP3Beta, ALS и LightFM превосходит SLIM и Prod2Vec по времени выполнения, тогда как Prod2Vec требует меньше всего памяти. Чтобы часто переобучать модель, мы установили максимально приемлемое время выполнения составило 3 часа, что соответствует ALS, LightFM и RP3Beta. От стоимости В перспективе допустимое использование памяти было установлено на уровне 110 ГБ, что соответствует всем моделям. выполнено. Мы можем заметить, что наши модели можно обучать даже на одном ноутбуке.

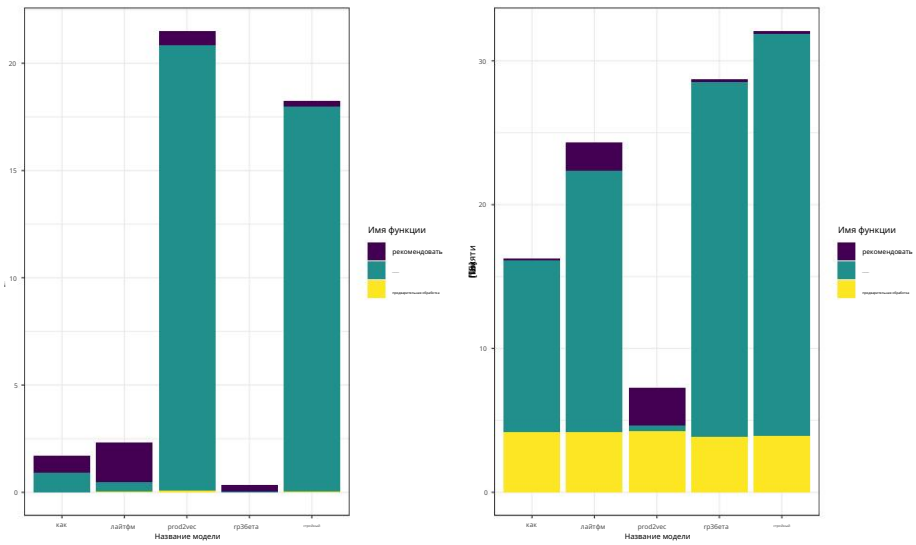


Рисунок 8. Время выполнения и использование памяти. Лучше всего рассматривать его в цвете. Источник: собственное исследование.

5.5. Краткое содержание

Лабораторные исследования позволили нам выбрать модели для онлайн-тестирования с помощью OLX. пользователи. Во-первых, мы выбрали модель RP3Beta для онлайн-тестирования, потому что это лучший подход. относительно точности и времени выполнения. Относительно высокое потребление памяти по-прежнему достаточно низкий для тестирования модели в производстве. Исследование влияния этого относительно небольшое разнообразие метода может стать интересным направлением будущих исследований.

Machine Translated by Google

SLIM — вторая лучшая модель с точки зрения точности @ 10. Тем не менее, мы не выбрали этот метод для онлайн-оценки из-за его низкой эффективности и высокого совпадения с уже выбранной моделью RP3Beta. Повышение эффективности реализации SLIM может изменить это решение.

Среди других моделей наибольшая точность@10 достигается у ALS. Кроме того, ALS уже был внедрен и использовался для рекомендаций по вакансиям в OLX; поэтому мы решили включить его в онлайн-сравнение.

Несмотря на то, что Prod2Vec и LightFM выдают самые разнообразные рекомендации, мы решили не тестировать их онлайн из-за значительно худшей точности@10 и эффективности у Prod2Vec.

6. Онлайн-оценка

Чтобы оценить эффективность выбранных методов, мы провели два онлайн-A/B -теста с пользователями OLX Jobs. Мы сосредоточились на пользователях, которые недавно просмотрели объявления о работе на нашей платформе. Мы отправили им электронные письма с десятью рекомендованными объявлениями о работе, а когда пользователь установил приложение OLX, ему также было отправлено push-уведомление с одним рекомендуемым объявлением о работе. Затем мы наблюдали, подавал ли пользователь заявку на какую-либо работу в течение следующих 48 часов (конвертированный пользователь).

6.1. Влияние модели БАС

Первый эксперимент был направлен на ответ на вопрос: увеличивает ли внедрение рекомендательного метода количество пользователей, претендующих на работу? В этом тесте мы разделили пользователей на две группы: контрольную группу без рекомендаций и группу с рекомендациями, сформированными с помощью модели ALS. Лишь 10% пользователей попали в контрольную группу, поскольку модель ALS уже была внедрена и использовалась для рекомендаций по работе в OLX еще до теста. Мы предпочли собрать минимальную выборку, проведя более расширенное тестирование, чтобы наблюдать долгосрочное влияние на наших пользователей. Тест проводился в течение 25 дней в марте 2021 г. Результаты представлены в таблице 11. Статистическую значимость преимущества БАС над контрольной группой исследовали с помощью критерия хи-квадрат, значение - оказалось равным 0. Этот эксперимент подтвердил нашу гипотезу о том, что рекомендации влияют на активность: процент конвертированных пользователей после получения сообщения значительно увеличился на (16,83% - 15,98%)/15,98% 5,3%.

Наблюдаемый рост означает, что сотни дополнительных пользователей ежедневно подают заявки на работу через OLX Jobs.

Таблица 11. Сравнение активности с рекомендациями и без них.

Вариант	Пользователи	% совершивших конверсию пользователей
контроль	129 308 1	15,98%
БАС	170 262	16,83%

Источник: собственное исследование.

6.2. Сравнение моделей ALS и RP3Beta

Во втором эксперименте мы проверяли, превзошла ли лучшая модель лабораторных испытаний БАС. Мы разделили наших пользователей на три группы, получая рекомендации от разных рекомендательных систем: ALS, RP3Beta и смешанного варианта. Мы создали половину рекомендаций с помощью ALS, а другую половину — с помощью RP3Beta. Второй эксперимент проводился в течение 28 дней в марте и апреле 2021 года. Результаты представлены в таблице 12.

Здесь рассчитывается процент конвертированных пользователей относительно всех пользователей; однако изменение системы рекомендаций не коснулось пользователей, не открывавших сообщение. Мы решили их отфильтровать, поскольку их поведение добавляло к нашим данным ненужный шум. В таблице 13 представлено влияние на пользователей, открывших сообщение. Преимущество RP3Beta над ALS статистически значимо (значение p 10–6). Разница между вариантами RP3Beta и ALS+RP3Beta не является статистически значимой (значение p составляет 0,19).

Machine Translated by Google

Таблица 12. Сравнение различных методов рекомендаций.

Вариант	Пользователи	% совершивших конверсию пользователей
БАС	343 892	15,25%
РП3Бета	345 273	15,40%
БАС+RP3Бета	343 896	15,30%

Источник: собственное исследование.

Таблица 13. Сравнение различных методов рекомендаций для пользователей, открывших сообщение.

Вариант	Пользователи	% совершивших конверсию пользователей
БАС	44 775	19,66%
РП3Бета	46 097	20,94%
БАС+RP3Бета	45 469	20,59%

Источник: собственное исследование.

Из пользователей, открывших сообщение, мы наблюдали 1,28 п.п. (процентных пунктов) больше конвертированных пользователей из RP3Beta, чем из ALS. Поскольку уровень открытия сообщений составил 13,2%, это эквивалентно увеличению на 1,28 × 0,132 0,17 п.п. всех целевых пользователей. На основе По результатам первого эксперимента, БАС дает увеличение на 0,85 п.п. по сравнению с контрольной группой. Это означает что замена ALS на RP3Beta увеличивает наше влияние на всех целевых пользователей примерно 0,17/0,85 = 20%.

6.3. Обсуждение

Мы доказали, что отправка рекомендаций о вакансиях из модели ALS увеличивает количество пользователей, откликнувшихся на объявления о работе, увеличилось более чем на 5%. Более того, RP3Beta значительно превосходит БАС.

Кроме того, мы провели еще один онлайн-тест A/B, сравнив модель RP3Beta с вариант, в котором мы заменили половину рекомендаций RP3Beta на рекомендации созданный на основе гибридной рекомендательной системы глубокого обучения, используемой в OLX (https://tech.olx.com/item2vec-neural-item-embeddings-to-enhance-recommendations-1fd948a6f293, по состоянию на 29 мая 2024 г.). Мы решили не описывать этот тест в этой статье, так как данная работа основное внимание уделяется подходам совместной фильтрации. Мы также хотели обеспечить воспроизводимость наших результатов в опубликованном наборе данных. К сожалению, мы не смогли опубликовать все функции. используется нашей внутренней моделью. Мы не заметили никаких улучшений по сравнению с RP3Beta, хотя вторая модель использует дополнительные знания о объявлениях о работе. Следовательно, мы подтвердил, что правильно определенная и настроенная простая модель может превзойти более передовые глубокие решения. Это согласуется с результатами Dacrema et al. [64] и Анелли и др. [30], которые сравнили несколько ненейронных подходов, включая RP3Beta, с более продвинутые нейронные модели.

Эффективность модели RP3Beta позволяет OLX Jobs обучать ее с нуля раз в день с очень низкими затратами (на процессор). Современная совместная фильтрация Модель в OLX Jobs в настоящее время используется для генерации электронных писем и отправки рекомендаций для наши пользователи. Недавно мы адаптировали эту модель для предоставления рекомендаций на месте в режиме реального времени. учитывая недавнее взаимодействие пользователя с элементами [67]. Мы также предложили обучение P3, чтобы ранговая модель (P3LTR) [68], обобщение RP3Beta, где мы узнаем о важности отношений пользователь-элемент на основе типов прошлых взаимодействий пользователя и элемента и временных меток.

В этой работе мы восполнили пробел в исследованиях, связанный с отсутствием сравнений классические подходы совместной фильтрации в сфере работы, предоставляя комплексные лабораторные исследования на реальном наборе данных и онлайн-оценка с участием миллионов пользователей. Кроме того, мы сделали общедоступным исходный код для автономной оценки и набор данных. доступный. Мы призываем других исследователей разработать новые методы рекомендаций для помочь пользователям быстрее найти идеальную работу.

Методы, рассматриваемые в этой работе, требуют только триплетов (пользователь, элемент и временная метка). в качестве входных данных для обучения и оценки; следовательно, их можно применять в нескольких областях,

Machine Translated by Google

даже для наборов данных, состоящих из десятков миллионов взаимодействий. Мы ожидаем хороших результатов в области объявлений, основываясь на наших первоначальных результатах, полученных в других категориях OLX.

7. Резюме

В этой статье рассматривается тема рекомендаций о вакансиях в онлайн-объявлениях на примере OLX Jobs. Мы провели обширную оценку подходов, представляющих различные семейства рекомендательных методов.

В ходе лабораторных исследований мы продемонстрировали, что RP3Beta, SLIM и ALS работают значительно лучше, чем LightFM и Prod2Vec. Мы также обнаружили, что Prod2Vec предлагает самое необычайное разнообразие, которое может понадобиться при работе над новизной рекомендаций. Более того, мы проверили, насколько схожи рекомендации, генерируемые разными моделями.

Мы также провели оценку, отправив миллионы сообщений онлайн-пользователям. A/B-тесты проводились с использованием ALS и RP3Beta. Результаты этих тестов показали , что отправка рекомендаций о вакансиях, сгенерированных этими моделями, увеличивает количество пользователей, связывающихся с рекламодателями из OLX Jobs. Благодаря доступности RP3Beta в производственной среде мы повысили эффективность на 20 % (с учетом реакции пользователей).

При оценке методов с точки зрения объявлений модель, рекомендовавшая наиболее популярные (еще не встречавшиеся) предметы, оказалась непригодной для рекомендаций по вакансиям. Также для сравнения была добавлена рекомендательная модель, основанная на случайных рекомендациях, чтобы продемонстрировать преимущество исследуемых методов.

Еще одним важным результатом этого исследования является набор данных, который позволит проводить дальнейшие исследования по рекомендациям для классифицированных потребностей.

Одним из недостатков нашей работы является отсутствие сравнения с глубокими нейронными сетями (DNN) или нейронными сетями на графах (GNN). Для таких методов сложно удовлетворить наши временные ограничения без более дорогой инфраструктуры (например, с графическим процессором). Ослабление наших ограничений или сравнение наиболее эффективных DNN и GNN было бы интересно для будущей работы. Еще одним недостатком является то, что сравнение проводится только на одном наборе данных. Размер набора данных оказывает существенное влияние на выбор метода. Было бы интересно протестировать одни и те же методы на наборах данных разного размера, либо на подмножествах исходного набора данных, либо на других наборах данных из предметной области (например, с разных рынков, на которых работает OLX). Кроме того, было бы интересно изучить корреляцию между разнообразием и онлайн-результатами, чтобы разработать офлайн-метрику, которая лучше соответствовала бы бизнес-целям.

Вклад авторов: Концептуализация, РК; методология, РК и ТГ; программное обеспечение, РК и ВД; валидация, РК, ВД, ТГ и АФ; формальный анализ, РК и ТГ; расследование, РК и ТГ; ресурсы, РК; курирование данных, РК и ВД; письменное – подготовка оригинального проекта, РК, АФ и ТГ; написание – рецензирование и редактирование, РК, А.Ф. и ТГ; визуализация, РК и ТГ; наблюдение, ТГ и ФП; администрация проекта, РК Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: Набор данных: https://www.kaggle.com/datasets/olxdatascience/olx-jobs-interactions , доступ 29 мая 2024 г.; исходный код: https://github.com/rob-kwiec/olx-jobs-рекомендации , доступ 29 мая 2024 г.

Конфликты интересов: авторы Роберт Квецинский, Агата Филиповска и Вячеслав Дубров работали в компании OLX Group. Оставшийся автор заявляет, что исследование проводилось при отсутствии каких-либо коммерческих или финансовых отношений, которые могли бы быть истолкованы как потенциальный конфликт интересов.

Рекомендации

1. Дэвис, М. Что такое онлайн-объявления? Журнал Big Easy, 4 января 2021 г. 2.
Джонс, С. Интернет-объявления; Проект Pew Internet & American Life: Вашингтон, округ Колумбия, США, 2009 г.
3. Гурвиль, Дж.; Соман, Д. Чрезмерный выбор и тип ассортимента: когда и почему разнообразие имеет неприятные последствия. Отметка. наук. 2005, 24, 382–395.
[Перекрестная ссылка]

Machine Translated by Google

4. Чернев А. Когда больше значит меньше, а меньше значит больше: роль наличия и ассортимента идеальных точек в потребительском выборе. Дж. Потребление. Рез. 2003, 30, 170–183. [Перекрестная ссылка]

5. Боутрайт, П.; Нуньес, Дж. Сокращение ассортимента: подход, основанный на атрибутах. Дж. Марк. 2001, 65, 50–63. [Перекрестная ссылка]

6. Гомес-Урибе, К.; Хант, Н. Рекомендательная система Netflix. АКМ Транс. Менеджер. Инф. Сист. 2015, 6, 1–19. [Перекрестная ссылка]

7. Смит, Б.; Линден, Г. Два десятилетия рекомендательных систем на Amazon.com. IEEE Интернет-вычисления. 2017, 21, 12–18. [Перекрестная ссылка]

8. Аггарвал, СС Рекомендательные системы: Учебник, 1-е изд.; Springer: Берлин/Гейдельберг, Германия, 2016. 9. Ли, С.; Караханна, Э. Системы онлайн-рекомендаций в контексте электронной коммерции B2C: обзор и будущие направления. Дж. доц. Инф. Сист. 2015, 16, 72–107. [Перекрестная ссылка]

10. Твардовский Б. Моделирование контекстной информации в рекомендательных системах с учетом сеанса с помощью нейронных сетей. В материалах 10-й конференции ACM по рекомендательным системам, Бостон, Массачусетс, США, 15–19 сентября 2016 г.; стр. 273–276. [Перекрестная ссылка]

11. Алотаиби С. Обзор систем рекомендации вакансий. Межд. Дж. Физ. наук. 2012, 7, 5127–5142. [Перекрестная ссылка]

12. Абель, Ф.; Бенчур, А.; Кольсдорф, Д.; Ларсон, М.; Паловичс, Р. RecSys Challenge'16: Материалы конкурса рекомендательных систем; Ассоциация вычислительной техники: Нью-Йорк, штат Нью-Йорк, США, 2016 г.

13. Абель, Ф.; Делджу, Ю.; Элахи, М.; Колсдорф, Д. RecSys Challenge'17: Материалы конкурса рекомендательных систем 2017; Ассоциация вычислительной техники: Нью-Йорк, штат Нью-Йорк, США, 2017 г.

14. Ли, С.; Ши, Б.; Ян, Дж.; Ян, Дж.; Ван, С.; Чен, Ф.; Он, В. Глубокое понимание работы в LinkedIn. В материалах 43-й Международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, виртуально, 25–30 июля 2020 г.; стр. 2145–2148. [Перекрестная ссылка]

15. Чжу, Г.; Чен, Ю.; Ван, С. Персонализированные рекомендации для курсовой работы с поддержкой графического сообщества и междоменные данные Интеграция. Устойчивое развитие 2022, 14, 7439. [CrossRef]

16. Лачич Э.; Райтер-Хаас, М.; Ковальд, Д.; Редди Даредди, М.; Чо, Дж.; Лекс, Э. Использование автокодировщиков для рекомендаций заданий на основе сеанса. свидания. Модель пользователя. Адаптация пользователя. Взаимодействуйте. 2020, 30, 617–658. [Перекрестная ссылка]

17. Паццани, МЮ; Биллсус, Д. Системы рекомендаций на основе контента. В книге «Адаптивный Интернет: методы и стратегии Интернета». Персонализация; Springer: Берлин/Гейдельберг, Германия, 2007 г.; стр. 325–341. [Перекрестная ссылка]

18. Чен, Р.; Хуа, Ц.; Чанг, Ю.С.; Ван, Б.; Чжан, Л.; Конг, Х. Обзор рекомендательных систем на основе совместной фильтрации: от традиционных методов к гибридным методам на основе социальных сетей. Доступ IEEE 2018, 6, 64301–64320. [Перекрестная ссылка]

19. Ван, Х.; Он, Х.; Ван, М.; Фэн, Ф.; Чуа, Т. С. Совместная фильтрация нейронных графов. В материалах 42-й Международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, Париж, Франция, 21–25 июля 2019 г.; стр. 165–174. [Перекрестная ссылка]

20. Шафер, Дж.Б.; Франковски, Д.; Херлокер, Дж.; Сен, С. Рекомендательные системы совместной фильтрации. В Адаптивной сети; Springer: Берлин/Гейдельберг, Германия, 2007 г.; стр. 291–324. [Перекрестная ссылка]

21. Такач, Г.; Пиласи, И.; Немет, Б.; Тикк, Д. Масштабируемые подходы к совместной фильтрации для больших рекомендательных систем. Дж. Мах. Учиться. Рез. 2009, 10, 623–656. [Перекрестная ссылка]

22. Джавахир, Г.; Сомсзор, М.; Косткова П. Сравнение неявной и явной обратной связи от онлайн- сервиса музыкальных рекомендаций. В материалах 1-го международного семинара по неоднородности и слиянию информации в рекомендательных системах, Барселона, Испания, 26 сентября 2010 г.; стр. 47–51. [Перекрестная ссылка]

23. Тамм Ю.М.; Дамдинов Р.; Васильев А. Метрики качества в рекомендательных системах: единообразно ли мы рассчитываем метрики? В материалах пятнадцатой конференции ACM по рекомендательным системам, Амстердам, Нидерланды, 27 сентября – 1 октября 2021 г.; стр. 708–713. [Перекрестная ссылка]

24. Каминскас, М.; Бридж, Д. Разнообразие, случайность, новизна и охват: обзор и эмпирический анализ целей сверхточности в рекомендательных системах. АКМ Транс. Взаимодействуйте. Интел. Сист. 2016, 7, 1–42. [Перекрестная ссылка]

25. Кунавер, М.; Позрл, Т. Разнообразие в рекомендательных системах, Обзор. Информационная система. 2017, 123, 154–162. [Перекрестная ссылка]

26. Ге, М.; Дельгадо-Баттенфельд, К.; Яннах, Д. За пределами точности: оценка рекомендательных систем по охвату и случайности. В материалах четвертой конференции ACM по рекомендательным системам, Барселона, Испания, 26–30 сентября 2010 г.; стр. 257–260. [Перекрестная ссылка]

27. Могенет А.; Фам, ТАН; Казама, М.; Конг, Дж. Прогнозирование онлайн-производительности систем рекомендации вакансий с помощью автономной оценки. В материалах 13-й конференции ACM по рекомендательным системам, Копенгаген, Дания, 16–20 сентября 2019 г.; стр. 477–480. [Перекрестная ссылка]

28. Анелли, Фольксваген; Беллогин, А.; Ди Ноя, Т.; Яннах, Д.; Помо, К. Алгоритмы рекомендаций Top-N: поиск новейших достижений. В материалах 30-й конференции ACM по моделированию, адаптации и персонализации пользователей, Барселона, Испания, 4–7 июля 2022 г.; стр. 121–131. [Перекрестная ссылка]

29. Дакрема, МФ; Больо, С.; Кремонези, П.; Яннак, Д. Тревожный анализ воспроизводимости и прогресса в рекомендациях Системные исследования. АКМ Транс. Инф. Сист. 2021, 39, 1–49. [Перекрестная ссылка]

30. Анелли, Фольксваген; Беллогин, А.; Ди Ноя, Т.; Помо, К. Новый взгляд на сравнение нейронной совместной фильтрации и матричной факторизации. В материалах пятнадцатой конференции ACM по рекомендательным системам, Амстердам, Нидерланды, 27 сентября – 1 октября 2021 г.; стр. 521–529. [Перекрестная ссылка]

Machine Translated by Google

31. Чжу Дж.; Дай, К.; Су, Л.; Ма, Р.; Лю, Дж.; Кай, Г.; Сяо, Х.; Чжан, Р. БАРС: На пути к открытому сравнительному анализу рекомендательных систем. В материалах 45-й Международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, Мадрид, Испания, 11–15 июля 2022 г.; стр. 2912–2923. [Перекрестная ссылка]
32. Донг, Ю.; Ли, Дж.; Шнабель, Т. Когда новее не значит лучше: действительно ли глубокое обучение приносит пользу рекомендациям неявной обратной связи? В материалах 46-й Международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска, Тайбэй, Тайвань, 23–27 июля 2023 г.; стр. 942–952. [Перекрестная ссылка]
33. Анелли, Фольксваген; Малитеста, Д.; Помо, К.; Беллогин А.; Ди Шашио, Э.; Ди Нойа, Т. Оспаривание мифа о совместной фильтрации графов: обоснованный анализ, основанный на воспроизводимости. В материалах 17-й конференции ACM по рекомендательным системам, Сингапур, 18–22 сентября 2023 г.; стр. 350–361. [Перекрестная ссылка]
34. Корень Ю.; Белл, Р.; Волинский, К. Методы матричной факторизации для рекомендательных систем. Компьютер 2009, 42, 30–37. [Перекрестная ссылка]
35. Кула, М. Встраивание метаданных для рекомендаций по холодному запуску пользователей и элементов. arXiv 2015, arXiv:1507.08439.
36. Рендл, С.; Фройденталер, К.; Гантнер, З.; Шмидт-Тиме, Л. БПР: Байесовский персонализированный рейтинг на основе неявной обратной связи. arXiv 2012, arXiv:1205.2618
37. Уэстон, Дж.; Бенджио, С.; Усунье, Н. Всаби: Масштабирование аннотаций изображений до большого словарного запаса. В материалах двадцать второй Международной совместной конференции по искусственному интеллекту, Барселона, Испания, 16–22 июля 2011 г.; стр. 2764–2770.
38. Уэстон, Дж.; Да, Х.; Вайс, Р.Дж. Учимся ранжировать рекомендации с потерей статистики K-порядка. В материалах 7-й конференции ACM по рекомендательным системам, Гонконг, Китай, 12–16 октября 2013 г.; стр. 245–248. [Перекрестная ссылка]
39. Ху, Ю.; Корен, Ю.; Волинский, К. Совместная фильтрация наборов данных неявной обратной связи. В материалах восьмой конференции IEEE 2008 г. Международная конференция по интеллектуальному анализу данных, Пиза, Италия, 15–19 декабря 2008 г.; стр. 263–272. [Перекрестная ссылка]
40. Такач, Г.; Пиласи, И.; Тикк, Д. Применение метода сопряженных градиентов для совместной фильтрации с неявной обратной связью. В материалах пятой конференции ACM по рекомендательным системам, Чикаго, Иллинойс, США, 23–27 октября 2011 г.; стр. 297–300. [Перекрестная ссылка]
41. Нин, Х.; Карипис, Г. СЛИМ: Разреженные линейные методы для рекомендательных систем Top-N. В материалах 11-й Международной конференции IEEE по интеллектуальному анализу данных 2011 г., Ванкувер, Британская Колумбия, Канада, 11–14 декабря 2011 г.; стр. 497–506. [Перекрестная ссылка]
42. Кристоффель Ф.; Паудель, Б.; Ньюэлл, К.; Бернштейн, А. Блокбастеры и желтофиоль: точные, разнообразные и масштабируемые рекомендации с помощью случайных блужданий. В материалах 9-й конференции ACM по рекомендательным системам, Вена, Австрия, 16–20 сентября 2015 г.; стр. 163–170. [Перекрестная ссылка]
43. Купер, К.; Ли, Ш.; Радзик Т.; Сиантос, Ю. Случайные блуждания в рекомендательных системах: точные вычисления и моделирование. В материалах 23-й Международной конференции по Всемирной паутине, Нью-Йорк, штат Нью-Йорк, США, 7–11 апреля 2014 г.; стр. 811–816. [Перекрестная ссылка]
44. Грбович, М.; Радосавлевич, В.; Джурич, Н.; Бхамидипати, Н.; Савла, Дж.; Бхагван, В.; Шарп, Д. Электронная коммерция в вашем почтовом ящике: рекомендации по продуктам в масштабе. В материалах 21-й Международной конференции ACM SIGKDD по обнаружению знаний и интеллектуальному анализу данных, Сидней, Новый Южный Уэльс, Австралия, 10–13 августа 2015 г.; стр. 1809–1818. [Перекрестная ссылка]
45. Баркан, О.; Кенигштейн, Н. ITEM2VEC: Встраивание нейронных элементов для совместной фильтрации. В материалах 26-го Международного семинара IEEE по машинному обучению для обработки сигналов (MLSP) 2016 г., Виетри-суль-Маре, Италия, 13–16 сентября 2016 г.; стр. 1–6. [Перекрестная ссылка]
46. Миколов Т.; Суцкевер И.; Чен, К.; Коррадо, Г.; Дин, Дж. Распределенные представления слов и фраз и их композиционность. В материалах 26-й Международной конференции по нейронным системам обработки информации — Том 2, Ред-Хук, Нью-Йорк, США, 5–10 декабря 2013 г.; стр. 3111–3119.
47. Де Рюйт, К.; Бхулаи, С. Системы рекомендации вакансий: обзор. arXiv 2021, arXiv:2111.13576.
48. Кула, М. LightFM — репозиторий Github. 2023 г. Доступно онлайн: https://github.com/lyst/lightfm. (по состоянию на 29 мая 2024 г.).
49. Фредериксон, Б. ALS — репозиторий Github. 2022 г. Доступно онлайн: https://github.com/benfred/implicit/blob/main/implicit/. als.py (по состоянию на 29 мая 2024 г.).
50. Леви, М.; Гризель, О. СЛИМ — репозиторий Github. 2014. Доступно онлайн: https://github.com/Mendeley/mrec/blob/master/. mrec/item_similarity/slim.py (по состоянию на 29 мая 2024 г.).
51. Квадрана, М. СЛИМ — репозиторий Github. 2021 г. Доступно онлайн: https://github.com/MaurizioFD/RecSys2019_DeepLearning_. Evaluation/blob/master/SLIM_ElasticNet/ SLIMElasticNetRecommender.py (по состоянию на 29 мая 2024 г.).
52. Бернардис, К. RP3Beta — репозиторий Github. 2021 г. Доступно онлайн: https://github.com/MaurizioFD/RecSys2019_DeepLearning_. Evaluation/blob/master/GraphBased/ RP3betaRecommender.py (по состоянию на 29 мая 2024 г.).
53. Паудель, Б.; Кристоффель, Ф.; Ньюэлл, К.; Бернштейн А. Обновляемые, точные, разнообразные и масштабируемые рекомендации для интерактивных технологий. Приложения. АКМ Транс. Взаимодействуйте. Интел. Сист. 2016, 7, 1–34. [Перекрестная ссылка]
54. Рехůрек, Р.; Сойка, П. Программная среда для тематического моделирования в крупных корпорациях. В материалах семинара LREC 2010 «Новые вызовы для структур НЛП», Валлетта, Мальта, 22 мая 2010 г.; стр. 45–50, ISBN 2-9517408-6-7. Доступно онлайн: http://is.muni.cz/publication/884893/en. (по состоянию на 29 мая 2012 г.).
55. Мэн, З.; Маккриди, Р.; Макдональд, К.; Оунис, И. Исследование стратегий разделения данных для оценки моделей рекомендаций. В материалах 14-й конференции ACM по рекомендательным системам, виртуально, 22–26 сентября 2020 г.; стр. 681–686. [Перекрестная ссылка]
56. Шани, Г.; Гунавардана А. Оценка рекомендательных систем. В Справочнике по рекомендательным системам; Спрингер: Бостон, Массачусетс, США, 2011 г.; стр. 257–297. [Перекрестная ссылка]

Machine Translated by Google

57. Фридман М. Использование рангов, чтобы избежать предположения о нормальности, подразумеваемого при дисперсионном анализе. Варенье. Стат. доц. 1937, 32, 675–701. [Перекрестная ссылка]

58. Фридман М. Сравнение альтернативных критериев значимости проблемы ранжирования. Анна. Математика. Стат. 1940, 11, 86–92. [Перекрестная ссылка]

59. Иман, Р.Л.; Давенпорт, Дж. М. Приближения критической области статистики Фридмана. Коммун. Статистические методы 1980, 9, 571–595. [Перекрестная ссылка]

60. Гарсия, С.; Эррера, Ф. Расширение «Статистического сравнения классификаторов по нескольким наборам данных» для всех парных сравнений. Дж. Мах. Учиться. Рез. 2008, 9, 2677–2694.

61. Гарсия, С.; Фернандес А.; Луенго, Дж.; Эррера, Ф. Расширенные непараметрические тесты для множественных сравнений при планировании экспериментов в области вычислительного интеллекта и интеллектуального анализа данных: экспериментальный анализ мощности. Инф. наук. 2010, 180, 2044–2064. [Перекрестная ссылка]

62. Демшар Дж. Статистическое сравнение классификаторов по нескольким наборам данных. Дж. Мах. Учиться. Рез. 2006, 7, 1–30.

63. Неменьи П. Множественные сравнения без распределения. Кандидат наук. Диссертация, Принстонский университет, Принстон, Нью-Джерси, США, 1963 г.

64. Дакрема, МФ; Кремонези, П.; Яннах, Д. Действительно ли мы добиваемся большого прогресса? Тревожный анализ последних подходов к нейронным рекомендациям. В материалах 13-й конференции ACM по рекомендательным системам, Копенгаген, Дания, 16–20 сентября 2019 г.; стр. 101–109. [Перекрестная ссылка]

65. Уилкоксон Ф. Индивидуальные сравнения методами ранжирования. Биом. Бык. 1945, 1, 80–83. [Перекрестная ссылка]

66. Виджаймина, М.; Кавита, К. Обзор мер сходства при анализе текста. Мах. Учиться. Прил. Межд. Дж. 2016, 3, 19–28. [Перекрестная ссылка]

67. Квечинский Р.; Мельничак, Г.; Гурецкий, Т. Сравнение рекомендаций для работы в режиме реального времени и пакетной обработки. IEEE Access 2023, 11, 20553–20559. [Перекрестная ссылка]

68. Квечинский Р.; Гурецкий, Т.; Филиповска, А. Изучение важности ребер в рекомендациях на основе двудольных графов. В материалах 17-й конференции по информатике и интеллектуальным системам (FedCSIS) 2022 г., София, Болгария, 4–7 сентября 2022 г.; стр. 227–233. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Блокчейн и гомоморфное шифрование для безопасности данных и

Статистическая конфиденциальность

Рахул Радж 1,*, Йешем Курт Пекер [1] и Зейнеп Делал Мутлу [2]

[1] Школа компьютерных наук TSYS, Государственный университет Колумбуса, Колумбус, Джорджия 31907, США;
[2] peker_yesem@columbusstate.edu Независимый исследователь, 1200 Брюссель, Бельгия; zeynepdelalmutlu@gmail.com * Переписка: raj_rahul@students.columbusstate.edu

Аннотация: В этом исследовании предлагается система на основе блокчейна, которая использует полностью гомоморфное шифрование для обеспечения безопасности данных и статистической конфиденциальности, когда данные передаются третьим лицам для анализа или исследовательских целей. Предлагаемая система не только обеспечивает безопасность данных при передаче, хранении и использовании, но также гарантирует конфиденциальность и вычислительную целостность для простых статистических вычислений. Это достигается за счет использования атрибутов технологии блокчейн, обеспечивающей доступность и целостность данных, в сочетании с гомоморфным шифрованием, обеспечивающим конфиденциальность используемых данных. Вычисления выполняются на смарт-контрактах, находящихся в блокчейне, что обеспечивает вычислительную целостность. Предлагаемая система реализована на блокчейне Zama и выполняет статистические операции, включая среднее значение, медиану и дисперсию зашифрованных данных. Результаты показывают , что на блокчейне можно выполнять полностью гомоморфные вычисления. Несмотря на то, что текущие ограничения вычислений на блокчейне не позволяют запускать систему для больших наборов данных, технология доступна, и с развитием более эффективных гомоморфных операций на блокчейнах предлагаемая система предоставит окончательное решение для обеспечения столь желаемой безопасности. свойства в приложениях, включая конфиденциальность данных и статистическую информацию, конфиденциальность и целостность при хранении, передаче и использовании.

Ключевые слова: конфиденциальность данных; целостность данных; вычислительная целостность; гомоморфное шифрование; блокчейн; статистическая конфиденциальность; безопасность данных; статистическая конфиденциальность

check for updates

Цитирование: Радж Р.; Курт Пекер, Ю.; Мутлу, ZD Blockchain и Гомоморфное шифрование для безопасности данных и статистической конфиденциальности. Электроника 2024, 13, 3050. https://doi.org/10.3390/electronics13153050 .

Академический редактор: Симеоне Марино

Поступила: 8 июня 2024 г.
Пересмотрено: 20 июля 2024 г.
Принято: 26 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Защита конфиденциальности пользователей и конфиденциальности конфиденциальных данных жизненно важна во всех системах, где пользовательские данные хранятся, передаются или обрабатываются. Часто данные необходимо анализировать в деловых целях или на благо общества. Например, данные переписи населения фиксируют информацию о населении страны, штата, города или других четко определенных географических территорий. Во многих случаях они включают конфиденциальную информацию, такую как возраст, раса, пол, религиозные убеждения и долгосрочное состояние здоровья. Данные переписи населения являются основными данными, используемыми для понимания социальных, экономических и демографических условий на местном и национальном уровне , а также для предоставления информации, которая необходима правительствам для разработки политики, планирования и управления государственными услугами, а также распределения финансирования. Защита данных переписи имеет важное значение от их сбора до обработки. Одной из мер защиты является предоставление доступа к данным только уполномоченным лицам. Шифрование является распространенным механизмом реализации этой защиты. Данные шифруются при хранении и передаются между авторизованными пользователями. Однако существует риск непреднамеренных ошибок со стороны авторизованных пользователей или того, что авторизованный пользователь станет злонамеренным, что поставит под угрозу безопасность и конфиденциальность данных, а также целостность анализа. В некоторых случаях данные необходимо предоставлять третьим сторонам (например, исследователям), чтобы можно было провести правильный и объективный анализ. В этих случаях для защиты конфиденциальности пользователей данные деидентифицируются или анонимизируются перед их передачей. Однако эти методы не всегда сохраняют конфиденциальность пользователей должным образом. Помимо

Machine Translated by Google

угрозы: когда становятся доступными большие наборы данных, информация пользователя может быть утечка из-за более сложных атак. особенно когда доступны большие наборы данных. Исследование 2016 года «Разоблачено! Обзор атак на частные данные», Дворк и др. представляет обзор распространенных атак на совокупные данные. Он описывает реконструкцию и отслеживание атак. Атаки реконструкции включают в себя аппроксимацию конфиденциальных характеристик отдельных лиц в наборе данных. Атаки по отслеживанию направлены на определение того, включены ли данные целевого человека в набор данных или нет [1]. За исследованием Дворка последовало множество исследований, посвященных атакам на конфиденциальность данных и механизмам защиты от них [2–6].

Два термина, связанных с безопасностью данных в статистическом смысле, — это статистическая конфиденциальность и статистическая конфиденциальность. Статистическая конфиденциальность – это «искусство разработки механизма конфиденциальности, который преобразует конфиденциальные данные в данные, одновременно полезные и неконфиденциальные» [7]. Целью методологий статистической конфиденциальности является «обеспечение широкого обмена данными в различных контекстах данных и областях, где желательно или требуется, чтобы личность людей или конфиденциальные атрибуты были защищены (например, перепись, здравоохранение, геномные данные, социальные сети)» [8].

Статистическая конфиденциальность означает «защиту информации отдельных статистических единиц» или, что эквивалентно, «защиту данных, собранных для статистических целей» [9 Статистическая конфиденциальность тесно связана со статистической конфиденциальностью. Несмотря на то, что в литературе не существует четкого различия между этими терминами, основываясь на определениях, упомянутых в этой статье, мы считаем, что статистическая конфиденциальность означает защиту статистических единиц, т. е. данных, собранных о физических или юридических лицах, и обеспечение одобренного использования их. Статистическая конфиденциальность — это более широкая концепция безопасности, которая включает преобразование конфиденциальных данных в неконфиденциальные и полезные данные. Статистическая конфиденциальность является одним из основополагающих принципов официальной статистики, одобренных Генеральной Ассамблеей ООН [10]. Это также фундаментальный принцип европейской статистики, определяющий принципы, правила и процедуры защиты конфиденциальных данных, в то же время разрешающий их использование в статистических целях [11].

В первой публикации «Инновации в федеральной статистике» [12] обсуждаются проблемы и необходимость новой парадигмы в федеральной статистической системе. Глава 5 публикации под названием «Защита частной жизни и конфиденциальности при предоставлении доступа к данным для исследовательских целей» посвящена защите конфиденциальности информации, собранной для статистического использования. Обсудив последствия для безопасности объединения нескольких наборов данных для анализа, включая правовую основу конфиденциальности и конфиденциальности, авторы делают два основных вывода: «Поскольку федеральные статистические агентства продвигаются вперед в объединении нескольких наборов данных, они должны одновременно заниматься количественной оценкой и контролем риск потери конфиденциальности» и «Методы повышения конфиденциальности и анализ статистических данных, сохраняющий конфиденциальность, потенциально могут позволить использовать источники данных частного сектора для федеральной статистики». С этими выводами они рекомендуют: «Федеральным статистическим агентствам следует внедрить современные технологии баз данных, криптографии, сохранения и повышения конфиденциальности».

В этом исследовании мы предлагаем систему, которая обеспечивает безопасность данных, статистическую конфиденциальность и конфиденциальность, используя методы и инструменты исключительно из математики и информатики . Мы изучаем использование системы на основе блокчейна, которая позволяет выполнять статистические вычисления над зашифрованными данными, используя полностью гомоморфное шифрование в блокчейне с использованием смарт-контрактов. Предлагаемая система обеспечивает сквозную безопасность данных с момента их передачи владельцу данных до момента получения исследователем результатов статистических вычислений. Обеспеченная безопасность включает целостность и конфиденциальность данных при передаче, хранении и использовании. Мало того, что отдельные точки данных никогда не раскрываются в виде открытого текста, но и результаты доступны только авторизованным пользователям в виде открытого текста. Кроме того, поскольку вычисления выполняются в блокчейне, вмешательство в алгоритмы статистических вычислений практически невозможно. Более того, запрос статистики, отличной от той, которая одобрена в смарт-контракте, невозможен. Таким образом, предлагаемая система обеспечивает статистическую конфиденциальность и конфиденциальность, необходимые для различных приложений, включая федеральные статистические системы. В разделе 2 мы приводим краткое описание недавних исследований, в которых блокчейн и гомоморфное шифрование используются для обеспечения безопасности данных. В

Machine Translated by Google

В разделе 3 мы описываем предлагаемую нами систему и реализацию концепции системы. В разделе 4 мы представляем результаты проверки концепции и, наконец, в разделе 5 мы даем обсуждение результатов.

2. Сопутствующая работа

Было проведено множество исследований, в которых пытались обеспечить безопасность и целостность данных не только тогда, когда они находятся в состоянии покоя, но и когда они используются, с помощью различных средств, включая гомоморфное шифрование и технологию блокчейна. В одном из таких исследований предлагается система парадигмы совместного обучения данных для совместного использования данных медицинских изображений, в которой реестр блокчейна обеспечивает децентрализацию моделей федеративного обучения без использования центрального сервера, а гомоморфное шифрование снимает проблемы, связанные с обменом необработанными данными [13].]. Их система гомоморфно шифрует градиенты и передает их через блокчейн.

Лян и др. [14] предлагают интеграцию блокчейна и гомоморфного шифрования для решения проблем защиты авторских прав. Их исследование предлагает математическую модель на основе гомоморфного шифрования внутри блокчейна, которая защищает транзакции, а также обеспечивает целостность и конфиденциальность данных за счет использования смарт-контрактов.

Яджи и др. [15] предложили систему, которая использует схемы шифрования Гольдвассера-Микали и Пайе для сравнительного оценочного исследования с акцентом на методы конфиденциальности данных с использованием технологии блокчейн для приложений ИИ. Исследование показало, что атак на блокчейн, таких как коллизия, прообраз и атаки на кошелек, можно избежать посредством шифрования блоков с использованием предложенных схем шифрования Гольдв

Мутлу и др. [16] предложили систему, которая использует технологию блокчейна и гомоморфное шифрование и позволяет третьим лицам (исследователям) выполнять линейную регрессию зашифрованных данных. Они используют алгоритм Палье для расчета суммы, необходимой для линейной регрессии, с помощью смарт-контрактов. Владелец данных шифрует данные с помощью открытого ключа исследователя и отправляет их в смарт-контракт, где выполняется расчет. Затем к зашифрованному результату может получить доступ исследователь, который может расшифровать его в своей системе, используя свой закрытый ключ.

Ванин и др. [17] предлагают модель защиты личных медицинских записей (PHR), в которой используется файловая система межпланетного протокола, основанная на распределенных хэш-таблицах (DHT) вместе с блокчейном. Метаданные PHR хранятся в блокчейне и распространяются по сети, а данные PHR хранятся вне цепочки через сеть IPFS. Они используют два элемента: Data Steward (DS), который отвечает за хранение PHR от имени человека, и Shared Data Vault (SDV), которое представляет собой временную область хранения IPFS, где медицинские учреждения могут получить доступ к PHR с согласия человека. . Зашифрованные данные доступны общественности через статистические порталы, где они могут выполнять операции с данными, используя гомоморфное шифрование, для получения значимых результатов. Для этой цели они используют библиотеку Microsoft Simple Encrypted Arithmetic Library (SEAL), которая реализует алгоритм BFV на языке JavaScript.

Умар и др. [18] предложили модель электронного голосования с использованием алгоритма Пайе. Как только избиратель отдает свой голос, он шифруется гомоморфно. Затем создается новый блок транзакции , который содержит зашифрованные бюллетени, псевдонимный адрес избирателя и администратора, временную метку создания блока, хэш предыдущего блока транзакции , а также хэш текущий блок. Затем новый блок транзакции добывается с использованием механизма консенсуса. После майнинга новый блок фиксируется в реестре блокчейна. Этот процесс продолжается до конца выборов, после чего администратор подсчитывает зашифрованные голоса с помощью алгоритма Пайе, который дает окончательную сумму, которую затем можно расшифровать для определения результатов.

Шреста и др. [19] провели исследование, в котором анализируются проблемы безопасности, связанные с Интернетом вещей (IoT). Одной из основных проблем, на которую было обращено внимание, является конфиденциальность данных. В этом исследовании также изучается возможность интеграции Интернета вещей с блокчейном и гомоморфным шифрованием. Преимущества включают неизменяемость данных, незабываемость, устранение единых точек отказа и конфиденциальность данных.

Machine Translated by Google

Кальдарола и др. [20] предложили Neural Fairness Protocol — механизм консенсуса , который объединяет алгоритм цифровой подписи на основе эллиптических кривых (ECDSA) и криптографические методы. Этот протокол повышает анонимность и подотчетность в транзакциях блокчейна, в частности, за счет разработки надежного порогового алгоритма ECDSA. Помимо первоначального применения, это нововведение демонстрирует потенциал для повышения безопасности различных криптовалют, таких как Litecoin и Mastercoin.

Предлагаемую систему в этом исследовании можно сравнить с работой, предложенной Ляном [14], Яджи [15] и Мутлу [16] в том смысле, что она использует технологию блокчейна вместе с гомоморфным шифрованием для обеспечения конфиденциальности данных, пока они используются. Кроме того, система, предложенная Мутлу [14], использует частичное гомоморфное шифрование смарт-контракта, что аналогично архитектуре предлагаемой нами системы. Однако в вышеупомянутых исследованиях используется алгоритм Палье, который является частично гомоморфным, тогда как в этом исследовании используется полностью гомоморфная система, которая позволяет не только складывать, но и умножать зашифрованные данные, что позволяет безопасно выполнять более широкий диапазон вычислений.

3. Материалы и методы.

В этом разделе мы описываем систему, которая обеспечивает безопасный обмен и анализ данных между владельцами данных и третьими лицами. Владельцами данных являются юридические или физические лица, имеющие доступ к простым данным. Их также можно рассматривать как датчики, способные обрабатывать данные и передавать их. Владельцы уполномочены или дают согласие на анализ данных при условии сохранения конфиденциальности данных. Третьими сторонами в системе являются исследователи. Это могут быть физические или юридические лица, уполномоченные проводить определенные виды анализа данных. У них может не быть разрешения на просмотр данных в открытом виде.

3.1. Описание предлагаемой системы

Система шифрует данные, используя гомоморфную схему шифрования, прежде чем они покинут владельца данных. Зашифрованные данные хранятся в смарт-контракте в блокчейне. Анализы, которые можно выполнить на основе данных, также хранятся в смарт-контракте. Когда исследователю необходимо провести анализ данных, предполагая, что тип анализа доступен в смарт-контракте, анализ выполняется на зашифрованных данных с использованием гомоморфных вычислений. Результаты гомоморфных вычислений также шифруются. Чтобы получить результат на своей стороне, исследователь запрашивает повторное зашифрование результата с помощью ключа исследователя, чтобы его можно было безопасно передать исследователю и расшифровать только он. Здесь следует отметить, что данные открытого текста никогда не раскрываются в течение всего процесса, даже во время повторного шифрования. Когда исследователь отправляет свой открытый ключ в смарт-контракт для повторного шифрования, смарт-контракт инициирует протокол распределенного дешифрования, посредством которого части результата расшифровываются различными валидаторами, повторно шифруются с использованием открытого ключа исследователя, а затем объединяются для получения конечный перезашифрованный результат [21]. На рисунке 1 показан обзор системы, в которой главными действующими лицами являются владелец данных и исследователь.

Machine Translated by Google

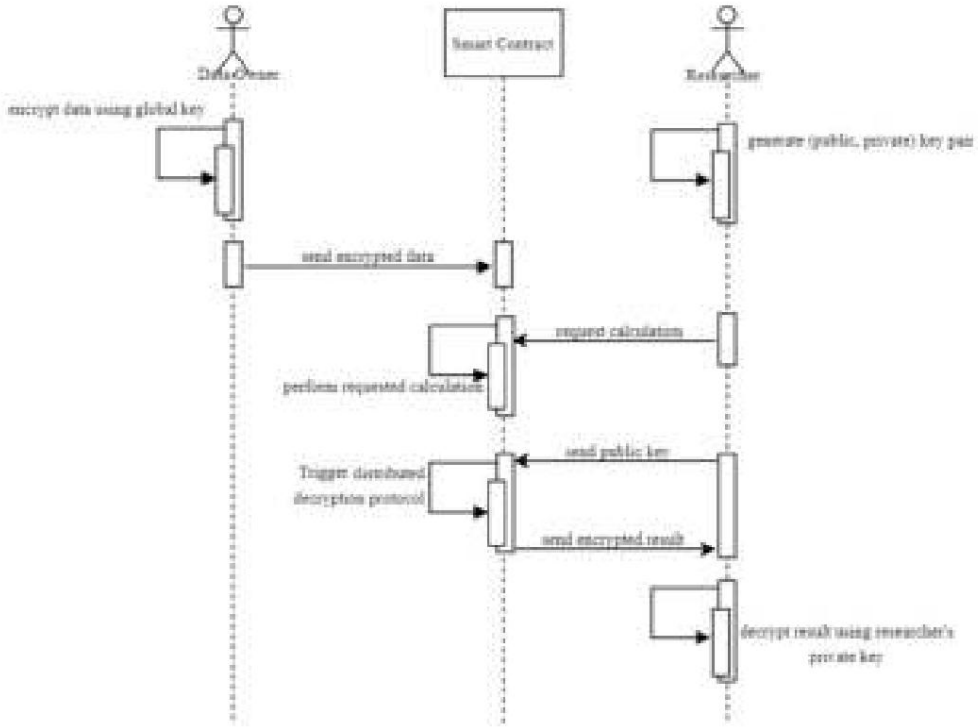


Рисунок 1. Обзор предлагаемой системы.

3.2. Реализация

Предлагаемая система реализована на блокчейне Zama, блокчейне на основе виртуальной машины Ethereum (EVM), который поддерживает вычисления над зашифрованными значениями [18]. Как и у Ethereum, у которого есть валюта, известная как Ether, у Zama есть собственная валюта, называемая ZAMA. Основная идея Zama — предоставление конфиденциальных смарт-контрактов. Это означает, что данные, отправленные или полученные от смарт-контракта, зашифрованы и не могут быть прочитаны, если передача данных будет перехвачена. Zama использует асимметричное шифрование и, следовательно, использует два ключа: открытый и закрытый. Открытый ключ, также известный как глобальный ключ, публично хранится в цепочке и используется каждым пользователем для шифрования своих данных и выполнения вычислений с этими зашифрованными данными. Глобальный ключ генерируется на этапе установки первоначальными валидаторами и безопасно повторно передается при изменении набора валидаторов [21]. Это позволяет смешивать зашифрованные данные от нескольких пользователей и по нескольким смарт-контрактам. Закрытый ключ используется для расшифровки данных и не принадлежит ни одному пользователю. Вместо этого Zama использует пороговый протокол. В пороговом протоколе части закрытого ключа распределяются между узлами валидации в сети, и определенное количество валидаторов должно сотрудничать для расшифровки данных. В случае с Zama для выполнения расшифровки необходимо участие не менее трети валидаторов [18]. Этот метод повышает безопасность за счет децентрализации, предотвращая несанкционированное использование ключа. Zama использует вероятностную схему шифрования, что означает, что зашифрованное значение одного открытого текста не всегда будет одинаковым. Например, рассмотрим массив, состоящий из 4 элементов [2, 4, 1, 4]. Гипотетически, как только элементы этого массива будут зашифрованы, он станет [d5048, e2314, x3215, o9849]. Обратите внимание, что число 4 дважды появляется в массиве открытого текста по индексу 1 и индексу 3, но его зашифрованная версия не имеет повторений, а значения по индексу 1 и индексу 3 различны. Эта схема еще больше повышает безопасность, не позволяя злоумышленникам получить информацию об открытом тексте tt на основе шаблонов в зашифрованном тексте.

фф

Предлагаемая система использует библиотеки для гомоморфных вычислений, предоставленные Zama. После компиляции контракта создается двоичный интерфейс приложения (ABI). ABI определяет функции, доступные в смарт-контракте, а также их параметры и типы возвращаемых значений, позволяя пользователям взаимодействовать со смарт-контрактом. Владелец данных и исследователь имеют свои собственные приложения JavaScript, которые используют ABI и методы.

Machine Translated by Google

предоставляется библиотекой Ethers для связи со смарт-контрактом. Владелец данных отправляет данные в смарт-контракт, а исследователь запрашивает анализ у смарт- контракта. На рисунке 2 показано взаимодействие участников со смарт-контрактом.

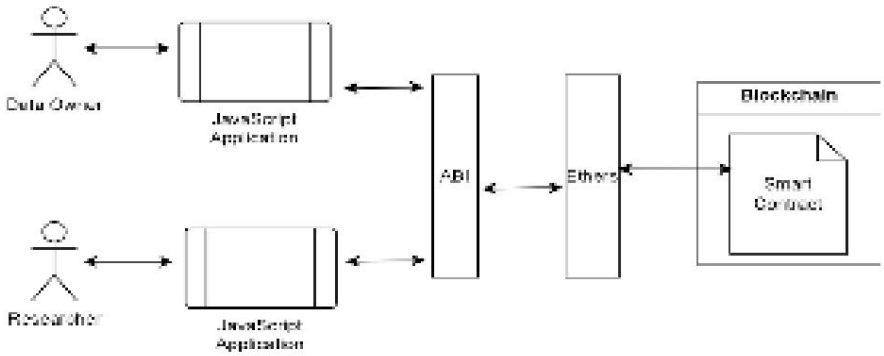


Рисунок 2. Взаимодействие со смарт-контрактом.

Ниже приведен список инструментов и технологий, которые использовались для внедрения системы:

- Solidity: язык программирования, который используется для написания смарт-контрактов на различных платформы, включая Ethereum.
- Node.js: это среда выполнения JavaScript, которая позволяет выполнять код JavaScript вне веб-браузера. Для этой системы использовалась версия 18.18.0.
- Remix: это интегрированная среда разработки (IDE) на основе браузера, используемая для разработка и внедрение смарт-контрактов.
- MetaMask: это криптовалютный кошелек, который используется для подключения и взаимодействия с блокчейн.
- Visual Studio Code: это интегрированная среда разработки, которая поддерживает разработку различных программных продуктов. языки, включая JavaScript.
- Ethers.js: это библиотека JavaScript, которая позволяет взаимодействовать со смарт-контрактами. Для этой системы использовалась версия 6.10.0.
- FhEVM: это библиотека, предоставляемая Zama, которая позволяет создавать конфиденциальные смарт-контракты на EVM с использованием Solidity. Для этой системы использовалась версия 0.4.0.
- FhEVMjs: это библиотека JavaScript, предоставляемая Zama, которая позволяет взаимодействовать со смарт-контрактами, включая шифрование данных и генерацию открытых/закрытых кл Для этой системы использовалась версия 0.4.0.

В рамках данного исследования проводится проверка концепции предлагаемой системы. Из-за ограничений библиотек для гомоморфных вычислений на Zama и работы с общедоступным блокчейном, как будет подробно описано в разделе 3.2.1, реализация включала вычисления простых описательных статистических данных, таких как среднее значение, медиана и дисперсия. Доказательство концепции было проверено на небольшом количестве точек данных в диапазоне чисел от 1 до 50.

3.2.1. Проблемы

При внедрении системы в Solidity и Zama были выявлены следующие ограничения и проблемы :

- Типы данных: fhEVM поддерживает только беззнаковые целые числа, которые являются 8-битными, 16-битными, 32-битными или 64-битными. Кроме того, методы, предоставляемые fhEVM для вычислений с зашифрованными данными, такие как сложение, вычитание, умножение и деление, возвращают зашифрованное целое число без знака. В случае деления существует дополнительное ограничение: зашифрованное целое число можно разделить только на целое число открытого текста. Это означает, что в настоящее время невозможно разделить два зашифрованных целых числа. Даже если бы существовала возможность такого разделения, результат всегда был бы зашифрованным целым числом из-за его возвращаемого типа; следовательно, невозможно работать с десятичными числами. Более

Machine Translated by Google

не подписаны; следовательно, невозможно обрабатывать и отрицательные целые числа. Эти ограничения ограничивают количество анализов, которые можно выполнить на смарт-контракте.

- Лимит газа: вычисления в блокчейне выполняются на газе — единице затрат, связанной с выполнением транзакции или вычислений в сети. Газ используется для оплаты валидаторам ресурсов, которые они используют для проведения транзакций. Лимит газа означает максимальное количество газа, которое можно потратить на транзакцию или вычисление. Установка ограничения на потребление газа важна, поскольку она гарантирует, что транзакции будут иметь заранее определенный верхний предел потребления ресурсов, чтобы защититься от сценариев, в которых выполнение кода может войти в бесконечный цикл. В настоящее время на Zama devnet установлен лимит газа в 10 000 000 [22]. Однако гомоморфные вычисления довольно затратны с точки зрения ресурсов, а поскольку эти вычисления выполняются на смарт-контракте, они потребляют много газа. Потребление газа зависит от операций, которые выполняются с зашифрованными данными, и увеличивается с увеличением сложности алгоритма и бит типа данных (32-битные операции будут потреблять больше газа по сравнению с 8-битными операциями). Текущий лимит газа ограничивает функциональность с точки зрения количества точек данных, которые могут быть обработаны.
- Solidity: В отличие от традиционных языков программирования, таких как Python и Java, Solidity имеет различные ограничения. Одна из них заключается в том, что структура данных, используемая для хранения значений пар ключей (сопоставлений), весьма ограничена с точки зрения функциональности, и невозможно перебирать ключи или значения. Это затрудняет реализацию различных алгоритмов, таких как вычисление моды (наиболее частое значение в массиве).

3.2.2. Расчет среднего значения

Для набора значений {x1, x2, . . . , xn} среднее арифметическое, обозначаемое $\bar{x}$, представляет собой сумму значений, деленную на количество значений, как показано в уравнении (1).

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n} \quad (1)$$

В разработанной системе в смарт-контракте рассчитывается только числитель — сумма зашифрованных значений . Операция деления выполняется на стороне исследователя . Это связано с ограничением, упомянутым в разделе 3.2.1 , согласно которому деление всегда будет возвращать целое число и, следовательно, терять точность результата. Подводя итог, смарт-контракт отправляет исследователю зашифрованную сумму, а также зашифрованное количество элементов.
Эти два значения затем расшифровываются исследователем, и для получения результата выполняется окончательное деление.

3.2.3. Расчет медианы

Вычисление медианы отсортированных данных — простая задача. Предполагая, что данные состоят из n значений, медиана — это средний элемент, если n нечетное, и среднее значение двух средних элементов, если n четное. Формула для медианы отсортированного набора данных с n точками данных приведена в уравнении (2).

$$\text{Мед} = \begin{cases} \frac{n+1}{2}\text{-й элемент if n нечетный} \\ \left(\frac{n}{2}\text{-й элемент} + \left(\frac{n}{2}+1\right)\text{-й элемент}\right) \text{ if n четный} \end{cases} \quad (2)$$

В системе дифференциация при вычислении медианы просто выполняется с помощью оператора if. Если n нечетное, медиана рассчитывается на смарт-контракте и возвращается исследователю в зашифрованном виде. Если n четное, сумма двух средних элементов рассчитывается в смарт-контракте и отправляется исследователю. Затем результат расшифровывается исследователем и делится на 2 для получения медианы. Таким образом, исследователь принимает решение о проведении деления исходя из количества элементов.

Приведенный выше расчет медианы предполагает, что данные отсортированы. Однако самой большой проблемой при вычислении медианы несортированных данных является реализация алгоритма сортировки в смарт-контракте, поскольку сортировка зашифрованных данных — нетривиальная задач

Machine Translated by Google

предлагаемая система использует алгоритм пузырьковой сортировки для сортировки данных. Этот алгоритм был выбран из-за его простоты. Он работает путем сравнения двух соседних значений в массиве и замены чисел на основе результата этого сравнения, что упрощает реализацию. Более того, реализация пузырьковой сортировки не требует какой-либо дополнительной структуры данных. Это позволяет писать на Solidity, учитывая ограничения языка программирования, а также библиотеки fhEVM.

3.2.4. Расчет дисперсии

Формула расчета дисперсии для набора значений {x1, x2, . . . , xn} определяется уравнением (3), где x представляет собой среднее значение данных.

$$B = \frac{\sum_{i=1}^{n} (xi - \bar{x})^2}{1\,n - 1} \tag{3}$$

Как обсуждалось в разделе 3.2.2, невозможно вычислить точное среднее значение любой последовательности в смарт-контракте, поскольку функция деления в библиотеке fhEVM всегда возвращает зашифрованное целое число. Чтобы исключить вычисление среднего значения в формуле дисперсии, мы вывели эквивалентную формулу, которая не включала среднее значение или какой-либо оператор деления, кроме основной операции деления на последнем шаге. Пошаговый вывод формулы, начиная с уравнения (3), показан ниже.

$$B = \frac{\sum_{i=1}^{n} (xi - \bar{x})^2}{1\,n - 1}$$

$$= \frac{\sum_{i=1}^{n} xj\,(xi - \frac{\sum_{j=1}^{n} xj}{n})^2}{n - 1}$$

$$= \frac{\sum_{i=1}^{n} (xi^2 - 2xi\frac{\sum_{j=1}^{n} xj}{n} + \frac{(\sum_{j=1}^{n} xj)^2}{n^2})}{n - 1}$$

$$= \frac{\sum_{i=1}^{n} xi^2 - 2\sum_{i=1}^{n} xi\frac{\sum_{j=1}^{n} xj}{n} + n\frac{(\sum_{j=1}^{n} xj)^2}{n^2}}{n - 1}$$

$$= \frac{n\sum_{i=1}^{n} xi^2 - 2\sum_{i=1}^{n} xi\sum_{j=1}^{n} xj + \frac{1}{n}(\sum_{j=1}^{n} xj)^2}{n(n - 1)}$$

$$= \frac{n\sum_{i=1}^{n} xi^2 - 2\sum_{i=1}^{n} xi\sum_{j=1}^{n} xj + (\sum_{j=1}^{n} xj)^2}{n(n - 1)} \tag{4}$$

Этот подход исключил из алгоритма вычисление среднего значения и, в свою очередь, также исключил деление. Однако, как показано в уравнении (4), полностью исключить деление невозможно . Учитывая эти ограничения, в смарт-контракте реализован только расчет числителя. Зашифрованный результат числителя отправляется исследователю вместе с зашифрованным значением n, где и производится окончательное деление. При более внимательном рассмотрении уравнения (4) можно заметить, что в самом числителе есть три основных компонента. Кроме того, числитель также включает в себя вычитание. Как обсуждалось в разделе 3.2.1, Zama предоставляет только целые числа без знака, а это означает, что любой результат ниже нуля подорвет точность результата. Чтобы решить эту проблему, эти три компонента рассчитываются отдельно, а их результаты сохраняются в отдельных переменных. Затем первый компонент добавляется к третьему, а затем второй компонент вычитается из результата сложения. Это гарантирует, что результат не упадет ниже нуля.

Machine Translated by Google

3.3. Гарантии безопасности предлагаемой системы

Предлагаемая система обеспечивает требования безопасности по всем принципам информационной безопасности. Он обеспечивает гарантии безопасности при передаче, хранении и использовании данных. Далее мы опишем, как конфиденциальность, целостность и доступность достигаются в предлагаемой системе.

- Конфиденциальность: данные шифруются перед отправкой и остаются зашифрованными с этого момента. Данные шифруются при передаче, хранении и использовании, что обеспечивает статистическую конфиденциальность, а также статистическую конфиденциальность. Зашифрованные данные передаются в смарт- контракт, где они хранятся, что позволяет выполнять над ними вычисления. Результат этих вычислений также шифруется. Прежде чем результат будет передан исследователю , он повторно шифруется после внутреннего расшифрования в блокчейне с использованием протокола распределенного дешифрования. Эта операция не видна на блоках цепочки и, следовательно, не видна пользователям. Обман системы для получения ключа дешифрования потребует сотрудничества одной трети узлов блокчейна, что маловероятно для блокчейнов. • Целостность данных: каждый блок в блокчейне содержит хэш предыдущего
блока, образуя цепочку. Вмешательство в один блок изменяет хэш этого блока и всех блоков, следующих за ним. Поскольку создание цепочки с обновленными блоками является практически невыполнимой задачей, данные в блокчейне защищены от несанкционированного доступа. Это обеспечивает целостность данных. • Целостность вычислений: Алгоритмы вычислений, разрешенные для данных, реализованы в смарт-
контракте. Поскольку смарт-контракты являются частью блокчейна, их несанкционированное изменение невозможно. Это обеспечивает целостность вычислений и, следовательно, доверие к алгоритмам, реализованным для вычислений. • Авторизация: Авторизацию в предлагаемой системе можно рассматривать с двух точек зрения. Одним из них является контроль того, какой анализ может быть
выполнен на данных. Поскольку функции для анализа доступны в смарт-контракте в предлагаемой системе, и только авторизованные пользователи/субъекты могут изменять функциональность смарт-контракта, исследователь ограничен использованием функций, доступных в смарт-контракте. Предполагая, что в смарт-контракте реализованы только утвержденные функции, исследователь не может выполнять несанкционированный анализ данных. Другой аспект авторизации в системе касается того, кто может отправлять данные в смарт-контракт и кто может запрашивать анализ данных. Приложения блокчейна можно запрограммировать для контроля доступа к функциям, предусмотренным в смарт-контракте. Каждый пользователь уникально идентифицируется своим открытым ключом, и этот ключ можно использовать для определения его доступа к функциям смарт-контракта.

- Доступность: Доступность — это функция, предоставляемая самой цепочкой блоков из-за ее децентрализованной природы. Каждый узел в сети имеет собственную копию реестра, поэтому даже если один узел станет недоступным, вся система останется доступной.
Это устраняет единственную точку отказа и повышает отказоустойчивость системы.

4. Результаты

Мы запустили экспериментальную реализацию системы, в которой смарт-контракт имеет функцию расчета среднего значения, медианы и дисперсии набора данных. Мы записали время, необходимое для выполнения различных операций для точек данных 4, 8, 10, 16, 24 и 32. Эти операции включают установку данных (т. е. отправку данных в смарт- контракт), расчет среднего значения, сортировку данных. (что необходимо для нахождения медианы), расчет медианы и расчет дисперсии. Все эксперименты проводились с 16-битным типом данных. Каждый набор данных был сгенерирован случайным образом с баллами в диапазоне от 0 до 50.
Каждый созданный набор данных сразу же отправлялся в смарт-контракт. Каждый размер набора данных был протестирован восемь раз, чтобы гарантировать надежность результатов. Испытания проводились в разное время и в разных географических точках с разной связью. В таблице 1 показано среднее время восьми испытаний для каждого размера набора данных.

Machine Translated by Google

Таблица 1. Время выполнения операций от запроса до результата в миллисекундах.

Количество Точки данных (n)	Отправка Данные	Расчет	Сортировка данных	медиана Расчет	Дисперсия Расчет
4	8088.66679	7512.111693	9063.634214	6312.926094	18,109,45269
8	6642.155596	8792.629253	22330.79229	7679.880587	28,161,12535
10	7243.960528	10486.1769	34,412,11508	7358.729393	
16	7111.459872	11948.44701			
24	9990.13181	16 596,62563			
32	13 358,13821	21 479,25259			

Как показано в Таблице 1, размеры некоторых наборов данных не подходили для определенных целей. операций из-за ограничения на газ, установленного Замой. Например, в процессе сортировки выполнялся на наборах данных, содержащих 4, 8 и 10 точек данных, процесс дисперсии был применяется только к наборам данных, содержащим 4 и 8 точек данных. Отправка наборов данных в блокчейн и расчет среднего значения можно выполнить для всех наборов данных.

На рис . 3 показано сравнение времени каждой операции в зависимости от количества точек данных. Он также сравнивает время для различных операций. В целом, поскольку количество точек данных увеличивается, увеличивается и время, необходимое для операции. Некоторыми исключениями из этого являются отправка данных для меньшего количества точек данных и расчет медианы. Мы приписываем это связано с тем, что операции, выполнение которых требует короткого времени, более подвержены на них влияет объем транзакций, а также качество сети возможность подключения. Как показано на графике, дисперсия и сортировка занимают больше времени, чем установка данные, расчет среднего значения и расчет медианы. Расчет медианы занимает минимум времени поскольку он включает в себя несколько простых операций по сравнению с другими операциями, независимо от размер данных.

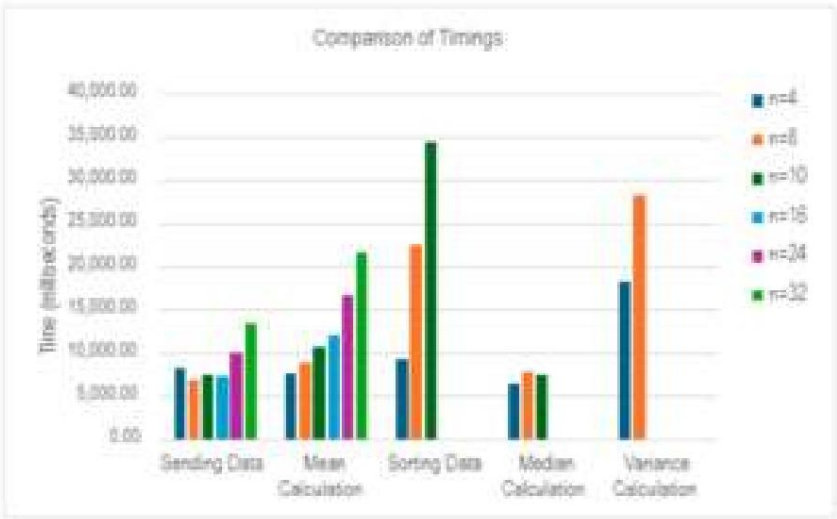


Рисунок 3. Сравнение времени операций по количеству точек данных.

Мы также записывали использование газа при каждой операции для каждого теста и проверяли, что сортировка данных и расчет отклонений достигли предела газа, установленного Zama, в 10 точках данных и 8 точек данных соответственно. В таблице 2 показаны средние значения восьми испытаний для каждого показателя.

установить размер.

Machine Translated by Google

Таблица 2. Использование газа на производстве.

Количество Точки данных (n)	Отправка Данные	Иметь в виду Расчет	Сортировка Данные	медиана Расчет	Дисперсия Расчет
4	290 491,5	577 618	1 367 854	169 517	5 022 035
8	565 458,875	1 124 706	6 007 343	169 517	8 213 723
10	709 171,5	1 398 250	9 565 005	169 517	
16	1 166 246,625	2 218 883			
24	1 835 907,625	3 313 060			
32	2 574 607,375	4 407 237			

Неудивительно, что расход газа был постоянным для заданного количества точек данных. за ту же операцию. На рисунке 4 показано сравнение использования газа с количеством точек данных для каждой операции, а также сравнение использования газа с различные операции.

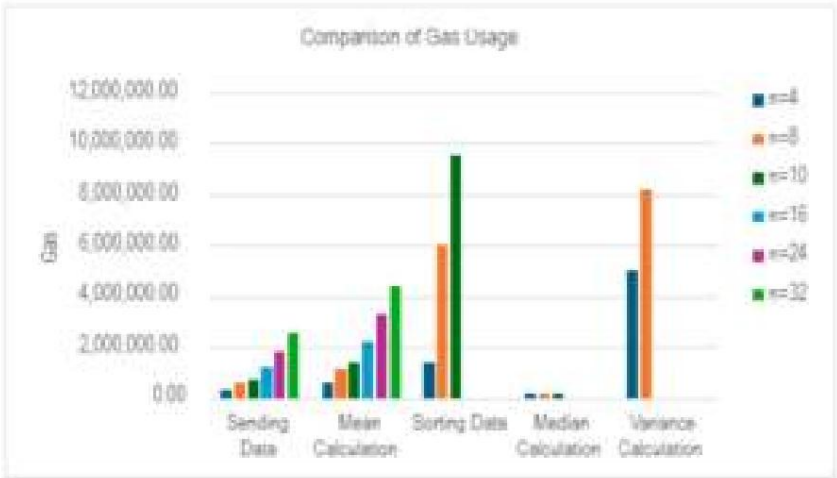


Рисунок 4. Сравнение использования газа операциями по количеству точек данных.

Как показано на рисунке 4 , потребление газа увеличивалось с увеличением количества точек данных, за исключением медиана. Медианный расчет включал в себя простую операцию нахождения индекса средний элемент, требующий всего одного сложения и одного деления независимо от числа точек данных. Следовательно, потребление газа было одинаковым для всех трех размеров набора данных. Расчет медиана требует сортировки данных. Однако, поскольку сортировка данных требует слишком много газа, невозможно рассчитать более чем для 10 точек данных. Аналогичным образом требуется расчет отклонения около 10 миллионов газа для восьми точек данных. Добавление дополнительной точки данных потребует более 10 миллионов газа, что превышает лимит газа, установленный Замой. Как есть видно из графика и описания реализации операций в Раздел 3: операции, требующие большего количества вычислений, требуют больше газа.

5. Обсуждение

В этом исследовании была предложена система, которая обеспечивает безопасность данных, статистическую конфиденциальность и конфиденциальность с использованием методов и инструментов исключительно из области математики и информатики. Система использует полностью гомоморфное шифрование в блокчейне и обеспечивает сквозную безопасность данных с момента, когда они покидают владельца данных, до тех пор, пока их не отправит исследователь получает результаты вычислений. Данные защищены при передаче, хранении, а также в использовать. Кроме того, предлагаемая система также обеспечивает вычислительную целостность. Со всем Благодаря предоставляемой ею защите система является идеальным решением для обеспечения безопасности данных. требования.

Machine Translated by Google

Проверка концепции предлагаемой системы осуществляется на публичном блокчейне Zama. При реализации исследователь может запросить безопасные вычисления среднего, медианы и дисперсии. Представлены результаты по времени выполнения необходимых операций над различным количеством точек данных, а также ограничения работы с библиотеками гомоморфного шифрования и общедоступным блокчейном. Хотя это исследование успешно демонстрирует возможность защиты данных и вычислений в блокчейне с использованием полностью гомоморфного шифрования, существуют определенные проблемы и области, которые требуют улучшения.

Полностью гомоморфные операции довольно дороги с точки зрения ресурсов, и существуют ограничения на количество вычислений, которые могут быть выполнены в блокчейне до превышения лимитов газа и стоимости. Это особенно важно при запуске приложения в основной сети, где за транзакции взимается комиссия. Это ограничение можно частично смягчить в частной или разрешенной сети блокчейна. Такой блокчейн предоставляет те же гарантии, что и общедоступный блокчейн, и дает больше контроля над тем, кто может использовать приложение.

Будущие исследования в этой области могут включать более сложные алгоритмы в различных дисциплинах, от статистики до машинного обучения, и включать более широкий диапазон данных по мере появления новых ресурсов и развития технологий. Более того, традиционные алгоритмы, разработанные для работы с открытыми текстовыми данными, такие как поиск и сортировка, необходимо тщательно пересмотреть для поддержки гомоморфных вычислений с учетом возможностей доступных библиотек. Еще одним возможным направлением будущих исследований может стать разработка структуры, которая поддерживает полностью гомоморфные вычисления на основе этого смарт-контракта и обеспечивает более широкий диапазон функциональных возможностей, преодолевая при этом ограничения текущей системы, такие как разрешение деления двух зашифрованных чисел и включение десятичных или плавающих чисел. номера точек.

Вклад автора: Концептуализация, YKP; Курирование данных, RR, YKP и ZDM; Формальный анализ, YKP и ZDM; Расследование, PP, ЮКП и ЗДМ; Методология, PP и YKP; Администрация проекта, YKP; Ресурсы, PP, ЮКП и ЗДМ; Программное обеспечение, PP и ЗДМ; Надзор , ЮКП; Валидация, RR, YKP и ZDM; Визуализация, YKP и ZDM; Письменность — первоначальный проект, RR и YKP; Написание – рецензирование и редактирование, RR, YKP и ZDM. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: оригинальные материалы, представленные в этом исследовании, включены в статью. Дополнительные запросы можно направить соответствующему автору(ам).

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Дворк, К.; Смит, А.; Стейнке, Т.; Ульман, Дж. Разоблачен! Обзор атак на частные данные. Анну. Преподобный Стат. Прил. 2017, 4, 61–84. [Перекрестная ссылка]

2. Китамура, К.; Ирван, М.; Ямагучи, Р.С. Тестовые атаки на анонимность и индикаторы уязвимости для раскрытия «характеристик пациента» в медицинских статьях. В материалах семинаров Европейского симпозиума IEEE по безопасности и конфиденциальности 2022 года (EuroS&PW), Генуя, Италия, 6–10 июня 2022 года; стр. 186–193. [Перекрестная ссылка]

3. Ниссим, К. Конфиденциальность: от реконструкции базы данных к юридическим теоремам. В материалах 40-го симпозиума ACM SIGMOD-SIGACT-SIGAI по принципам систем баз данных, в PODS'21, виртуальное мероприятие, Китай, 20–25 июня 2021 г.; Ассоциация вычислительной техники: Нью-Йорк, штат Нью-Йорк, США, 2021 г.; стр. 33–41. [Перекрестная ссылка]

4. Ван Ю.-Р.; Цай, Ю.-К. Защита обмена данными для обеспечения конфиденциальности в Financial Vision. Прил. наук. 2022, 12, 7408. [CrossRef]

5. Агарвал, С. Срей и Мишра, конфиденциальность данных и моделей. В «Ответственном искусственном интеллекте: реализация этических и объективных алгоритмов»; Спрингер Международное издательство: Чам, Швейцария, 2021 г.; стр. 153–170. [Перекрестная ссылка]

6. Соколовская А.; Кочарев Л. Интеграция технических и правовых концепций конфиденциальности. IEEE Access 2018, 6, 26543–26557. [Перекрестная ссылка] 7. Кифер, Д.; Лин, Б.-Р. Аксиоматический взгляд на статистическую конфиденциальность и полезность. Дж. Прив. Конфиденциальность 2012, 4. [CrossRef]

8. Славкович, А.; Симан, Дж. Конфиденциальность статистических данных: песня о конфиденциальности и полезности. Анну. Преподобный Стат. Прил. 2023, 10, 189–218. [Перекрестная ссылка]

9. С.З. Статистическая конфиденциальность – Руководство для составителей РСМТУ 2010 г. – Статистическая информация ООН. Доступно онлайн: https://unstats.un.org/.wiki/display/M2CG/C.3.++Статистика+конфиденциальность (по состоянию на 4 июня 2024 г.).

Machine Translated by Google

10. СООН — Фундаментальные принципы национальной официальной статистики. Доступно онлайн: https://unstats.un.org/fpos/. (доступ на 5 июня 2024 г.).
11. Статистическая конфиденциальность и защита персональных данных — Евростат. Доступно онлайн: https://ec.europa.eu/eurostat/web/. микроданные/статистическая конфиденциальность и защита персональных данных (по состоянию на 18 июня 2024 г.).
12. Национальные академии наук, техники и медицины; Отдел поведенческих и социальных наук и образования; Комитет национальной статистики; Группа по совершенствованию федеральной статистики для политических и социальных исследований с использованием нескольких источников данных и современных методов оценки; Гроувс, РМ; Харрис-Кожетин, бакалавр (ред.) Инновации в федеральной статистике: объединение источников данных при защите конфиденциальности; The National Academies Press: Вашингтон, округ Колумбия, США, 2017 г. [CrossRef]
13. Кумар, Р.; Кумар, Дж.; Хан, А.А.; Али, Х.; Бернар, СМ; Хан, РУ; Цзэн, С. Агрегация моделей медицинских изображений на основе блокчейна и гомоморфного шифрования, сохраняющая конфиденциальность. Вычислить. Мед. График изображения. 2022, 102, 102139. [CrossRef] [ПабМед]
14. Лян, В.; Чжан, Д.; Лей, Х.; Тан, М.; Ли, К.-Ц.; Зомая, А. Блокчейн с авторским правом Circuit: гомоморфный подход на основе блокчейна Шифрование для защиты IP-каналов. IEEE Транс. Экстрен. Вершина. Вычислить. 2020, 9, 14:10–14:20. [Перекрестная ссылка]
15. Яджи, С.; Бангера, К.; Нилима, Б. Сохранение конфиденциальности в блокчейне на основе системы частичного гомоморфного шифрования для приложений искусственного интеллекта. В материалах 25-й Международной конференции IEEE по высокопроизводительным вычислениям (HiPCW) 2018 г., Бангалор, Индия, 17–20 декабря 2018 г.; стр. 81–85. [Перекрестная ссылка]
16. Мутлу, З.Д.; Пекер, Ю.К.; Айдын, А. Сельчук. Линейная регрессия с сохранением конфиденциальности на основе блокчейна. Дж. Сообщество миллиметровых волн. Оптим. Модель. 2023, 3, 45–49. Доступно онлайн: https://www.jomcom.org/index.php/1/article/view/81. (по состоянию на 5 июня 2024 г.).
17. Ванин Ф.Н.; Поликарпо, LM; Риги, РД; Черт возьми, С.М.; да Силва, В.Ф.; Голдим, Дж.; да Коста, Калифорния. Модель сквозной защиты данных на основе блокчейна для обмена личными медицинскими записями: подход полностью гомоморфного шифрования. Датчики 2023, 23, 14. [Перекрестная ссылка] [ПабМед]
18. Умар, Б.; Оланийи, О.; Олажиде, Д.; Дого, Э. Пайе ChainNode на основе криптосистемы для безопасного электронного голосования. Передний. Блокчейн 2022, 5, 927013. [CrossRef]
19. Шреста, Р.; Ким, С. Интеграция Интернета вещей с блокчейном и гомоморфным шифрованием: сложные проблемы и возможности. Адв. Вычислить. 2019, 115, 293–331. [Перекрестная ссылка]
20. Кальдарола, Ф.; д'Атри, Г.; Занардо, Э. Протокол блокчейна Neural Fairness с использованием лотереи эллиптических кривых. Математика 2022, 10, 3040. [ПерекрестнаяСсылка]
21. fhevm/fhevm-whitepaper.pdf на main·zama-ai/fhevm·GitHub. Доступно онлайн: https://github.com/zama-ai/fhevm/blob/. main/fhevm-whitepaper.pdf (по состоянию на 4 июня 2024 г.).
22. Оценка газа|0,3|fhEVM. Доступно онлайн: https://docs.zama.ai/fhevm/v/0.3-2/how-to/gas. (по состоянию на 4 июня 2024 г.).

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Моделирование анализа сборки каналов в CRN на основе Стратегия приоритетного планирования с зарезервированной очередью

Цяньюй Сюй [1], Суопин Ли [2,*], Джаафар Габер [3] и Юйчжоу Хань [2]

[1] Школа электротехники и электроники, Сианьский авиационный университет, Сиань 710077, Китай; 202202001@xaau.edu.cn

[2] Школа электротехники и информационной инженерии, Технологический университет Ланьчжоу, Ланьчжоу 730050, Китай; 19119376252@163.com

[3] Факультет компьютерных наук и вычислительной техники, Технологический университет Бельфор-Монбельяр, 90010 Бельфор, Франция; gaber@utbm.fr

* Переписка: lsuop@lut.edu.cn

Аннотация: В сетях когнитивного радио сборка каналов позволяет вторичным пользователям (SU) расширять пропускную способность сети и улучшать использование спектра. Стратегии планирования, основанные только на классификации гетерогенных услуг, не могут гарантировать приоритет доставки жизненно важных эластичных услуг в особых сценариях, таких как экстренное спасение. Поэтому в этой работе предлагается стратегия приоритетного планирования с зарезервированной очередью (Ps-rq). Статический коэффициент определяется для классификации SU на эластичные службы и службы реального времени на основе типа сообщения, а динамический коэффициент определяется для дифференциации высокоприоритетных эластичных служб на основе достоверности информации, корреляции сообщений и размера сообщения. Пользователи с высоким приоритетом в прерванных эластичных службах помещаются в зарезервированную очередь для обеспечения их обслуживания. Соответственно, представлены алгоритм планирования и процесс ди Проводится анализ цепи Маркова в непрерывном времени и выводятся все возможные переходные состояния, триггерные события, скорости перехода и условия перехода системы, начиная с общего состояния. Кроме того, получены оценочные показатели производительности системы. Примеры исследований и результаты моделирования доказывают, что предлагаемая стратегия может повысить пропускную способность сети, снизить вероятность блокировки и вероятность принудительного завершения для вторичных пользователей, а также заметно повысить производительность высокоприоритетных эластичных услуг. Кроме того, мы проанализируем характеристики Ps-rq путем всестороннего сравнения с четырьмя другими схемами. Эксперимент доказывает, что стратегия Ps-rq может эффективно улучшить качество обслуживания жизненно важных эластичных сервисов на основе обеспечения справедливого планирования.

Ключевые слова: когнитивные радиосети; сборка каналов; стратегия приоритетного планирования; резервная очередь; цепь Маркова; использование спектра; условия перехода

check for updates

Цитирование: Сюй, К.; Ли, С.; Габер, Дж.; Хан, Ю. Моделирование анализа сборки каналов в CRN на основе стратегии приоритетного планирования с Зарезервированная очередь. Электроника 2024, 13, 3051. https://doi.org/10.3390/электроника13153051

Академические редакторы: Дюк Тай Ле и Сайед Мухаммад Раза

Поступила: 11 июня 2024 г.
Пересмотрено: 29 июля 2024 г.
Принято: 30 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Традиционные сети беспроводной связи в основном основаны на статическом распределении спектра. Однако лицензированный пользователь всегда занимает фиксированную полосу частот, что приведет к чрезмерной трате спектра [1]. Сети когнитивного радио (CRN) позволяют вторичным пользователям (SU) исследовать и использовать неиспользуемый спектр первичных пользователей (PU) посредством измерения спектра для улучшения использования спектра [2,3].

Улучшение использования спектра требует от систем когнитивного радио гибкого планирования и распределения доступных ресурсов спектра [4,5]. Динамический доступ к спектру (DSA) может гарантировать, что SU получат доступ к пробелам в лицензируемой полосе спектра [6]. Гибкость DSA отражается в том, что текущие SU могут выйти из канала после обнаружения сигналов PU , а затем адаптация спектра может быть выполнена посредством передачи обслуживания спектра или динамического выбора канала. Разработка правил планирования ресурсов на основе DSA, соответствующих среде приложения, может помочь повысить производительность вторичной сети [7].

Machine Translated by Google

В CRN внедрение технологии сборки каналов и технологии адаптации спектра в процессе DSA может значительно повысить эффективность обслуживания SU.
Технология сборки каналов позволяет SU связывать или объединять один свободный канал с другими для увеличения пропускной способности вторичной сети. В сочетании с технологией адаптации спектра текущие SU могут адаптивно регулировать количество собранных каналов в соответствии с деятельностью других PU и SU, чтобы обеспечить гибкое планирование. Однако моделирование процесса DSA на основе распределения ресурсов является сложным, поэтому необходимо анализировать и оценивать производительность вторичных сетей, чтобы помочь разработать правила работы

Производительность обслуживания SU неотделима от метода планирования и распределения ресурсов. Поток услуг в беспроводной связи можно разделить на услуги реального времени (например, аудио и видео) и гибкие услуги (например, данные) на основе чувствительности к задержке. Служба реального времени обычно имеет более высокий приоритет, чем эластичная служба, в то время как эластичная служба имеет меньшую чувствительность к задержке, а скорость ее передачи можно адаптивно регулировать в соответствии с сетевой средой.

Исследователи предложили различные методы распределения ресурсов для гетерогенного трафика для улучшения использования спектра [8–14]. Основной метод заключается в объединении доступных ресурсов спектра в пул спектра в соответствии с динамической сетевой средой, а затем в их распределении по требованию. Среди них реф. [10–13] исследовали проблему планирования ресурсов многоканальных когнитивных радиосетей с гетерогенным трафиком на основе сборки каналов и использовали математические методы для создания модели анализа для оценки производительности вторичных сетей. Предыдущие исследования [11] доказали, что организация переменного количества объединенных каналов для эластичных услуг и фиксированного количества объединенных каналов для услуг реального времени в зависимости от операционной среды может помочь повысить эффективность использования спектра.

В [15] авторы предположили, что принятие процесса цепи Маркова с непрерывным временем (CTMC) может обеспечить метод моделирования для анализа производительности CRN со сборкой каналов, который имеет то преимущество, что позволяет избежать проблемы синхронизации времени PU и CУ на одном канале. В [11] авторы предложили стратегию сборки каналов с учетом адаптации спектра и гетерогенных услуг, а CTMC использовался для оценки производительности вторичной сети. Однако прерванное ПУ СУ не может продолжить обслуживание после выхода из канала, что заведомо увеличивает вероятность сброса СУ. Поэтому исх. [12] предложили создавать очереди для разнородных услуг и применять алгоритм планирования очередей для повышения пропускной способности сети на основе расстановки приоритетов. Предыдущие исследования [12,16] доказали, что создание очередей для SU может повысить пропускную способность сети и снизить вероятность блокировки SU.
С учетом множества факторов, влияющих на пропускную способность канала, включая изменяющуюся во времени беспроводную линию связи, конечный буфер, классы трафика и схему адаптивной модуляции и кодирования (AMC) , в [17] была предложена стратегия сборки каналов с очередями, а также Взаимодействие между параметрами системы исследовалось путем создания модели анализа CTMC. В [18] на основе сборки каналов был предложен адаптивный алгоритм аренды спектра , а производительность DSA трафика SU с приоритетом доступа обсуждалась путем разработки модели CTMC. Вышеупомянутое исследование показывает, что внедрение технологии сборки каналов для гетерогенных услуг может улучшить использование свободного спектра. Кроме того, создание CTMC может помочь смоделировать динамику планирования ресурсов спектра среди пользователей в задаче распределения ресурсов. Вышеупомянутое исследование представляет собой более строгую и усовершенствованную методологию математического анализа производительности вторичной системы.

В CRN с услугами нескольких типов распределение ресурсов на основе классификации услуг может повысить надежность передачи важной информации. Однако улучшения обычно приводят к проблеме нехватки низкоприоритетных услуг из-за отсутствия обслуживания в течение длительного времени. Схемы планирования ресурсов, которые могут облегчить проблему голодания пользователей с низким приоритетом, были предложены в [19–21], а производительность вторичной системы оценивалась путем создания моделей марковского анализа с очередями. В [20] была принята технология привязки каналов с предотвращением голодания и невытесняющий приоритет.

Machine Translated by Google

Модель очередей M/G/1 использовалась для описания процесса передачи спектра абонентов SU, чтобы улучшить использование ресурсов. Алгоритм планирования циклического приоритета (RRP), сочетающий статическое и динамическое распределение ресурсов, был предложен в [21]. Этот алгоритм получил возможности передачи для пользователей с низким приоритетом путем временной приостановки очереди с высоким приоритетом. Целью этого является минимизация нехватки низкоприоритетных услуг в CRN. Фактически, алгоритм RRP в [21] использует планирование опроса очереди, но когда трафик одной очереди велик, планирование пользователей только на основе опроса очереди значительно снизит производительность системы [22]. Планирование разнородных услуг, основанное на разделении типов услуг, действительно может предложить более удобную схему планирования приоритетов для распределения каналов и планирования приоритетов в CRN во время инициализации системы . Однако предыдущее исследование не смогло удовлетворить требования к передаче высокоприоритетных эластичных пользователей в практических приложениях и, таким образом, не смогло обеспечить фундаментальное решение проблемы нехватки услуг, вызванной приоритетами планирования.

Следовательно, это мотивирует исследование нашего исследования. В этом документе исследуется планирование и распределение ресурсов с классификацией услуг для CRN, которым требуются услуги ранжирования в особых сценариях. Некоторым важным эластичным услугам необходимо уделять более высокий приоритет в особых сценариях применения CRN, включая экстренное спасение, дроны [23] и БПЛА. Поэтому, учитывая, что схемы планирования с иерархическими сервисами часто приводят к проблеме нехватки низкоприоритетных сервисов из-за длительных периодов необслуживания сервисов, целью данного исследования является повышение эффективности доставки важной информации на основе обеспечения справедливого механизма распределения. насколько это возможно.

В отличие от наших предыдущих исследований [9], в этой статье рассматривается более общая схема планирования, основанная на гетерогенных сервисах. Целью данного исследования является предложение более комплексной схемы планирования и распределения ресурсов для CRN, которая может гарантировать качество обслуживания высокоприоритетных эластичных услуг. Более конкретно, на основе планирования опроса предоставляется новая зарезервированная очередь для эластичных услуг с высоким приоритетом, которые прерываются PU, а самый высокий приоритет предоставляется для SU в зарезервированной очереди. На основе технологии сборки каналов и адаптации спектра анализируется процесс динамического доступа к спектру гетерогенных услуг и исследуется распределение ресурсов спектра между SU посредством разработки CTMC. Основные достижения этого исследования заключаются в следующем: 1. В этом исследовании предлагается стратегия приоритетного

планирования с зарезервированной очередью (стратегия Ps-rq). Представлены системная модель и схема классификации услуг предлагаемой стратегии . Комбинируя технологию сборки каналов и адаптации спектра, создается пропорциональная схема распределения приоритетов, основанная на стратегии Ps-rq.

2. Процесс потока ресурсов вторичной сети отображается на CTMC. Используя методы стохастического анализа, сложные действия пользователя в CRN индивидуально связаны с переходами состояний. Предоставляются событие перехода, скорость перехода и условие перехода, которые могут вызвать изменения в состояниях системы, тем самым предлагая теоретическую основу для повышения производительности вторичных систем посредством сборки каналов.

3. Приведена процедура уменьшения размерности многомерных цепей Маркова.
На основе этого исследуются и численно моделируются аналитические результаты работы вторичной сети. Мы представляем сравнение Ps-rq с четырьмя предыдущими схемами, и посредством числовой проверки доказано, что предлагаемая стратегия обеспечивает качество обслуживания для прерванных важных эластичных услуг, сохраняя при этом справедливое планирование.

2. Стратегия приоритетного планирования с зарезервированной очередью (стратегия Ps-rq)

В этом разделе представлена схема классификации гетерогенных услуг для CRN с услугами нескольких типов. На этой основе предлагается стратегия приоритетного планирования с зарезервированной очередью (Ps-rq). Цель состоит в том, чтобы обеспечить справедливое планирование для CRN с многотипными услугами, которое может гарантировать надежность доставки важной информации, одновременно облегчая проблему нехватки очередей с низким приоритетом.

Machine Translated by Google

2.1. Модель системы и предположения

Централизованные сети когнитивного радио (CRN) состоят из двух типов радиостанций: первичного пользователя (PU) и вторичного пользователя (SU). Планирование пользователя контролируется соответствующей базовой станцией. Центральный узел отвечает за взаимодействие с окружающей средой , а также за управление распределением спектра и решениями по планированию для других SU через канал управления. Центральный узел в централизованных CRN может повысить производительность системы за счет принятия эффективных решений и снижения помех между пользователями. В отличие от распределенных CRN, централизованные CRN устраняют необходимость дальнейшего обсуждения таких аспектов, как межпользовательские помехи и эгоизм узлов.

Предположим, что имеется M (M Z) PU, совместно использующих полосу спектра, и каждая PU имеет доступ только к одному каналу. По сравнению с SU, PU имеют абсолютный приоритет доступа к каналу. SU оснащены средствами обнаружения спектральной энергии, при условии, что SU обладают идеальной сенсорной способностью. А необходимое время обнаружения настолько мало, что ПУ не достигнут канала за это время. Как только PU будет обнаружен, текущий SU немедленно освободит собранные каналы и выйдет, одновременно выполнив адаптацию спектра . Если предположить, что на адаптацию спектра затрачивается меньше времени, то переключение каналов не повлияет на передачу SU.

2.2. Классификация гетерогенных услуг

При поддержке системной модели в этом разделе предлагается гетерогенная схема классификации услуг для удовлетворения требований планирования для CRN с услугами нескольких типов, включая данные, изображения, аудио и видео. Например, задержка доставки или потеря пакетов жизненно важной информации пациентов во время неотложной медицинской помощи может быть опасной для жизни, и эта часть информации в основном представляет собой эластичную услугу. Однако предыдущее исследование [12] , основанное на классификации гетерогенных услуг, не учитывало важные эластичные услуги с высоким приоритетом доставки в особых сценариях. Приоритет передачи и надежность важных эластичных услуг вряд ли могут быть обеспечены. На основе классификации гетерогенных услуг многотипные услуги в CRN также классифицируются по динамическим факторам и выделяются эластичные услуги с высоким приоритетом. Подробности следующие:

- Статический фактор: тип сообщения. SU в CRN классифицируются на эластичные услуги (SUe) и услуги реального времени (SUr) в соответствии с чувствительностью к задержке. Эластичные услуги характеризуются короткой длиной кадра, интенсивным потоком трафика и высокой устойчивостью к задержкам. Услуги реального времени требуют своевременности передачи данных. В зависимости от допуска задержки приоритет услуги определяется следующим образом:

$$\text{Приоритет: SUr} > \text{SUe};$$

- Динамический фактор: эластичные услуги в CRN переклассифицируются на основе трех динамических факторов, включая достоверность информации, корреляцию сообщений и размер сообщения.

Срок действия информации (Val) относится к времени, оставшемуся до крайнего срока. Сообщения с меньшим оставшимся временем имеют более высокий приоритет.

$$\text{Вал} = \frac{\text{Время, оставшееся до крайнего срока}}{\text{Срок доставки информации}}; \quad (1)$$

Корреляция сообщений (Cor) относится к типу сообщения, которое необходимо доставить в первую очередь.

$$\begin{matrix}\text{Кор} = 1, & \text{Актуальность важной информации} \\ \text{Кор} = 0, & \text{Другие}\end{matrix}; \quad (2)$$

Размер сообщения (Size) обратно пропорционален приоритету сообщения.

$$\text{Размер} = \frac{\text{Размер сообщения SU}}{\text{Средняя длина сообщения}}; \quad (3)$$

Machine Translated by Google

Размер сообщения SU;

Средняя длина сообщения

После объединения трех вышеупомянутых динамических факторов эластичные услуги классифицируются на высокие, средние и низкие приоритеты. Цель состоит в том, чтобы идентифицировать эластичные сервисы с высоким приоритетом (SUh e), а затем установить их приоритет перед услугами реального времени. Приоритет обслуживания предлагается следующий:

$$\text{Приоритет}: \text{СУ}_e > \text{СУр} > \text{СУэ}$$

Кроме того, трехмерное (3D) представление приоритетов эластичных сервисов представлено на рисунке 1. Нечеткое управление Такаги-Сугено используется для определения приоритетов эластичных сервисов [24]. Входное значение системы нечеткого вывода находится в диапазоне от 0 до 1. Выход системы является приоритетным и делится на три нечетких множества, домены которых равны [-1, 1]. По приспособленности правилом «Если-то» принято устанавливать 27 нечетких правил, а по приоритетности эластичные сервисы делятся на высокие, средние и низкие приоритеты. Например, если Val = 0, Cor = 1 и Size = 1, то Priority = 1, что соответствует важной информации с коротким оставшимся временем и небольшим размером сообщения. Из входных и выходных переменных используются гауссовы и трапециевидные функции принадлежности для получения трехмерного приоритетного представления эластичных

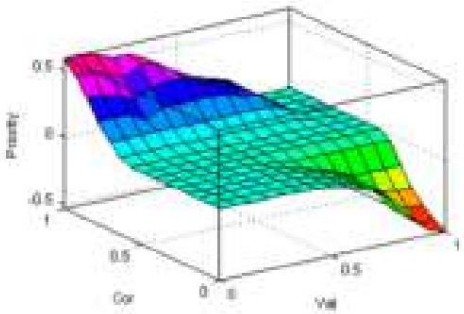


Рисунок 1. Трехмерное приоритетное представление эластичных сервисов.

2.3. Стратегия Ps-rq На

основе описанной выше схемы классификации гетерогенных услуг трафик с многотипными услугами, поступающий в систему, обслуживается стратегией приоритетного планирования с зарезервированной очередью (стратегия Ps-rq). Как показано на рисунке 2, стратегия Ps-rq состоит из блока планирования информации и блока передачи данных. Пользователи, входящие во вторичную систему (SU), делятся на эластичную службу (SUe) и службу реального времени (SUr) в соответствии со статическим коэффициентом через классификатор пакетов I. (Классификаторы пакетов используются для анализа типов входящих пакетов.) Для SUe и SUr организованы разные физические очереди , которые обозначаются Le и Lr .

Если во время передачи обслуживание SU прерывается PU , будет выполнена адаптация спектра. Учитывая, что SUr требует высокой своевременности и обладает устойчивостью к джиттеру, в случае неудачной адаптации спектра прерванный SUr будет отброшен и освободит собранные каналы, а не вернется обратно в То есть у Lr нет петли обратной связи.

В отличие от SUr , SUe предъявляет низкие требования к своевременности и может содержать важные L -сигналы. Таким образом, для Le организуется петля обратной связи . Прерванное SUe будет повторно введено в классификатор пакетов II и классифицировано по высоким, средним и низким приоритетам на основе динамического фактора. Цель состоит в том, чтобы отделить эластичные сервисы с высоким приоритетом, пометить их как SUh и , поместить в зарезервированную очередь L e . При этом прерванная эластичная услуга со средним приоритетом (SUm e) и эластичная услуга с низким приоритетом (SUl e) снимаются со своих меток и возвращаются в конец очереди Le (петля обратной связи). Предполагая, что каждая очередь изоморфна, вторичные пользователи в очереди ожидают планирования в соответствии с порядком обслуживания в порядке очереди (FCFS). Приоритет очереди установлен на L > Lr > Le . SU разных типов могут выстраиваться только в соответствующую очередь. Обслуживание SU будет заблокировано, если в очереди не останется места. Если он заполнен, когда прерванные SUe и SUh входят в соответствующую очередь, услуга будет вынуждена завершиться.

Machine Translated by Google

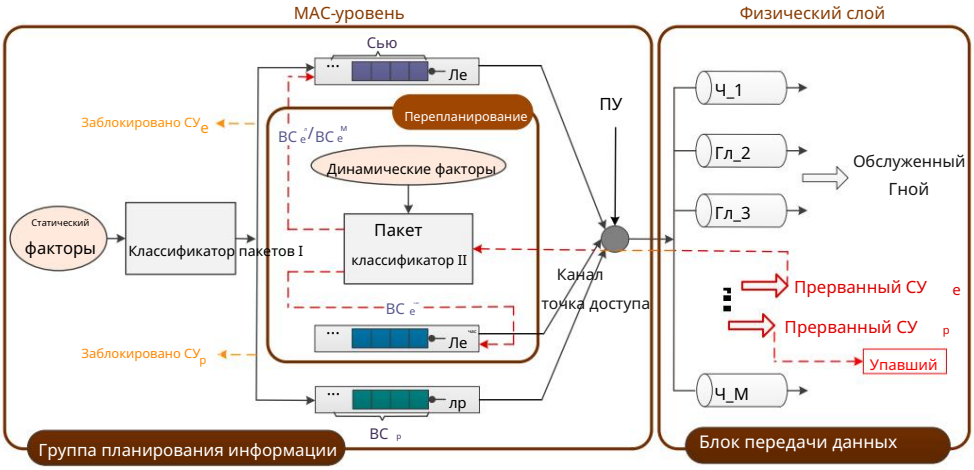


Рисунок 2. Стратегия Ps-rq.

2.4. Распределение каналов на основе стратегии Ps-rq

В процессе планирования применяется технология объединения каналов для повышения скорости обслуживания SU путем объединения одного или нескольких каналов в SU. При этом используется технология адаптации спектра, позволяющая адаптивно распределять собранные каналы. Для CRN с услугами нескольких типов распределение каналов на основе стратегии Ps-rq выглядит следующим образом: SUr и отмеченный SUh объединяются с фиксированным количеством каналов для обеспечения эффективности передачи . Точнее, можно запустить службу SUr . Соответственно, SUh путем каналов , где a Z , может запустить услугу путем агрегирования каналов, где b Z . Отличие в том, что SUe может гибко регулировать количество собираемых каналов на основе технологии адаптации спектра. Быть *. В частности, обслуживание SUe можно запустить путем агрегирования w каналов, а скорость обслуживания можно повысить за счет увеличения агрегированных каналов до верхней границы v, где w, v Z и w v. Для удобства запуска обслуживания SUr и отмеченного SUh, настройка параметров в этом исследовании основана на a = b < w. Предположим, что скорость обслуживания СУ при внедрении технологии сборки каналов возрастает линейно, скорость обслуживания СУ при сборке каналов увеличивается в раз .

Каналы распределяются по SU на основе пропорционального распределения каналов с приоритетом. В соответствии с приоритетом очереди L > Lr > Le , если имеется N свободных свободных каналов, определите , пусты ли три очереди. Если ни один из них не пуст, свободный канал выделяется в очереди для SUh , SUr и SUe в пропорции 1 : α : β, где α, β 1. Здесь значение α , β можно определить по формуле требованиям приоритета и допуска задержек SUh , SUr и SUe . Затем оставшиеся каналы выделяются SUh в соответствии с приоритетом, за которым следуют SUr и SUr .

Если только две очереди не пусты, процесс распределения аналогичен описанному выше. Если хотя бы один не пуст, в эту очередь выделяются все свободные каналы. Более того, если еще есть свободные каналы, оставшиеся каналы выделяются текущему SUe . Вышеупомянутый процесс распределения реализован в MATLAB, а пропорциональное распределение каналов с приоритетом показано на схеме 1. Здесь количество свободных каналов по-прежнему определяется устройствами SU, получающими доступную информацию о свободных каналах с помощью технологии обнаружения энергии спектра для достижения оппортунистического доступа.

Machine Translated by Google

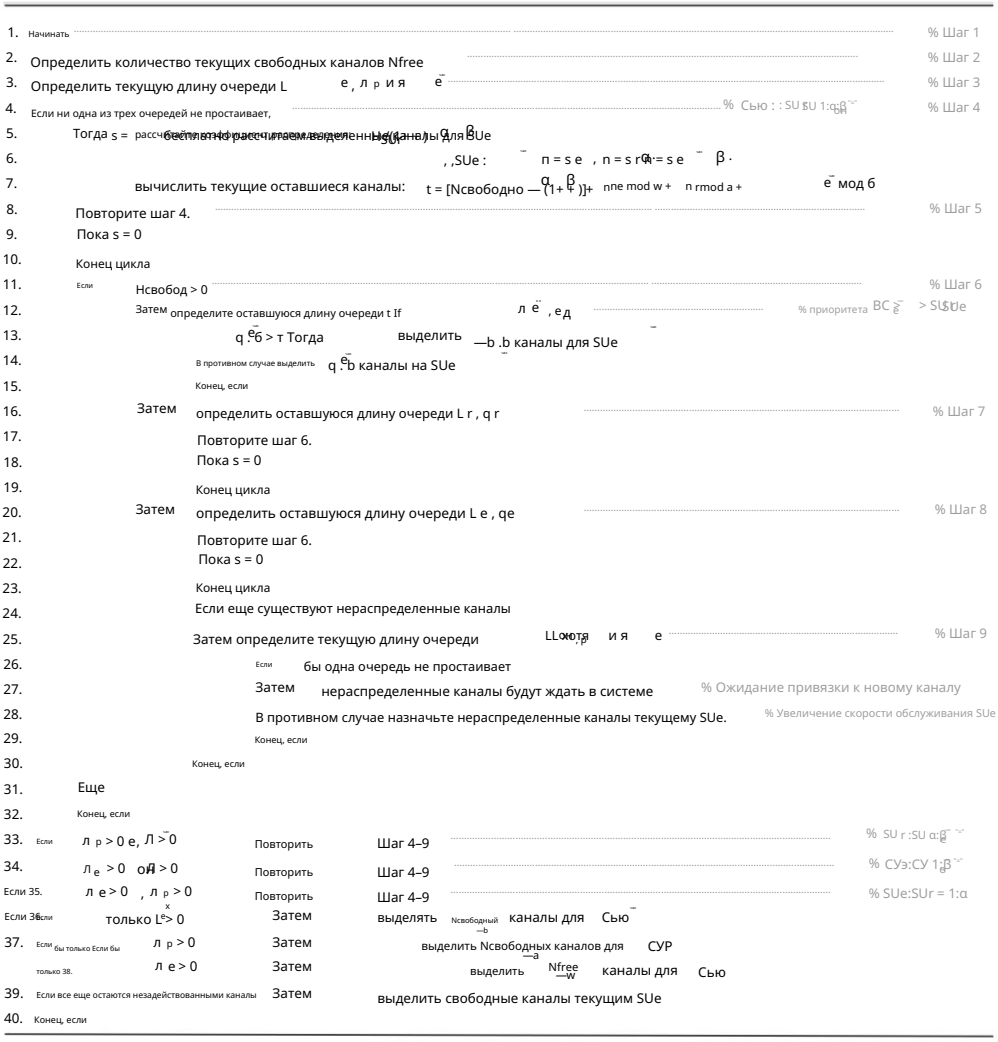


Схема 1. Пропорциональное распределение каналов с приоритетом.

3. Процесс динамического доступа к каналу

В CRN поведение прибытия и ухода PU и SU приводит к изменению состояние системы. Для анализа процесса движения ресурсов SU в различных состояния системы, процесс динамического доступа к каналу, основанный на предложенной стратегии Ps-rq, В этом разделе исследовано путем классификации действий пользователей в CRN на четыре события: Прибытие PU, вылет PU, прибытие SU и отправление SU.

• Событие A: прибытие ПУ

SU в CRN классифицируются на эластичные услуги и услуги реального времени в соответствии с чувствительность к задержке. Эластичные услуги имеют характеристики короткой длины рамы, тяжелые поток трафика и высокая устойчивость к задержкам. Сервисы реального времени требуют своевременности данных. передача инфекции. В зависимости от допуска задержки приоритет услуги определяется следующим образом:

При поступлении ПУ, если в системе есть свободный канал, он будет выделен вновь прибывшая ПУ. При этом услуги других SU не пострадают. Но если в системе нет свободных каналов, ПУ прервет любой текущий канал SU из-за более высокого приоритета доступа к данным. Прерванное СУ выйдет из канала и выполнить адаптацию спектра. То есть действующее СУэ с максимальным количеством каналов будет передавать канал прерванному SU до тех пор, пока не будут собраны каналы SUe не может поддерживать услугу, тогда действующее SUe со вторым максимальным номером каналов пожертвуют свой канал . . . до тех пор, пока все SUe не смогут обеспечить необходимое количество каналов прерываемого СУ. Тогда адаптация не удалась, а это значит, что все

Machine Translated by Google

текущий SU не может передать канал для прерванного SU. SUe , который провалил
адаптация войдет в классификатор пакетов II и отметит приоритет через нечеткий контроллер. Если
SUe принадлежит SUh e , он войдет в зарезервированную очередь Le ; в противном случае он вернется обратно в
конец очереди Ле . Если очередь заполнена, обслуживание SU будет вынуждено прекратиться.

Если PU вытесняет канал SUr или отмеченного SUh e , адаптация спектра будет
выполняться таким же образом. Услуга продолжится после успешной адаптации;
в противном случае прерванный SUr или SUh e будет вынужден прекратить и освободить собранный
канал.

В частности, каналы, освобожденные принудительно ликвидированным СУх, или СУр не может быть
свободный канал может использоваться для текущих, чьи собранные каналы не использовались для
запуска обслуживания SU в очереди из-за установки a = b < w. Однако
достичь верхней границы с помощью технологии сборки каналов.
каналы, освобожденные принудительно завершенным SUe, могут использоваться для организации очереди SUh e и СУР
В частности, свободный канал предпочтительно выделяется для подчиняясь и Lr e . Распределение каналов SU в очереди происходит по схеме 1. Если в очереди нет SU,
правило обслуживания в порядке очереди (FCFS). Служба может быть запущена только тогда, когда —
очередь использует свободный канал, эти каналы будут объединены с другими SU, подчиняясь Правилу 1.
В системе существуют бесплатные каналы. Если в очереди не используется свободный канал, минимально
Правило 1: Свободный канал будет предпочтительно присоединен к текущему SUe с
агрегируются каналы до тех пор, пока SUe не будет собрано до верхней границы, v. Если есть
канал можно собрать только если каналы еще свободны, остальные каналы будут подключены к текущему SUe с
в и обслуживание можно запустить
когда — в системе существуют свободные каналы. Если оба в очереди не связаны, v.
второй минимальный агрегированный канал до тех пор, пока все текущие SUe не достигнут верхнего уровня
используйте свободный канал, канал можно собрать в . Аналогично, услуга Мы представляем блок-схему, как показано на рисунке
3, чтобы показать динамический доступ к каналу.
из может быть запущен только тогда, когда в системе уже существуют свободные каналы. Если
нет процесса, основанного на стратегии Ps-rq по прибытии PU аналогично и другие мероприятия, включая ПУ
SU в очереди использует свободный канал, канал будет собран в текущий процесс доступа SU.
отправление, прибытие SU и отправление SU также могут использовать блок-схему для описания канала
в системе, подчиняющейся Правилу 1.

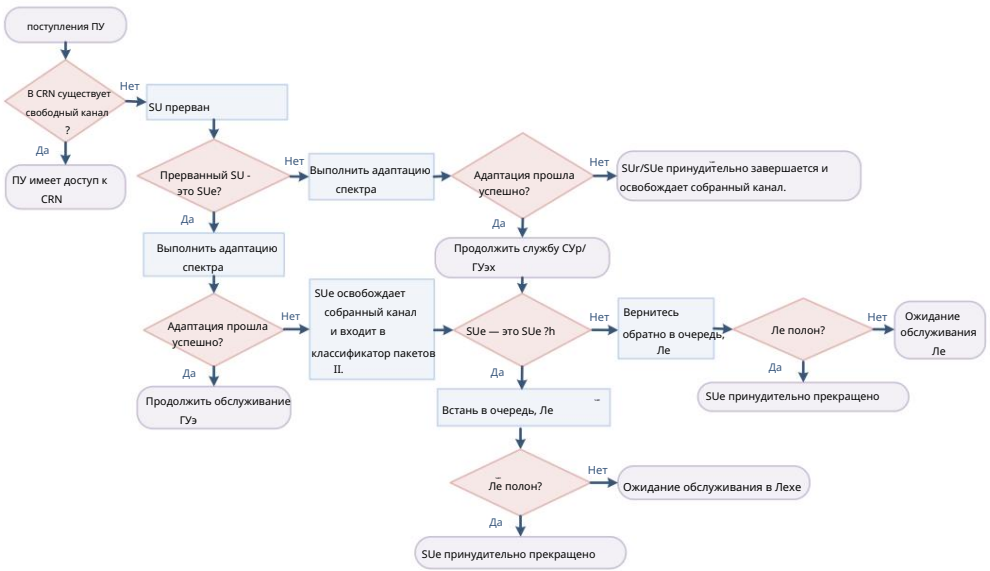


Рисунок 3. Процесс динамического доступа к каналу на основе стратегии Ps-rq по прибытии PU.

Рисунок 3. Процесс динамического доступа к каналу на основе стратегии Ps-rq по прибытии PU.

• Событие C: прибытие SU

Событие B: отъезд ПУ

В системе есть два вида прихода СУ, а точнее, Уход ПУ освобождает канал. .
Если в систему поступает SU и в системе имеется b свободных каналов, то свободный канал может
используется для SU, ожидающих в очереди. Если SU в очереди не используют свободный канал,
быть использован для текущего SU , собранные каналы которого не достигли цели.
затем служба вновь прибывшего обслуживание верхней границы с технологией сборки каналов.
можно начинать. Аналогично осуществляется
В частности, свободный канал предпочтительно выделяется SUh e в Л e , подчиняясь правилу
вновь прибывшие могут быть запущены, когда количество каналов, существующих в
обслуживания в порядке очереди (FCFS). Служба СУ e можно запустить только тогда, когда b 1 свободен
система достигает минимального значения w или a . Если свободные каналы в системе не
могут существовать в системе. Если СУ в очереди не используйте свободный канал, канал
поддерживать услугу текущая будет выполнять адаптацию спектра для пожертвования, которую
можно собрать в SUr в Lr , и услугу SUr можно запустить только тогда, когда - 1 свободно
канал для вновь прибывших SU. Если все текущие не могут предоставить SUr в очереди , не используйте
каналы существуют в системе. Если оба СУ e свободный канал,
канал можно собрать до SUe в Le . Аналогично, службу SUe можно запустить только
необходимого количества каналов, вновь поступившее СУ будет помещено в конец соответствующей
очереди. Если очередь заполнена, обслуживание SU будет заблокировано. Однако, поскольку все время
заполнено, вход в очередь происходит из прерванного ,
обслуживание будет вынужден завершиться, то есть заблокированных .

• Событие D: вылет SU

Machine Translated by Google

когда в системе уже существует w - 1 свободный канал. Если ни один SU в очереди не использует свободный канал, этот канал будет объединен с текущими SU в системе, подчиняясь Правилу 1.

• Событие C: прибытие SU.

В системе существует три типа прибытия SU, включая SUh e , SUe и SUr . SU, пришедшее в систему, — Если это SUh и в системе имеется b свободных каналов, то можно запустить обслуживание вновь прибывшего SUh . Аналогично, обслуживание вновь прибывших SUe или SUr может быть запущено, когда количество существующих в системе каналов достигнет минимального значения w или a. Если свободные каналы в системе не могут поддерживать услугу, текущее SU выполнит адаптацию спектра, чтобы передать канал вновь прибывшим SU. Если все текущие SU не могут обеспечить необходимое количество каналов, вновь прибывшее SU будет помещено в конец соответствующей очереди. Если очередь заполнена, обслуживание SU будет заблокировано. Однако, поскольку все SUh e , входящие в очередь, заполнены, служба SUh будет вынуждена

Л час e h приходят из прерванного SUe , когда L e e завершить, то есть заблокированного СУ нет e .

• Событие D: вылет SU

Вылет СУ освобождает все агрегированные каналы. То есть уход SUe освобождает k (w k v) каналов; уход СУР освобождает каналы ; и уход СУх e освобождает b- каналы. Для распределения каналов, освобождаемых SU, используется схема 1 , где – сумма количества каналов, освобождаемых SU, и количества неиспользуемых Nсвободных каналов в системе. Если после того, как все SU, ожидающие в очереди, все еще остаются свободные каналы , свободный канал будет выделен текущему SUe , подчиняясь Правилу 1.

4. Анализ CTMC и меры QoS

В процессе доступа к каналу действия пользователя вызывают изменение состояния системы. Цепи Маркова с непрерывным временем (CTMC) используются для установки соответствующей связи между возможным переходом состояния системы и запускающими событиями, вызывающими изменение состояния системы, что может отображать поток ресурсов спектра между многотипными SU. Таким образом, анализ производительности вторичной системы можно выполнить с помощью CTMC. Деятельность SU можно отобразить в многомерном вероятностном пространстве. Затем можно создать математическую модель предлагаемой стратегии Ps-rq для получения показателей QoS посредством стационарного решения и оценить производительность вторичной системы.

4.1. Планирование динамического анализа и моделирование CTMC Обратите внимание, что общее состояние выглядит следующим образом:

$$x = lw, \ldots, lv, ga, gb, lpu, leq, lrq, lhq$$

где lm (w m v) – количество СУэ с m агрегированными каналами, ga – количество СУр с агрегированными каналами, gb – количество СУх с b агрегированными каналами, lpu – количество ПУ в системе и leq, lrq и lhq — текущие длины очереди Le , Lr и L e соответственно.

час

Далее множество возможных состояний системы можно обозначить как

$$S = lw, \ldots, lv, ga, gb, lpu, leq, lrq, lhq \quad 0; б(x) \quad M; лек \quad Qэкв; lrq \quad Qrq; lhq \quad Qhq; \sum_{я=1}^{v\ w} ilw+i < w, \text{если } leq > 0 \tag{4}$$

Здесь Qeq, Qrq и Qhq — полные емкости Le , Lr и L e соответственно. если leq > 0, это означает, что если Le не пусто, то час $\sum_{я=1}^{v\ w} ilw+i < ш,$

суммирование максимального числа каналов

Machine Translated by Google

Сумма, которую все текущие SUe могут пожертвовать, всегда будет меньше минимального агрегированного числа. b(x) — общее количество используемых каналов в состоянии x , которое можно описать следующим образом:

$$b(x) = lpu + aga + bgb + \sum_{m=w\ млм}^{в} \quad (5)$$

Кроме того, M b(x) — количество свободных каналов в системе (т. е. Nfree, полагаемое в схеме 1).

Чтобы проанализировать скорость перехода из общего состояния x в другие состояния, мы предполагаем, что приход пользователя в систему подчиняется распределению Пуассона, а уход — экспоненциальному распределению. Обратите внимание, что скорость поступления PU равна λp , а скорость отправления — μp. Скорости поступления SUr и SUe равны λsr и λse, а скорости отправления — μsr и μse соответственно. Более того, планирование с циклом обратной связи SUe может привести к тому, что внешнее поступление услуг, входящих в Le и L , больше не будет подчиняться распределению Пуассона. Чтобы упростить наш анализ, мы предполагаем, что настройка контура обратной связи мало влияет на услуги в Le и L , и предположим, что экзогенное поступление SU все службы подчиняются распределению Пуассона.

На основании вышеизложенных предположений путем анализа процесса доступа к каналу устанавливается модель CTMC, основанная на стратегии Ps-rq. Разрабатываются все возможные целевые состояния, виды деятельности, скорости перехода и условия перехода системы, начиная с общего состояния x .

• Событие A: прибытие ПУ

По прибытии PU переходы состояний вторичной системы, начиная с общего состояния x, до всех возможных триггерных событий, выводятся, как показано на рисунке 4, а условия перехода для всех возможных состояний перечислены на рисунке 5. За исключением состояния t1 на рисунке Как видно на рисунке , в системе нет свободных каналов, которые можно было бы напрямую выделить прибывшим ПУ. Таким образом, существует M b(x) = 0 в состояниях от t2 до t35.

Прибытие PU может опережать обслуживание любых SU. Для состояния t4 в состояние t14 и из состояния t16 в состояние t17 прерывания обслуживания на рисунке 5 это указывает на то, что устройства SU в этих состояниях не могут выполнять адаптацию спектра, т. е. m (w < m v), lm = 0. Поскольку Lr не имеет петли обратной связи, обслуживание прерванного SUr будет принудительно прекращено, например, из состояния t30 в состояние t32.

В состоянии t3 и состоянии t15 могут быть два случая, когда SU не выполняют адаптацию спектра: в системе нет вторичного пользователя, который может выполнить адаптацию, то есть m (w < m v), lm = 0, либо СУ в системе не могут осуществить адаптацию спектра и m (h = m = v), lm > 0. Аналогичные условия должны выполняться для состояний t18, t30 и t33.

Кроме того, поскольку a = b < w, a = b w 1. То есть каналы, освобождаемые входом прерванное SUe не может запустить обслуживание SUe в Le , но обслуживание SUr и SUh h Lr и L освобождается принудительно e завершенным$_e$ может быть запущен, например, из состояния t6 в состояние t14. Это то же самое, что и каналы, SUe , каналы освобождаются прерванным или завершенным SUr например, в состоянии от t21 до состояния t29. Тем не и SUh запускают любое обслуживание SU в очереди. • Событие B: отъезд ПУ e менее, они не могут поддерживать

При выходе ПУ переходы состояний вторичной системы, начиная с общего состояния x до всех возможных триггерных событий, выведены на рисунке 6, а условия перехода для всех возможных состояний приведены на рисунке 7.

В этом случае, если очередь L не пусто и b = 1, то сервис SUh e в очереди можно запустить напрямую. Если b > 1, то требуется M b(x) = b 1, то есть в системе существует b 1 каналов, например, в состоянии t1. Это то же самое, что и SUr, например, в состоянии t2. Поскольку a = b < w, т. е. w/ = 1, обслуживание SUe в Le можно начать только при условии M b(x) = w 1, например, в состоянии t3.

Machine Translated by Google

Рис. 4. Переход системы из общего состояния x во все достижимые состояния по прибытии ПУ.

Machine Translated by Google

Состояние	Условия перехода	Состояние	Условия перехода
т1	M xb > 0)(	т18	M xb = ;0)(> ;0 eq Qeq l = ; или (lvm = wm), м > 0 (lvm wm), м "=" 0 ;
т2	M xb = ;0)(k > ;0 k = max{ }w+1,...,v ;	т19	M xb = ;0)(лв > ;1 экв Qeq l = ; m= 2w v;1 = ;0 (wm <),lvm м = ;0
т3	M xb = ;0)(lw > ;0 экв Qeq l <; или (), wm lvm м 0 ( ), wm lvm м 0 ;	.....	
т4	M xb = ;0)(л ш > ;1 экв Qeq l < ; m = 2w v;1 л rq = lhq = ;0 wm lvm ( ), = ;0 м	т20	M xb = ;0)(лв > ;1 л экв Кэк; ЛК ш ;1 я буду ;0 rq hq w ( )(lwv ш + ;1)1 = ;wm lvm ;(<) м
.....		т21	M xb = ;0)(л экв Кэк; л квартира > 0 итого w b;1 вм<лвм), = ;0 м
т5	M xb = ;0)(лв > ;1 экв Qeq l <; = лк ш ;1 1 << vw ; Ж лвв)(+ ш ;1)1 (wm <),lvm м = = ;0 rq hq 0	.....	
т6	M xb = ;0)(экв Qeq l <; л квартира > 0 итого w b;1 (wm <),lvm м = ;0	т22	M xb = ;0)(л экв Кэк; л > ;0 rq w a;1 = ;0 l (wm <),lvm м = ;0
.....		.....	
т7	M xb = ;0)(экв Qeq l <; л rq > ;0 w a;1 lhq = ;0 (wm <),lvm м = ;0	т23	M xb = ;0)(л ш > ;0 л экв Кэк; л rq > ;0 л квартира > 0 итого ш 1 + ба ; вм(лвм), = ;0 м
.....		.....	
т8	M xb = ;0)(лв > ;0 экв Qeq l <; л rq > ;0 lhq > ;0 ш 1 + ба ;	т24	M xb = ;0)(экв Qeq l =; лхк > ;0 лв > ;1 1 << vw ; wb 1 +<+ = < abbwk ; вм<лвм), = ;0 м
.....		.....	
т9	M xb = ;0)(экв Qeq l <; лхк > ;0 лв > ;1 1 << vw ; wb 1 +<+ = < abbwk ; вм лвм), = ;0 м	т25	M xb = ;0)(экв Qeq l =; лхк > ;0 wb = 1 (<)lwvb ш ;1 lq = lw ш > ;1 1 (wm <),lvm м = ;0 1< < vw ;
.....		т26	M xb = ;0)(экв Qeq l =; л rq > ;0 lw > ;1 wb 1 + = < awk ; лхк = ;0 (wm <),lvm м = ;0 1 << vw ;
т10	M xb = ;0)(экв Qeq l <; л _____ > ;0 вб лвввб = <) ;1 ш = лк ш ;1 lw > ;1 (wm <),lvm м = ;0 1 << vw ;	.....	
т11	M xb = ;0)(экв Qeq l <; л rq > ;0 lw > ;1 wb 1 + = < awk ; lhq = ;0 1 << vw ; (wm <),lvm м = ;0	т27	xbM = ;0)(экв Qeq l =; л rq > ;0 ;0 1 (+= < ;1) wb lwva hq = лв > ;1 lq = ш ;1 (<),lvmwm м = ;0 1 << vw ;
.....		т28	xbM = ;0)(; экв=Ql экв л ш > ;1 л rq > ;0 л квартира > 0 итого ш 1 --++=- wkba ; (<),lvmwm м = ;0 1 << vw ;
т12	M xb = ;0)(экв Qeq l <; л rq > ;0 wb = 1 (<)lwva ш ;1 лхк = ;0 lq = ш ;1 лв > ;1 (wm <),lvm м = ;0 1 << vw ;	.....	
т13	M xb = ;0)(экв Qeq l <; л ш > ;1 л rq > ;0 л _____ > ;0 ш 1 --++=- wkba ; 1 << vw ; вм лвм (<), = ;0 м	т29	xbM = ;0)(экв Qeq l =; lhq > ;0 л q > ;0 lq lw = ш ;1 w 1 (=)(lwvba ш 1); lvm(wm), = ;0 1 << vw ;
.....		т30	xbM = ;0)(га > ;0 или (), лвмвм м 0 ( ), лвмвм м 0 ;
т14	M xb = ;0)(экв Qeq l <; л > ;0 л > ;0 lq = ш ;1 л ш > ;1 w (л++=)(lwvba ш 1); 1 << vw ; л(лвмвм), = ;0 м	т31	xbM = ;0)(г а > ;0 ak = ;1 1 << vkw ; > = 0 ш лвмвм), = ;0 м
.....		.....	
т15	M xb = ;0)(gb > ;0 hq Qhq l < ; или (), wm lvm м 0 ( ), wm lvm м 0 ;	т32	xbM = ;0)(га > ;0 = лк ш ; а (= ;) лвв ш лв > ;0 (<),lvmwm м = ;0 1 << vw ;
т16	M xb = ;0)(gb > ;0 л < итого квартира ; bk = ;1 1 << vkw ; лв > ;0 (wm <),lvm м = ;0	т33	xbM = ;0)(gb > ;0 hq Qhq l =; или (), лвмвм м 0 лвмвм < = м 0 ;
.....		т34	xbM = ;0)(> ;0 b g hq Qhq l =; bk = ;1 1 << vkw ; л ш > ;0 (<),lvmwm м = ;0
т17	M xb = ;0)(gb > ;0 hq Qhq l <; = лк ш ; б (= ;) лвв ш 1 << vw ; л ш > ;0 (wm <),lvm м = ;0	.....	
		т35	xbM = ;0)(gb > ;0 л итого квартира ; = лк ш ; б (= ;) лвв ш лв > ;0 (<),lvmwm м = ;0 1 << vw ;

Рис. 5. Условия перехода системы из общего состояния x во все достижимые состояния на ПУ прибытие.

Machine Translated by Google

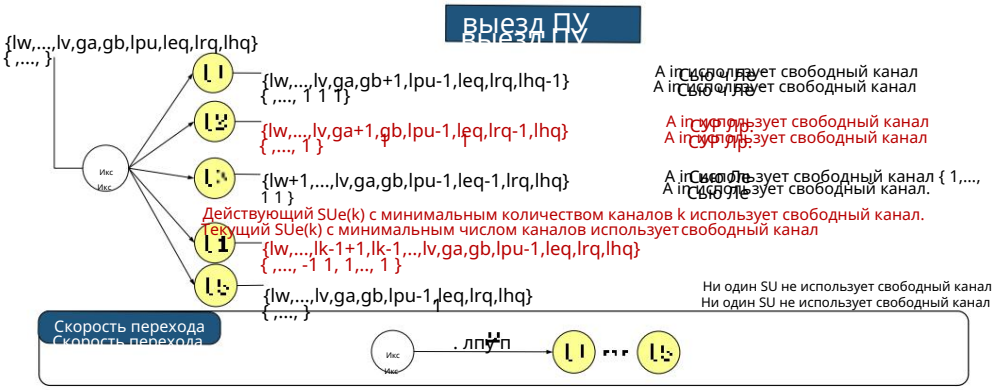


Рис. 6. Переход системы из общего состояния x во все достижимые состояния при выходе ПУ.

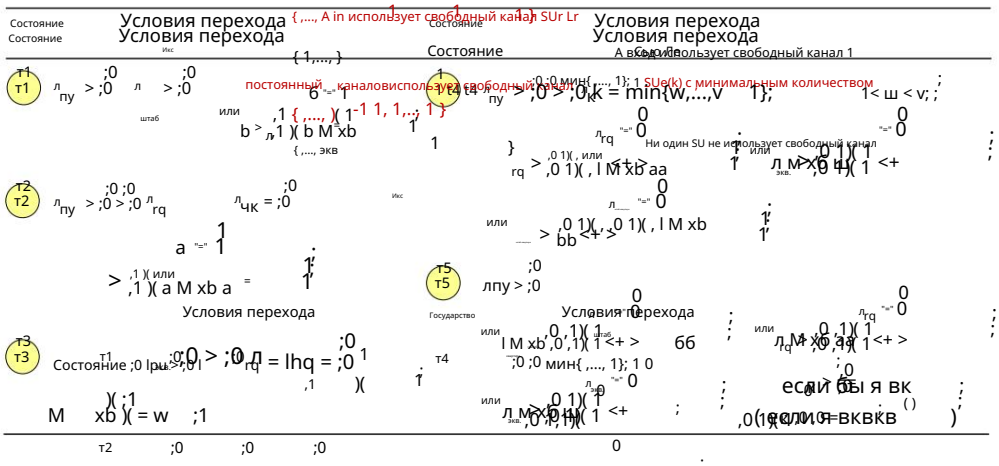


Рис. 7. Условия перехода системы из общего состояния x во все достижимые состояния на ПУ выезд.

• Событие C: прибытие SU.

По прибытии СУ происходят переходы состояний вторичной системы, начиная с общее состояние x для всех возможных событий запуска определяется, как показано на рисунке 8, а условия перехода для всех возможных состояний приведены на рисунке 9.

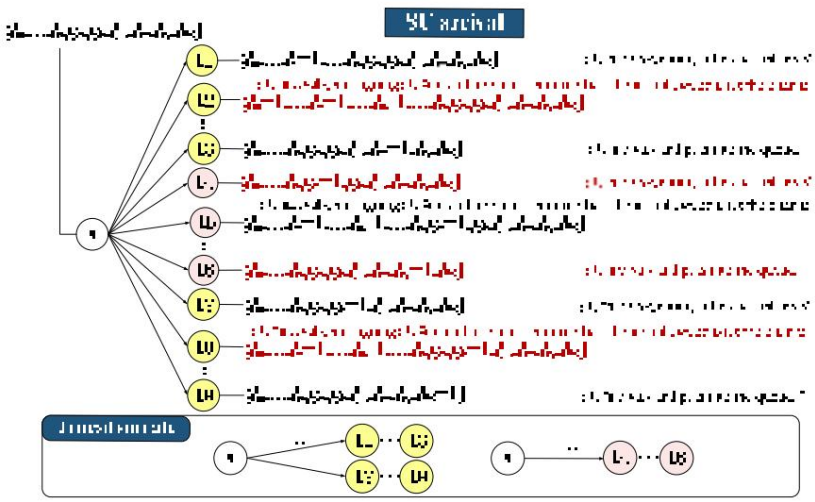


Рис. 8. Переход системы из общего состояния x во все достижимые состояния по прибытии SU.

Machine Translated by Google

Состояние	Условия перехода	Состояние	Условия перехода
т1	M xb)(> ш;	т6	rq Qrq l <; если Mvw xb( ) < ( )м лв ам ; если w Mvm xb)(a
т2	lm > ;0 m = max{w +1,..., v}; < вв ; mh + M xb)(= w;	т7	M xb)(> b;
.....			
т3	экв Qeq l <; Mvw xb)(если w ( ) м < ; если w Mvm xb)(w	т8	lm > ;0 m = max{w +1,..., v}; ш < v; mh + M xb = b;)(
т4	M xb)(> a;	.....	
т5	lm > ;0 m = max{w+1,...,v}; ш < v; mh + M ;)(xb)(= a;	т9	hq Qhq l < если Mvw xb ;)(+ ( )м лев бм < ; если w Mvm xb)(b <
.....			

Рис. 9. Условия перехода системы из общего состояния x во все достижимые состояния на SU.

Условие обслуживания SUe может быть запущено, когда его поступление равно сумме количество свободных каналов, существующих в системе, и каналов, которые могут быть освобождены действующими SU, не меньше нижней границы SUe, т. е. M b(x) + (m w)lm w. В противном случае он будет помещен в очередь Le при условии, что leq < Qeq, например, в состоянии t3. СУР и СУ приходят таким же образом, т.е. в состояние t6 и состояние t9.

• Событие D: вылет SU

При выходе из СУ происходят переходы состояний вторичной системы, начиная с общее состояние x для всех возможных событий запуска определяется, как показано на рисунке 10, а условия перехода для всех возможных состояний приведены на рисунке 11.

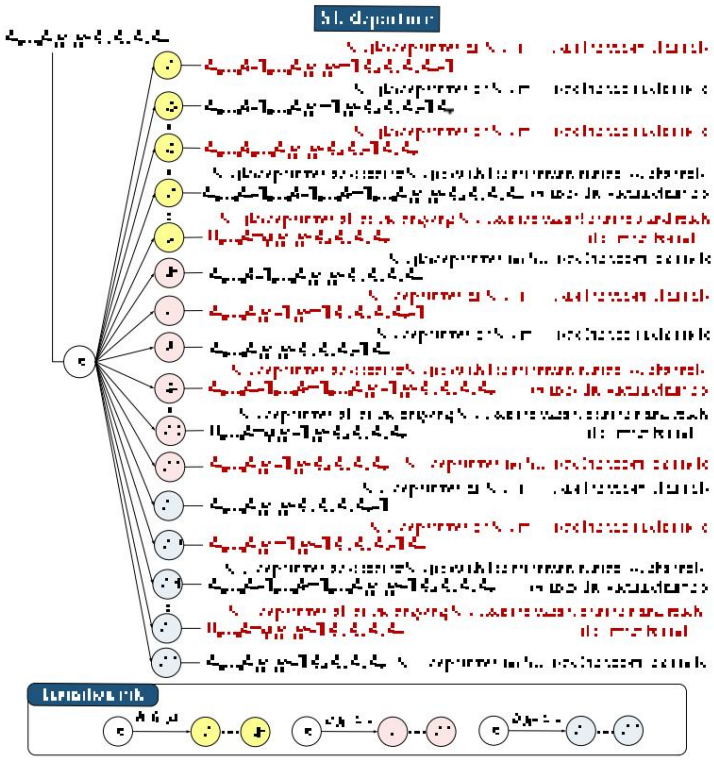


Рис. 10. Переход системы из общего состояния x во все достижимые состояния при выходе СУ.

Machine Translated by Google

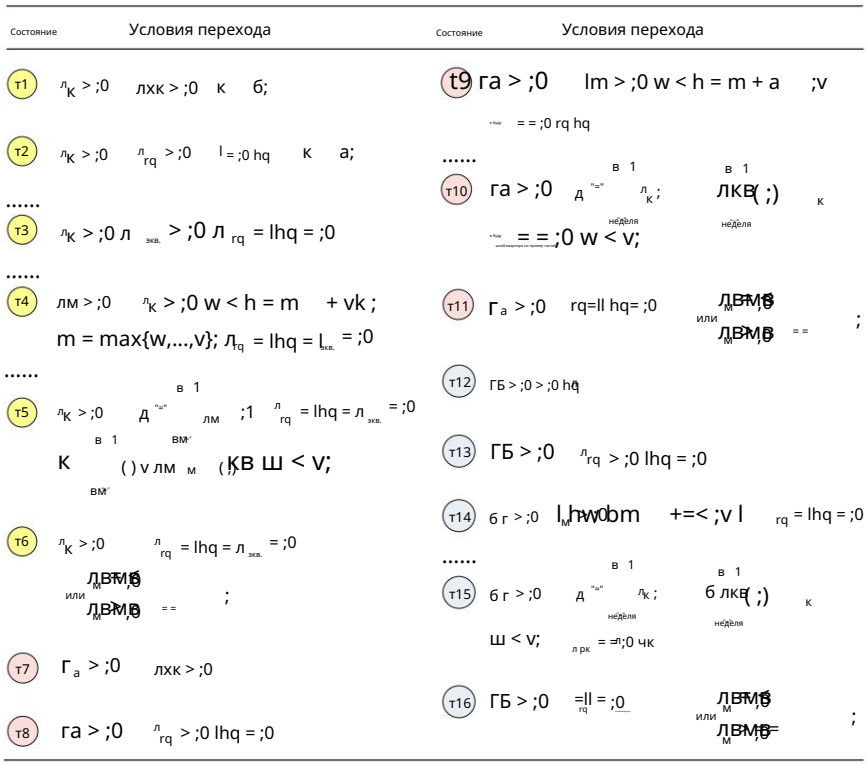


Рис. 11. Условия перехода системы из общего состояния x во все достижимые состояния на выезд СУ.

При этом общее количество свободных каналов, используемых действующими SU, не должно быть больше, чем количество каналов, выпущенных уходящими SU, например, условие k m=1w (v m)lm (v k) в состоянии t5. Это похоже на состояние t10 и t15.

4.2. Метрики качества обслуживания

Благодаря моделированию CTMC любая возможная активность пользователя может вызвать переход от любое достижимое состояние ti в другое достижимое состояние tj , как показано на рисунках 4–11. qij (i, j S и i/ = j) представляют собой общую скорость перехода, где S — набор допустимых состояний системы. Пусть π(x) — стационарная вероятность системы, которую можно вычислить по глобальным уравнения баланса и условие нормировки следующим образом.

$$\pi(x)Q = 0, \quad x\ \pi(x) = 1 \quad (6)$$

где

π(x) — вектор стационарной вероятности, а Q — матрица скорости перехода, созданная скоростью перехода qij. Диагональные элементы матрицы скорости перехода Q, qii могут найти как qii = qij. Таким образом, метрики QoS вторичной системы, основанные на предложенную стратегию Ps-rq можно получить.

• Пропускная способность сети означает среднее количество услуг SU, выполненных в CRN. за единицу времени (единица: пакет/единица времени) [12]. Пусть Ce , Cr — мощности SUe , SUr , соответственно, мы можем получить:

$$Kp = \sum_{x\varepsilon S} a\gamma a\mu sr\pi(x), \quad (7)$$

$$Ce = \sum_{x\varepsilon S} \sum_{k=sh}^{в} (klk + bgb)\mu se\pi(x) \quad (8)$$

где x S — множество всех возможных состояний систем, изображенных на рисунках 4–11.

Machine Translated by Google

• Использование спектра означает отношение среднего количества используемых каналов к общему количеству доступных каналов [25], которое можно определить по формуле:

$$U = \sum_{x \varepsilon S} \pi(x) \frac{б(x)}{M} \quad (9)$$

где b(x) — общее количество доступных каналов состояния x, M — общее количество каналов в системе, а S — набор всех возможных переходных состояний, представленных на рисунках 4–11.

• Вероятность блокировки относится к вероятности того, что вновь прибывший SU не сможет быть обслужен из-за того, что все каналы в системе заняты. Поскольку в этом исследовании для SU установлена очередь ожидания , обслуживание вновь прибывшего SU будет заблокировано, когда соответствующая очередь будет заполнена. Следует отметить, что SUh h – это входящие в L SUe прерванные PU, а не вновь прибывшие в систему SU.

Следовательно, у СУ нет случая блокировки.

Вновь поступившее SUe будет заблокировано, если выполняются оба следующих условия: (1) сумма общего количества свободных каналов в системе и общего количества каналов , которые могут быть предоставлены текущим SUe , недостаточна для запуска. обслуживание SUe (т.е. нижняя граница сборки каналов не достигнута) и (2) очередь Le заполнена. Таким образом, вероятность блокировки SUe , P b ,

можно получить:

$$П_{б}^{e} = \sum_{Ae} \pi(x) \quad (10)$$

где заданное Ae — это условие срабатывания, при котором SUe блокируется:

$$Ae = x \quad S, leq = Qeq, M \quad b(x) + \sum_{к=ш+1}^{в} (k \quad w) lk < w$$

Аналогично, вновь прибывший SUr будет заблокирован, если выполняются оба следующих условия : (1) сумма общего количества свободных каналов в системе и общего количества каналов, которые могут быть предоставлены текущим SUr, недостаточна. для запуска службы SUr и (2) очередь Lr заполнена. Тогда вероятность блокировки SUr ,P может быть определена по формуле:

$$П_{б}^{p} = \sum_{Ap} \pi(x) \quad (11)$$

где заданное Ae — это условие срабатывания, при котором SUe блокируется:

$$Ar = x \quad S, lrq = Qrq, M \quad b(x) + \sum_{к=ш+1}^{в} (k \quad w) lk < a$$

• Вероятность принудительного завершения означает вероятность того, что обслуживаемые SU будут вынуждены завершить работу из-за прибытия PU. В качестве набора петли обратной связи SUe SUe , прерванный PU, может вернуться в очередь Le и L e через классификатор пакетов II. Если соответствующая очередь заполнена, служба будет вынуждена завершить работу.

Соответственно, обслуживание SUh вынуждено прекращаться при выполнении следующих трех условий: (1) в системе нет свободных каналов; (2) текущий SUe не может выполнить адаптацию спектра для совместного использования каналов с прерванными SU и (3) очередь h заполнена. Точнее, состояния t33 – t35 на рисунке 4 указывают, что PU вытесняет любые SUh L

e , e

с b агрегированными каналами, а СУх e вынужден прекратить действие. Принудительное прекращение

Machine Translated by Google

вероятность $П_{ф}^{eh}$, можно рассчитать по соотношению среднего принудительного завершения вероятности SUh к скорости поступления SUh e , λse, следующим образом:

$$П_{ф}^{ага} = \pi(x)_{Б_{е}^{час}} \frac{\lambda pbgb}{(M \quad lpu)\lambda se} \tag{12}$$

ч , где $Б_{е}$ представляет собой набор условий, при которых SUh вынужден прекратить работу:

$$Б_{е}^{час} x \quad S, M = b(x), lpu < M, gb > 0, lhq = Qhq, \text{ или } \quad m(w = \begin{matrix} m(w < m \quad v), lm = 0 = \\ m = v), lm > 0 \end{matrix}$$

Аналогично, услуга SUe принудительно прекращается, когда выполняются следующие три условия: (1) после вытеснения количество оставшихся каналов SUe недостаточно для запуска службы, то есть SUe с w агрегированными каналами упрежденный; (2) текущий SUe не может выполнить адаптацию спектра для совместного использования каналов с прерванными SU и (3) очередь Le заполнена. На этом основании состояния с t18 по t29 на рисунке 4 указывают, что PU вытесняет любое SUe с w собранными каналами, и SUe вынуждено завершить работу. Вероятность принудительного завершения SUe , P может быть получена по формуле: f ,

$$П_{ф}^{e} = \pi(x)_{Быть} \frac{\lambda pwlw}{(M \quad lpu)\Lambda e} \tag{13}$$

где Λe представляет собой скорость прибытия SUe , которая не заблокирована: Λe = λse(1 P $_{б}^{e}$) и Be — набор условий, при которых SUe вынужден завершить работу:

$$Be = x \quad S, M = b(x), lpu < M, lw > 0, leq = Qeq \text{ или } \quad m(w = \begin{matrix} m(w < m \quad v), lm = 0 \\ m = v), lm > 0 \end{matrix}$$

Поскольку для очереди Lr не установлен цикл обратной связи , при выполнении следующих двух условий обслуживание SUr принудительно прекращается: (1) в системе нет свободных каналов для продолжения обслуживания и (2) продолжающееся SUe не может пожертвовать для этого каналы. На основании этого состояния t30 – t32 на рисунке 4 указывают , что PU вытесняет любой SUr с агрегированными каналами, и SUr вынужден завершить работу. Вероятность принудительного завершения SUr может $П_{ф}^{p}$, быть получена следующим образом:

$$П_{ф}^{p} = \pi(x)_{Бр} \frac{\lambda paga}{(M \quad lpu)\Lambda r} \tag{14}$$

где Λr представляет собой скорость поступления незаблокированного SUr : Λr = λsr(1 P — $_{б}^{p}$) и Br набор условий, при которых SUr вынужден завершиться:

$$Br = x \quad S, M = b(x), lpu < M, ga > 0 \text{ или } \quad m(w = \begin{matrix} m(w < m \quad v), lm = 0 \\ m = v), lm > 0 \end{matrix}$$

5. Анализ эффективности предлагаемой стратегии Ps-rq 5.1. Сравнение и анализ характеристик модели.

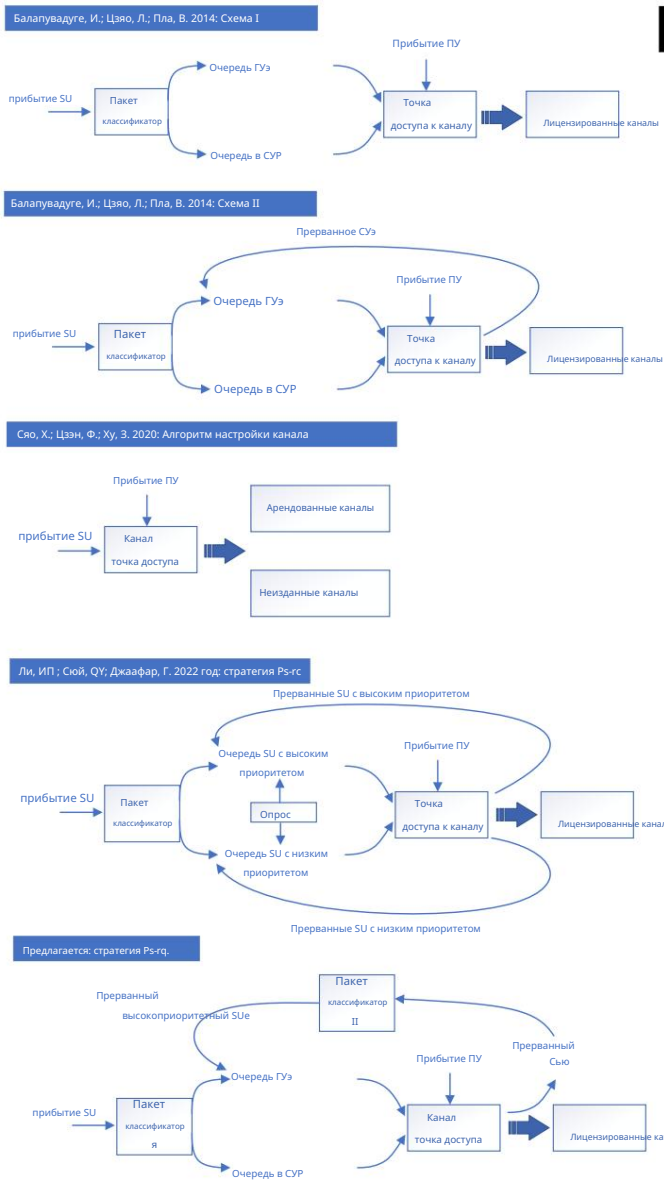
В этом разделе мы сравниваем предложенную стратегию Ps-rq с двумя классическими схемами распределения каналов (ссылка [12], 2014 г.), а также двумя новыми (ссылка [18], 2020 г. и ссылка [9], 2022), с целью проиллюстрировать характеристики предлагаемой стратегии. Сравнение характеристик модели показано на рисунке 12.

Machine Translated by Google

. erormance nayss oe ropose s-rq raegy

5.1. Сравнение и анализ характеристик моделей

В этом разделе мы сравниваем предложенную стратегию Ps-rq с двумя классическими схемами распределения каналов (ссылка [12], 2014 г.), а также двумя новыми (ссылка [18], 2020 г.). и ссылка [9], 2022 г.), с целью проиллюстрировать характеристики предлагаемой стратегии. Сравнение характеристик модели показано на рисунке 12.



Тип трафика	Повышение потенциала	Достоинства	Ограничения
Гетерогенный сервис	/	· Простая схема · Задержка критически важных приложений	· Высокая вероятность принудительного завершения · Применять только тогда, когда PU не активен
Гетерогенный сервис	Обратная связь	· Сокращение случаев принудительного прекращения вероятность · Приложения, устойчивые к задержкам	·Очередь SUe вызывает проблемы с обслуживанием
/	Спектр в лизинг	· Простая схема · Сильная реализуемость	·Пропускная способность сети зависит от масштаба арендуемого спектра. · Гетерогенные услуги не рассматривается
Приоритетное обслуживание	Опрос, петля обратной связи	· Обеспечить заранее определенные высокоприоритетные услуги	·Приоритет должен быть определен во время инициализации сети. ·Трудно адаптироваться к меняющейся сетевой среде.
Гетерогенный сервис	Обратная связь	· Облегчение голодания проблема СУэ · Честный график	· Виды услуг разделены на детали ·Высокая сложность планирования

Рисунок 12. Сравнение характеристик моделей [9,12,18].

Упомянутые исследовательские модели, как показано на рисунке 12, направлены на повышение пропускной способности вторичных сетей. Используя метод случайного анализа и создавая соответствующую цепь Маркова с непрерывным временем для систематического анализа, выводятся метрики производительности для оценки эффективности исследовательских схем. Однако упомянутые исследования столкнулись с проблемами в различении типов услуг, как показано в [9,18], или столкнулись с ограничениями в реализации приложений дифференцированного обслуживания из-за ограничений сетевой среды, например, в [12].

Эти ограничения послужили источником вдохновения для стратегии, изложенной в этой статье. Стратегия Ps-rq, предложенная в этой статье, направлена на обеспечение более справедливой дифференциации службы, решая проблему голода. Для этого мы используем классификаторы пакетов I. и II для выполнения двух уровней дифференциации услуг. Целью классификатора пакетов I является разделить поступающие SU на SUe и SUr . Учитывая, что если SUe прервется ПУ все стоит в очереди на переобслуживание, это неизбежно приведет к проблеме голодания другого очередь. Таким образом, модель, предложенная в этой статье, учитывает требования к обслуживанию частично прерванный высокоприоритетный SUe и включает классификатор пакетов 2 для обеспечения справедливого предоставление услуг при предложении дифференцированных услуг, что представляет собой примечательный аспект

Machine Translated by Google

нашей исследовательской структуры. Однако сложность предложенной стратегии планирования по-прежнему сложно обойти в свете уточнения классификации услуг. Далее мы проведем подробный анализ эффективности исследовательской стратегии Ps-rq, представленной в этой статье.

5.2. Численное моделирование и анализ

Чтобы проиллюстрировать производительность вторичной системы на основе предложенной стратегии сборки каналов, в этом разделе представлены результаты численного моделирования эластичных сервисов и сервисов реального времени на основе MATLAB. Результаты моделирования основаны на процедуре уменьшения размерности многомерных цепей Маркова. Для проверки разработанной модели мы представляем результаты моделирования производительности вторичной системы при различных скоростях поступления первичных пользователей. Основные параметры численного моделирования задаются следующим образом.

Число ПУ в системе задано равным 6 (M = 6), а численный эксперимент проводится в CRN с шестью каналами, разделяющими полосу спектра. Скорость поступления ПУ задается равной λp = 1, а скорость ухода – μp = 0,5. Скорость прибытия и скорость отправления эластичного сервиса задаются как λse = 2 и μse = 1 соответственно, а скорость поступления сервиса реального времени задаются как λsr = 1 и μsr = 1. Обратите внимание, что единицей измерения этих параметров является «сеансов за единицу времени». Присвоив этим параметрам определенные значения, такие как скорость 0,5 сеансов в секунду и длину потока в битах, пропускную способность можно определить количественно в кбит/с или Мбит/с.

Динамические параметры системы a, b, w, v, Qeq, Qrq, Qhq) в процессе доступа к спектру задаются как a = b = 1, w = 2, v = 4, Qeq = 2, Qrq = 1 и Qhq = 1. Коэффициент пропорционального распределения составляет α : β = 1 : 1. На основании приведенных выше настроек системы можно указать общее состояние x системы в разделе 4.1 :

$$(\,l2\,l3\,l4\,g1\,g1\,{}^{3\;2\;1}_{6\;6}\qquad \text{лпу}\,{}^{6\;\;2\;\;\;1\;\;\;1}_{\text{Кек}\;\;\text{КРК}\;\;\text{Кхк})} \tag{15}$$

где li (i = 2, 3, 4) представляет собой SUe , который агрегирует i каналы, а ga, gb представляют собой SUr с агрегированными каналами или SUh с b агрегированными каналами (a = b = 1); верхний индекс каждой позиции указывает верхний предел количества пользователей на соответствующей позиции. 2 позиция, например l3, которая указывает, что существует не более двух SUe, собирающих три 6 каналов в системе одновременно, lpu, что означает, что в системе существует не более шести PU. 2 одновременно и Qeq, что указывает на то, что максимальная емкость очереди Le равна 2.

Численное моделирование основано на параметризованном общем состоянии. Процедура уменьшения размерности многомерных цепей Маркова представляется следующим образом.
Поскольку общее состояние, предполагаемое в этой статье, является многомерным, напрямую сгенерировать матрицу скорости перехода Q сложно. Поэтому мы рассматриваем понижение маркеров многомерности. Теоретически размер матрицы Q составляет 2592 × 2592, что определяется произведением верхних индексов уравнения (15). Возьмите любое общее состояние I : (1 0 0 1 1 1 0 0 0) , существующее в системе, чтобы показать процесс перехода на более раннюю версию следующим образом.
Состояние I : (1 0 0 1 1 1 0 0 0) указывает, что в системе имеется действующий SUe с двумя агрегированными каналами, действующий SUr с одним агрегированным каналом, действующий SUh с одним агрегированным каналом и действующий PU. Следовательно, в этом состоянии в системе существуют пять потоков . Соответственно, состояние I : (1 0 0 1 1 1 0 0 0) можно обозначить как состояние I5, , где первая позиция индекса представляет собой общее количество сосуществующих потоков в системе, а вторая позиция счет идет по порядку, а верхний предел представляет собой сумму возможных сочетаний состояний с пятью потоками в системе. Если в этом состоянии в системе происходит приход пользователя, например приход PU, то соответствующий переход состояния будет (1 0 0 1 1 1 0 0 0) (1 0 0 1 1 2 0 0 0). В этом случае в системе сосуществуют шесть потоков и переходный процесс обозначается как I5, I6, .

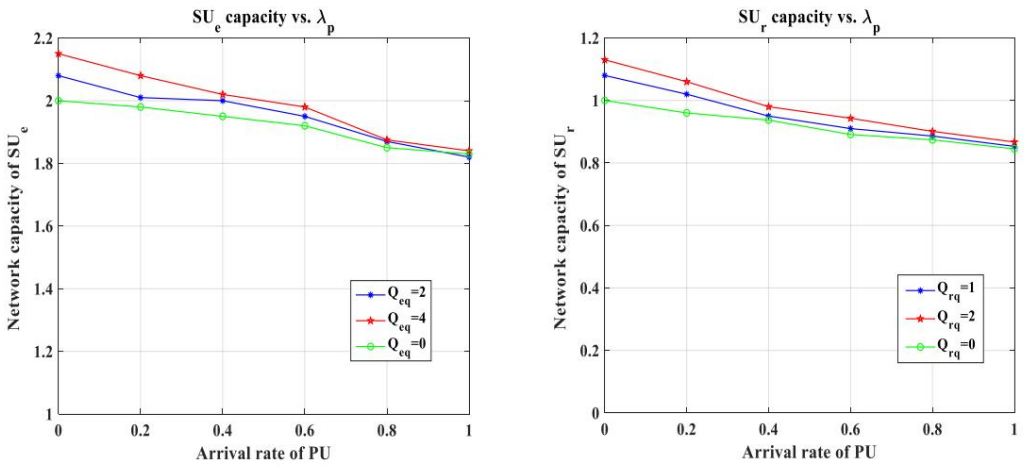
Если уход пользователя произошел в

Machine Translated by Google

система в этом состоянии, например SUh е отъезд, то соответствующий переход состояния
это (1 0 0 1 1 1 0 0 0) (0 1 0 1 0 1 0 0 0). В этом случае в потоке сосуществуют четыре потока.
системе и переходный процесс обозначается как I5, I4, . Согласно рисункам 4–11,
возможные переходные события и скорости перехода системы в любое общее состояние могут
можно получить, и затем можно разработать матрицу скорости перехода Q. Кроме того,
можно решить установившуюся вероятность π(x) и определить показатели производительности вторичного
системы, включая пропускную способность сети, использование спектра, вероятность блокировки и принудительное
вероятность завершения может быть дополнительно смоделирована. Мы используем MATLAB для решения задачи Маркова.
модели и запускать симуляции. На этой основе представлены четыре исследования и канал
анализируется производительность сборки на основе предложенной стратегии Ps-rq.

Случай 1: пропускная способность вторичной сети в зависимости от скорости поступления PU

На рисунке 13 показана зависимость между пропускной способностью сети SUe и SUr и
скорость поступления ПУ. Как видно из рисунка, емкость вторичной сети
уменьшается с увеличением скорости поступления ПУ. При количестве ПУ, входящих в
системы увеличивается, абсолютный приоритет доступа к данным уменьшит общее количество доступных
оставшиеся каналы в системе, поэтому количество SU, которые могут принимать услуги в системе
система также соответственно уменьшится. В результате пропускная способность сети SUe и SUr
уменьшается с увеличением скорости поступления ПУ.



(а) (б)

Рисунок 13. (а) Пропускная способность сети SUe, (б) пропускная способность сети SUr.

рисунке 13. (а) Пропускная способность сети SUe, (б) Пропускная способность сети на .

Однако из рисунка 13 видно, что увеличение общей емкости очереди
количества SU увеличивает пропускную способность вторичной сети. При λp = 0,2 емкость
очередь SUe , Qeq, увеличивается от 0 до 4, а пропускная способность вторичной сети увеличивается на
емкость очереди SUr , Qrq, увеличивается с 0 до 4, что составляет увеличение на 9,43%.
вторичной сети. После увеличения емкости очереди количество SU
потоков, которые могут сосуществовать в системе, будут увеличиваться, что увеличивает вероятность того, что SU
сможет принять услугу. Сравнивая рисунок 13, нетрудно обнаружить, что
спектра с градиентным спуском пропускная способность сети SUr меньше, чем у SUe . Это потому что
имеет более высокий приоритет обслуживания, чем SUe (за исключением прерванного SU,
вновь прибывшего SUr , чем для SUe .

Случай 2: использование спектра в зависимости от скорости поступления SUe

Взаимосвязь между использованием спектра и скоростью прибытия SU показана на рисунке.
Рисунок 14 со скоростью поступления SUe по оси x. Из рисунка видно, что
Использование спектра увеличивается с увеличением скорости поступления. Когда Qrq = 2, Qeq = 4,
коэффициент использования спектра составляет до 92,3%. С увеличением скорости поступления SU,
Спектр простоя используется полностью. Когда использование спектра становится насыщенным

Случай 2: использование спектра в зависимости от скорости поступления SUe

Зависимость между использованием спектра и скоростью поступления SU такова:
Рисунок 14 со скоростью поступления по оси X. Это видно из цифры 4,8%;
Использование спектра увеличивается с увеличением скорости поступления. в емкости
коэффициент использования спектра составляет до 92,3%. С увеличением
скорости поступления SU, свободный спектр используется полностью. При использовании
насыщение происходит постепенно, градиент будет уменьшаться. Кроме того, нетрудно, что SUr
что увеличение емкости очереди может увеличить емкость вторичной сети, запустите службу для
Использование труб соответственно улучшается.

Machine Translated by Google

Взаимосвязь между использованием спектра и скоростью поступления SU показана на рисунке 14, где скорость поступления отложена по оси x. Из рисунка видно, что использование спектра увеличивается с увеличением скорости поступления. Когда "=" "=" , коэффициент использования спектра составляет до 92,3%. С увеличением скорости поступления SU свободный спектр используется полностью. Когда использование спектра постепенно станет насыщенным, градиент будет уменьшаться. Кроме того, нетрудно обнаружить, что увеличение спецификация может увеличить емкость вторичной сети, поэтому использование спектра Использование трубопроводов соответственно улучшается.

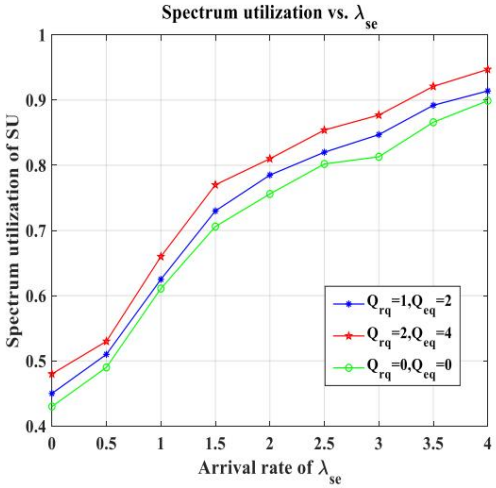


Рисунок 14. Использование спектра SU.

Случай 3: вероятность блокировки в зависимости от скорости поступления PU

На рисунке 15 показана связь между вероятностью блокировки SU и скоростью поступления PU. В целом вероятность блокировки SU увеличивается с увеличением скорости поступления PU. Свободные каналы используются полностью, когда количество PU в системе увеличивается за счет абсолютного приоритета доступа к данным, что приводит к увеличению вероятности блокировки SU. Поскольку SUr имеет более высокий приоритет, чем SUe , вероятность SU. Поскольку имеет более высокий приоритет, чем блокировка, вероятность доступа к данным ниже, чем у SUe . , вероятность блокировки доступа к данным ниже, чем у .

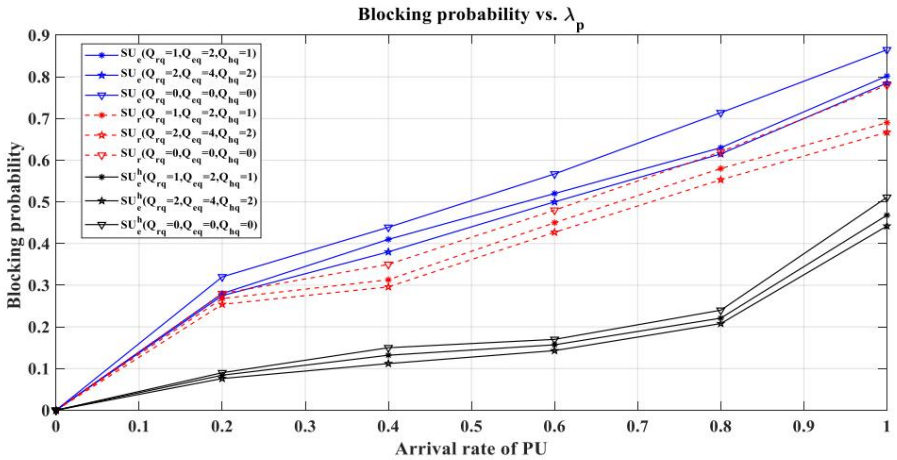


Рисунок 15. Вероятность блокировки SU.

Похоже, что обе кривые относительно стабильны, и разница в вероятности блокировки составляет не более 0,1. Это иллюстрирует тот факт, что предлагаемая стратегия Ps-rq может способность не более 0,1. Это иллюстрирует тот факт, что предложенная стратегия Ps-rq может обеспечить стабильную производительность SUe и SUr на основе уменьшения пробы блокировки. возможности высокоприоритетных пользователей SUe . Внедрение дифференцированных услуг обеспечивает более вероятность высокоприоритетных пользователей. Внедрение дифференцированных услуг обеспечивает справедливое компромиссное планирование, не вызывая чрезмерного голодания пользователей с низким приоритетом. отправлен (SUe). Одновременно доступных свободных, чем существующих вторичных пользователей, уменьшается. отключению пользователей. Однако доступных свободных каналов, существующих PU, по-прежнему приводит к блокировке SUr и SUe постепенно приближается к 1.

Сравнивая вероятность блокировки SUe и SUr при разных мощностях очереди и постепенно приближаясь к 1 , мы можем обнаружить, что большая пропускная способность улучшает снижение блокировки . вероятность.

Сравнивая вероятность блокировки при разной емкости очереди, мы можем обнаружить, что большая емкость улучшает снижение вероятности блокировки. Например, вероятность блокировки снижается на 10,82% при увеличении емкости очереди. И когда скорость поступления PU равна "=" = 4 по сравнению с

Machine Translated by Google

Например, для SUr при скорости поступления PU λp = 0,6 вероятность блокировки снижается на 10,82% при емкости очереди Qrq = 2, Qeq = 4 по сравнению с Qrq = 0, Qeq = 0. А для SUe при λp = 0,6 вероятность блокировки уменьшается на 11,04% при той же стоимости. Это объясняет, что система может предоставлять услуги более потоков с увеличением емкости очереди, что дает больше возможностей для спекулятивный доступ SU. Для сравнения можно заметить, что увеличение очереди емкость позволяет существенно повысить производительность SUe . Причина в том, что обратная связь цикл установлен для обслуживания SUe в предлагаемой стратегии Ps-rq и большей емкости очереди. устанавливается при настройке системы.

Случай 4: вероятность принудительного завершения в зависимости от скорости поступления PU

На рисунке 16 показана взаимосвязь между вероятностью принудительного завершения SU. и скорость поступления PU. Как показано на рисунке, вероятность принудительного завершения SU увеличивается с увеличением скорости поступления PU, что также определяется доступом к данным приоритет ПУ. Поскольку для прерванного SUh установлена зарезервированная очередь. e , принудительное прекращение вероятность SUh e является самым низким на рисунке 16, например, при максимальном значении разницы, λp = 0,8, при Qrq = 1, Qeq = 2, Qhq = 1 вероятность принудительного завершения SUh e на 29,23% ниже, чем у SUe и на 42,5% ниже, чем у SUr . Однако из-за настройке контура обратной связи для SUe , SUr имеет более высокую вероятность принудительного завершения, чем Сью . Таким образом, прерванное SUe может вернуться в очередь и не будет вынуждено прекращается до тех пор, пока очередь не заполнится.

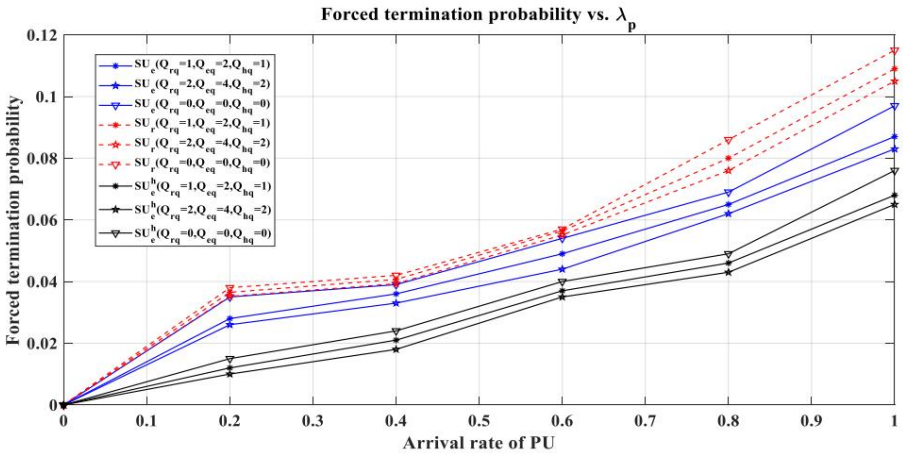


Рисунок 16. Вероятность принудительного завершения SU.

Службы реального времени более чувствительны к задержкам и потере пакетов из-за небольшого дрожания. не оказывает серьезного влияния на качество голосовых/изображений в реальном времени. Поэтому для SUh не обратная петля принята для . При этом общая занятость канала может быть снижена за счет увеличения вероятности принудительного завершения SUr . В целом, способом, но за счет увеличения вероятности принудительного завершения . Принятая предложенная стратегия Ps-rq может дифференцировать приоритеты обслуживания разнородных услуг. и обеспечить для них справедливое планирование. Исходя из этого, качество обслуживания первоочередных сервиса, SUh e , то, что прерывается PU, можно улучшить. высокоприоритетное эластичное обслуживание, превышающее PU, может быть еще раз продемонстрировать превосходство предлагаемой стратегии в повышении

Чтобы дополнительно продемонстрировать превосходство предложенной стратегии в улучшении стратегии Ps-rq в отношении вероятности принудительного завершения, как производительности SU, проводится сравнение между стратегией Ps-rq и результатами показано на рисунке 17. сравнения без Ps, которые все получены при начальных условиях Qeq = 2 и Qrq = 1. Как можно рq относительно вероятности принудительного завершения, как показано на рисунке 17. Из рисунка видно, что для SUe или SUr , когда стратегия не принята, Все результаты обработки произошли при начальных условиях = 2 и = 1. Вероятность принудительного завершения для SU значительно увеличивается. Анализ производительности Из e стратегии, предложенной в этой статье, поэтому нет никакого сравнения СУ видно, является ли ситуация, когда стратегия не принята, специфичной для вероятность принудительного завершения SUh , если не принимать на вооружение Ps-rq. На рисунке показано вероятность принудительного завершения для SU значительно возрастает. Анализ производительности показывает, что увеличение вероятности принудительного завершения для SUe больше, чем для SUr, является специфичным для стратегии, предложенной в этой статье, таким образом, сравнение в отсутствие Ps-rq не проводится . Особенно когда скорость поступления ПУ невелика, у СУэ больше вероятность принудительного завершения, если не принять Ps-rq. Из рисунка видно, что прирост вероятности принудительного завершения при больше, чем при

в отсутствие Ps-rq. Особенно когда скорость прибытия ПУ невелика, имеет

больше возможностей для доступа к каналу. Из предложенной стратегии очевидно, что отличный результат можно объяснить повышением производительности .

Machine Translated by Google

вероятность принудительного завершения для SU значительно возрастает. Анализ эффективности специфичен для стратегии, предложенной в этой статье; таким образом, не проводится сравнение вероятности принудительного завершения без применения Ps-rq. Из рисунка видно, что прирост вероятности принудительного завершения при больше, чем при

в отсутствие Ps-rq. Особенно когда скорость прибытия ПУ невелика, имеет

больше возможностей для доступа к каналу. Из предложенной стратегии видно, что отличный результат можно объяснить улучшенной возможности доступа к каналу отсутствуют. Из предложенной стратегии очевидно, что производительностью SUn . $e^{\cdot}$.

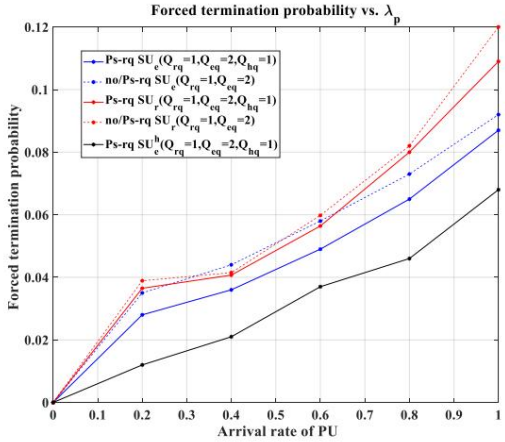


Рисунок 17. Сравнение вероятности принудительного завершения с Ps-rq и без него.

Рисунок 17. Сравнение вероятности принудительного завершения с Ps-rq и без него.

6. Выводы

В этом исследовании разрабатывается и предлагается стратегия приоритетного планирования с зарезервированной очередью (стратегия Ps-rq) для CRN. Эта стратегия использует схему классификации услуг, сочетающую статические факторы (тип сообщения) и динамические факторы (достоверность информации, корреляция сообщений и размер сообщения), которые могут дифференцировать эластичные услуги с высоким приоритетом на основе классификации гетерогенных услуг во вторичных услугах. Установив новую зарезервированную очередь для жизненно важных эластичных сервисов, предлагаемая стратегия может улучшить приоритет обслуживания высокоприоритетных эластичных сервисов, прерываемых PU. Кроме того, для определения приоритетов эластичных сервисов используется нечеткое управление Такаги-Сугено, и представлено трехмерное приоритетно На этой основе обеспечивается пропорциональная схема распределения каналов с приоритетом, сочетающаяся с технологией сборки каналов и технологией адаптации спектра на основе стратегии Ps-rq. Разделив действия пользователя в CRN на четыре события, включая прибытие PU, отправление PU, прибытие SU и отправление SU, исследуется поток ресурсов во вторичной сети и дается процесс динамического доступа к каналу на основе страт При создании CTMC процесс потока ресурсов во вторичной сети отображается в многомерном пространстве состояний, и выводятся все возможные переходные состояния, скорости перехода и условия перехода, вызванные четырьмя видами событий, которые начинаются с общего состояния. После понижения многомерных маркеров стационарное решение системы может быть получено путем получения матрицы скорости перехода, а также ряда индексов оценки производительности, таких как пропускная способность вторичной сети, использование спектра, вероятность блокировки и вероятность принудительного завершения. полученный.

Кроме того, мы сравниваем предложенную стратегию Ps-rq с четырехканальными схемами сборки, чтобы продемонстрировать ее пригодность для конкретных сценариев, связанных с передачей эластичного трафика. Объединив процедуру уменьшения размерности многомерных цепей Маркова, мы численно моделируем аналитические результаты показателей производительности с использованием MATLAB и предоставляем четыре варианта исследования. Результаты моделирования показывают, что в том же измерении с увеличением емкости буфера пропускная способность сети SUe увеличивается на 4,8%, а вероятность блокировки снижается на 11,04%, в то время как емкость сети SUr увеличивается на 9,43%, а вероятность блокировки вероятность снижается на 10,82%. А вероятность принудительного завершения высокоприоритетных эластичных сервисов, рассматриваемых в этой статье, на 29,23% ниже, чем у SUe , и на 42,5% ниже, чем у SUr . Численные результаты показывают, что предложенная стратегия может помочь улучшить приоритет обслуживания высокоприоритетных эластичных услуг , которые прерываются PU в CRN, и снизить вероятность принудительного завершения на основе обеспечения справедливого планирования.

Machine Translated by Google

Вклад автора: QX изучила литературу, разработала концепцию исследования, предоставила математические модели, разработала алгоритм, выполнила численное моделирование, взяла на себя ответственность за подготовку первоначального проекта и отредактировала рукопись; С.Л. разработал концепции исследования, усовершенствовал методологию систематических исследований и анализа и руководил завершением доработки статьи; JG и YH дали ценные предложения по доработке и проверили вывод формул и грамматику английского языка. Все авторы прочитали и согласились с опубликованной версией рукописи

Финансирование: Это исследование частично финансировалось Национальным фондом естественных наук Китая (61663024) и проектом докторского фонда естественных наук Сианьского авиационного университета (2022KY0201).

Заявление о доступности данных. Данные, подтверждающие эту статью, взяты из ранее опубликованных исследований и наборов данных, которые цитировались.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Федеральная комиссия по связи. Отчет Целевой группы по политике в области спектра. ET-документ; № 02-135; Федеральные коммуникации Комиссия: Вашингтон, округ Колумбия, США, 2002 г.
2. Хуан, XL; Ли, YX; Гао, Ю. Доступ к спектру на основе Q-Learning для передачи мультимедиа по когнитивным радиосетям. IEEE Транс. Когн. Коммун. Сеть. 2021, 7, 110–119. [Перекрестная ссылка]
3. Шрикар, Д.; Нелла, А.; Мамиди, Р. Новая интегрированная система когнитивной радиоантенны UWB-зондирования и 8-элементной связи MIMO. Электроника 2023, 12, 330. [CrossRef]
4. Музаффар, Мьюзикл; Шарки, Р. Обзор зондирования спектра в современных когнитивных радиосетях. Телекоммуникации. Сист. Модель. Анальный. Дес. Менеджер. 2024, 2, 85. [CrossRef]
5. Ахмад, В.; Радзи, Н.; Самиди, Ф.С. Технология 5G: на пути к динамическому совместному использованию спектра с использованием когнитивных радиосетей. IEEE Access 2020, 8, 14460–14488. [Перекрестная ссылка]
6. Азалы, Н.; Бадран, EF Повышение производительности динамического доступа к спектру посредством резервирования каналов для когнитивного радио Сети. Вирел. Перс. Коммун. 2021, 118, 2867–2883. [Перекрестная ссылка]
7. Лей, Дж. Агрегация и фрагментация каналов для транспортных потоков; Шпрингер: Берлин/Гейдельберг, Германия, 2020.
8. Лю, Х.; Цзя, М. Интеллектуальное распределение ресурсов спектра на основе совместной оптимизации в гетерогенном когнитивном радио. IEEE Транс. Экстрен. Вершина. Вычислить. Интел. 2020, 4, 5–12. [Перекрестная ссылка]
9. Ли, СП; Сюй, QY; Джаафар, Г. Моделирование и анализ производительности сборки каналов на основе стратегии Ps-rc с очередями с приоритетами в CRN. Вирел. Коммун. Моб. Вычислить. 2022, 17, 6384261. [CrossRef]
10. Ма, Л.; Сюй, Ю.; Фу, Ю. Схема планирования ресурсов для агрегации спектра в гетерогенных когнитивных радиосистемах. Сети. Китайская коммуна. 2015, 12, 100–111.
11. Цзяо, Л.; Ли, ФЮ; Пла, В. Моделирование и анализ производительности сборки каналов в многоканальных когнитивных радиосетях с адаптацией спектра. IEEE Транс. Вех. Технол. 2012, 61, 2686–2697. [Перекрестная ссылка]
12. Балапувадуге, И.; Цзяо, Л.; Пла, В. Сборка каналов с очередями на основе приоритетов в сетях когнитивного радио: стратегии и оценка производительности. IEEE Транс. Вирел. Коммун. 2014, 13, 630–645. [Перекрестная ссылка]
13. Эсеного, Э.; Шривастава, В.М. Стратегии сборки двух гетерогенных каналов в сетях когнитивного радио: перформанс Анализ. Межд. Дж. Инж. Технол. Иннов. 2017, 7, 98–116.
14. Чжан, В.; Сан, Ю.; Дэн, Л. Динамическое распределение спектра для гетерогенных когнитивных радиосетей с несколькими каналами. Система IEEE. Дж. 2019, 13, 53–64. [Перекрестная ссылка]
15. Чжу, Х. Анализ когнитивного доступа к радиоспектру с оптимальным резервированием каналов. IEEE-сообщество. Летт. 2007, 11, 304–306. [Перекрестная ссылка]
16. Эсеного, Э.; Мамбу, Э.Н. Интеграция режима массового обслуживания в политику агрегации когнитивных радиоканалов: результат Оценка. Ж. Коммун. 2017, 1–5. [Перекрестная ссылка]
17. Эсеного, Э.; Шривастава, Стратегия сборки каналов VM в сетях когнитивного радио: подход на основе очередей. Межд. Дж. Коммун. Антенна Пропаг. IRCEAP 2017, 7, 31–47. [Перекрестная ссылка]
18. Сяо, Х.; Цзэн, Ф.; Ху, З. Динамическая адаптивная к потоку аренда спектра с агрегацией каналов в сетях когнитивного радио. Датчики 2020, 20, 3800. [CrossRef] [ПабМед]
19. Эль-Тухи, А.Т.; Арслан Х. Повышение производительности SU с низким приоритетом с использованием зарезервированных каналов в CRN. IEEE Вирел. Коммун. Летт. 2020, 9, 513–517. [Перекрестная ссылка]
20. Байракдар, Мэн; Кэлхан, А. Улучшение использования передачи спектра для приоритетных пользователей когнитивного радио путем использования Объединение каналов с уменьшением голода. Межд. Дж. Электрон. Коммун. 2017, 71, 181–191. [Перекрестная ссылка]
21. Суганти, Н.; Минакши, С. Эффективный алгоритм планирования, использующий систему массового обслуживания для минимизации голодания не в реальном времени Вторичные пользователи в сети когнитивного радио. Класт. Вычислить. 2018, 25, 1–12. [Перекрестная ссылка]
22. Ифтихар, М.; Маткур, Х.; Имран, М. Новая структура системы массового обслуживания G/M/1, основанная на механизме планирования и опроса для анализа нескольких классов самоподобного и LRD-трафика. Вирел. Сеть. 2016, 22, 1269–1284. [Перекрестная ссылка]

Machine Translated by Google

23. Хан, М.; Ахмад, Дж.; Саддик А.Е. Drone-HAT: гибридный преобразователь внимания для распознавания сложных действий в видео наблюдения с помощью дронов. В материалах семинаров конференции IEEE/CVF по компьютерному зрению и распознаванию образов (CVPR), Сиэтл, Вашингтон, США, 19–21 июня 2024 г.; стр. 4713–4722.

24. Тахерхани, Н.; Пьер, С. Приоритизация и планирование сообщений для контроля перегрузки в транспортных сетях Ad Hoc. Вычислить. Сеть. 2016, 108, 15–28. [Перекрестная ссылка]

25. Хан, В.; Ли, Дж.; Тиан, З. Стратегии динамического зондирования для эффективного использования спектра в сетях когнитивного радио. IEEE Транс. Вирел. Коммун. 2011, 10, 3644–3655. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Многопутевая маршрутизация на основе глубокого обучения с подкреплением для LEO Сети мегасозвездий

Чи Хан [1], Вэй Сюн [1,2,*] и Жунхуань Юй [1,2]

[1] Национальная ключевая лаборатория обнаружения космических целей, Университет космической инженерии, Пекин 101400, Китай; hanchi@hgd.edu.cn (Швейцария); yrh1983@163.com (РЮ)
[2] Школа космической информации, Университет космической инженерии, Пекин 101400, Китай *
Адрес для переписки: xiongwei@hgd.edu.cn

Аннотация: Расширение сетей мегагруппировок (MCN) представляет собой многообещающее решение для достижения глобального покрытия Интернета. Для удовлетворения растущего спроса на спутниковые услуги многопутевая маршрутизация позволяет одновременно устанавливать несколько путей передачи, обеспечивая параллельную передачу потоков. Тем не менее, мобильность спутников и изменяющиеся во времени состояния каналов представляют собой проблему для обнаружения оптимальных путей и планирования трафика при многопутевой маршрутизации. Учитывая негибкость традиционных алгоритмов маршрутизации на основе статического глубокого обучения (DRL) при работе с изменяющимися во времени топологиями созвездий, многопутевая маршрутизация на основе DRL (DMR), реализуемая графовой нейронной сетью (GNN), предлагается в качестве средства улучшения производительность передачи MCN. DMR разделяет задачу стохастической оптимизации многопутевой маршрутизации при ограничениях трафика и пропускной способности на две подзадачи: обнаружение многопутевой маршрутизации и планирование многопутевого трафика. Во-первых, предлагается алгоритм обнаружения многопутевого маршрута на основе минимального количества переходов (MHMRD) для расчета нескольких доступных путей между всеми узлами источника и назначения.

Во-вторых, схема планирования многопутевого трафика (GMTS) на основе GNN предлагается как средство динамического планирования трафика на каждом доступном пути для каждого потока данных на основе информации о состоянии ISL и потребности в трафике. Результаты моделирования показывают, что предложенная схема может быть масштабирована до созвездий с различными конфигурациями без необходимости повторного обучения и повысить пропускную способность, коэффициент завершения и задержку на 42,64%, 17,39% и 3,66% по сравнению с алгоритмом «первый кратчайший путь» (SPF) соответственно.

Ключевые слова: спутниковая сеть; многопутевая маршрутизация; глубокое обучение с подкреплением; планирование движения; количество прыжков

check for updates

Цитирование: Хан, К.; Сюн, В.; Ю, Р. Многопутевая маршрутизация на основе глубокого обучения с подкреплением для сетей мегасозвездий LEO. Электроника 2024, 13, 3054. https://doi.org/10.3390/electronics13153054 .

Академический редактор: Стефано Сканцио

Поступила: 12 июля 2024 г.
Пересмотрено: 25 июля 2024 г.
Принято: 31 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария.
Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons
Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Характеристики сетей мегасозвездий (MCN), такие как широкое покрытие территории , низкая задержка и высокая пропускная способность, дают им уникальные преимущества в удаленной области связи и услугах, чувствительных к задержкам (например, экстренная связь и мониторинг стихийных бедствий) [1] . Вышеупомянутые преимущества привели к ускоренному развитию MCN, которые, как ожидается, будут выполнять множество услуг, включая связь, дистанционное зондирование, навигацию и позиционирование. Более того, MCN могут стать ключевым компонентом инфраструктуры интегрированной сети «космос-воздух-земля» (SAGIN) [2–4].

Многолучевая маршрутизация является перспективным подходом для удовлетворения растущего спроса на услуги спутниковой связи . Многопутевая маршрутизация позволяет одновременно устанавливать несколько путей передачи между узлами источника и назначения для каждого потока данных, что может увеличить пропускную способность сети [5]. В отличие от однопутевой маршрутизации, многопутевая маршрутизация позволяет оптимально использовать доступную полосу пропускания, тем самым уменьшая перегрузку, улучшая пропускную способность и повышая отказоустойчивость. Это достигается за счет возможности параллельной передачи одной услуги по нескольким п

Machine Translated by Google

Многопутевая маршрутизация повышает устойчивость спутниковых сетей с ограниченными ресурсами и позволяет противостоять сценариям отказа узла или канала. Однако высокая динамика топологий спутниковых сетей и состояний каналов передачи создает проблемы для обнаружения многопутевой маршрутизации и планирования многопутевого трафика [6–8].

Обнаружение многопутевой маршрутизации является первой проблемой, возникающей из-за меняющейся во времени природы спутниковой топологии и трафика. Такие факторы, как изменения положения спутников и сложная электромагнитная обстановка, могут привести к снижению скорости передачи лазерной линии или даже к ее сбою. В результате сбои каналов могут возникать часто в MCN, а не быть исключительным событием, как в наземных сетях. Частое возникновение пропусков или сбоев канала препятствует возможности механизма широковещательной передачи своевременно обнаруживать и поддерживать маршруты, как это имеет место в наземных сетях. Это приводит к увеличению накладных расходов, связанных с широковещательной передачей и управлением маршрутами. Следовательно, протоколы специальной маршрутизации на основе топологии, такие как ADOV [9], неоптимальны для сетей MCN. Кроме того, из-за мобильности и энергетических ограничений спутников LEO установленные связи часто не поддерживаются стабильно. Топологические изменения в процессе обнаружения путей могут привести к быстрому истечению срока действия обнаруженных путей, при этом соответствующие пути перестанут существовать на момент передачи данных. Вышеупомянутые факторы требуют обнаружения многопутевой маршрутизации, чтобы иметь возможность своевременно обнаруживать сбои каналов и разумно устанавливать несколько альтернативных путей. Некоторые схемы в настоящее время используют вычисление пути в наземном центре управления сетью (NCC) [10]. NCC повышает производительность передачи по сети за счет всестороннего рассмотрения распределения наземного пользовательского трафика и использования каналов, а также полного использования резервирования каналов, создаваемого ячеистой топологией спутниковой сети, для одновременного планирования нескольких параллельных путей передачи для каждого потока.

Планирование многопутевого трафика в MCN является второй проблемой. Сложные электромагнитные условия, изменения положения спутников и другие факторы приводят к колебаниям качества связи. Как следствие, традиционное планирование многопутевого трафика на основе статической маршрутизации больше не применимо в MCN. Кроме того, изменяющийся во времени характер наземного пользовательского трафика и колебания качества связи делают спутниковую сеть неспособной точно идентифицировать модели трафика, а для детального планирования трафика между несколькими маршрутами часто не хватает эффективной информации. Глубокое обучение с подкреплением (DRL) продемонстрировало значительные преимущества в описании пространственно-временных особенностей, решений о потоках и других связанных областях. Значительное количество ученых предпочли использовать DRL для целей адаптивного планирования многопутевого потока, о чем свидетельствуют [11–14]. Однако изменяющиеся во времени спутниковые сети и нестабильное качество межспутниковых линий связи (ISL) требуют агрегирования нескольких временных интервалов для принятия решений в моделях, использующих DRL. По мере увеличения размера созвездия размерность модели резко увеличивается, что приводит к значительному увеличению вычислительных усилий и снижению масштабируемости модели . Графовые нейронные сети (GNN) способны выявлять особенности и закономерности в данных с графовой структурой, выполнять реляционные рассуждения и комбинаторное обобщение данных с графовой структурой [15,16], и их можно применять без дополнительной настройки модели для созвездий различных топологий. и размеры [17,18].

Недавно DRL продемонстрировал значительный потенциал в изучении временных и пространственных характеристик трафика с целью принятия решений о маршрутизации. Исследователи использовали DRL для адаптивного распределения трафика с целью замены статической сегментации трафика. Однако модели на основе DRL требуют возможности суммировать непрерывные топологические снимки для принятия решений в динамической спутниковой среде. Спутниковые сети имеют многомерные и разреженные пространства состояний, а размерность моделей DRL резко увеличивается с размером созвездия и уровнем спроса на трафик, что, в свою очередь, приводит к плохой масштабируемости модели. Графовые нейронные сети (GNN) облегчают вывод реляционных данных и обобщение комбинаторных структур внутри информации, структурированной на графе. Это позволяет применять модель DRL к спутниковым группировкам разных размеров без необходимости дополнительных модификаций. В свете вышеупомянутого анализа мы предлагаем многопутевую маршрутизацию (DMR) на основе DRL, встроенную в GNN, для улучшения производите

Machine Translated by Google

Проблема многопутевой маршрутизации изначально моделируется как задача стохастической оптимизации с целью максимизации эффективности сети при ограничениях трафика и пропускной способности . Чтобы получить оптимальное решение, исходная задача разделяется на две подзадачи: обнаружение многопутевой маршрутизации и планирование многопутевого трафика. Что касается исходной подзадачи, предлагается полностью использовать резервированную топологию спутниковой сети с введением алгоритма обнаружения многопутевой маршрутизации на основе минимального количества переходов (MHMRD). MHMRD предназначен для вычисления количества доступных путей между всеми узлами источника и назначения. На этой основе вторая подзадача решается путем разработки многопутевого планирования трафика на основе GNN (GMTS). GMTS моделирует процесс планирования многопутевого трафика как процесс принятия решений Маркова (MDP), который динамически планирует трафик по каждому доступному маршруту для каждого потока данных на основе информации о состоянии канала спутниковой сети и потребности в трафике. Результаты моделирования на Iridium и OneWeb показывают, что предлагаемая схема не зависит от переобучения для группировок разного размера. Кроме того, пропускная способность сети, средняя задержка и скорость завершения потока превысили показатели базовой схемы.

Основные положения данной статьи резюмируются следующим образом:

- Проблема многопутевой маршрутизации в спутниковых сетях моделируется как задача стохастической оптимизации при ограничениях трафика и пропускной способности. Целью является максимизация эффективности сети. Проблема разделена на подзадачу обнаружения многопутевой маршрутизации и подзадачу планирования многопутевого трафика. • Для решения подзадачи обнаружения многопутевой маршрутизации был предложен алгоритм MHMRD как средство планирования нескольких путей с минимальным количеством переходов для каждого потока между любой парой узлов в сети на основе информации о состоянии канала. • Подзадача планирования многопутевого трафика моделируется как MDP, при этом предлагается схема GMTS для динамического планирования доли трафика на каждом доступном пути для каждого потока данных. GMTS — это масштабируемое решение, которое можно применять к различным созвездиям.

Остальная часть этой статьи организована следующим образом. В разделе 2 представлены некоторые связанные с этим работы. В разделе 3 мы приводим модель системы и формулировку задачи. В разделе 4 описывается многопутевая маршрутизация на основе DRL . В разделе 5 представлены моделирование и основные результаты. Наконец, мы суммируем результаты в разделе 6.

2. Сопутствующие работы

2.1. Многопутевая маршрутизация

В спутниковых сетях существует несколько маршрутов с одинаковым количеством переходов от начала до конца. Чтобы повысить надежность передачи данных и расширить возможности балансировки сетевой нагрузки, исследователи предложили использовать многопутевую маршрутизацию в спутниковых сетях, где каждый поток делится на несколько подпотоков для параллельной передачи по разным путям. Чтобы распределить трафик по нескольким путям, традиционный метод многопутевого доступа с равной стоимостью (ECMP) выполняет статическое разделение трафика на основе информации о сообщении. Однако подход ECMP не учитывает ограничения сетевых параметров, таких как пропускная способность и задержка, и подвержен перегрузкам при высокой загрузке сети.

Для достижения динамического управления потоками для многопутевой маршрутизации была предложена архитектура спутниковой сети на основе программно-конфигурируемой сети (SDN) для центрального управления топологией и управления трафиком [18–21]. В сетевой архитектуре на основе SDN плоскость данных и плоскость управления разделены, контроллер SDN определяет маршрутизацию на основе параметров сети и требований QoS, а спутник отвечает за пересылку данных.

В [22] авторы попытались объединить сетевое кодирование и многопутевую маршрутизацию, чтобы повысить эффективность передачи данных в спутниковых сетях. Однако подход централизованного управления подвержен проблемам с большими задержками. В [23] для уменьшения задержки управления и повышения стабильности передачи авторы применили распределенный подход к управлению маршрутизацией. В [24] авторы предложили многопутевую кооперативную маршрутизацию на основе сетевого кодирования (NCMCR). Чтобы оптимизировать пропускную способность, каждый поток

Machine Translated by Google

в NCMCR передается по множеству непересекающихся каналов. Однако вышеупомянутые схемы многопутевой маршрутизации не учитывают параметры каналов различных путей пересылки и, следовательно, не способны адаптивно выполнять разумное разделение трафика. В [1] авторы разработали схему адаптивной балансировки трафика между несколькими временными путями путем объединения разреженных и избыточных механизмов сетевого кодирования. Этот подход позволяет достичь детерминированных гарантий задержки в сценариях с ограниченными ресурсами узла и многопользовательской конкуренцией. Тем не менее, все вышеупомянутые схемы маршрутизации используют статические топологии, что не позволяет им адаптироваться к спутниковым сетям с динамически изменяющимися топологиями.

2.2. Интеллектуальная

маршрутизация Изменяющаяся во времени среда спутниковой сети и динамические межзвездные связи усложняют маршрутизацию спутниковой сети. Чтобы гарантировать качество обслуживания сети в условиях увеличения размеров пользователей и группировок, многопутевая маршрутизация требует более эффективного планирования пути и более точных стратегий управления трафиком. Обучение с подкреплением — это интеллектуальный инструмент, который поддерживает принятие решений путем взаимодействия с окружающей средой [25–27], и ученые предложили многочисленные схемы интеллектуальной маршрутизации, основанные на обучении с подкреплением. В [25] авторы разработали контролируемую систему глубокого обучения для построения таблиц маршрутизации. В [28] авторы применили глубокое обучение на основе графов в спутниковых сетях с помощью архитектуры нейронной сети, называемой нейронной сетью графовых запросов. Тем не менее, такие контролируемые схемы маршрутизации на основе обучения ограничены из-за их неспособности обобщать и адаптироваться к различным созвездиям. Чтобы решить эту проблему, авторы [29] предложили модель процесса принятия решений Маркова (MDP) для маршрутизации спутниковой сети, используя многоагентное глубокое обучение с подкреплением (DRL) для удовлетворения разнообразных требований к качеству обслуживания (QoS). Однако политика фиксированной цели делает алгоритмы маршрутизации на основе DRL негибкими, поскольку они неспособны адаптироваться к динамически изменяющимся сетям. Очевидно, что традиционные статические конфигурации целей неспособны отразить различную важность различных показателей
Чтобы решить эту проблему, авторы [17,30] интегрировали графовую нейронную сеть (GNN) в интеллектуальные функции DRL, что позволяет своевременно корректировать показатели производительности сети путем прогнозирования тенденции достижения целей оптимизации алгоритмов маршрутизации. Это позволяет интеллекту изучать оптимальные пути, которые могут адаптироваться к различным изменениям окружающей среды. В [31,32] авторы предложили различные архитектуры DRL, каждая из которых динамически корректирует веса каналов в соответствии с нагрузкой ключевы
В [33] авторы расширили возможности TCP, чтобы объединить соответствующую управляющую информацию и гибко поддерживать связь между подпотоками транспортного уровня и контроллером SDN. Была предложена архитектура MPTCP (scMPTCP), сотрудничающая с SDN, которая выбирает маршруты для новых подпотоков на основе доступной пропускной способности каждого маршрута.

и позволяет избежать узких мест других подпотоков. Он также может адаптироваться к изменениям сетевой нагрузки.
В [34] авторы предложили контроллер DQN, рассматривающий модель потерь на трассе, основанную на марковской модели процесса принятия решений для выбора сети и адаптивного распределения ресурсов в гетерогенных сетях.

Тем не менее, изучение вышеупомянутого исследования показывает, что существующие исследования на основе DRL нелегко масштабировать из-за присущей им связи с входными состояниями и ограниченной топологии, которая согласуется с обучающими данными. Чтобы разрешить вышеупомянутое противоречие, предлагаемая многопутевая маршрутизация на основе DRL, встроенная в GNN, может легко определить оптимальное действие в пространстве непрерывных действий. DMR разделяет проблему многопутевой маршрутизации на отдельные компоненты — обнаружение многопутевого трафика и планирование многопутевого трафика — для уменьшения сложности решения. Одновременно с этим целевая политика, принятая DMR, может быть изменена в соответствии с преобладающими условиями окружающей среды, тем самым повышая гибкость системы по отношению к различным топологиям.

Machine Translated by Google

3. Модель системы и постановка задачи 3.1. Сценарий многопутевого доступа

Сценарий многопутевой маршрутизации в MCN показан на рисунке 1, где спутниковая сеть разделена на взаимно независимые плоскости данных и плоскости управления через SDN.
В наземной сети данные в буфере отправки передаются отправителю через маршрутизаторы через несколько переходов и объединяются в несколько подпотоков. Каждый подпоток может выбрать оптимальный путь передачи в соответствии с текущим состоянием сети, а затем сходиться в приемнике.

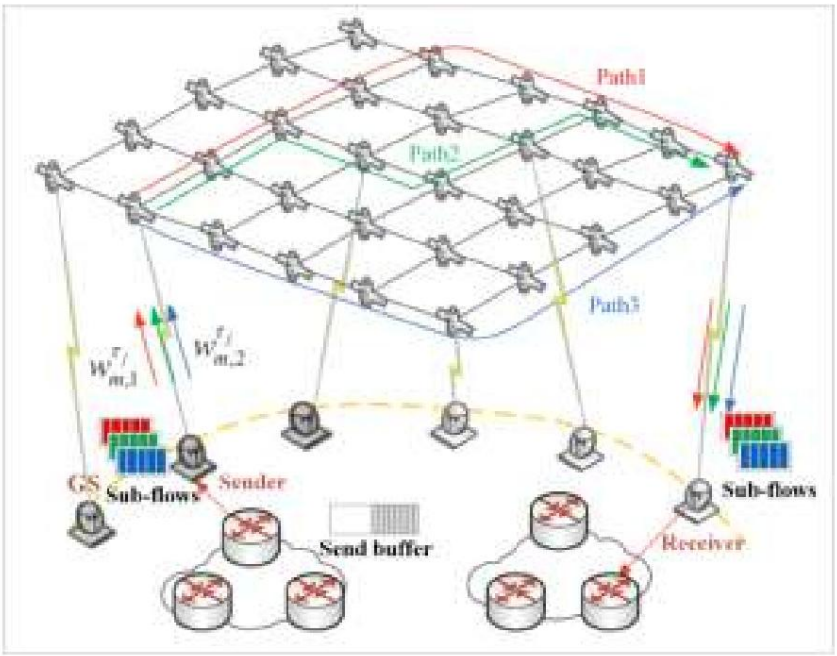


Рисунок 1. Многопутевая маршрутизация в сети мегасозвездия.

Плоскость данных включает группировку NSNP/NP/F Walker Delta, которая распределена равномерно и симметрично, и наземную станцию. Плоскости орбит NP равномерно распределены вдоль экватора, тогда как спутники NS распределены равномерно в каждой плоскости. Каждый спутник может быть подключен к четырем межспутниковым каналам связи (ISL), состоящим из двух внутриплоскостных ISL и двух межплоскостных ISL. В плоскости управления центр управления сетью берет на себя функции унифицированного управления топологией, планирования трафика и управления путями. Из-за мобильности сетей мегагруппировок LEO межплоскостные ISL подвержены частым сбоям и перестройкам, а топология находится в состоянии постоянного изменения. Следовательно, NCC должен постоянно обновляться, чтобы соответствовать развивающейся топологии, что представляет собой серьезную проблему для планирования многопутевой маршрутизации и планирования трафика.

3.2. Многопутевая маршрутизация и модель трафика

MCN моделируется как пространственно-временной граф G = {(V, E, T, N)}, где V = S G — узел, состоящий из всех спутников и наземных шлюзов, а E = ES EG — множество ISL ES. и набор наземных каналов связи EG, а T — вектор временных интервалов длиной N. Как показано на рисунке 1, ячеистая топология MCN допускает существование нескольких потенциальных маршрутов между любой заданной парой узлов. Предположим, что количество потоков пакетов во временном интервале t равно M (т. е. Pt = { (si , di) |i = 1, 2, · · ·, M}). Другими словами, переменная M представляет количество подпотоков, как показано на рисунке 1, где три подпотока в отправителе ожидают передачи по нескольким путям. Количество доступных путей между

исходный узел si и целевой узел di — это Lsi ,di , обозначаемый как Rt = p $^{T}_{1,1,1,L,M,L}$ $^{tt\cdots,p\cdots,p}$.

Вот, пм ,л ^{T} — l-й путь между m-й парой узлов источника и назначения, состоящий

Machine Translated by Google

$p_{m,l}^{т} \varepsilon 1,\varepsilon 2,\cdots,\varepsilon j,\cdots,\varepsilon nm,l$, где nm,l — общий прыжок = всех ISL на этом пути (т. е. количество текущего канала).

Пусть $q_{\varepsilon j}$ — задержка в очереди канала εj, t — задержка обработки пакета, а y — расстояние канала. Тогда задержка l-го пути передачи m-й пары узлов источник-назначение может быть выражена следующим образом:

$$\eta_{м,х}^{т} = \sum_{j=1}^{нм,х} \left(\varepsilon_j^{т} t + r + q_{\varepsilon j} \right) \frac{y}{c} \quad (1)$$

где c — скорость света в вакууме, а nm,l — количество переходов текущего канала.

Каждый исходный узел может разделить трафик на подпотоки Lsi ,di для параллельного прохождения по путям Lsi ,di . Пусть ωt будет коэффициентом трафика каждого подпотока. Тогда сумма Lsi ,di между любой парой узлов источника и назначения подпотоки должна быть равна единице, что можно выразить следующим образом:

$$\sum_{л=1}^{л} \omega_{м,л}^{т} = 1, \quad m = 1, 2, \cdots, M \quad (2)$$

Следовательно, трафик канала εj в момент времени t может быть выражен как

$$f_{\varepsilon j}^{т} = \sum_{м=1}^{M} \sum_{л=1}^{л} b_{m,l}^{т} \delta_{m,l}^{т} \omega \quad (3)$$

где $б_{м}^{т}$ и $\delta_{м,л}^{т}$ — требования к полосе пропускания m-й пары к узлам источника и назначения; — коэффициент потери пакетов на пути. Учитывая, что максимальная пропускная способность межзвездного канала равна BW, трафик ISL не должен превышать максимальный предел пропускной способности, который можно выразить следующим образом:

$$f_{\varepsilon j}^{т} \le BW, \quad \varepsilon j \in p_{м,л}^{т} \quad (4)$$

Следовательно, пропускная способность между m-й парой узлов может быть выражена как

$$f = \max f_{m,l}^{т}(\varepsilon j), \quad \varepsilon j \in p_{м,л}^{т} \quad (5)$$

3.3. Постановка проблемы

Чтобы измерить задержку и пропускную способность ISL, функция полезности MCN определяется как

$$ты\left(\bar{f}t, \bar{d}t \right) = \beta 1 \log \bar{f}t + \beta 2 \log \bar{d}t \quad (6)$$

где β1 и β2 — коэффициенты важности пропускной способности и задержки соответственно, $\bar{f}t$ и $\bar{d}t$ — средние пропускная способность и задержка β1 и β2 удовлетворяют β1 + β2 = 1 и текущему временному интервалу соответственно, что можно выразить как следует:

$$\bar{фут} = \sum_{м=1}^{M} \sum_{л=1}^{|фм|} \sum_{у=1}^{|пк|} \frac{пак}{т}, \quad t \in [1, T] \quad (7)$$

$$\bar{d}t = \frac{\sum_{l=1}^{M} \sum_{y=1}^{|пк|} \int_{[1,T]} d_{m,l}^{т}(t)\, dt}{\sum_{l=1}^{M} \sum_{y=1}^{|фм|} |пк|}, \quad ку \in m=1 \quad (8)$$

где | фм| — количество потоков данных между m-й парой узлов, |pk | — количество пакетов для k-го потока данных, pak — размер пакета, а ку {0, 1} — двоичная переменная , которая измеряет, был ли успешно отправлен у-й пакет.

Machine Translated by Google

Следовательно, оптимизация многопутевой маршрутизации спутниковой сети может быть выражена как следующая задача оптимизации:

$$
\begin{aligned}
&P0: \text{макс } U\ f, \quad \bar{\ }_{A}, \quad t \quad [1, T] \\
&C1: \bar{\ }d = \frac{1}{H} \sum_{i=1}^{H} \bar{\ }_{дт} \\
&C2: \text{ж} = \frac{1}{H} \sum_{i=1}^{H} \bar{\ } \quad \text{футы} \\
&st\ C3: f \epsilon j\ L \quad BW, \quad \epsilon j \quad p \quad {}^{T}_{м,\ л} \\
&C4: \quad \omega t \quad \sum_{l=1} l, \quad m \quad [1, M]\ m,l \\
&C5: d\ t \quad d\ t\ m,l_{мой'} \quad 1 \quad l \quad y \quad L, \quad m \quad [1, M]
\end{aligned}
\tag{9}
$$

где ограничения C1 и C2 обозначают среднюю задержку и пропускную способность спутниковой сети соответственно, ограничение C3 предусматривает, что пропускная способность канала εj в любой данный момент t не может превышать предел полосы пропускания, ограничение C4 означает, что общий объем связи между любой парой узлов — это совокупный результат прохождения нескольких подпотоков по нескольким путям, и, наконец, ограничение C5 описывает ограничения задержки всех возможных путей между каждой парой узлов. Видно, что P0 — NP-трудная задача с неровным пространством решений. P0 состоит из двух частей — обнаружения многопутевой маршрутизации и планирования многопутевого трафика — которые сходятся медленно и могут попасть в локальный оптимум, если решить их с помощью стохастической оптимизации. Чтобы уменьшить сложность решения, в этой статье проблема разделена на две подзадачи: обнаружение маршрутизации и планирование трафика, чтобы повысить эффективность алгоритма.

4. Многопутевая маршрутизация на основе DRL

В этой статье мы предлагаем многопутевую маршрутизацию на основе DRL (DMR), как показано на рисунке 2, которая сначала использует алгоритм обнаружения многопути на основе минимального количества переходов для поиска нескольких путей с приемлемыми задержками, а затем использует алгоритм GMTS для определения схемы планирования трафика на основе состояния сети созвездия.

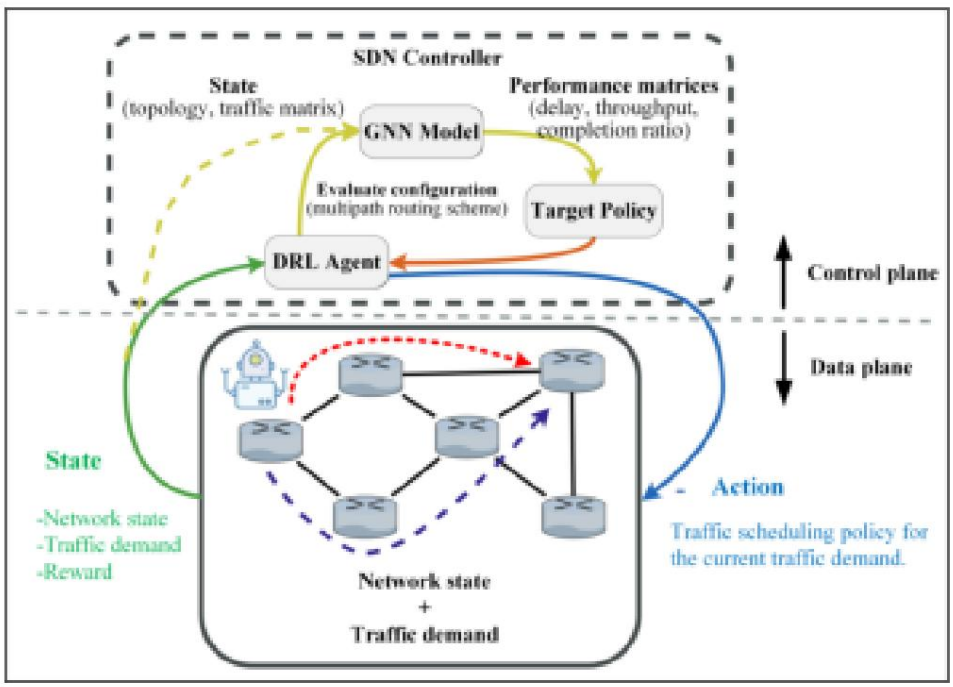


Рисунок 2. Общая структура многопутевой маршрутизации на основе DRL.

Machine Translated by Google

4.1. Обнаружение многопутевой маршрутизации

Целью подзадачи обнаружения многопутевой маршрутизации является определение оптимального пути задержки и неоптимального пути задержки. В сетях MCN задержка распространения, которая зависит от количества переходов на пути, представляет собой существенную составляющую сквозной задержки. Следовательно, многопутевые пути в первую очередь рассматриваются как оптимальные и субоптимальные пути внутри сети. Подзадачу обнаружения многопутевой маршрутизации можно выразить следующим образом:

$$\begin{matrix} P1 : \min d\,t_{m,l}, & t \quad T, m \quad [1, M], l \quad [1, L] \\ st\ d\,t \quad d\,t_{m,l} & _{мой'} \quad 1 \quad l \quad y \quad L \end{matrix} \tag{10}$$

Первоначально определяется путь с минимальным количеством переходов между любой парой узлов. На этой основе NCC определяет неоптимальный путь передачи (т. е. резервный путь) для каждого потока в порядке убывания трафика. Это связано с тем, что централизованный трафик с большей вероятностью вызовет перегрузку канала, чем децентрализованный крошечный трафик. В созвездии NSNP/NP/F дельты Ваклера , как показано на рисунке 3, наклонение равно α, а разность прямых восхождений восходящего узла (RAAN) соседних орбит равна Ω = 2π/NP. Разность фаз между соседними спутниками на одной орбите равна Φ = 2π/NS, а разность фаз между соседними спутниками на соседних орбитах равна υ = 2πF/NSNP, а u1 и u2 являются аргументом широты (или фазового угла) значения спутников источника и назначения соответственно. Аргумент широты представляет собой угол между восходящим узлом и спутником и в основном определяет положение спутника на орбите, где ζ(u1) и ζ(u2) указывают разницу долгот источника и спутника назначения по отношению к соответствующему восходящему спутнику. узлы соответственно. В созвездии Дельта Уокера подсчет сквозных переходов включает как поперечные межплоскостные переходы Hh , так и внутриплоскостные переход

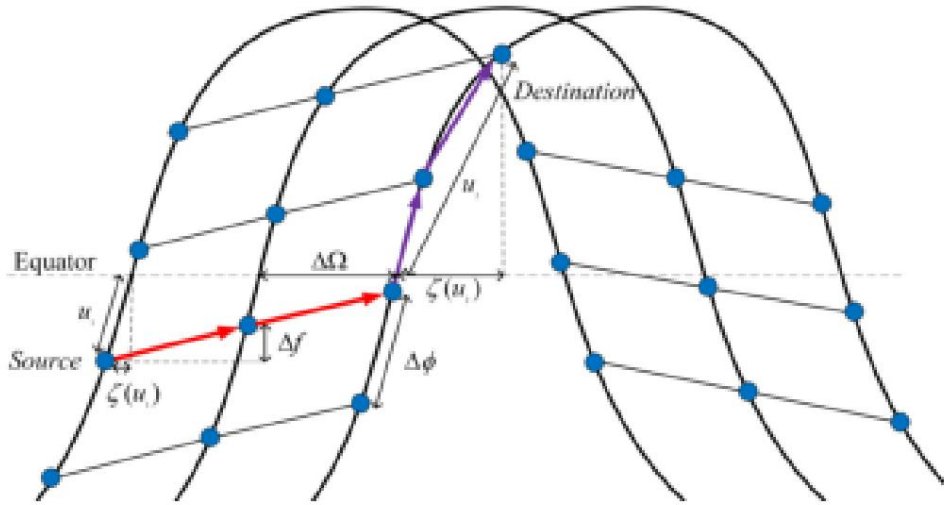


Рисунок 3. Трек подспутниковой точки сетей мегагруппировок.

4.1.1. Межплоскостные переходы

Hh Число межплоскостных переходов зависит от разницы RAAN Y0 орбит, на которых расположены узлы источника и назначения. Учитывая пару спутников-источника и пункта назначения (s, d), Y0 рассчитывается следующим образом:

$$Y0 = (Y2 \quad Y1) \bmod 2\pi \quad [0, 2\pi] \tag{11}$$

где Y1 и Y2 — значения RAAN s и d соответственно. Если спутник назначения находится к западу от спутника-источника, то разница RAAN составляет 2π Y0. Со времени PAAH

Machine Translated by Google

разница между соседними орбитами постоянна (т. е. Ω = 2π/NP), межплоскостные скачки в западном и восточном направлениях можно выразить следующим образом:

$$Ч_{ч} = \frac{2\pi \quad Y0}{\Omega} \tag{12}$$

$$Ч_{ч} = \frac{Y0}{Ом} \tag{13}$$

где x = sgn(x) x + 1/2 обозначает целое число, ближайшее к x, H — количество прыжков между орбитальными плоскостями на запад, а H — количество прыжков между плоскостями на восток.

4.1.2. Внутриплоскостные

скачки Hv Внутриплоскостные скачки зависят от разности фазовых углов u спутника. Каждый скачок внутри плоскости увеличивает фазовый угол на Φ, а каждый скачок между плоскостями приводит к увеличению фазового угла на f . Следовательно, фазовый угол спутника назначения можно выразить следующим образом:

$$u2 = u1 + f \cdot H + \frac{\Phi \cdot Ч_{в}}{y} \tag{14}$$

где H — количество переходов внутри плоскости в восточном направлении. Чтобы вычислить Hv, необходимо сначала исключить разность фаз f · H , вызванную Hh, из разности фаз u = u2 u1, где u1 и u2 — аргумент значений широты спутников источника и назначения. , соответственно. Чтобы различать разность фаз из-за распространения на восток и на запад, u можно выразить отдельно следующим образом:

$$u = (u2 \quad u1 \quad H \cdot f) \bmod 2\pi \tag{15}$$

$$и = (и2 \quad и1 + H \cdot f) \bmod 2\pi \tag{16}$$

где f — изменение фазового угла из-за скачков между плоскостями. Как показано на рисунке 3, трасса подспутниковой точки разделена на восходящий сегмент (с юго-запада на северо-восток) и нисходящий сегмент (с северо-запада на юго-восток), что позволяет передавать данные в обоих направлениях. Следовательно, в данной работе рассчитываются внутриплоскостные прыжки в четырех направлениях, которые можно выразить следующим образом:

$$ЧАС_{в} = \frac{ты}{Ф} \tag{17}$$

$$ЧАС_{в} = \frac{y}{Ф} \tag{18}$$

$$Ч_{в} = \frac{2\pi \quad ты}{Ф} \tag{19}$$

$$Ч_{в} = \frac{2\pi \quad y}{Ф} \tag{20}$$

где u и u обозначают разность фаз из-за распространения на запад и восток , а H обозначают между соседними орбитами соответственно. Между тем, разность $_{в}$, $ЧАС_{в}$, $Ч_{в}$, $_{в}$ фаз H обусловлена скачками по орбите в сторону северо-запада, северо-востока, юго-запада и юго-востока соответственно.

Machine Translated by Google

Межплоскостные прыжки Hh и внутриплоскостные прыжки Hv были рассчитаны ранее. Следовательно, минимальное сквозное количество переходов в созвездии наклонной орбиты может быть равно выражается следующим образом:

$$Ч = \text{мин} \begin{matrix} Н_{час} + Ч_{в} \\ Н_{час} + Ч_{в} \\ Ч_{час} + Ч_{в} \\ Ч_{час} + Ч_{в} \end{matrix} \quad (21)$$

На основе получения кратчайшего пути с количеством переходов выбирается несколько доступных путей. генерируется. Приняв вычисление неоптимального пути между парами узлов (s, d) как Например, во-первых, частота занятости $F^{т}_{а, б}$ каждого канала и минимальное количество переходов набор путей $R^{т}_{а, б}$ вычисляются. При этом порог частоты занятости канала ξε равен определенный. Вес , м $_{а, б}$ [r1,r2] звена (a, b) устанавливается, если $F^{т}_{а, б}$ $_{а, б}$ $_{ξε}$; иначе, мт [r2,r3], где r1 < r2 < r3 — случайные числа. Чтобы предотвратить взаимосвязь несопоставимые пути, NCC устранит связи в $R^{т}_{а, б}$ из сети. Впоследствии NCC определит оптимальные возможные пути для (s, d) среди оставшихся ссылок, используя принцип наименьшего количества сквозных переходов. Вышеупомянутый процесс повторяется непрерывно до L возможных путей между любыми двумя узлами в сети были идентифицированы. Конкретный процесс MHMRD проиллюстрирован в алгоритме 1. вычислительная сложность алгоритма 1 равна O(N log N).

Алгоритм 1: Обнаружение многопутевого маршрута на основе минимального количества переходов (MHMRD).

Вход: {(V, E, T, N)}, пара узлов источник-назначение P.

Выход: Кандидатские пути R

```
1 Инициализация: R =
2 для i = 1: N делать
3     Вычислите путь с минимальным количеством переходов R.  т через уравнение (21);
4     Вычислить другие доступные пути;
5     для (s, d) в P do
6         для l = 1: L  1 делать
7             Вычислите частоту присутствия F  т  каждого ISL;
                                               а, б
8             для (a, b) в R  я  не делаю
9                 Ф т  Ф т  + 1;
                   а, б  а, б
10            конец
11            для (a, b) в E  т  делать
12                если Ф т  ξε , то
                       а, б
13                    м  [r1,r2];
                       а, б
14                конец
15                если Ф т  > ξε , тогда
                       а, б
16                    м  [r2,r3];
                       а, б
17                конец
18            конец
19            для e в R  т  делать
                         с, д
20                Удалите край с буквы E.  т ;
21            конец
22            Вычислите возможные пути R через путь с минимальным количеством переходов;
23        конец
24 конец
25 конец
```

Machine Translated by Google

4.2. Планирование многопутевого трафика

На основании завершения обнаружения многопутевого пути второй подзадачей многопутевой маршрутизации для MCN является планирование многопутевого трафика, которое можно выразить следующим образом.

$$\text{P2: макс } U\left(f_{л}, \bar{}_{д}, t \quad T\right)$$
$$\text{ул. } \sum_{л=1} f_{м,л} m = 1, \cdots, M \tag{22}$$
$$f_{\varepsilon j}^{т} \quad BW, \quad \varepsilon j \quad p_{м,л}^{т}$$

Поскольку решение о распределении трафика в спутниковых сетях зависит от текущего узла и состояний окружающих узлов в сети, независимо от исторического состояния, подзадачу планирования многопутевого трафика можно описать как марковский процесс принятия решений (MDP). Возникновение сбоев в ISL или узлах приводит к ухудшению функциональности спутниковой сети, что приводит к частым изменениям топологии сети. Учитывая, что GNN обладает превосходными возможностями обобщения для топологий различных размеров, в этой статье предлагается использовать планирование многопутевого трафика (GMTS) на основе GNN, как показано на рисунке 4. Агент DRL использует алгоритм оптимизации проксимальной политики (PPO), который представляет собой алгоритм актер-критик. На каждом этапе MDP модель актера выбирает решение по планированию трафика, а модель критика присваивает этому решению оценку на основе обратной связи с вознаграждением окружающей среды. Процесс повторяется непрерывно с целью оптимизации задачи планирования потока и максимизации совокупного вознаграждения.

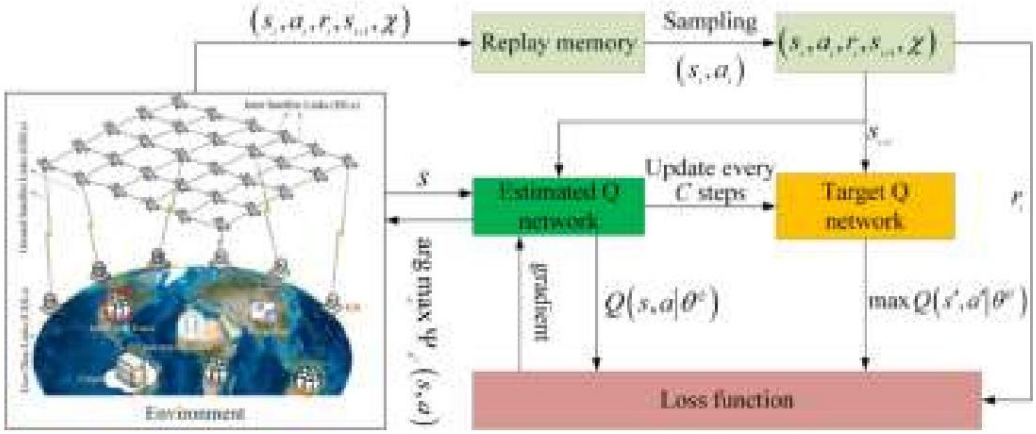


Рисунок 4. Структура предлагаемого алгоритма планирования многопутевого трафика (GMTS) на основе GNN.

4.2.1. Состояние

Разделив время на несколько временных интервалов, матрицу спроса на трафик в каждом временном интервале можно обозначить как TR = [b1, b2, · · ·, bM], где bi обозначает пропускную способность между i-й парой узлов. Согласно уравнению (5), пропускная способность j-го канала равна fj .
Следовательно, остаточная полоса пропускания i-го канала, обозначаемая ci , может быть выражена через полосу пропускания канала BW:

$$ci = BW \quad fi, \quad i \quad [1, L] \tag{23}$$

Учитывая, что максимальное количество переходов между парами узлов равно Hmax, j -й путь между i-й парой узлов можно обозначить как pi,j = {e1,e2, · · ·,eHmax }, а соответствующая матрица пути может быть равна обозначается как Pt = [p1,1, · · ·, p1,L, · · ·, pM,L]. Следовательно, вектор состояния во временном интервале t можно выразить следующим образом:

$$st = [Ct, TRt, Gt, Pt], st \quad S \tag{24}$$

где S — пространство состояний спутниковой сети.

Machine Translated by Google

4.2.2. Действие

На основе информации в пространстве состояний NCC разделяет сетевой трафик между каждой парой узлов на L возможных путей, которые вычисляются заранее. Следовательно, действие определяется как вектор , состоящий из доли трафика на путях-кандидатах, который можно выразить следующим образом:

$$\text{ttt at} = \omega \cdot \cdot \cdot, \omega \cdot \cdot \cdot, \omega\ 1,1, \quad 1,Л\ , \quad М,Л \tag{25}$$

где доля потока на каждом пути должна удовлетворять уравнению (2).

4.2.3. Награда

После выполнения действия в состоянии st в момент времени t агент получит обратную связь для оценки эффективности действия. При планировании многопутевого трафика вознаграждением является целевая функция подзадачи P2, которую можно выразить следующим образом:

$$\text{рт} = \sum_{м=1}^{М} \sum_{л=1}^{л} \text{ты}\ f_{m,l}\ ,\ \bar{}d_{m,l} \tag{26}$$

где $f_{m,l}$ — средняя пропускная способность на l-м пути между m-й парой узлов, а $\bar{}d_{m,l}$ — средняя задержка на l-м пути между m-й парой узлов. Агент будет стремиться завершить планирование многопутевого трафика с более высокой пропускной способностью и меньшей задержкой.
Поскольку принятие решения разделено на несколько этапов, совокупное вознаграждение Rt , основанное на определенном одном мгновенном вознаграждении, может быть выражено следующим образом:

$$\text{Рт} = \sum_{k=0}^{k} \rho\ r_{t+k+1} \tag{27}$$

где ρ [0, 1] — коэффициент дисконтирования. Учитывая, что в каждом состоянии существует несколько возможных действий , средний выигрыш для каждого состояния и средний выигрыш для каждого действия состояния можно выразить следующим образом:

$$\text{Гμ}(s) = \text{Eμ}(\ Gt\ |st = s) \tag{28}$$

$$\Psi\mu(s, a) = \text{Eμ}(\ Gt\ |st = s, at = a) \tag{29}$$

Здесь Ψμ(s, a) максимизируется, когда агент DRL находит оптимальную политику μ, , преобразующую процесс обновления Ψμ(s, a) с количеством итераций к форме Беллмана; другими словами, мы имеем

$$\Psi\mu(s, a) = \text{Eμ}\ r_t + \rho\Psi\mu(s_{t+1}, a_{t+1})\ s_t = s, a_t = a \tag{30}$$

После того как ДХО сходится к устойчивому уровню, соответствующее оптимальное действие aμ можно выразить следующим образом:

$$a\mu = \arg\max_a \Psi\mu\ (s, a) \tag{31}$$

4.3. Учебный процесс ГМТС

Псевдокод процесса обучения GMTS показан в алгоритме 2 с вычислительной сложностью O(L). Сначала требуется инициализация μ(s|θ μ) и Q s, a|θ Q , где θ μ и θ Q — гиперпараметры сети акторов и критической сети соответственно. На этапе обучения буфер воспроизведения опыта используется для хранения данных взаимодействия с окружающей средой, включая состояния, действия и награды. Чтобы предотвратить переобучение, сгенерированные выборки (st , at ,rt ,st+1, χ) сохраняются в буфере воспроизведения опыта на каждом этапе принятия решения. На этом этапе χ {0, 1} используется для определения того, завершено ли обучение модели на текущем этапе. Выборка (si , ai ,ri ,si+1 , χ) будет извлекаться через указанные интервалы времени в

Machine Translated by Google

в соответствии с порогом Bmax буфера воспроизведения с целью обучения сети.
На этом этапе целевое значение Q можно выразить следующим образом:

$$Q\ yi = ri + \rho(1 \quad \chi)\ Q\ si+1\ ,\ ai+1\ |\theta \tag{32}$$

Алгоритм 2: Процесс обучения многопутевому планированию трафика на основе GNN (GMTS).

Входные данные: Bmax, сеть критиков, сеть актеров и целевая сеть. Выходные данные: aμ

```
для i = 1: L do
    Сбросить начальное состояние
    как s0 для t = 1: T
        do Вычислить действие в , вознаградить rt и новое состояние st
        +1 Сохранение (st , at ,rt ,st+1, χ) в буфере
        воспроизведения. Обновить состояние st до st+1, пока
            |B|    Bmax do для j
                = 1: NB do Рассчитайте число pi в соответствии с уравнением (34)
                Рассчитайте ωi согласно уравнению (36)
                Рассчитайте целевое значение Q yi в соответствии с уравнением (32).
                Обновите вероятность перехода выборки в соответствии с уравнением
                  (35).

        конец Рассчитайте потери сети критиков и сети акторов согласно уравнениям (38) и
          (39)
        Обновите параметр сети актеров и сети критиков на основе градиентного нисходящего
          подхода. конец
конец
```

В алгоритме GMTS ошибка временной разницы (TD-ошибка) используется как средство оценки значимости переходов. Ошибка TD между целевым значением Q и текущим значением Q может быть выражена следующим образом:

$$\delta i \text{ знак равно } yi - Q\ si\ ,\ ai\ |\theta^{\ Q} \tag{33}$$

Соответствующая пропорциональная расстановка приоритетов равна

$$\text{пи} = \delta i + \eta \tag{34}$$

где η обозначает чрезвычайно малое количество, а η добавляется, чтобы образцы можно было отбирать, даже если ошибка TD равна нулю. Применяется метод стохастической выборки, и вероятность перехода выборки i равна

$$P(i) = \quad k \ \frac{\text{Пи}^{\ \alpha}}{p \qquad \alpha_{\text{к}}} \tag{35}$$

где α определяет степень использования приоритетов. Когда α = 0, метод выборки деградирует до случайной выборки. Веса выборки важности (IS) для (si , ai ,ri ,si+1 , χ) равны

$$\omega i = \quad \frac{\text{Пи})}{\text{Пмин}}^{\ \beta} \tag{36}$$

Machine Translated by Google

где β используется для смещенного и несмещенного управления. Учитывая скорость обучения ϑ Q, сеть актеров θ Q можно обновить следующим образом:

$$\theta Q \quad \theta Q + \vartheta Q \delta i \quad \theta Q \, Q \, s, a|\theta \qquad (37)$$

Следовательно, функцию потерь сети критиков и сети акторов можно выразить следующим образом:

$$Lq = \frac{1}{\text{Примечание}} \sum_{я=1}^{\text{Вмакс}} \omega i \delta \, я^{2} \qquad (38)$$

$$L\mu = \quad Q \, s, a|\theta \qquad (39)$$

где NB — размер буфера воспроизведения B.

4.4. Рабочий процесс

DMR. Многопутевую маршрутизацию на основе DRL (DMR), предложенную в этой статье, можно разделить на две фазы: обнаружение многопутевого пути на основе минимального числа переходов (MHMRD) и планирование многопутевого трафика на основе GNN (GMTS), как показано на рисунке 5. На первом этапе центр управления сетью сначала планирует несколько оптимальных путей между произвольными парами узлов на основе принципа минимального количества переходов в соответствии с топологией сети. В частности, на основе периодичности и предсказуемости спутниковых орбит сеть делится на несколько временных интервалов в соответствии с интервалами временных интервалов, и топология сети остается стабильной в каждом дискретном временном интервале. В каждом временном интервале MHMRD вычисляет L доступных путей с минимальным количеством переходов для каждой услуги на основе порядка приоритетов, определенного MHMRD на основе сквозного объема услуги. После завершения расчета для текущего временного интервала он переходит к следующему временному интервалу. На втором этапе центр управления сетью генерирует коэффициенты разделения для каждой услуги на разных путях на основе модели, полученной в результате обучения GMTS, и отправляет их соответствующим спутникам в подходящее время перед началом каждого временного интервала.

С одной стороны, в этой статье используется приоритетное воспроизведение опыта, чтобы ускорить сходимость модели путем выборки образцов из пула воспроизведения опыта в соответствии с определенным приоритетом. С другой стороны, из-за динамической природы спутниковых сетей топология сети и матрицы трафика изменяются во времени, и традиционные сети DNN не могут эффективно обрабатывать такие меняющиеся входные данные. Поэтому в этой статье мы принимаем GNN для структур графов переменного размера и соответственно конфигураций маршрутов. Поскольку GNN обладают лучшими возможностями обобщения для динамических сетевых структур разных масштабов, их можно использовать для агрегирования элементарных признаков без указания входных измерений.

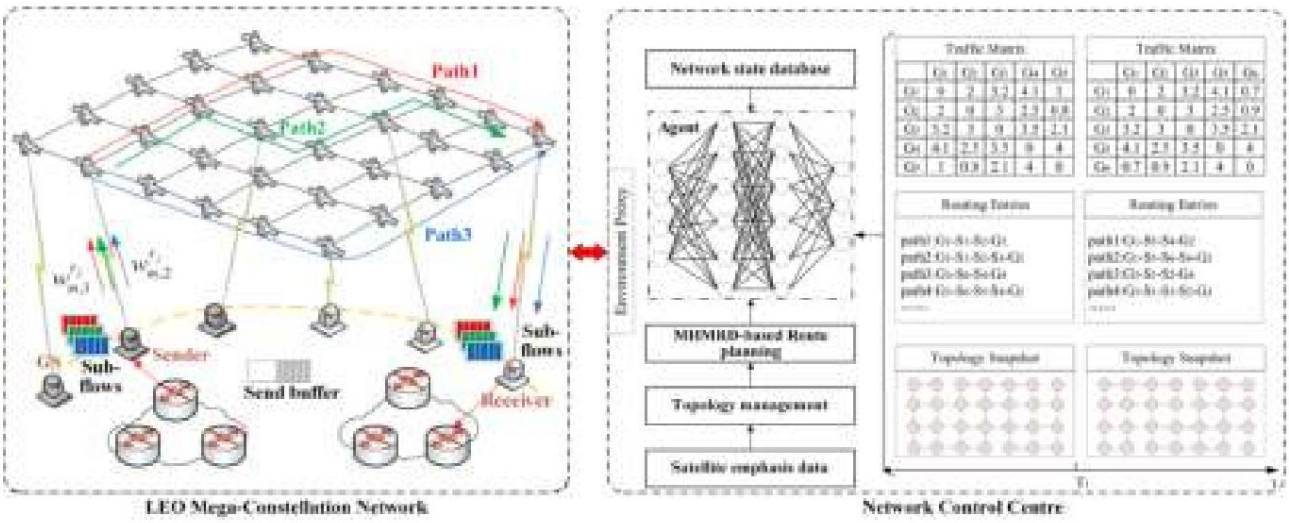


Рисунок 5. Многопутевая маршрутизация на основе DRL для сетей мегагруппировок LEO.

Machine Translated by Google

5. Оценка эффективности

5.1. Настройка моделирования

В этой статье мы реализуем вычисление маршрутизации и планирование многопутевого трафика для Мегасозвездия LEO в NS3 для анализа производительности предложенного алгоритма DMR. Чтобы проанализировать масштабируемость многопутевой маршрутизации в различных созвездиях, это В статье проводятся тесты на двух созвездиях разного размера. Первое созвездие было созвездие Иридиум [35] с 66 спутниками на полярной орбите, а вторым был Группировка OneWeb [36] с 648 спутниками, содержащая в общей сложности 18 орбитальных плоскостей. с 36 спутниками, равномерно распределенными в каждой орбитальной плоскости. Параметры двух Созвездия показаны в Таблице 1. Для проверки эффективности многолучевого распространения маршрутизации при различных объемах трафика, набор данных о трафике был создан на основе плотность наземного трафика, где количество пар узлов источник-назначение составляло 50. предложенный алгоритм GMTS был реализован на основе Python 3.9 и Pytorch 1.14, и соответствующие параметры приведены в таблице 2.

Таблица 1. Параметры созвездия.

Созвездие	НП	НС	НГ	ε	час	Э
Иридий	6	11	16	90	780 км	20
OneWeb	18	36	16	53	550 км	20

Таблица 2. Параметры обучения.

	Параметр	Ценить
Параметр сети	Нг	24
	Lq	100
	пак	1 КБ
Параметр обучения	Примечание	32
	Вмакс	100
	α	0,6
	β	0,5
	ρ	0,8

Производительность предлагаемого алгоритма обнаружения многопутевого пути (MHMRD) с минимальным количеством переходов, который распределяет сквозной трафик в равных пропорциях. для предварительного расчета нескольких путей было протестировано в первую очередь. На основе этого алгоритм GMTS обучался с использованием топологии созвездия наклонной орбиты и матриц трафика [37], которые были направлены в NCC после обучения и протестированы по ключевым показателям, таким как средний пропускная способность, задержка и скорость завершения потока спутниковой сети. Следующее В данной работе использовались алгоритмы сравнения:

- Алгоритм поиска кратчайшего пути (SPF): SPF использует среднюю задержку передачи в качестве метрика канала с целью минимизации общей задержки.
- Многолучевая кооперативная маршрутизация на основе сетевого кодирования (NCMCR) [24]: NCMCR предназначен для решения проблем маршрутизации, возникающих из-за часто меняющейся топологии. и потенциально редкое и прерывистое соединение. Алгоритм NCMCR принимает Преимущество предсказуемости относительного движения спутников с изменяющейся во времени топологией, моделируемой в виде пространственно-временной карты.
- Многопутевая магистральная маршрутизация на основе Ant для балансировки нагрузки (AMBRLB) [38]: по порядку Чтобы преодолеть проблемы переполнения трафика и накладных расходов, балансировка нагрузки на основе ant Предложен алгоритм многопутевой маршрутизации магистральной сети в MANET. При инициировании передачи от узла-источника к узлу назначения, оптимизация муравьиной колонии (ACO) алгоритм используется для идентификации нескольких путей с наибольшей вероятностью. успеха.

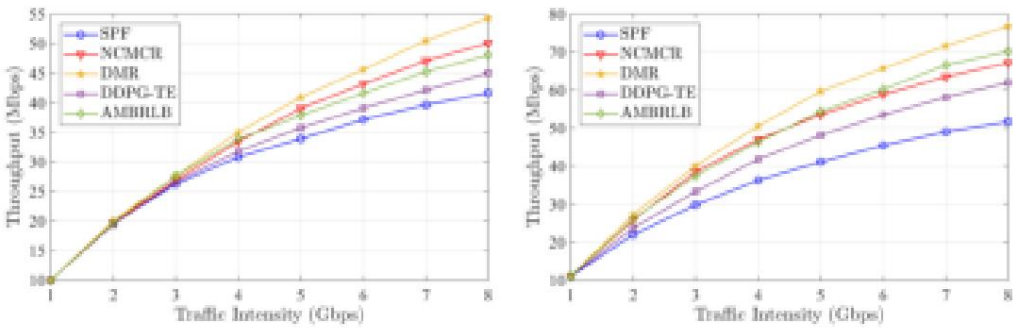
Machine Translated by Google

• Глубокая детерминированная политика градиентного проектирования трафика (DDPG-TE) [39]: этот метод использует алгоритм DDPG для динамического распределения пропорций трафика по различным путям передачи. Следует отметить, что модель неприменима к произвольным спутниковым группировкам, поскольку размеры пространств состояний и действий ограничены размером входной топологии и матриц трафика.

5.2. Результаты и анализ

5.2.1. Пропускная

способность Средняя пропускная способность MCN в группировках Iridium и OneWeb показана на рисунке 6, где каждая точка данных представляет собой среднее значение пропускной способности по пяти последовательным топологическим снимкам. Можно заметить, что в контексте многопутевой маршрутизации по мере увеличения объема входного трафика большая часть трафика распределялась по нескольким путям для пересылки на основе разных политик. Это привело к общему увеличению средней пропускной способности сети. Тем не менее, рост пропускной способности постепенно замедлился. Предложенный алгоритм DMR продемонстрировал превосходную пропускную способность при различных размерах пользователей. DMR увеличил среднюю пропускную способность на 30,26% и 8,45% по сравнению с SPF и NCMCR соответственно, когда общий объем услуг достиг верхнего предела. На рисунке 6b показана пропускная способность многопутевой маршрутизации в созвездии OneWeb. Можно заметить, что DMR использовала ту же модель, что и для группировки Iridium, в то время как другие алгоритмы были переобучены на основе характеристик новой сети. Можно заметить, что алгоритм DMR продолжал демонстрировать превосходную производительность по сравнению с другими эталонными алгоритмами в контексте изменений топологии и размера созвездия, что отражает улучшенную способность DMR к обобщению. При интенсивности трафика 8 Гбит/с DMR продемонстрировал улучшение средней пропускной способности на 42,64% и 9,55% по сравнению с SPF и NCMCR соответственно.



(а) (б)

Рисунок 6. Средняя пропускная способность на разных созвездиях с различной интенсивностью трафика. (а) Средняя пропускная способность для Iridium. (б) Средняя пропускная способность OneWeb.

5.2.2. Средний коэффициент завершения потока

На рисунке 7 показан средний коэффициент завершения потока для различной интенсивности трафика в группировках Iridium и OneWeb. По мере увеличения интенсивности трафика некоторые узлы сети постепенно оказывались перегруженными. Об этом свидетельствует тенденция снижения средних коэффициентов полноты всех пяти схем маршрутизации. При интенсивности трафика 8 Гбит/с средний коэффициент завершения потока предложенной схемы DMR увеличился на 17,39% и 11,52% на группировках Iridium и OneWeb соответственно по сравнению с S Между тем, в созвездии OneWeb средняя скорость завершения потока алгоритма DMR улучшилась на 1,98%, 2,58% и 6,49% по сравнению с NCMCR, AMBRLB и DDPG-TE соответственно. Следовательно, можно видеть, что средний коэффициент завершения потока в более крупной группировке OneWeb был постоянно выше, чем в меньшей группировке Iridium, даже при различной интенсивности трафика. Это демонстрирует преимущество крупномасштабных группировок при работе с высокими нагрузками. Таким образом, можно сделать вывод, что

Machine Translated by Google

предлагаемая схема DMR демонстрирует превосходную способность к обобщению для группировок различного масштаба с более высоким коэффициентом завершения потока, чем базовая схема.

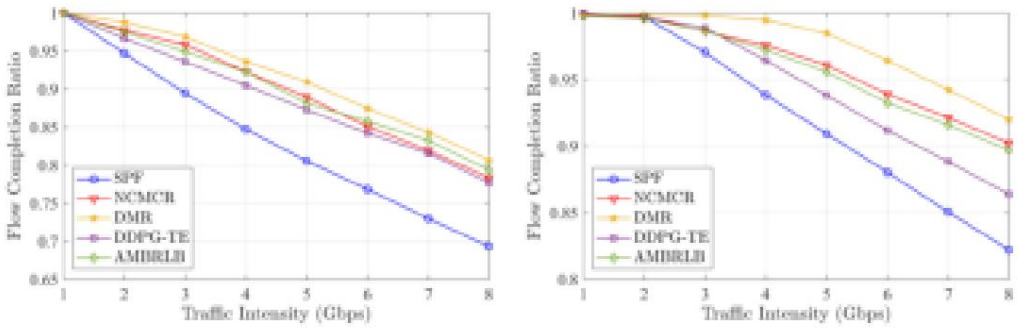


Рисунок 7. Средний коэффициент заполнения потока различных группировок с различной интенсивностью трафика. (а) Средний коэффициент завершения потока для созвездия Иридиум. (б) Средний коэффициент завершения потока для группировки OneWeb.

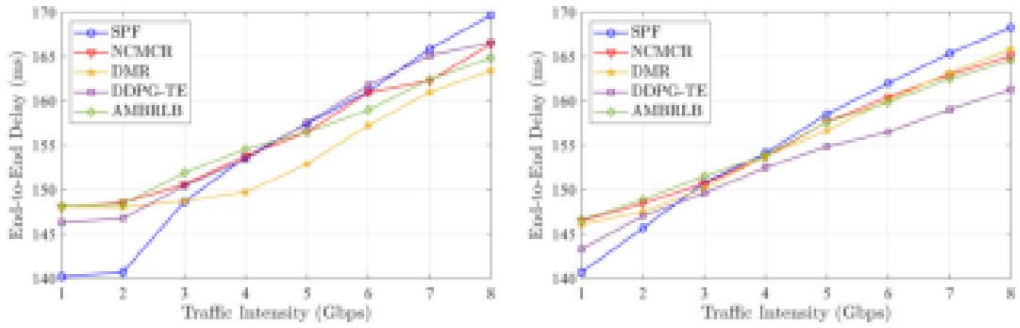
5.2.3. Средняя сквозная задержка На

рисунке 8 показана средняя сквозная задержка в зависимости от интенсивности трафика в созвездиях Iridium и OneWeb. По мере увеличения нагрузки на сеть задержка в очередях также возрастала, при этом по путям-кандидатам с большим количеством переходов передавалось больше потоков. Следовательно, средняя сквозная задержка постепенно увеличивалась для всех сравниваемых схем. В созвездии Iridium SPF достиг самой низкой сквозной задержки при низкой загрузке сети. Это связано с тем, что схема всегда передает пакеты по кратчайшему пути, когда все каналы сети находятся в нормальном состоянии. По мере увеличения нагрузки на сеть, когда общий объем обслуживания достигает верхней границы, SPF не может динамически изменять путь, чтобы избежать перегруженных путей, вызванных возросшей нагрузкой. Следовательно, в созвездии Iridium схема DMR снизила среднюю сквозную задержку на 3,67%, 1,78%, 1,86% и 0,86% по сравнению с SPF, NCMCR, DDPG- TE и AMBRLB соответственно. При распространении многопутевой маршрутизации на более крупные группировки OneWeb задержка для DMR приблизилась к задержке NCMCR, AMBRLB и DDPG-TE, когда интенсивность трафика достигла верхнего предела. Это связано с тем, что и схема DMR, и схема DDPG-TE учитывают динамическое распределение трафика по разным путям. В целом средняя сквозная задержка DMR при высокой нагрузке находилась между задержками SPF и DDPG-TE. Основная цель многопутевой маршрутизации — повысить пропускную способность и устойчивость сети при нарушении топологии. В небольших группировках многопутевая маршрутизация предлагает дополнительные возможности маршрутизации, которые можно использовать для своевременного обхода перегруженных узлов, тем самым уменьшая задержку.

Однако в крупномасштабных группировках всегда имеется достаточное количество альтернативных маршрутов, и улучшение показателей сквозной задержки незначительно.

Результаты экспериментов показывают, что схема SPF с кратчайшим путем демонстрирует меньшую среднюю задержку в сценариях с низкой нагрузкой. Однако SPF продемонстрировал самую низкую среднюю пропускную способность и средний коэффициент завершения потока, а производительность SPF быстро ухудшалась в условиях высокой нагрузки. Напротив, предложенная схема DMR продемонстрировала лучшую среднюю пропускную способность и коэффициент завершения потока при различных нагрузках, а также продемонстрировала лучшую способность к обобщению по созвездиям разных размеров и топологий без повторного обучения.

Machine Translated by Google



(а) (б)

Рисунок 8. Средняя сквозная задержка для разных группировок с различной интенсивностью трафика. (а) Средняя сквозная задержка для группировки Iridium. (b) Средняя сквозная задержка для группировки OneWeb.

6. Выводы

В этом документе предложено решение для спутниковых сетей на основе глубокого обучения (DRL) на основе многопутевой маршрутизации (DMR). Цель заключалась в решении проблемы многопутевой маршрутизации передачи в сетях мегагруппировок LEO. Предлагаемый подход DMR разделяет многопутевую маршрутизацию на две подзадачи: выбор многопутевого маршрута и планирование многопутевого трафика. В контексте подзадачи выбора многопутевого пути в этой статье предложена централизованная схема маршрутизации MHMRD, которая генерирует набор доступных путей для сквозных узлов. Это достигается путем сбора информации о состоянии узла на основе текущего минимального сквозного количества переходов и использования канала. В подзадаче планирования многопутевого трафика предлагается планирование многопутевого трафика на осн Каждый узел маршрутизации контролируется агентом DQN, а GNN используется для решения проблемы топологически различных топологий спутниковых сетей. Оптимальная модель планирования многопутевого трафика была получена посредством итеративного обучения. Предложенная схема DMR была окончательно проверена на мегасозвездиях двух разных размеров в среде мод Результаты моделирования показывают, что предложенная схема DMR продемонстрировала превосходные характеристики с точки зрения средней пропускной способности сети и среднего коэффициента завершения потока по сравнению с базовой схемой. Кроме того, схема DMR эффективно обеспечивает маршрутизацию спутниковой сети при балансировке нагрузки, сохраняя при этом минимальное увеличение затрат на задержку. Более того, схема DMR, разработанная в этой статье, продемонстрировала хорошую масштабируемость и адаптируемость в контексте развивающихся топологий группировок. Его также можно распространить на спутниковые сети с меняющейся топологией без необходимости дополнительного обучения.

Следует отметить, что по мере того, как размер группировки продолжает расширяться, вычислительные требования для онлайн-маршрутизации на основе DQN также возрастают. В будущем мы намерены изучить потенциал внедрения таких решений, как развертывание нескольких контроллеров, для повышения оперативности сети и дальнейшей проверки ее эффективности в реальных сцен

Вклад авторов: Концептуализация, CH; методология, CH; программное обеспечение, RY; проверка, WX; формальный анализ, CH; расследование, WX; ресурсы, WX; курирование данных, CH; письменное – подготовка первоначального черновика, CH; написание — рецензирование и редактирование, CH; визуализация, CH; надзор, WX; администрирование проекта, WX и RY; приобретение финансирования, WX и RY. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Данное исследование финансировалось проектом Национального фонда ключевых лабораторий (грант № 614201003022207) и Проектом новой инженерии по аэрокосмическим дисциплинам (грант № 145AXL250004000X).

Заявление о доступности данных: Доступ к соответствующим данным и исходному коду этого документа можно получить, обратившись по адресу hanchi@hgd.edu.cn.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Machine Translated by Google

Сокращения

В данной рукописи использованы следующие сокращения:

ДХО	Глубокое обучение с подкреплением
Многопутевая маршрутизация на основе DMR DRL	
MHMRD Алгоритм обнаружения многопутевой маршрутизации на основе минимального количества переходов	
Нейронная сеть GNN Graph	
Планирование многопутевого трафика на основе GMTS GNN	
Сеть MCN Megaconstellation	
Интегрированная сеть SAGIN «Космос-воздух-земля»	
ISL	Межспутниковая связь
SDN	Программно-определяемая сеть
RAAN Прямое восхождение восходящего узла	
НКЦ	Центр управления сетью
ЛЕО	Низкая околоземная орбита
Сеть DQN Deep Q	

Рекомендации

1. Цзян, Х.; Хуанг, Ю.; Ли, Дж.; Он, Х.; Чен, С.; Ян, Ф.; Ян, Дж. Пространственно-временная маршрутизация, избыточное кодирование и многопутевое планирование для детерминированной передачи по спутниковой сети. IEEE Транс. Коммун. 2023, 71, 2860–2875. [Перекрестная ссылка]
2. Кодхели О.; Лагунас, Э.; Матуро, Н.; Шарма, СК; Шанкар, Б.; Монтойя, JFM; Дункан, JCM; Спано, Д.; Хацинотас, С.; Киселев, С.; и другие. Спутниковая связь в новую космическую эпоху: обзор и будущие вызовы. IEEE-сообщество. Выж. Репетитор. 2021, 23, 70–109. [Перекрестная ссылка]
3. Рахманов А.; Уайзман, Ю. Сжатие данных GNSS с целью ускорения связи с автономными транспортными средствами. Дистанционное зондирование. 2023, 15, 2165. [CrossRef]
4. Коррейя, SD; Перес, Р.; Матуш-Карвалью, Ж.; Лейтхардт, VRQ µJSON, облегченная схема сжатия для встроенной передачи данных GNSS на узлах Интернета вещей. В материалах 5-й конференции по облакам и Интернету вещей (CIoT) 2022 г., Марракеш, Марокко, 28–30 марта 2022 г.; стр. 232–238.
5. Сюэ, К.; Ли, В.; Ю, Л.; Шан, Дж.; Чен, Х.; Лу, С. CEPO: Управляемая моделью структура бесшовного роуминга для беспроводной сети Mesh Сеть с многопутевым TCP. IEEE Транс. Коммун. 2019, 67, 1284–1296. [Перекрестная ссылка]
6. Лю, Х.; Ма, Т.; Цинь, Х.; Чжоу, Х.; Чжао, Л. Совместная многопутевая маршрутизация с поддержкой DRL для сверхплотных спутниковых сетей на низкой околоземной орбите. В материалах GLOBECOM 2023, Куала-Лумпур, Малайзия, 4–8 декабря 2023 г.; стр. 5961–5966.
7. Ли, Р.; Чжан, Дж.; Чжэн, С.; Ван, К.; Ван, П.; Чжан, Х. Алгоритм маршрутизации мега-созвездий LEO, основанный на сегментации областей. В материалах конференции IEEE по беспроводной связи и сетям 2023 г. (WCNC), Глазго, Великобритания, 26–29 марта 2023 г.; стр. 1–6.
8. Ли, Ю.; Чжан, К.; Яо, Х.; Гао, Р.; Синь, Х.; Ю, Ф.Р. Стигмергия и иерархическое обучение для оптимизации маршрутизации в многодоменных совместных спутниковых сетях. IEEE J. Выберите. Районы Коммун. 2024, 42, 1188–1203. [Перекрестная ссылка]
9. Перкинс, СЕ; Ройер, Э.М. Специальная дистанционная векторная маршрутизация по требованию. В материалах WMCSA'99. Второй семинар IEEE по мобильным вычислительным системам и приложениям, Новый Орлеан, Луизиана, США, 25–26 февраля 1999 г.; стр. 90–100.
10. Цзян Ф.; Чжан, К.; Ян, З.; Юань, П. Многопутевая маршрутизация на основе пространственно-временного графа в устойчивых к сбоям наблюдениях за Землей Спутниковые сети. IEEE Транс. Аэросп. Электрон. Сист. 2019, 55, 2592–2603. [Перекрестная ссылка]
11. Хуанг Ю.; Ян, Д.; Фэн, Б.; Тиан, А.; Донг, П.; Ю, С.; Чжан, Х. Алгоритм многопутевой маршрутизации с поддержкой GNN для пространственных Временные переменные спутниковые сети на околоземной орбите. IEEE Транс. Вех. Технол. 2024, 73, 5454–5468. [Перекрестная ссылка]
12. Хуан Ю.; Цзян, Х.; Чен, С.; Ян, Ф.; Ян, Дж. Интеллектуальный подход к планированию многопутевого трафика, стимулируемый феромонами Спутниковые сети LEO. IEEE Транс. Вирел. Коммун. 2022, 21, 5889–5902. [Перекрестная ссылка]
13. Тиан А.; Фэн, Б.; Чжоу, Х.; Хуанг, Ю.; Суд, К.; Ю, С.; Чжан Х. Эффективное объединенное совместное кэширование на основе DRL для мобильных устройств Граничные сети. IEEE Транс. Сеть. Серв. Менеджер. 2023, 20, 246–260. [Перекрестная ссылка]
14. Хуан, Х.; Юань, Т.; Цяо, Г.; Рен, Ю. Глубокое обучение с подкреплением для управления мультимедийным трафиком в программно определяемых системах. Сеть. Сеть IEEE. 2018, 32, 35–41. [Перекрестная ссылка]
15. Лей, К.; Го, П.; Ван, Ю.; Ву, Х.; Чжао, В. Решение проблем маршрутизации с помощью нейронной сети внимания с остаточным графом ребер. Нейрокомпьютинг 2022, 508, 79–98. [Перекрестная ссылка]
16. Чжэн, Х.; Хуанг, В.; Ли, Х.; Ли, Г. Исследование технологии обобщенной интеллектуальной маршрутизации на основе графовой нейронной сети. Электроника 2022, 11, 2952. [CrossRef]
17. Вэй, Х.; Чжао, Ю.; Сюй, К. G-Маршрутизация: гибкая онлайн-маршрутизация на основе графовых нейронных сетей. Сеть IEEE. 2023, 37, 90–96. [Перекрестная ссылка]
18. Русек, К.; Суарес-Варела, Дж.; Алмасан, П.; Барлет-Рос, П.; Кабельос-Апарисио, А. RouteNet: Использование графовых нейронных сетей для сетевого моделирования и оптимизации в SDN. IEEE J. Выберите. Районы Коммун. 2020, 38, 2260–2270. [Перекрестная ссылка]
19. Би, Ю.; Вешать.; Сюй, С.; Ван, Х.; Лин, К.; Ю, З.; Сан, П. Программно-конфигурируемые космически-наземные интегрированные сети: архитектура, Проблемы и решения. Сеть IEEE. 2019, 33, 22–28. [Перекрестная ссылка]

Machine Translated by Google

20. Ху, М.; Сяо, М.; Ху, Ю.; Кай, К.; Дэн, Т.; Пэн, К. Программно определяемая многоадресная рассылка с использованием сегментной маршрутизации в спутниковых сетях LEO. IEEE Транс. Моб. Вычислить. 2024, 23, 835–849. [Перекрестная ссылка]

21. Хан, З.; Сюй, К.; Чжао, Г.; Ван, С.; Ченг, К.; Ю, С. Модель изменяющейся во времени топологии для динамической маршрутизации в сетях группировок спутников на низкой околоземной орбите. IEEE Транс. Вех. Технол. 2023, 72, 3440–3454. [Перекрестная ссылка]

22. Джамбене, Г.; Луонг, Дания; Де Кола, Т.; Ле, Вирджиния; Мухаммад, М. Анализ подхода к блочному кодированию на уровне пакетов для Наземно-спутниковые мобильные системы. IEEE Транс. Вех. Технол. 2019, 68, 8117–8132. [Перекрестная ссылка]

23. Чен, З.; Чжоу, В.; Ву, С.; Ченг, Л. Адаптивный протокол многопутевой маршрутизации по требованию с поддержкой QoS для высокоскоростной сети MANET. IEEE Access 2020, 8, 44760–44773. [Перекрестная ссылка]

24. Тан, Ф.; Чжан, Х.; Ян, Л.Т. Многопутевая совместная маршрутизация с эффективным подтверждением для спутниковых сетей LEO. IEEE Пер. Моб. Вычислить. 2019, 18, 179–192. [Перекрестная ссылка]

25. Ли, Г.; Чжоу, Х.; Фэн, Б.; Чжан, Ю.; Ю. С. Эффективное обеспечение цепочек сервисных функций в наложенных сетях с использованием армирования. обучение. IEEE Транс. Облачные вычисления. 2022, 10, 383–395. [Перекрестная ссылка]

26. Като, Н.; Фадлулла, З.М.; Мао, Б.; Тан, Ф.; Акаши, О.; Иноуэ, Т.; Мизутани, К. Концепция глубокого обучения для управления гетерогенным сетевым трафиком: предложение, проблемы и перспективы на будущее. IEEE Вирел. Коммун. 2017, 24, 146–153. [Перекрестная ссылка]

27. Мао, Б.; Фадлулла, З.М.; Тан, Ф.; Като, Н.; Акаши, О.; Иноуэ, Т.; Мизутани, К. Маршрутизация или вычисления? Сдвиг парадигмы в сторону передачи пакетов в интеллектуальной компьютерной сети на основе глубокого обучения. IEEE Транс. Вычислить. 2017, 66, 1946–1960. [Перекрестная ссылка]

28. Гейер Ф.; Карл, Г. Обучение и создание протоколов распределенной маршрутизации с использованием глубокого обучения на основе графов. В материалах семинара 2018 г. по аналитике больших данных и машинному обучению для сетей передачи данных, Будапешт, Венгрия, 20 августа 2018 г.; стр. 40–45.

29. Лю, К.; Сюй, М.; Ян, Ю.; Гэн, Н. DRL-OR: Онлайн-маршрутизация на основе глубокого обучения для удовлетворения требований к различным типам услуг. В материалах конференции IEEE INFOCOM 2021 — конференции IEEE по компьютерным коммуникациям, Ванкувер, Британская Колумбия, Канада, 10–13 мая 2021 г.; стр. 1–10.

30. Алмасан, П.; Суарес-Варела, Дж.; Русек, К.; Барлет-Рос, П.; Кабельос-Апарисио, А. Глубокое обучение с подкреплением и графовые нейронные сети: изучение варианта использования оптимизации маршрутизации. Вычислить. Коммун. 2022, 196, 184–194. [Перекрестная ссылка]

31. Вс, П.; Го, З.; Ли, Дж.; Сюй, Ю.; Лан, Дж.; Ху, Ю. Включение масштабируемой маршрутизации в программно-конфигурируемых сетях с глубоким обучением с подкреплением на критических узлах. IEEE/ACM Транс. Сеть. 2022, 30, 629–640. [Перекрестная ссылка]

32. Ким, Г.; Ким, Ю.; Лим, Х. Маршрутизация на основе глубокого обучения в программно-определяемых сетях. IEEE Доступ 2022, 10, 18121–18133. [Перекрестная ссылка]

33. Цзян, З.; Ву, К.; Ли, Х.; Ву, Дж. scMPTCP: Многопутевая передача с поддержкой SDN для спутниковой сети с распознаванием нагрузки. Доступ IEEE 2018, 6, 19823–19832. [Перекрестная ссылка]

34. Де Сантис, Э.; Джузеппи, А.; Пьетрабисса, А.; Каппони, М.; Делли Присколи, Ф. Интеграция спутников в 5G: глубокое усиление Обучение выбору сети. Max. Интел. Рез. 2022, 19, 127–137. [Перекрестная ссылка]

35. Созвездие Иридиум. Доступно онлайн: https://celestrak.org/NORAD/elements/table.php?GROUP=iridium. (доступ 30 июль 2024 г.).

36. Созвездие Oneweb. Доступно онлайн: https://celestrak.org/NORAD/elements/table.php?GROUP=oneweb. (доступ 30 июль 2024 г.).

37. Архив дорожного движения Рабочей группы МАВИ. Доступно онлайн: https://mawi.wide.ad.jp/mawi/ (по состоянию на 30 июля 2024 г.).

38. Фрэнсис Энтони Селви, П.; Маникандан, Многопутевая магистральная маршрутизация на основе MSK Ant для балансировки нагрузки в MANET. ИЭПП Коммун. 2017, 11, 136–141. [Перекрестная ссылка]

39. Чен, Б.; Сан, П.; Чжан, П.; Лан, Дж.; Купить.; Шен, Дж. Организация трафика на основе глубокого обучения с подкреплением в гибридной сети IP/SR. Китайская коммуна. 2021, 18, 204–213. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Проектирование компенсатора помех на основе расширенного LMI для Линейные системы с изменяющимися параметрами

Мустафа Инчи 1 и Юсуф Алтун 2,*

1 Центр дизайна Volta Motor Industry and Trade Inc., завод Кумайери, 81700 Дюздже, Турция; mustafa@volta.com.tr

2 Факультет компьютерной инженерии, Университет Дюздже, 81620 Дюздже, Турция * Адрес для переписки: yusufaltun@duzce.edu.tr

Аннотация: В этой статье представлено новое представление расширенного линейного матричного неравенства (LMI) для проектирования контроллера с обратной связью по состоянию и динамического компенсатора возмущений с прямой связью для систем с линейным изменением параметров (LPV). Улучшенные LMI являются выпуклыми и конечномерными без какого-либо итерационного подхода. Проекты основаны на новой предложенной эквивалентной ограниченной вещественной лемме (BRL) посредством расширения матрицы для систем LPV и неопределенных линейных систем в условиях изменяющихся во времени параметрических неопределенностей (TVPU). Этот расширенный BRL дает более низкие консервативные результаты, чем существующие методы, с точки зрения надежной стабильности. Соответственно, в дополнение к регулятору с обратной связью по состоянию разрабатывается динамический компенсатор возмущений . В этой статье основное внимание уделяется разработке компенсаторов помех в дополнение к разработке контроллеров с обратной связью по состоянию. Динамические матрицы компенсатора изменяются в зависимости от изменяющихся во времени параметров LPV или неопределенной системы в процессе эксплуатации, при условии, что возмущения и параметры измеримы или наблюдаемы. Компенсатор может быть спроектирован так, чтобы ослаблять помехи/шумы или улучшать о Наконец, представлены численные результаты и результаты моделирования, подтверждающие как эффективность, так и меньшую консервативность предлагаемых LMI.

Ключевые слова: прямая связь LPV; компенсатор помех; контроль H ; расширенный БРЛ; надежный контроль

check for updates

Цитирование: ¨Инчи, М.; Алтун, Ю. Конструкция компенсатора помех На основе расширенного LMI для линейного Системы с изменяющимися параметрами. Электроника 2024, 13, 3055. https://doi.org/10.3390/электроника13153055

Академический редактор: Лука Патане

Поступила: 5 июля 2024 г.
Пересмотрено: 26 июля 2024 г.
Принято: 29 июля 2024 г.
Опубликовано: 1 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/ 4.0/).

1. Введение

Надежный анализ производительности систем LPV, которые представляют собой особый класс нелинейных систем, и линейных систем с изменяющимися во времени параметрическими неопределенностями (TVPU) привлек значительное внимание. За последние два десятилетия для этих систем было предложено множество LMI, основанных на критериях устойчивости Ляпунова, поскольку эффективность методов на основе LMI хорошо известна, например [1–4]. Большинство из них основаны на стандартной квадратичной функции Ляпунова, известной как квадратичная устойчивость, где используется одна квадратичная функция Ляпунова, как в [5]. Например, численное сравнение проводится с точки зрения консервативности квадратичных условий устойчивости в [6]. Однако многие исследования показали, что зависящая от параметра матрица Ляпунова (PDLM) преодолевает консервативность, полученную из одной квадратичной функции Ляпунова для систем, подобных системам, описанным в [6–13]. Однако это не подходит для систем LPV или системы непрерывного времени (CTS) под TVPU без дополнительного метода, поскольку PDLM или параметризованный LMI (PLMI) вызывает проблему невыпуклой оптимизации. Однако PDLM уменьшает консерватизм, вытекающий из условий квадратичной устойчивости. Для решения этой проблемы были предложены некоторые подходы, такие как метод сетки [14], s-процедура [15], метод мультивыпуклости [15,16] и методы из [17,18]. Тем не менее, сеточный метод справедлив только для локальных условий эксплуатации, противоречащих всем допустимым траект Кроме того, другие подходы включают дополнительные условия LMI, некоторые ограничения или сложность помимо существующих LMI для получения выпуклого решения при использовании PDLM. Кроме того, эти подходы не позволяют разработать многоцелевой контроль.

Machine Translated by Google

в [12]. Более того, контроллер с обратной связью по состоянию, спроектированный с использованием этих методов, зависит от функции Ляпунова, ограниченной свойством симметричности.

В последнее время, чтобы обеспечить меньшую консервативность для неопределенных и LPV-систем, матрицы Ляпунова в зависимости от параметров или PLMI исследуются без дополнительных условий LMI. Соответственно, для решения невыпуклых проблем были предложены «расширенные» или «расширенные» LMI, которые стали очень популярны в течение последних двух десятилетий. S-переменные и LMI с слабой переменной (SV-LMI) также известны, поскольку для LMI обычно используются слабые переменные [19]. Преимущество расширенных LMI заключается в том, что более низкие консервативные результаты получаются для очень больших наборов параметров, поскольку они не содержат умножения матрицы Ляпунова на матрицы пространства состояний системы в отличие от обычных LMI, как в [5]. Этот случай также позволяет нам проверить робастную работу или стабильность, и благодаря разделению для систем получается выпуклая формулировка LMI. Кроме того, расширенные LMI оказались ценными инструментами для разработки многоцелевых контроллеров с меньшей консервативностью, как в [20]. Короче говоря, подход с расширенным LMI представляет собой выпуклое решение и менее консервативные результаты для LPV или неопределенных линейных систем без каких-либо дополнительных условий, несмотря на PDLM, что приводит к невыпуклости или бесконечным LMI в обычных конструкциях. Например, условие устойчивости предложено в [8,21], но конструкция управления H не рассматривается. В [20] представлено очевидное преимущество расширенных LMI перед классическими LMI при проектировании многокритериальных регуляторов H2/D-стабильности, но эта формулировка не может быть распространена на синтез управления H . В [22] предлагается управление H2 через резервную переменную, но управление H не рассматривается. В [23] предложены более низкие консервативные дилатационные условия LMI при рационально варьируемых параметрах, но они не подходят для CTS с системами TVPU и LPV. Помимо этих исследований, улучшения управления с обратной связью по состоянию в [24–31] с использованием расширенных LMI более примечательны в литературе с точки зрения консервативности и/или простоты. Эти исследования менее консервативны, чем стандартная конструкция контроллера с обратной связью по состоянию в [1]. Однако ни один из них не полностью подходит для систем LPV и CTS в условиях TVPU, тогда как они улучшены для CTS в условиях неизменяемых во времени параметрических неопределенностей (TIPU). Кроме того, предлагаемый расширенный LMI превосходит TIPU с точки зрения консервативности. Соответственно, данная конкретная тема до сих пор остается открытой в литературе, особенно для КТС с системами ТВПУ и ЛПВ. Таким образом, первый вклад этой статьи заключается в том, что новый расширенный BRL предлагается для систем LPV и CTS с TVPU. Впоследствии это расширяется для решения проблемы управления H с обратной связью по состоянию .
Поэтому менее консервативные результаты получены для систем LPV или CTS с ТВПУ согласно существующим в литературе результатам.

Компенсаторы или контроллеры прямой связи используются для отслеживания опорного сигнала или для ослабления внешних или внутренних возмущений, которые в основном предсказуемы или измеримы, как в [32,33]. Производительность системы улучшается за счет дополнительного формирования управляющего сигнала в соответствии с измеренными значениями возмущения или эталонными значениями. В частности, для многих систем реализованы компенсаторы возмущений, например [34]. Компенсатор возмущений используется вместе с контроллером с обратной связью, как в [35,36], поскольку конструкция с обратной связью гарантирует стабильность объекта. Однако компенсатор не обеспечивает устойчивости неустойчивой системы. Есть два основных метода. Во-первых, они разрабатываются одновременно , как в [37,38]. Во втором случае определенная производительность предварительно достигается с помощью контроллера обратной связи, а затем часть прямой связи проектируется, как в [39]. Данная статья основана на втором подходе. Для систем LTI, таких как те, что использовались в приведенных выше исследованиях, было представлено множество компенсаторов или контроллеров с прямой связью . Имеется несколько работ по методу H , особенно для систем LPV, например [40–45]. В [40] для устранения возмущений вибрационной системы был разработан H - регулятор пониженного порядка.
В [41] конструкция динамического контроллера с прямой связью H была рассмотрена путем получения обратной системы. В [42] переключаемое устойчивое управление с прямой связью было представлено для класса неопределенных линейных систем, определенных в классической форме дробно-линейного преобразования. Однако ни один из них не подходит для систем LPV. В [43,44] используются статические конструкции с прямой связью LPV , но они не являются динамическими контроллерами. Динамический контроллер с прямой связью был разработан для систем LPV в [45], но существует множество LMI, поскольку некоторые дополнительные LMI

Machine Translated by Google

условия необходимы из-за невыпуклости использования PDLM. В дополнение
LMI основаны на общем BRL, поэтому LMI являются несколько консервативными из-за
умножение матрицы Ляпунова. Поэтому упреждающий регулятор или компенсатор на основе коэффициента
усиления L2 редко обсуждается, особенно для систем LPV. Более того,
конструкция динамического упреждающего регулятора или компенсатора не использовалась в вышеизложенном
исследования, достигающие преимуществ расширенного LMI. Поэтому второй вклад
В данной статье используется динамический компенсатор прямой связи LPV, основанный на коэффициенте усиления L2.
конструкция контроллера для достижения преимуществ расширенного подхода LMI, который, согласно
к литературному обзору до сих пор не сообщалось в литературе. В [45], которая ближе
В данной статье исследование является довольно сложным из-за возникновения бесконечных LMI. В
Кроме того, конструкция основана на стандарте BRL, который дает более консервативные результаты.

Короче говоря, в этом исследовании предлагаются решения на основе неитеративных LMI для разработки
контроллер обратной связи по состоянию LPV и динамический компенсатор прямой связи LPV. Они есть
удобен как для ЛПВ, так и для КТС с ТВПУ (линейными системами с параметрической неопределенностью).
Основные положения статьи заключаются в следующем:

(1) Расширенный BRL на основе неитеративного LMI предлагается для CTS с TVPU и
Система LPV, обеспечивающая преимущества слабой переменной и расширенной LMI. Затем
новый контроллер с обратной связью по состоянию создан на основе нового эквивалента БРЛ. С
По своему превосходству он имеет лучший результат среди недавних исследований в [5,24–30] в
считают, что он менее консервативен, как показано в Таблице 1, и примерно 10%
улучшение было достигнуто от лучшего результата среди них.

(2) Все предлагаемые решения не требуют каких-либо дополнительных условий LMI, как в [45] , ввиду
к предлагаемому расширенному BRL, даже если используется PDLM, что приводит к невыпуклости или
бесконечности LMI. Они проще и менее консервативны, чем обычные.
подходы в литературе и менее консервативны, чем аналогичные исследования в литературе, что
доказано численными тестами. Таким образом, LMI обеспечивают как простоту, так и
меньшая консервативность.

(3) Предлагается динамический компенсатор прямой связи LPV на основе расширенного BRL.
для систем LPV с измеримыми или наблюдаемыми помехами или шумом. Поэтому,
компенсатор работает онлайн с изменяющимися во времени параметрическими неопределенностями
неопределенные системы или изменяющиеся во времени параметры систем LPV.

Таблица 1. Сравнение достигнутых γmin.

	[5]	[24]	[25]	[26]	[27]	[28]	[29]	[30]	Предлагаемая теорема 1
γmin	1,557	1,498	1,478	1,476	1,251	1,241	1,240	1,241	1.118

2. Предварительные сведения

В этом разделе даются некоторые определения и леммы, проливающие свет на предлагаемую
леммы и теоремы.

Определение 1. Система, описанная в (1), называется системой LPV. Здесь x(t) $\Re^{nw}$ представляет собой
состояния, u(t) $\Re^{nu}$ представляет собой входные сигналы управления, z(t) $\Re^{nz}$ представляет собой управляемые
выходы, y(t) $\Re^{ny}$ представляет измеренные выходы, ω(t) $\Re^{nw}$ представляет помехи
и все матрицы имеют соответствующий размер и зависят от параметров α , изменяющихся со временем.
Наборы параметров определяются формулой (2) где R и D обозначают наборы параметров для α и
его производная соответственно. Кроме того, матрица A(α) является матрицей состояний, которая определяет
связь между текущим состоянием и скоростью его изменения, B1(α) — входная матрица возмущения , B2(α) —
входная матрица управления, C1(α) — матрица, связывающая вектор состояния x(t) с тот
управляемые выходы z(t), D11(α) — матрица прямой связи для входного возмущения ω(t) непосредственно
связывающий вход возмущения ω(t) с управляемым выходом z(t), D12(α) представляет собой матрицу прямой связи
для управляющего входа u(t), непосредственно связывающего управляющий вход с управляемым выходом z(t), C2(α) равна
матрица измеренного выхода, определяющая взаимосвязь между состоянием и измеренным выходом.

Machine Translated by Google

Положим, D21(α) — это матрица прямой связи для входа возмущения, непосредственно связывающая вход возмущения ω(t) с измеренным выходным сигналом y(t) , а D22(α) — матрица прямой связи для управляющего входа u(t), напрямую связывающая управляющий вход к измеряемому выходу y(t).

$$\begin{aligned} \dot{x}(t) &= A(\alpha(t))x(t) + B1(\alpha(t))\omega(t) + B2(\alpha(t))u(t) \\ z(t) &= C1(\alpha(t))x(t) + D11(\alpha(t))\omega(t) + D12(\alpha(t))u(t) \\ y(t) &= C2(\alpha(t))x(t) + D21(\alpha(t))\omega(t) + D22(\alpha(t))u(t) \end{aligned} \tag{1}$$

$$R = \{\alpha \in \Re^n : \alpha_i \in [\underline{\alpha}_i, \overline{\alpha}_i], \ i = 1, \ldots, n\}, \quad D = \{\alpha \in \Re^n : \alpha_i \in \{\underline{\alpha}_i, \overline{\alpha}_i\}, \ i = 1, \ldots, n\} \tag{2}$$

Кроме того, замкнутую форму системы можно записать как (3), где Z обозначает выходы z и y. Кроме того, матрицы A, B, C и D рассматриваются как матрицы замкнутого цикла.

$$\begin{aligned} \dot{x}(t) &= A(\alpha(t))x(t) + B(\alpha(t))\omega(t) \\ Z(t) &= C(\alpha(t))x(t) + D(\alpha(t))\omega(t) \end{aligned} \tag{3}$$

Предполагается, что параметры доступны онлайн (измеримы или вычислимы). Кроме того, параметры получаются из состояний нелинейной системы, хотя обычно она выглядит как неопределенная система. Если параметры не выводятся из состояний, в отличие от параметров системы LPV, (1) представляет собой линейную систему с изменяющимися во времени неопределенностями, как в примере 1.

Норму H_∞ для системы (3) можно вычислить с помощью общепринятой формулировки LMI, заданной леммой 1, которая в литературе хорошо известна как BRL. Соответственно, если $\gamma > 0$ — скаляр γ, мы определяем показатель эффективности по формуле (4), где $\omega(t) \in L_2^n[0, \infty)$ положительный экзогенное возмущение.

$$J(\omega, z) = \int_0^\infty z^T z - \gamma^2 \omega^T \omega \, d\tau \tag{4}$$

Лемма 1. Обозначим через G(α) устойчивую трансфер-матрицу системы ЛПВ в (3) для всех допустимых параметрические траектории. Тогда и только $\gamma^2 I - D^T(\alpha)D(\alpha) > 0$, $G(s)_\infty < \gamma$ и следующих $\alpha \in D$. тогда, когда γ- утверждения эквивалентны для всех $\alpha \in R$ и

(я) Если существует положительно определенная дифференцируемая симметричная матрица $P(\alpha) \in \Re^{n \times n}$, затем

$$\begin{aligned} &A^T(\alpha)P(\alpha) + P(\alpha)A(\alpha) + \dot{P}(\alpha) + C^T(\alpha)C(\alpha) \\ &+ (B^T(\alpha)P(\alpha) + D^T(\alpha)C(\alpha))^T (\gamma^2 I - D^T(\alpha)D(\alpha))^{-1} (B^T(\alpha)P(\alpha) + D^T(\alpha)C(\alpha)) < 0 \end{aligned} \tag{5}$$

(ii) Если существует положительно определенная дифференцируемая симметричная матрица $P(\alpha) \in \Re^{n \times n}$, затем

$$\begin{bmatrix} A^T(\alpha)P(\alpha) + P(\alpha)A(\alpha) + \dot{P}(\alpha) & P(\alpha)B(\alpha) & C^T(\alpha) \\ B^T(\alpha)P(\alpha) & -\gamma^2 I & D^T(\alpha) \\ C(\alpha) & D(\alpha) & -I \end{bmatrix} < 0 \tag{6}$$

Замечание 1. LMI, подобные приведенным в (6), неудобны для проектирования регуляторов для линейных систем с неопределенными параметрами политопного или аффинного типа или систем LPV, поскольку неравенство предполагает умножение матрицы Ляпунова на матрицы объекта. Следовательно, это дело приводит к невыпуклой оптимизации и бесконечным LMI. Тем не менее, были предложены некоторые методы устранения этой трудности, такие как метод сетки [14], s-процедура [15], метод многовыпуклости [11] и методы в [17,18]. Однако метод многовыпуклости и s-процедура приводят к множеству LMI, а метод сетки применим только для локальных условий эксплуатации в отличие от траекторий всех допустимых параметров. Кроме того, такие методы не позволяют спроектировать многоцелевой контроллер в [12]. Использование параметрически-независимой (фи

Machine Translated by Google

в (7), как и в [5,41,46], является обычным явлением, но это приводит к весьма консервативным результатам. И наоборот, когда функция Ляпунова изолируется от матриц объекта с помощью слабой переменной, что обычно называют расширенным или расширенным подходом LMI из-за увеличения размерности LMI, это позволяет уменьшить консерватизм для анализа и проектирования, поскольку в [23,24,27,31]. Следовательно, достижение расширенного состояния, эквивалентного BRL, является эффективным подходом к преодолению этой сложности в конструкциях устойчивого управления, таких как H2 и H . Соответственно, в лемме 3 для проектирования контроллера было предложено новое обозначение для вычисления H , включая TVPU.

Стандартный LMI в (7), полученный из леммы 1, обычно используется для L2-усиления для систем или конструкция H (K(α) = R(α)P ^{1}LPV в литературе [5].

$$\begin{matrix} A^{T}(\alpha)P + B2R(\alpha) + (\quad)^{T} & B1(\alpha)\ \text{ПКТ} & {}_{1}(\alpha) + P & T(\alpha)DT_{12} \\ \text{я} \quad \gamma^{2} & DT_{11}(\alpha) & \text{Я} & \end{matrix} \qquad 0 \tag{7}$$

Следующая лемма 2, основанная на лемме о проекции, полезна для вывода расширенные LMI.

Лемма 2. [21]: Если существуют симметричная матрица Ω и совместимые размерные матрицы Θ1, Θ2, то следующие условия эквивалентны:

я. Ω < 0 и Ω + Θ1ΘT + Θ2ΘT ii. Следующая $_{1}$ < 0
$_{2}$
задача LMI разрешима в соответствии с Q.

$$\begin{matrix} \text{Ом} & \Theta 2 + \Theta 1Q \\ \Theta T_{2} + QT\Theta T_{1} & Q \quad QT \end{matrix} \qquad 0$$

3. Новая лемма о расширенном ограниченном действительном веществе.

В этом разделе предлагается расширенный BRL для LPV и неопределенных систем. Соответственно, решение на основе LMI улучшается для разработки контроллера с обратной связью по состоянию H . Лемма 3 показывает расширенный BRL.

Лемма 3. Пусть G(α) обозначает трансфер-матрицу системы LPV в (3), тогда следующие два утверждения эквивалентны.

я. Тогда и только тогда, когда матрица A(α) устойчива для всех допустимых траекторий параметров и γ^{2} DT (α)D(α) I 0, производительность H для системы в (3) ограничена так, что индекс γ > 0. Таким образом, Gzω γ.

ii. Если существует положительно определенная, зависящая от параметра дифференцируемая симметричная матрица P(α) $\Re^{n\times n}$, вещественная положительная константа μ, удовлетворяющая условию P(α) 2μP(α) 0 , и совместимая размерная матрица Q, то задача LMI в (8) справедлива для всех α R и α D.

$$\begin{matrix} \dot{P}(\alpha) \quad 2\mu P(\alpha) & {}_{\text{я}}^{0} & P(\alpha) + A^{T}(\alpha)Q + \mu Q & C^{T}(\alpha) \\ & {}_{\gamma^{2}} & \text{Б}^{T}(\alpha)Q & DT(\alpha) \\ & & Q \quad QT & 0 \\ & & & \text{Я} \quad {}_{0} \end{matrix} \tag{8}$$

Доказательство леммы 3. ЛПВ-форма неравенства в утверждении i леммы 1 такова.

$$\begin{matrix} A^{T}(\alpha)P(\alpha) + P(\alpha)A(\alpha) + & \dot{P}(\alpha) + C^{T}(\alpha)C(\alpha) \\ + \text{Б}^{T}(\alpha)P(\alpha) + DT(\alpha)C(\alpha) & {}_{T\ \gamma^{2}}\text{Я} \quad DT(\alpha)D(\alpha) \end{matrix} \quad {}^{1}\ \text{Б}^{T}(\alpha)P(\alpha) + DT(\alpha)C(\alpha) \quad 0 \tag{9}$$

Machine Translated by Google

Неравенство в (9) модифицируется следующим образом, если в неравенство включить действительную постоянную переменную μ .

$$\begin{matrix} A^{T}(\alpha) + \mu I\ P(\alpha) + P(\alpha)(A(\alpha) + \mu I) + C^{T}(\alpha)C(\alpha) + \dot{P}(\alpha) \quad 2\mu\kappa P(\alpha) \\ + B^{T}(\alpha)P(\alpha) + D^{T}(\alpha)C(\alpha) \quad \gamma^{2} I \quad D^{T}(\alpha)D(\alpha) \quad 1 \quad B^{T}(\alpha)P(\alpha) + D^{T}(\alpha)C(\alpha) \quad 0 \end{matrix}$$

С помощью леммы Шура новое матричное неравенство получается следующим образом.

$$\begin{matrix} \Gamma 11(\alpha) & \Gamma 12(\alpha) \\ \Gamma 13(\alpha) & \Gamma 14(\alpha) \end{matrix} \quad 0$$

где

$$\begin{aligned} \Gamma 11(\alpha) &= A^{T}(\alpha) + \mu I\ P(\alpha) + P(\alpha)(A(\alpha) + \mu I) + C^{T}(\alpha)C(\alpha) + \dot{P}(\alpha) \quad 2\mu\kappa P(\alpha) \\ \Gamma 12(\alpha) &= P(\alpha)B(\alpha) + C^{T}(\alpha)D(\alpha), \\ \Gamma 13(\alpha) &= B^{T}(\alpha)P(\alpha) + D^{T}(\alpha)C(\alpha) \\ \Gamma 14(\alpha) &= D^{T}(\alpha)D(\alpha) \quad \gamma^{2} I \end{aligned}$$

Мы можем переписать неравенство следующим образом.

$$\begin{matrix} A^{T}(\alpha) + \mu I\ P(\alpha) + \begin{pmatrix} \\ B^{T}(\alpha)P(\alpha) \end{pmatrix}^{T} P(\alpha)B(\alpha)\ 0 \\ + \\ C^{T}(\alpha)C(\alpha) + \dot{P}(\alpha) \quad 2\mu\kappa P(\alpha) \quad C^{T}(\alpha)D(\alpha) \\ D^{T}(\alpha)C(\alpha) \quad D^{T}(\alpha)D(\alpha) \quad \gamma^{2} I \end{matrix} \quad 0 \qquad (10)$$

Это переписывается следующим образом.

$$\begin{matrix} C^{T}(\alpha)C(\alpha) + \dot{P}(\alpha) \quad 2\mu\kappa P(\alpha) & C(\alpha)\ TD(\alpha) \\ D^{T}(\alpha)C(\alpha) & D^{T}(\alpha)D(\alpha) \quad \gamma^{2} I \end{matrix} + \begin{matrix} A^{T}(\alpha) + \mu\kappa I \\ B^{T}(\alpha) \end{matrix} P(\alpha)\ 0 + \begin{matrix} P(\alpha) & A(\alpha) \\ 0 & +\mu\kappa I \end{matrix} B(\alpha) \quad 0 \qquad (11)$$

Далее мы можем получить неравенство (12), если применим лемму 2 к (11).

$$\begin{matrix} \varphi 11(\alpha) & \varphi 12(\alpha) & \varphi 13(\alpha) \\ & \varphi 22(\alpha) & \varphi 23(\alpha) \\ & & \varphi 33 \end{matrix} \quad 0 \qquad (12)$$

где

$$\begin{aligned} \varphi 11(\alpha) &= C^{T}(\alpha)C(\alpha) + \dot{P}(\alpha) \quad 2\mu P(\alpha), \\ \varphi 12(\alpha) &= C^{T}(\alpha)D(\alpha) \\ \varphi 13(\alpha) &= P(\alpha) + A^{T}(\alpha) + \mu I\ Q, \\ \varphi 22(\alpha) &= D^{T}(\alpha)D(\alpha) \quad {}^{2} I \\ \varphi 23(\alpha) &= B^{T}\gamma(\alpha)Q,\ \varphi 33 = \quad Q \quad QT \end{aligned}$$

Однако выражение в левой части (11) должно быть отрицательно определенным согласно лемме 2. Соответственно, перепишем выражение с учетом $\dot{P}(\alpha) \quad 2\mu P(\alpha) \quad 0$ следующим образом.

$$\dot{P}(\alpha) \quad 2\mu\kappa P(\alpha) \quad \begin{matrix} 0 \\ I \quad \gamma^{2} \end{matrix} + \begin{matrix} C^{T}(\alpha) \\ D^{T}(\alpha) \end{matrix} I\ C(\alpha)\ D(\alpha) \quad 0$$

Следовательно, получаем неравенство (13) с помощью леммы Шура.

$$\dot{P}(\alpha) \quad 2\mu\kappa P(\alpha) \quad \begin{matrix} 0 & C^{T}(\alpha) \\ -\gamma^{2} I & D^{T}(\alpha) \\ & I \end{matrix} \quad 0 \qquad (13)$$

Machine Translated by Google

Наконец, мы получим (8), если применим лемму Шура после переписывания неравенства (12) в виде следующую форму. Кроме того, условие (8) покрывает и условие (13).

$$\dot{P}(\alpha) \quad 2\mu P(\alpha) \; 0 \qquad P(\alpha) + A \quad {}^{T}(\alpha)Q + \mu Q \quad + \quad C^{T}(\alpha)$$
$$\text{я} \quad \gamma^{2} \qquad \text{Б}^{T}(\alpha)Q \qquad \text{ДТ}(\alpha) \qquad C(\alpha)\; D(\alpha)\; 0 \quad 0$$
$$Q \quad QT \qquad 0 \qquad \text{я}$$

Следовательно, доказательство завершено.

Все схемы можно использовать для параметров политопного или аффинного типа, сделав некоторые правила. Однако в примерах в этой статье используются изменяющиеся параметры аффинного типа , и, таким образом, все изменяющиеся во времени матрицы аффинно изменяются, например матрица Ляпунова в (14). Таким образом, аффинная квадратичная устойчивость получается на протяжении всей статьи, как и в [11].

$$P(\alpha) = P0 + \sum_{\text{я}=1}^{\text{н}} \alpha i Pi \,, \quad \alpha \dot{P}(\alpha) = \sum_{\text{я}=1}^{\text{н}} \dot{} \tag{14}$$

Теорема 1. Пусть система ЛПВ находится в (1) с изменяющимся во времени параметром, заданным в (2). Если существует положительный действительный скаляр$\dot{\mu}$, такой, что $\dot{L}(\alpha) \quad 2\mu L(\alpha) \; 0$, симметричная положительно определенная дифференцируемая матрица $L(\alpha)$ с совместимыми матрицами размерностей R, $F(\alpha)$, то LMI в (15) справедливо для (α) R и (α) D. Таким образом, существует регулятор H с обратной связью по состоянию $K(\alpha) = F(\alpha)R^{1}$.

$$\dot{L}(\alpha) \quad 2\mu L(\alpha) \; 0$$
$$\text{я} \quad \gamma^{2}$$
$$L(\alpha) + A(\alpha)R + B2F(\alpha) + \mu R \; B1(\alpha) \quad R \quad R \quad {}^{T} \tag{15}$$
$$C1(\alpha)R + D12F(\alpha) \qquad Д11(\alpha) \qquad 0 \qquad \text{Я} \qquad 0$$

Доказательство теоремы 1. Матрицы замкнутого контура в (16) получены так же, как и в (17), когда контроллер u = K(α)x подставляется в систему (1), где выход y(t) исключается, поскольку он считается обратной связью с состоянием.

$$\begin{aligned} . \; xcl(t) &= Acl(\alpha(t))xcl(t) + Bcl(\alpha(t))\omega(t), \\ zcl(t) &= Ccl(\alpha(t))xcl(t) + Dcl(\alpha(\tau))\omega(\tau) \end{aligned} \tag{16}$$

$$\begin{aligned} Acl(\alpha) &= A(\alpha) + B2K(\alpha), \; Bcl(\alpha) = B1(\alpha), \\ Ccl(\alpha) &= C(\alpha) + D12K(\alpha), \; Dcl(\alpha) = D11(\alpha) \end{aligned} \tag{17}$$

Если заменить матрицы пространства состояний в (17) на (8), получим следующее неравенство.

$$\dot{P}(\alpha) \quad 2\mu P(\alpha) \; 0$$
$$\text{я} \quad \gamma^{2}$$
$$P(\alpha) + QT \; (A(\alpha) + B2K(\alpha)) + \mu QT \; QTB1(\alpha) \quad Q \quad QT$$
$$C(\alpha) + Д12K(\alpha) \qquad Д11(\alpha) \qquad 0 \qquad \text{Я} \qquad 0$$

Однако существуют умножения неизвестных матриц. Чтобы получить LMI, мы применяем конгруэнтное преобразование к неравенству путем умножения транспонирования (18) и (18) соответственно . Таким образом, получается (15), где $L(\alpha) = Q \quad TP(\alpha)Q \quad 1$, $\dot{L}(\alpha) = Q \quad T \; \dot{P}(\alpha)\dot{Q} \quad 1$, $P = Q \quad 1$, $F(\alpha) = K(\alpha)P$. Следовательно, доказательство завершено.

$$\text{Диагностика Q} \quad {}^{1}{}_{\text{, я, вопрос}} \quad {}^{1}\text{, я} \tag{18}$$

Замечание 2. Предложенное решение дает три существенных преимущества. Во-первых, в отличие от (7), матрица Ляпунова может быть в виде матрицы, зависящей от параметра, без дополнительных подходов, таких как

Machine Translated by Google

как s-процедура, сетка и многовыпуклость, что снижает консервативность. Во-вторых, в отличие от общепринятой схемы (7) в [5], коэффициент усиления регулятора не зависит от матрицы Ляпунова, которая ограничена свойством симметричности. Вместо этого он меняется с несимметричной матрицей R для предложенного LMI в теореме 1. В-третьих, реальная слабая переменная μ дает приемлемые результаты даже для больших диапазонов параметров, поскольку большое увеличение диапазона влияет на значение производительности γ меньше , чем аналогичные значения в с точки зрения осуществимости и производительности, что подтверждается численными результатами.

Если мы рассмотрим CTS с TVPU, мы можем легко изменить LMI для проектирования независимого от параметров (статического или фиксированного) контроллера, как в теореме 1, предполагая, что неопределенные параметры как неизвестны во время разработки, так и не известны посредством онлайн-измерений, но известны как ограничено. Таким образом, некоторые матрицы преобразуются в матрицы, зависящие от параметров, тогда как некоторые матрицы преобразуются в матрицы, независимые от параметров, для получения формы LMI. Кроме того, теорему можно использовать для изменяющихся во времени параметров или неопределенностей аффинного или политопного типа, внеся некоторые правила. Однако в данной статье для проектов рассматриваются изменяющиеся во времени параметры аффинного типа. Для предлагаемого регулятора с обратной связью по состоянию были реализованы два численных примера. Все численные расчеты проводились с помощью парсера Yalmip [47] и решателя Sedumi [48], которые использовались на ПК (2 ядра ЦП 2,8 ГГц с 4 ГБ ОЗУ). В примере 1 представлена конструкция контроллера для CTS с TIPU, тогда как в примере 2 представлено управление по расписанию усиления для CTS с TVPU.

Пример 1. Обращается к контроллеру обратной связи по состоянию спутника в [25,49] . Система возникает из двух твердых тел, соединенных гибкой связью. Система аналогична (19). Неопределенные параметры моделируются как демпфирование f и пружина k. Диапазон k составляет [0,09 0,4], тогда как диапазон f составляет [0,0038 0,04]. Характеристики методов из [5,24–30] и предложенного решения в теореме 1 сравниваются по уровню γ , который напрямую связан с консервативностью. Производные параметров равны нулю; то есть неопределенности не дифференцируемы.

$$\dot{\begin{bmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{bmatrix}} = \begin{bmatrix}0&0&1&0\\ 0&0&0&1\\ -k&k&-f&f\end{bmatrix}\begin{bmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 0\\ 1\end{bmatrix}\omega + \begin{bmatrix}0\\ 0\\ 0\\ 0.1\end{bmatrix}u,$$

$$r = \begin{bmatrix}0&1&0&0\\ 0&0&0&0\end{bmatrix}\begin{bmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{bmatrix} + \begin{bmatrix}0\\ 0\end{bmatrix}\omega + \begin{bmatrix}0\\ 0.01\end{bmatrix}u \quad (19)$$

В таблице 1 показаны минимальные значения γ для результатов из [5,24–30] и предлагаемого решения в теореме 1 для этого примера. Результаты показывают, что предлагаемое решение дает гораздо лучшие результаты, чем другие методы. Коэффициент усиления обратной связи общего решения составляет K = -105 [6,1116 44,042 1,1054 81,096], тогда как коэффициент предлагаемого решения равен K = -[205,82 2224,6 58,393 2850,9]. Превосходство предлагаемого решения заключается в том, что оно менее консервативно, чем [5,24–30], как показано в таблице 1. Кроме того, на рисунке 1a показана взаимосвязь между μ и γ. На уровень γ практически никогда не влияют изменения μ после определенного значения для этой системы. Более того, предложенная теорема 1 дает лучший результат, даже если μ находится в диапазоне от 2,5 до 35, как показано на рисунке 1a.

Замечание 3. Предложенное решение для H -робастного управления может быть использовано для CTS с TVPU за счет получения преимуществ расширенных LMI/свободных переменных, но другие решения в [5,24–30] могут быть использованы только для CTS с TIPU. . Кроме того, LMI неитеративен. Таким образом, неитеративное решение с расширенным LMI включает в себя как TIPU, так и TVPU, обеспечивая преимущества расширенного LMI. В литературе о его существовании не сообщалось. Соответственно, в примере 1 представлена конструкция контроллера CTS с TIPU, тогда как в примере 2 представлена конструкция контроллера CTS с

Machine Translated by Google

ТВПУ. Наконец, пример 2 представляет собой только сравнение этой статьи со стандартом.
метод в [5], поскольку исследования в таблице 1 , за исключением стандартного дизайна в [1], не полностью справедливы для
ТВПУ. Характеристики проверяются путем увеличения неопределенного набора для TVPU в примере 2.

Пример 2. В этом примере управление LPV по планированию усиления разработано с использованием теоремы
1 для случайно сгенерированной системы LPV в (20), где диапазоны изменяющихся во
времени параметров α1 и α2 равны [0,01, 2] и [0,5, 1,5] соответственно. Кроме того, темпы изменения
параметры $\dot{\alpha}1$ и $\dot{\alpha}2$ равны [0, 3] и [0, 1] соответственно.

$$\begin{bmatrix} \dot{x}1 \\ \dot{x}2 \\ \dot{x}3 \end{bmatrix} = \begin{bmatrix} \alpha1 & \alpha1 & 0{,}8 \\ 0 & 0{,}2 & \alpha2 \\ 0{,}5 & \alpha2 & 0{,}1 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} + \begin{bmatrix} 0{,}1 \\ 0{,}2 \\ 0 \end{bmatrix} \omega + \begin{bmatrix} 1 \\ 0 \\ 0{,}1 \end{bmatrix} ты,$$
$$г = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} \omega + \begin{bmatrix} 0 \\ 0{,}8 \end{bmatrix} ты, \quad (20)$$

Как и в Таблице 2, улучшение существенно увеличивается по мере увеличения неопределенного набора.
по сравнению с обычным набором. Например, метод снижает уровень γ со скоростью
78,59% для α1 = [0,01 50], α2 = [0,1 190].

Таблица 2. Значения производительности γmin для ТВПУ.

	Значения γ								
	α1= [0,01 2], α2= [0,5 1,5]			α1= [0,01 5], α2= [0,1 19]			α1= [0,01 50], α2= [0,1 190]		
[5]	1,0280			6.3151			22,745		
Теорема 1	0,890 (мк = 3) 14,42% снизилось			2,2466 (мк = 2,5) 64,42% снизилось			4,869 (мк = 3,5) 78,59% снизилось		
	Слабые наборы переменных для осуществимости								
Теорема 1	μ = (0, 9 × 105]			μ = (0, 8 × 105]			μ = (0, 8 × 105]		
	Количество итераций — Количество переменных — Время решателя								
[5]	15	18	0,0174 с	28	18	0,126 с	49	18	0,215 с
Теорема 1	39	41	9,556 с	41	41	9,904 с	56	41	13,503 с

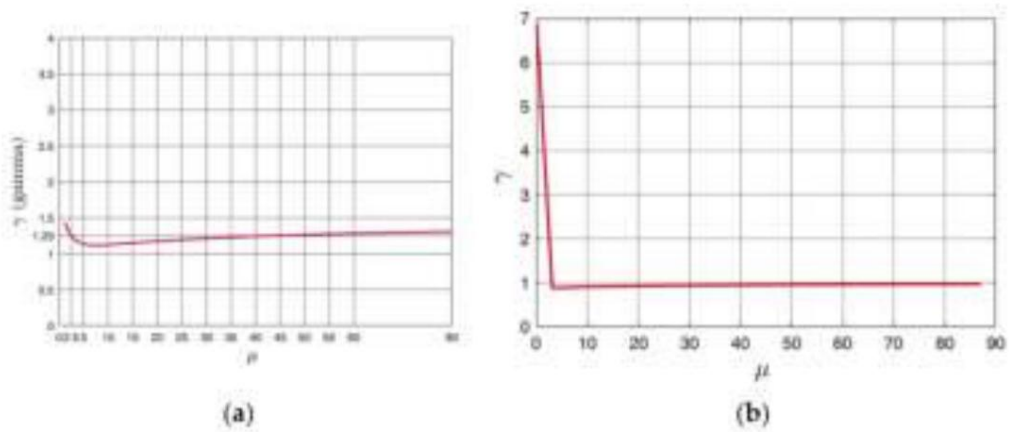


Рис. 1. Связь между уровнем γ и слабой переменной: (а) Пример 1; (б) Пример 2.

Кроме того, взаимосвязь между μ и γ показана на рисунке 1b. Следовательно, γ
уровень почти никогда не меняется путем изменения μ после определенного значения, поэтому его определение
легко, так что предлагаемое решение имеет лучшую производительность, чем обычное решение
даже если μ находится между 1 и 85.

Machine Translated by Google

С учетом этого получены результаты для предложенного решения в теореме 1 и общего плана из [5], где матрица Ляпунова фиксирована. При α1 = [0,01·2], α2 = [0,5·1,5] общий план дает уровень γ , равный 1,028, тогда как предложенный метод дает 0,890 при μ = 3. Соответственно, коэффициенты усиления обратной связи стандартного плана и проекта в теореме 1 равны K0 = 103 [23,383 170,568 269,482], K1 = [0,30251 2,2045 3,4833], K2 = [0,11503 0,84184 1,33] и K0 = [12,17 74,915 102,05], K1 = [0,9 135 3,0938 3,5387], K2 = [5,306 40,69 47,744] соответственно. Связь между μ и γ показана на рисунке 1b.

Судя по результатам, предложенное расширенное решение LMI менее консервативно, чем обычное LMI в [5] для проектирования планирования усиления, хотя оно имеет недостатки, которые заключаются в большем количестве итераций, количестве переменных и времени решателя. Осуществимость получена для большого неопределенного набора в Таблице 2, а слабая переменная μ обеспечивает допустимые результаты для большого набора (0, 8 × 105]. Хотя количество переменных LMI увеличилось с 18 до 41 по сравнению со стандартной в [5] было достигнуто улучшение до 78,59%. Наконец, хорошее улучшение и менее консервативные результаты достигаются для TVPU, даже если расширенные LMI являются более дорогостоящими, чем стандартные LMI. Примечательно, что сравнения в примере. сделано для независимой от параметра матрицы Ляпунова в [5]. Как подчеркнуто в замечании 1, при использовании PLDF используются такие методы, как метод многовыпуклости [11] и методы в [17,18] . Поэтому используются стандартные. Переменные LMI и сложность [5] существенно возрастают. Например, необходимо добавить еще одно условие LMI того же размера. Следовательно, предлагаемое условие LMI считается менее затратным, поскольку оно имеет одно условие LMI.

Два примера показали, что предлагаемое решение дает менее консервативные результаты. Кроме того, его можно в полной мере использовать для системы ЛПВ и неопределенной системы при ТВПУ, но остальные использовать нельзя, за исключением общего ЛМИ в [5]. С другой стороны, увеличение границ изменяющихся во времени неопределенностей или параметров увеличивает консервативность LMI. Однако конструкция, основанная на предложенном LMI, помимо более простой структуры, дает менее консервативные и более успешные результаты.

4. Проектирование динамического компенсатора возмущений для систем LPV.

Блок-схема компенсатора помех показана на рисунке 2 , где NCF — коэффициент нормализации или фильтр. Здесь динамический компенсатор прямой связи имеет вид (21).

$$\begin{aligned} \dot{x}c(t) &= Ac(\alpha)xc(t) + Bc(\alpha)\omega(t) \\ uc(t) &= Cc(\alpha)xc(t) + Dc(\alpha)\omega(t) \end{aligned} \quad (21)$$

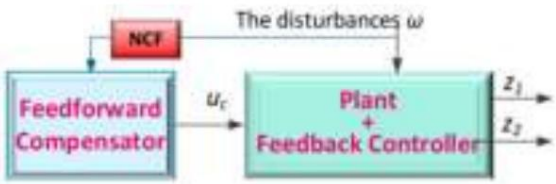


Рис. 2. Блок-схема компенсатора помех.

Соответственно, в теореме 2 представлена конструкция динамического компенсатора возмущений с прямой связью в соответствии с расширенным LMI в лемме 3.

Теорема 2. Пусть устойчивая система ЛПВ находится в (1) для всех допустимых траекторий параметров, набор которых такой же, как в (2). Если существуют вещественный скаляр μ > 0, положительно определенная дифференцируемая симметричная матрица Y(α) и совместимые матрицы M, N(α), то ЛМН в (22) и (23) с определениями в (24) имеют место для (α) R и (α) D. В таком случае динамический компенсатор прямой связи H в (21) получается из (25).

Machine Translated by Google

$$\begin{matrix}\Psi 11(\alpha) & \Psi 12(\alpha) & \Psi 13(\alpha) & \Psi 14(\alpha)\\ & \Psi 22(\alpha) & \Psi 23(\alpha) & \Psi 24(\alpha)\\ & & \Psi 33(\alpha) & \Psi 34\\ & & & \Psi 44\end{matrix}\quad 0 \tag{22}$$

$$Y(\alpha) = \begin{matrix}Y1(\alpha) & Y2(\alpha)\\ & Y3(\alpha)\end{matrix}\quad 0 \tag{23}$$

$$\begin{aligned}\Psi 11(\alpha) &= \begin{matrix}\dot{Y}1(\alpha) - 2\mu Y1(\alpha) & \dot{Y}2(\alpha) - 2\mu Y2(\alpha)\\ & \dot{Y}3(\alpha) - 2\mu Y3(\alpha)\end{matrix},\\ \Psi 12(\alpha) &= \begin{matrix}0 & Y1(\alpha) + R_1^T A^T(\alpha) + NT(\alpha)B_2^T + \mu K P_1^T\\ 0 & Y1(\alpha) + R_1^T A^T(\alpha) + NT(\alpha)B_2^T + \mu K P_2^T(\alpha)\end{matrix},\\ \Psi 13(\alpha) &= \begin{matrix}Y2(\alpha) + MT(\alpha)R_1^T C_1^T(\alpha) + HT(\alpha)DT_{12}\\ Y3(\alpha) + MT(\alpha)R(\alpha)_1^T C_1^T(\alpha) + HT(\alpha)DT_{12}\end{matrix},\\ \Psi 14(\alpha) &= \begin{matrix}p_1^T C_2^T + NT(\alpha)DT(\alpha) + \\ p_{12}^T C_2^T NT(\alpha)DT\end{matrix}_{22},\ \Psi 22(\alpha) = \begin{matrix}-\gamma^2 I & B1(\alpha) + DT(\alpha)B_\Phi^T\\ R1 & P_1^T\end{matrix},\\ \Psi 23(\alpha) &= \begin{matrix}B_\Phi^T(\alpha) & DT\,11(\alpha) + DT_\Phi(\alpha)DT_{12}\\ R1 & P_2^T(\alpha)\ 0\end{matrix},\ \Psi 24(\alpha) = \begin{matrix}DT\,21(\alpha) + DT_\Phi(\alpha)DT_{22}\\ 0\end{matrix},\\ \Psi 33(\alpha) &= \begin{matrix}-R2(\alpha) & R_2^T(\alpha), 0\\ I\end{matrix},\ \Psi 34 = \begin{matrix}0\\ 0\end{matrix},\ \Psi 44 = I\end{aligned} \tag{24}$$

$$\begin{aligned}Ac &= M(\alpha)R_1^{-1}(\alpha),\ Bc(\alpha) = BF(\alpha),\\ Cc &= N(\alpha)R_2^{-1}(\alpha),\ Dc = DF(\alpha)\end{aligned} \tag{25}$$

Доказательство теоремы 2. LMI в (8) умножается на следующую матрицу и транспонируется матрица слева направо, соответственно, когда мы применяем конгруэнтное преобразование.

$$\begin{matrix}Q & 1 & 0 & 0 & 0\\ 0 & I & 0 & 0\\ 0 & 0 & Q & 1 & 0\\ 0 & 0 & 0 & I\end{matrix}$$

Таким образом, согласно определениям (27) получается следующий LMI.

$$\begin{matrix}\dot{Y}(\alpha) & 2\mu Y(\alpha)\\ 0 & -\gamma^2 I\\ Y(\alpha) + A(\alpha)R + \mu R & B(\alpha)\\ C(\alpha)R & D(\alpha)\end{matrix}\quad \begin{matrix}R & R\\ 0 & I\end{matrix}^T\quad 0 \tag{26}$$

$$Y(\alpha) = Q\ TP(\alpha)Q^{-1},\ \dot{Y}(\alpha) = Q^T\dot{P}(\alpha)Q^{-1},\ P = K^{-1} \tag{27}$$

Форма пространства состояний от ω(t) до Z(t) такая же, как в (28), когда компенсатор в (21) реализуется в (1). Здесь вектор Z представляет собой выходные данные z и y, вектор X представляет состояния x в (1) и xc в (21), а ω — сигнал возмущения.

$$\begin{matrix}\dot{X}(\tau)\\ Z(\tau)\end{matrix}\ "=" \begin{matrix}Acl(\alpha) & Bcl(\alpha)\\ Ccl(\alpha) & Dcl(\alpha)\end{matrix}\quad \begin{matrix}x(\tau)\\ \omega(\tau)\end{matrix} \tag{28}$$

Соответственно, матрицы пространства состояний в (28) получаются как в (29) с реализацией. Следовательно, оптимальный регулятор достигается за счет минимизации усиления L2 от ω до Z.

$$\begin{aligned}Acl(\alpha) &= \begin{matrix}A(\alpha) & B2Cc(\alpha)\\ \text{До н.э} & \end{matrix}\ 0,\ Bcl(\alpha) = \begin{matrix}B1(\alpha) + B2Dc(\alpha)\\ \text{Д21} + \text{Д22Дк}(\alpha)\end{matrix},\\ Ccl(\alpha) &= \begin{matrix}C1(\alpha) & D12Cc(\alpha)\\ C2(\alpha) & D22Cc(\alpha)\end{matrix},\ Dcl(\alpha) = D11(\alpha) + D12Dc(\alpha)\end{aligned} \tag{29}$$

Machine Translated by Google

Определим матрицу Ляпунова, как в (23), а матрицу R(ξ) определим следующим образом.

$$R(\alpha) = \begin{matrix} P1 & P1 \\ R2(\alpha) & R2(\alpha) \end{matrix}$$

Соответственно, если матрица Ляпунова Y(α) в (23), ее производная, матрица R(α) и матрицы пространства состояний в (29) помещаются в (26) вместо A(α), B(α) , C(α) и D(α), получаем матричное неравенство. Если некоторые выражения определены так же, как в (30), мы получаем ЛМН в (22) с определениями в (24) и доказательство завершено.

$$M(\alpha) = Ac(\alpha)R2(\alpha),\ N = Cc(\alpha)R2(\alpha),$$
$$BF(\alpha) = Bc(\alpha),\ DF(\alpha) = Dc(\alpha) \quad (30)$$

Замечание 4. Предлагаемые LMI обеспечивают четыре существенных преимущества. Некоторые из них упомянуты в замечании 2. Прежде всего, матрицу Ляпунова можно обозначить в форме ЛПВ, что снижает консервативность, без дополнительных подходов, таких как сетка, s-процедура, многовыпуклость и условия в [17,18]. Например, в [45] существует множество LMI, поскольку из-за этих произведений используются условия многовыпуклости . Кроме того, LMI основаны на общем BRL в (6), поэтому LMI несколько консервативны из-за умножения матрицы Ляпунова. Однако предлагаемый расширенный BRL является лучшим среди доступных, как показано в таблице 1. Во-вторых, из-за произведений системных матриц с матрицей Ляпунова, как в (6), возникают бесконечные LMI и невыпуклая оптимизация, если матрица Ляпунова может быть в форме LPV. Однако для предложенной формулировки LMI в (8) такого случая не происходит. Таким образом, очевидно, что предлагаемое решение явл В-третьих, регуляторы LPV, разработанные PDLF, как в [45,50], обычно зависят от производной матрицы Ляпунова (матрицы Ac(α) регулятора). Однако операция производной вызывает проблему шума в практических приложениях реального времени. Предложенные конструкции не зависят от производной матрицы Ляпунова. Наконец, матрицы регулятора не зависят от матрицы Ляпунова, которая в обычных схемах ограничена условием симметричности. Вместо этого матрицы в предлагаемых конструкциях зависят от полной матрицы. Более того, он имеет простую структуру.

Замечание 5. Предлагаемый компенсатор не содержит регулятора обратной связи. Хорошо известно , что метод прямой связи применяется для улучшения характеристик отслеживания задания или снижения помех или шума, и он не влияет на стабильность, поэтому необходим контроллер с обратной связью. Однако для некоторых установок, например многих механических и электрических систем, конструкция с обратной связью не требуется, поскольку они являются стабильными системами. Поэтому компенсатор можно использовать не только для стабилизированных систем с регулятором обратной связи, но и для устойчивых систем с возмущениями. Кроме того, его можно использовать для улучшения отслеживания задания путем обмена опорными сигналами вместо входных воздействий на стабилизированные системы с контроллером обратной связи. Соответственно, для всех примеров в статье предполагается, что системы стабильны или стабилизированы с помощью контроллера с обратной связь Однако внутренняя стабильность упреждающего регулятора/компенсатора LPV должна обеспечиваться для всех изменяющихся во времени параметров, как в приведенных выше теоремах. В противном случае стабильная система становится нестабильной. Эта устойчивость достигается за счет разложения матрицы Ляпунова без необходимости использования какого-либо дополнительного метода устойчивости для всех предложенных теорем.

5. Моделирование исследований

В этом разделе был реализован численный пример конструкции компенсатора. Все расчеты численной оптимизации были выполнены с помощью Yalmip версии R20210331 [47] и Sedumi версии 1.30 [48], которые использовались на ПК (2 ядра процессора с частотой 2,8 ГГц и 4 ГБ оперативной памяти).

Пример 3. В этом примере показана конструкция компенсатора LPV для случайно сгенерированной системы LPV в (31). Изменяющиеся во времени параметры α1(t) и α2(t) ограничены соотношением (32). Хотя обычно она выглядит как неопределенная линейная система, ее считают системой LPV, если предположить, что

Machine Translated by Google

что измеримые или вычислимые параметры являются состояниями нелинейной системы. Система LPV является производной нелинейной системы, в отличие от неопределенной линейной системы.

$$\begin{aligned} x(t) &= x(t) + \begin{matrix} 1 + \alpha 1(t) & 3 \\ 6 & 4 \end{matrix} \quad \alpha 2(t)\, z(t) = 1\ 0{,}1\ x(t) + \begin{matrix} 0{,}5 + \alpha 1(\tau) \\ 0{,}8 \end{matrix} \omega(t) + \begin{matrix} 0{,}2 \\ 0{,}1 \end{matrix} \quad \text{и}(\tau) \\ y(t) &= 1\ 0\ x(t) \end{aligned} \tag{31}$$

$$\begin{matrix} 0{,}1 & \alpha 1(t) & 0{,}7 \text{ при} & 0{,}445 & . \alpha 1(t) & 1{,}82 \\ 1 & \alpha 2(t) & 5 \text{ при} & 11{,}7 & . \alpha 2(t) & 2{,}295 \end{matrix} \tag{32}$$

Когда система анализируется с учетом изменяющихся во времени параметров, получаются результаты, показанные на рисунках 3 и 4 . Различные изменяющиеся во времени параметры тестируются для анализа на рисунке 3. Однако для моделирования используется один из них, который показан на рисунке 3b. Рисунок 3a используется только для анализа (диаграмма Боде и отображение полюсов), поскольку предполагается, что изменение параметра линейное. Как правило, оба параметра вряд ли будут изменяться линейно. Изменение параметра на рисунке 3b происходит во время моделирования. Целью является наблюдение за реакциями на различные параметрические изменения. Для анализа диаграмма Боде системы показана на рисунке 4 с временем выборки 50 мс и временем выборки мод На рисунке 4a показана диаграмма Боде системы, предполагающая изменяющиеся во времени параметры, показанные на рисунке 3a, тогда как на рисунке 4b показана диаграмма Боде, предполагающая изменяющиеся во времени параметры, показанные на рисунке 3b. На рисунке видно, что изменяющиеся во времени параметры сильно влияют на нелинейность системы по диаграммам Боде выходо Судя по диаграммам Боде, возмущение более эффективно примерно при 3–6 рад/с.
Поэтому сигналы испытаний на помехи имеют низкую частоту.

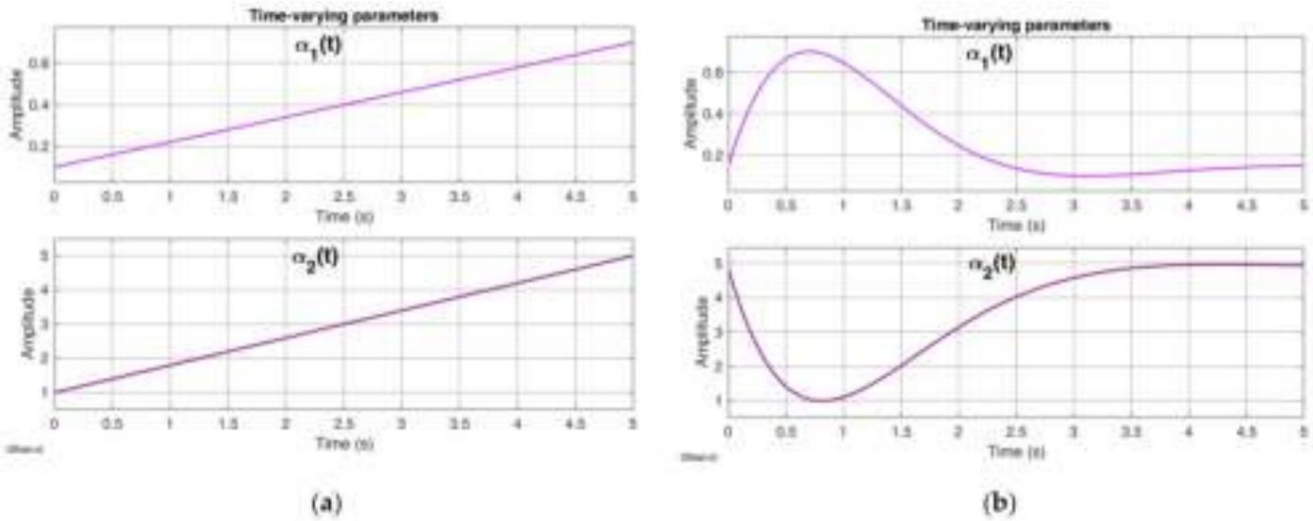


Рисунок 3. Изменяющиеся во времени параметры для: (a) анализа; (б) Анализ и моделирование.

Аналогичным образом, на рисунке 5 представлена полярная карта системы для различных изменяющихся во времени параметров на рисунке 3b, где время выборки составляет 50 мс. Из карты полюсов можно вывести два случая. Полюсы всех трансфер-матриц системы находятся в левой половине плоскости. Этот случай подтверждает устойчивость системы по всем изменяющимся во времени параметрам, указанным в замечании 5. В противном случае компенсатор применить нельзя, так как он не влияет на устойчивость. Во-вторых, параметры сильно влияют на систему.

Machine Translated by Google

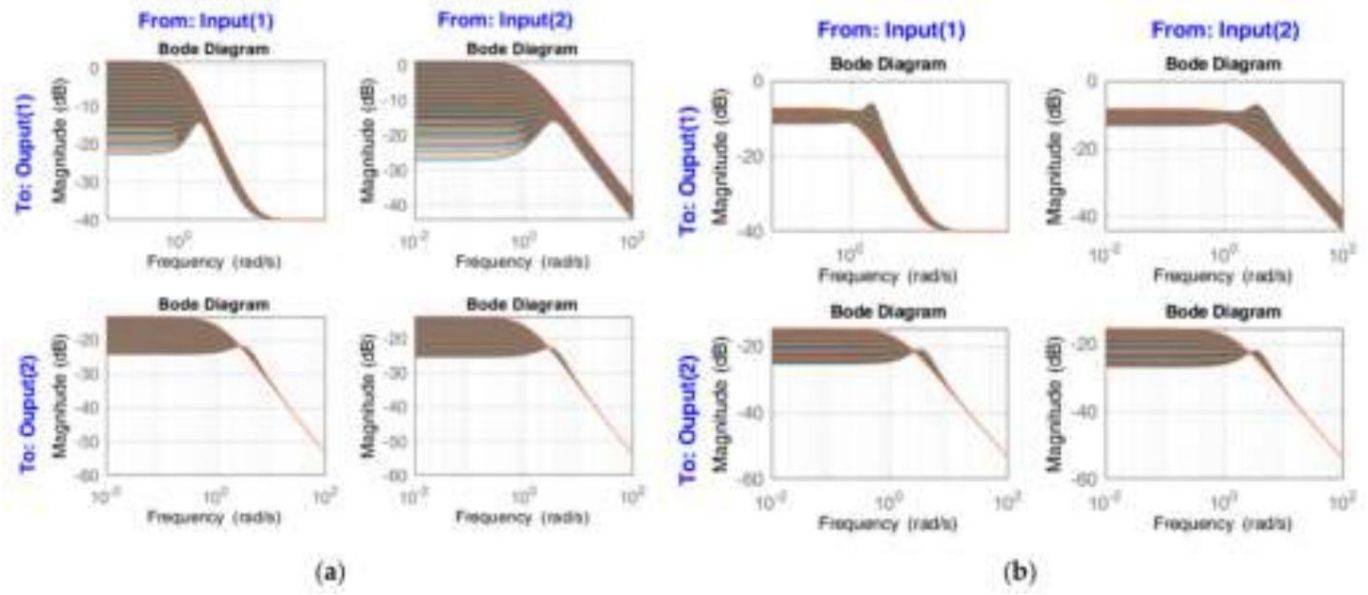


Рисунок 4. Диаграммы Боде системы для изменяющихся во времени параметров: (а) Рисунок 3а; (б) Рисунок 3б.

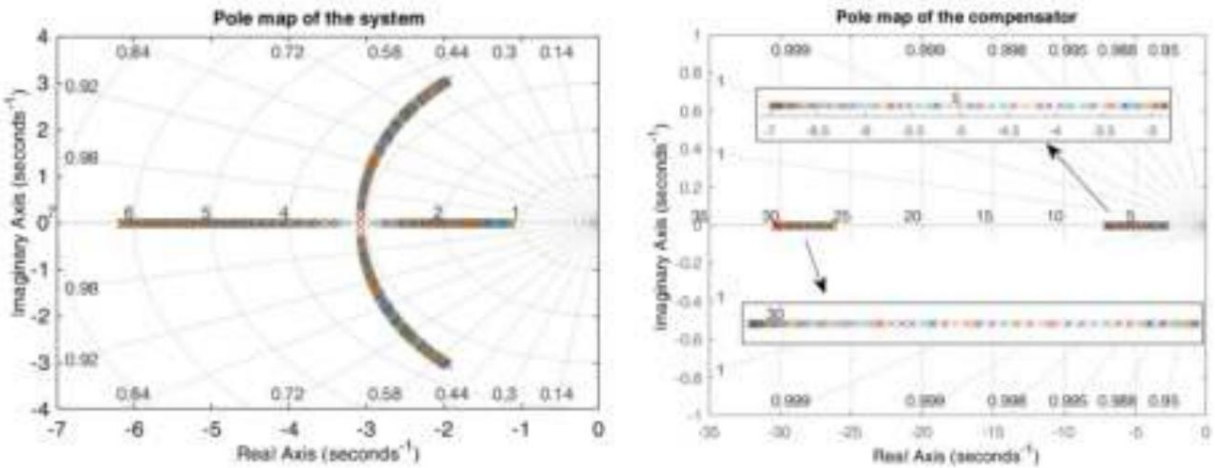


Рисунок 5. Полюсные карты системы и компенсатора.

На рисунке 6 показана блок-схема конструкции компенсатора LPV, где Wu и Wz представляют собой входной и выходной весовые фильтры соответственно. Однако для этого примера соответствующий компенсатор спроектирован без этих фильтров. Однако фильтры можно использовать по таким причинам, как структурные ограничения, упомянутые в примечании 6.

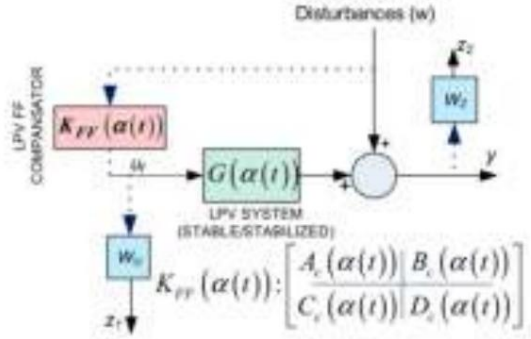


Рисунок 6. Структурная схема компенсатора LPV.

Machine Translated by Google

Замечание 6. Входы и выходы ограничены из-за механической/электрической структуры. Поэтому ограничения следует рассматривать в соответствии с задачей соответствующей области. Например, прогиб подвески не должен превышать верхний предел, тогда как для активных систем подвески транспортных средств прогиб шин должен быть минимизирован в максимально возможной степени. Существуют некоторые исследования входных и выходных ограничений на основе LMI, например [51]. Они регулируются с помощью NCF, как показано на рисунке 2. NCF может быть коэффициентом или фильтром, например фильтром нижних частот, который определяется в соответствии с поведением системы с учетом пределов частоты и усиления входов и выходов. Другой способ — это входные и выходные весовые функции для конструкции контроллера H , как показано на рисунке 6 и блок-схеме в [45]. Таким образом, решение о NCF может быть принято с учетом структуры системы. В частности, на высокой частоте вместо фиксированного NC может потребоваться фильтр NCF . Это связано с тем, что компенсатору это необходимо для регулирования предела привода на высокой частот

Матрицы получаются в (33), когда LMI в теореме 2 решаются для минимизации значения γ в соответствии с зависящими от параметра переменными Yi , BFi, DFi, Mi , Ni для i = 0, 1, 2 и независимыми от параметра переменными переменные R1 и R2. Таким образом, матрицы пространства состояний компенсатора LPV Aci, Bci, Cci и Dci в (21) для i = 0, 1 и 2 аффинно получаются из матриц в (33) по изменяющимся во времени параметрам α1(t) и α2(t) в (32) для всех t в ходе онлайн-операции. Следовательно, (34) показывает матрицы контроллера LPV в пространстве

$$M0 = \begin{matrix} 163{,}81 & 12{,}724 \\ 1{,}7999 & 202{,}82 \end{matrix}, M1 = \begin{matrix} 5{,}1074 & 22{,}914 \\ 11{,}194 & 36{,}474 \end{matrix},$$

$$M2 = \begin{matrix} 10{,}704 & 16{,}627 \\ 15{,}009 & 84{,}506 \end{matrix}, R1 = \begin{matrix} 7{,}1371 & -3{,}0487 \\ -20{,}361 & 92{,}818 \end{matrix},$$

$$P2 = \begin{matrix} 7{,}139 & -3{,}0494 & -20{,}361 \\ 92{,}826 & N1 = -111{,}95 \end{matrix}, N0 = \begin{matrix} 1175{,}1 & 1477{,}5 \end{matrix}, \quad (33)$$

$$364{,}76 \quad -0{,}07727 \quad 0{,}51139, N2 = \begin{matrix} 53{,}522 & 83{,}135 \end{matrix},$$

$$BC0 = \begin{matrix} 0{,}070224 & 0 \end{matrix}, BC1 = \begin{matrix} 0{,}46491 & 0 \end{matrix},$$

$$BC2 = ,$$

$$DC0 = -3{,}3213, DC1 = -4{,}6489, DC2 = -0{,}0003355$$

$$Ac(\alpha1, \alpha2) = \begin{matrix} 24{,}886 + 0{,}012541\alpha1 & 1{,}0907\alpha2 \\ 7{,}1544 & 0{,}49361\alpha1 \end{matrix} \quad \begin{matrix} 0{,}68047 & 0{,}24644\alpha1 + 0{,}14329\alpha2 \\ 0{,}54525\alpha2 & 2{,}42 + 0{,}37671\alpha1 \end{matrix} \quad 0{,}92828\alpha2$$

$$Bc(\alpha1) = \begin{matrix} 0{,}07727 + 0{,}070224\alpha1 \\ 0{,}51139 \end{matrix} \quad \begin{matrix} 0{,}46491\alpha1 \end{matrix} \quad (34)$$

$$Cc(\alpha1, \alpha2) = \begin{matrix} 131{,}53 & 4{,}9366\alpha1 \end{matrix} \quad 5{,}4537\alpha2 \quad 11{,}596 + 3{,}7674\alpha1 + 0{,}71644\alpha2$$

$$Dc(\alpha1, \alpha2) = \begin{matrix} 2{,}8863 & 4{,}6489\alpha1 \end{matrix} \quad 0{,}0003355\alpha2$$

Соответственно, на рисунке 7a показана диаграмма Боде компенсатора для изменяющихся во времени параметров на рисунке 3a, тогда как на рисунке 7b показана диаграмма Боде для изменяющихся во времени параметров на рисунке 3b для времени выборки 50 мс. Согласно картам полюсов и диаграммам Боде, компенсатор LPV существенно изменяется при изменении во времени параметро

Значение μ определяется как 0,42 путем поиска линии для получения хорошего ответа. Соответственно , значение индекса производительности γ составляет 0,037268. Это значение явно меняется в зависимости от системы и фильтров ввода-вывода. Это обеспечивает возможность работы с очень большими изменяющимися во времени параметрами, что доказано на примерах проектирования контроллеров с обратной связью по состоянию. На рисунке 8a также показана диаграмма Боде системы в сочетании с компенсатором от входа возмущения до выходов z и y с учетом изменяющихся во времени параметров, показанных на рисунке 3a, тогда как на рисунке 8b показана диаграмма Боде с учетом изменяющихся во времени параметров, показанных на рисунке 3b.

На рис . 9 показаны сигналы испытаний на воздействие помех, примененные для моделирования. Это ударный сигнал, как показано на рисунке 9a; Синусоидальный сигнал 5 рад/с, показанный на рисунке 9b; и сигнал белого шума , показанный на рисунке 9c.

Machine Translated by Google

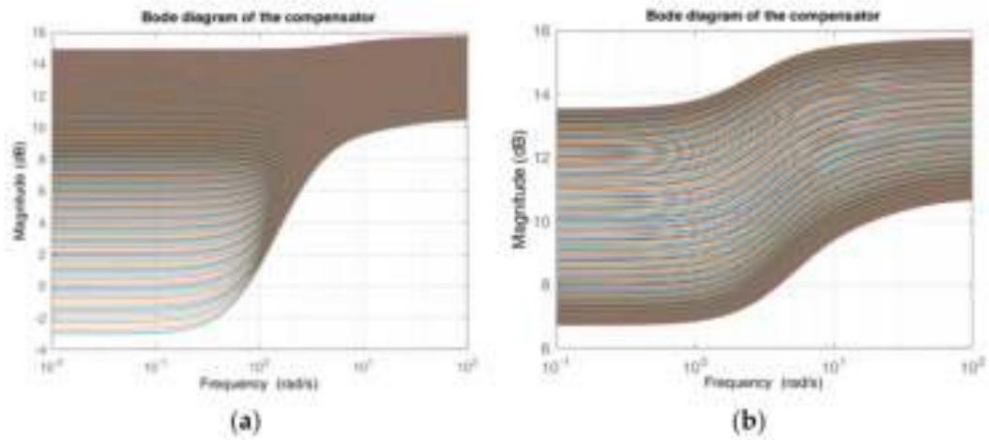


Рис. 7. Диаграммы Боде компенсатора изменяющихся во времени параметров: а — рис. 3а; (б) Рисунок 3б.

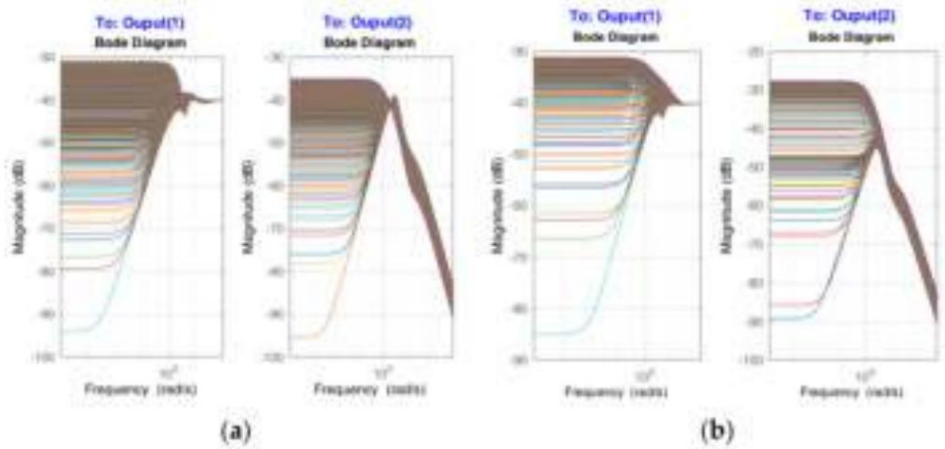


Рис. 8. Диаграммы Боде системы с компенсатором изменяющихся во времени параметров: а — рис. 3а; (б) Рисунок 3б.

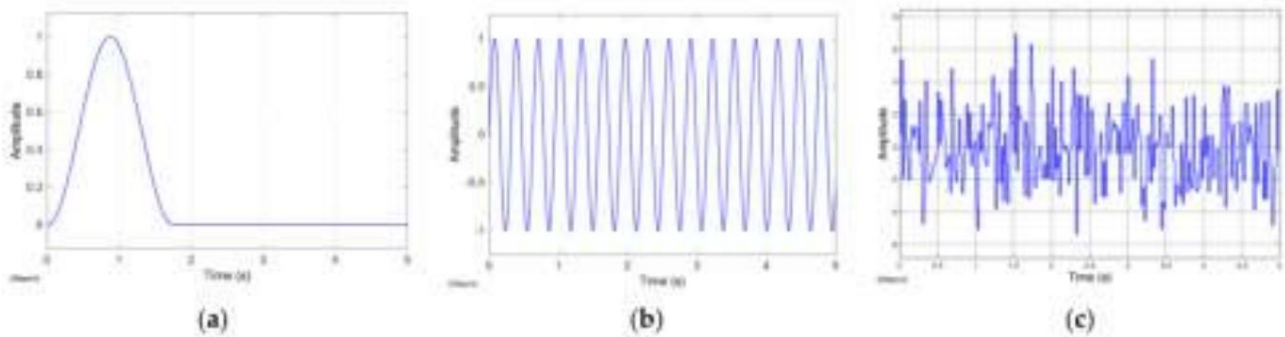


Рисунок 9. Применяемые испытательные помехи: (а) удар; (б) синусоидальный сигнал 5 рад/с; (с) сигнал белого шума.

Что касается результатов, на рисунке 10 показаны выходные данные z и y для сигнала внешнего возмущения на рисунке 9a. Как видно из результатов, разработанный компенсатор LPV эффективно ослабляет помехи примерно на 100% по сравнению со случаем без компенсатора.

Machine Translated by Google

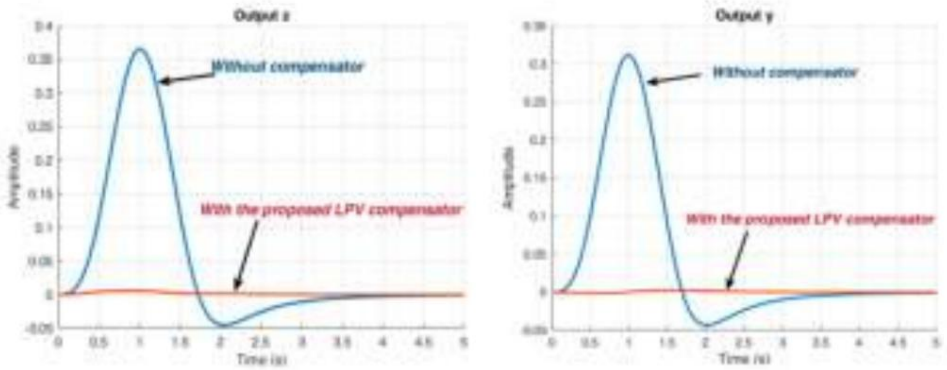


Рисунок 10. Выходы z и y для сигнала ударного возмущения.

Аналогичным образом, на рисунках 11 и 12 показаны выходные сигналы z и y для сигналов испытания на помехи на рисунках 9b и 9c соответственно. Сигналы, генерируемые параметрозависимым компенсатором от трех помех, показаны на рисунке 13.

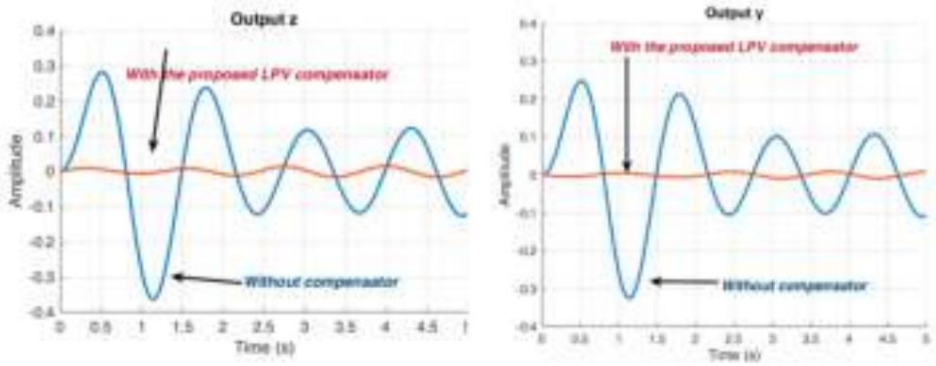


Рисунок 11. Выходные данные z и y для синусоидального возмущения 5 рад/с.

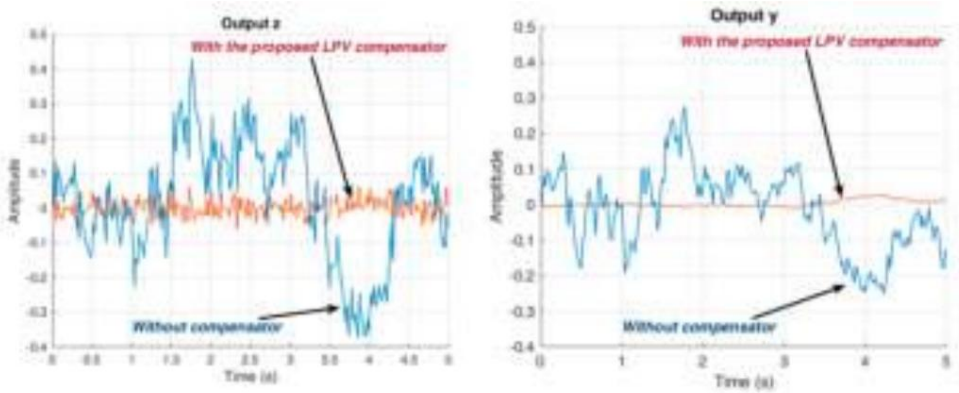


Рисунок 12. Выходные данные z и y для помех белого шума.

Система с упреждающим компенсатором LPV сравнивается с системой без компенсатора. Следует отметить, что система должна быть стабильной или стабилизированной контроллером с обратной связью. Как видно из результатов на рисунках 11 и 12, параметрозависимый компенсатор работает вполне успешно, аналогично результатам на рисунке 10. Таким образом, влияние внешних возмущений снижается почти на 100%. Таким образом, разработанный упреждающий компенсатор LPV имеет значительную эффективность в отношении помех. Наконец, конструкция компенсатора вполне удачна для системы LPV. В результате предложенный компенсатор вполне успешно подавляет измеримые и

Machine Translated by Google

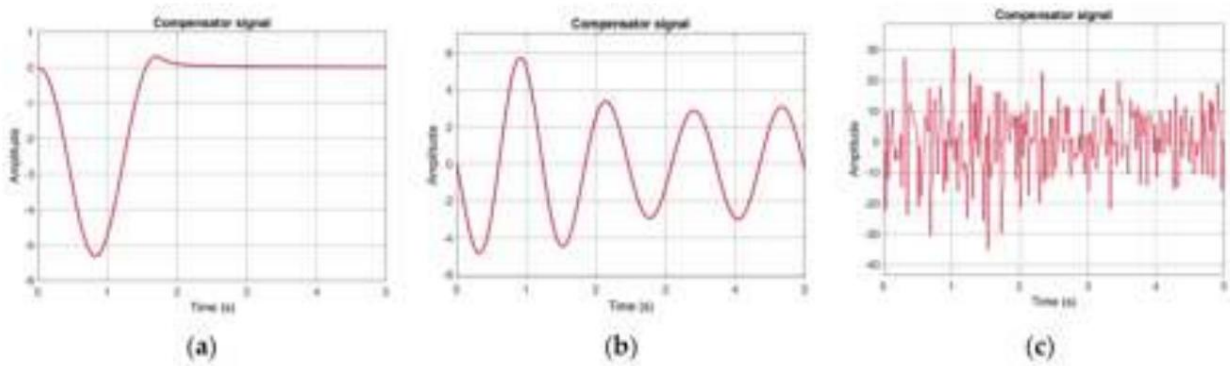


Рисунок 13. Сгенерированные сигналы силы для: (а) сигнала удара; (б) синусоидальный сигнал 5 рад/с; (с) сигнал белого шума.

В конструкции предлагаемого компенсатора ЛПВ на основе расширенного ЛМИ матрицу Ля-Пунова можно обозначить в виде ЛПВ, что снижает консервативность, без дополнительных подходов, таких как сетка, s-процедура, многовыпуклость в [11] и условия в [18,19]. Например, в [45] существует множество условий LMI для конструкции регулятора LPV , поскольку используются условия многовыпуклости из-за произведений A(α) × P(α) в (6). Поэтому предлагаемое решение является более простым и менее консервативным. В противном случае потребуются либо неэффективные решатели билинейного матричного неравенства (BMI), либо итерационные алгоритмы LMI.
Решения не требуют каких-либо итеративных алгоритмов LMI или BMI.

6. Выводы и будущие работы.

В данной статье основное внимание уделяется компенсатору с прямой связью от помех в дополнение к контроллеру с обратной связью по состоянию для системы LPV. Матрица Ляпунова отделяется от матриц системы путем введения слабой действительной переменной в планы на основе LMI; таким образом, обозначается PDLM. Таким образом, предлагаемые формулировки LMI как для контроллера с обратной связью по состоянию, так и для компенсатора с прямой связью являются конечномерными и не требуют дополнительных условий, таких как s-процедура, сетка и многовыпуклость, даже если используется PDLM. Таким образом, они имеют более простую структуру и меньшую консервативность, чем аналогичные конструкции с обратной связью по состоянию и прямой связью. Численные результаты работы регулятора с обратной связью по состоянию показывают, что решение менее консервативно и проще даже для очень больших диапазонов параметров. Результаты моделирования показывают, что как параметрозависимые, так и независимые от параметров компенсаторы с прямой связью, особенно для компенсатора, обладают значительной эффективностью против помех, поскольку эффекты внешних возмущений практически полностью устраняются . Таким образом, обоснованность и преимущества предлагаемых LMI эффективно достигаются с помощью числовых примеров и примеров моделирования. В результате предлагаемые решения LMI лучше подходят для систем с точки зрения простоты и консервативности.

Для дальнейших исследований первый из них можно рассматривать как получение оптимального вещественного скаляра μ с помощью методов метаэвристической оптимизации. Второй может одновременно спроектировать как управление с обратной связью, так и компенсатор с прямой связью для нестабильных систем. Третий — конструкция включения компенсатора прямой связи LPV через расширенный LMI. Кроме того, этот предлагаемый компенсатор с прямой связью с переключаемым компенсатором с прямой связью LPV можно сравнить по таким критериям, как производительность и сложность.

Вклад автора: Концептуализация, М.И. и Я; методология, М.И. и Я; программное обеспечение, М.И. и Я; проверка, да; формальный анализ, Я; расследование, М.И. и Я; ресурсы, М.И.; курирование данных, М.И.; письмо – подготовка первоначального черновика, М.И.; написание — обзор и редактирование, Я; визуализация, М.И. и Я; руководство Ю.А. Все авторы прочитали и согласились с опубликованным вариантом рукописи.

Финансирование: Данное исследование не получило внешнего финансирования.

Заявление о доступности данных: оригинальные материалы, представленные в исследовании, включены в статью, дальнейшие запросы можно направить соответствующему автору.

Machine Translated by Google

Конфликты интересов: Автор Мустафа Инчи работал в компании Volta Motor Industry and Trade Inc. Остальные авторы заявляют, что исследование проводилось в отсутствие каких-либо коммерческих или финансовых отношений, которые могли бы быть истолкованы как потенциальный конфликт интересов.

Рекомендации

1. Ким, К. Робастное статическое управление с обратной связью по выходу H для маршрутизаторов TCP/AQM на основе оптимизации LMI. Электроника 2024, 13, 2165. [Перекрестная ссылка]
2. Ким Ю.; Тран, ТВ; Ким, К.-Х. Надежное частотно-адаптивное бездатчиковое управление напряжением сети на основе EMPC для инвертора, подключенного к сети с LCL-фильтром. Электроника 2024, 13, 998. [CrossRef]
3. Немет, З.; Кучманн, М. Контроллер и наблюдатель на основе линейной матрицы для индукционной машины. Электроника 2022, 11, 3894. [CrossRef]
4. Ндабарушимана, Э.; Цинь, Н.; Ма, Л. Развязка помех для однофазного выпрямителя с широтно-импульсной модуляцией на основе Расширенный фильтр H-Infinity. Электроника 2023, 12, 2765. [CrossRef]
5. Бойд, С.; Эль Гауи, Л.; Ферон, Э.; Балакришнан, В. Линейные матричные неравенства в теории систем и управления; Общество промышленных и Издательство Applied Mathematics (SIAM): Филадельфия, Пенсильвания, США, 1994 г.; Том 15, ISBN 0898714850.
6. Де Оливейра, МС; Геромель, JC; Бернуссу, Дж. Расширенные характеристики норм H2 и H и параметризация контроллера для Системы дискретного времени. Межд. Дж. Контроль 2002, 75, 666–679. [Перекрестная ссылка]
7. Бара Г.И. Робастный анализ и управление параметрически-зависимыми неопределенными дескрипторными системами. Сист. Контрольная буква. 2011, 60, 356–364. [Перекрестная ссылка]
8. Цао, Ю.-Ю.; Лин З. Дескрипторный системный подход к робастному анализу устойчивости и синтезу контроллера. IEEE Транс. Автомат. Контроль 2004, 49, 2081–2084. [Перекрестная ссылка]
9. Де Каньи, Ж.; Камино, Дж. Ф.; Оливейра, RCLF; Перес, ПЛД; Свеверс, Дж. Управление дискретным временем H2 и H с планированием усиления Политопные изменяющиеся во времени системы. IET Теория управления Appl. 2010, 4, 362–380. [Перекрестная ссылка]
10. Оливейра, MCD; Геромель, JC; Бернуссу, Дж. Подход к оптимизации LMI к проектированию многокритериальных контроллеров для систем с дискретным временем. В материалах 38-й конференции IEEE по принятию решений и управлению (кат. № 99CH36304), Феникс, Аризона, США, 7–10 декабря 1999 г.; IEEE: Пискатауэй, Нью-Джерси, США, 1999 г.; Том 4, стр. 3611–3616.
11. Гахинет, П.; Апкарян П.; Чилали, М. Аффинные функции Ляпунова, зависящие от параметра, и вещественная параметрическая неопределенность. IEEE Пер. Автомат. Контроль 1996, 41, 436–442. [Перекрестная ссылка]
12. Пайплерс, Г.; Демёленэр, Б.; Свеверс, Дж.; Ванденберге, Л. Системы и контрольные буквы, расширенные характеристики LMI для Стабильность и производительность линейных систем. Сист. Контрольная буква. 2009, 58, 510–518. [Перекрестная ссылка]
13. Ку, К.-К.; Чен, Г.-В. Разработка контроллера с планированием усиления для дискретно-линейных систем с изменяющимися параметрами и мультипликативной функцией Шумы. Межд. J. Управление автомат. Сист. 2015, 13, 1382–1390. [Перекрестная ссылка]
14. Беккер Г.; Паккард, А. Устойчивая работа линейных параметрически изменяющихся систем с использованием параметрически зависимой линейной обратной связи. Сист. Контрольная буква. 1994, 23, 205–215. [Перекрестная ссылка]
15. Ферон Э.; Апкарян П.; Гаине, П. S-процедура анализа систем управления с параметрическими неопределенностями с помощью зависимых от параметра функций Ляпунова. В материалах Американской конференции по контролю 1995 г., Сиэтл, Вашингтон, США, 21–23 июня 1995 г.; IEEE: Пискатауэй, Нью-Джерси, США, 1995 г.; Том 1, стр. 968–972.
16. Вэй, Х.; Верхаген, М. Решения LMI для смешанной конструкции наблюдателя для обнаружения неисправностей H-/H для линейного изменения параметров Системы. Межд. Дж. Адаптируйся. Процесс управляющего сигнала. 2011, 25, 114–136. [Перекрестная ссылка]
17. Лим С. Анализ и управление системами с линейными параметрами; Стэнфордский университет: Стэнфорд, Калифорния, США, 1999 г.; ISBN 0599241535.
18. Оливейра, RCLF; Перес, PLD, зависящие от параметра LMI в робастном анализе: характеристика однородных полиномиально зависимых от параметра решений посредством релаксации LMI. IEEE Транс. Автомат. Контроль 2007, 52, 1334–1340. [Перекрестная ссылка]
19. Эбихара, Ю.; Посель, Д.; Арзелье, Д. Подход с использованием S-переменных к робастному управлению на основе LMI; Шпрингер: Берлин/Гейдельберг, Германия, 2015 г.; ISBN 978-1-4471-6605-4.
20. Эбихара, Ю.; Хагивара, Т. Новые характеристики расширенного LMI для синтеза многокритериальных контроллеров непрерывного времени. Автоматика 2004, 40, 2003–2009. [Перекрестная ссылка]
21. Сонг, Л.; Ян, Дж. Улучшенный подход к робастному анализу устойчивости и синтезу контроллеров для систем LPV. Межд. Дж. Робаст Нелинейное управление 2010, 21, 1574–1586. [Перекрестная ссылка]
22. Апкарян П.; Туан, HD; Бернуссу, Дж. Анализ в непрерывном времени, присвоение собственной структуры и синтез H2 с расширенными характеристиками линейных матричных неравенств (LMI). IEEE Транс. Автомат. Контроль 2001, 46, 1941–1946. [Перекрестная ссылка]
23. Эбихара, Ю.; Хагивара, Т. Расширенный подход LMI к робастному анализу производительности линейных, инвариантных во времени неопределенных систем. Автоматика 2005, 41, 1933–1941. [Перекрестная ссылка]
24. Трофино А.; Коутиньо, защитник; Барбоза, К.А. Улучшенные условия H2 и H для робастного анализа и синтеза управления линейными системами. Управление автоматом. 2005, 16, 427–434. [Перекрестная ссылка]
25. Шакед, У. Улучшенные представления LMI для анализа и проектирования систем непрерывного времени политопного типа. Неопределенность. IEEE Транс. Автомат. Контроль 2001, 46, 652–656. [Перекрестная ссылка]
26. Вэй Дж.; Ли, Л. Дальнейшее улучшение представлений LMI для анализа и проектирования систем непрерывного времени с неопределенностью политопного типа 1. В материалах 5-й Азиатской конференции по управлению 2004 г. (IEEE Cat. No. 04EX904), Мельбурн, Виктория, Австралия, 20 –23 июля 2004 г.; Том 2, стр. 1–7.

Machine Translated by Google

27. Саджади-Киа, С.; Джаббари, Ф. О расширении ограниченного действительного матричного неравенства. Межд. Дж. Контроль 2012, 85, 1593–1601. [Перекрестная ссылка]

28. Се, В. Эквивалентное LMI-представление ограниченной вещественной леммы для систем с непрерывным временем. J. Неравенства, приложение. 2008, 2008, 672905. [CrossRef]

29. Даббусси, К.; Зрида, Дж. Расширенный подход LMI к синтезу надежного H статического контроллера с обратной связью по состоянию. В материалах 16-й Международной конференции IEEE по электронике, схемам и системам 2009 г. (ICECS 2009), Ясмин Хаммамет, Тунис, 13–16 декабря 2009 г.; стр. 491–494.

30. Родригес, Луизиана; Оливейра, RCLF; Камино, Дж. Ф. Параметризированные LMI для надежного управления с обратной связью по состоянию политопных систем с непрерывным временем . Межд. Дж. Робастное нелинейное управление 2018, 28, 940–952. [Перекрестная ссылка]

31. Ху, К.; Джаймуха, И.М. Новая итеративная процедура, основанная на линейном матричном неравенстве, для управления с обратной связью по состоянию H2 и H политопных систем с непрерывным временем. Межд. Дж. Робастное нелинейное управление 2021, 31, 51–68. [Перекрестная ссылка]

32. Гусман, Дж.Л.; Хэгглунд, Т. Правила настройки для упреждающего управления на основе измеримых возмущений в сочетании с ПИД-регулированием: Обзор. Межд. Дж. Контроль 2021, 97, 2–15. [Перекрестная ссылка]

33. Чжу, В.; Чжан, З.; Чен, Дж.; Чжао, С.; Хуанг, С. Динамическое согласование данных для повышения производительности систем управления с прямой/обратной связью с шумом измерений. J. Управление процессами 2021, 108, 12–24. [Перекрестная ссылка]

34. Гарсиа-Маньяс, Ф.; Гусман, Дж.Л.; Родригес, Ф.; Беренгель, М.; Хэгглунд, Т. Экспериментальная оценка правил прямой настройки . Управление англ. Практика. 2021, 114, 104877. [CrossRef]

35. Ду, Ю.; Цао, В.; Она, Дж.; Ву, М.; Фанг, М. Подавление помех посредством прямой компенсации с использованием расширенного подхода, основанного на эквивалентных входных помехах. Дж. Франкл. Инст. 2020, 357, 10977–10996. [Перекрестная ссылка]

36. Хуан, М.; Дэн, Ю.; Ли, Х.; Ван, Дж. Ослабление пульсаций крутящего момента в СДСМ с использованием улучшенной устойчивой системы с двумя степенями свободы Контроллер через расширенный режим наблюдения за параметрами. ИСА Транс. 2022, 129, 558–571. [Перекрестная ссылка]

37. Адам, Э.Дж.; Маркетти, Дж. Л. Проектирование и настройка надежных контроллеров с прямой связью. Вычислить. хим. англ. 2004, 28, 1899–1911. [Перекрестная ссылка]

38. Пэн, К.; Сюй, Х.; Чжан Дж. Проект управления с двумя степенями свободы и улучшенным представлением H LMI. J. Неравенства, приложение. 2012, 2012, 147. [CrossRef]

39. Кёсе, И.Э.; Шерер, К.В. Надежное упреждающее управление с коэффициентом усиления L2 в неопределенных системах с использованием динамических IQC. Межд. Дж. Робастное нелинейное управление 2009, 19, 1224–1247. [Перекрестная ссылка]

40. Альма, М.; Мартинес, Джей-Джей; Ландау, И.Д.; Бюш, Г. Проектирование и настройка компенсаторов прямой связи H-бесконечности уменьшенного порядка для Активный контроль вибрации. IEEE Транс. Система управления. Технол. 2012, 20, 554–561. [Перекрестная ссылка]

41. Пэн, К.; Чжан, З.; Цзоу, Дж.; Ли, К.; Чжан Дж. Устойчивое управление с прямой связью и обратной связью на основе внутренней модели для системы LPV с возмущениями. Дж. Управление процессами 2013, 23, 1415–1425. [Перекрестная ссылка]

42. Юань, К.; Ву, Ф. Робастное управление с прямой связью с переключением H2 и H неопределенных систем LFT. Межд. J. Робастное нелинейное управление. 2016, 26, 1841–1856. [Перекрестная ссылка]

43. Флепс-Дезассе, М.; Бюнте, Т.; Сваричек, Ф.; Брембек, Дж. LPV Упреждающее управление полуактивными подвесками для повышения устойчивости крена. Управление англ. Практика. 2018, 78, 1–11. [Перекрестная ссылка]

44. Алтун Ю.; Гулез, К.; Мумку, Т.В. Синтез статического LPV-регулятора с прямой связью для линейных систем с переменными параметрами. В материалах 9-й Азиатской конференции по контролю (ASCC) 2013 г., Стамбул, Турция, 23–26 июня 2013 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2013 г.; стр. 1–4.

45. Алтун Ю.; Гулез, К. Синтез управления с упреждением с изменением линейных параметров с использованием функции Ляпунова, зависящей от параметра. Нелинейная Дин. 2014, 78, 2293–2307. [Перекрестная ссылка]

46. Венкатараман, Р.; Зайлер, П. Выпуклый LPV-синтез оценщиков и прямых связей с использованием двойственности и интегрального квадратичного уравнения Ограничения. Межд. Дж. Робастное нелинейное управление 2018, 28, 953–975. [Перекрестная ссылка]

47. Лёфберг Дж. Автоматическое робастное выпуклое программирование. Оптим. Методы 2012, 27, 115–129. [Перекрестная ссылка]

48. Лабит Ю.; Посель, Д.; Хенрион, Д. ИНТЕРФЕЙС SEDUMI 1.02: инструмент для решения проблем LMI с помощью SEDUMI. В материалах Международного симпозиума IEEE 2002 г. по проектированию систем автоматизированного управления, CACSD 2002 — Proceedings, Глазго, Великобритания, 20 сентября 2002 г.; стр. 272–277.

49. Гахинет, П.; Немировский А.; Лауб, Эй Джей; Чилали, М. Панель инструментов управления LMI. В материалах 33-й конференции IEEE по принятию решений и управлению 1994 г., Лейк-Буэна-Виста, Флорида, США, 14–16 декабря 1994 г.; The Mathworks, Inc.: Натик, Массачусетс, США, 1995; стр. 4–13.

50. Апкарян П.; Адамс, Р.Дж. Передовые методы планирования усиления для неопределенных систем. IEEE Транс. Система управления. Технол. 1998, 6, 21–32. [Перекрестная ссылка]

51. Чен, Х.; Го, К. Х. Ограниченное H - управление активными суспензиями: подход LMI. IEEE Транс. Контроль. Сист. Технол. 2005, 13, 412–421. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Адаптивный высокоэффективный драйвер светодиодной подсветки

Синьюй Сюй, Циюэ Чжо, Чуньхуэй Цзян, Цзычао Чжоу, Цзюньлун Тан и Ванхуэй Цзоу *

Школа физических и электронных наук, Университет науки и технологий Чанши, Чанша 410114, Китай * Адрес для переписки: zouwh@csust.edu.cn

Аннотация: Была предложена схема адаптивного высокоэффективного драйвера подсветки на светодиодах (LED) для решения проблемы дополнительных потерь мощности, вызванных изменением прямого напряжения светодиода. В этой схеме пиковый ток и рабочий цикл каждого канала светодиодов регулируются отдельно с помощью алгоритма адаптивного управления, чтобы минимизировать падение напряжения на линейном регуляторе тока (LCR) каждого канала светодиодов и уменьшить чрезмерные потери мощности в каждом канале светодиодов. и повысить общую энергоэффективность. Линейный стабилизатор тока, подходящий для адаптивного управления, разработан по технологии комплементарного металл-оксид-полупроводник (КМОП) 0,18 мкм с напряжением 5 В. Результаты моделирования показывают, что линейный регулятор тока может обеспечить линейно регулируемый ток канала в диапазоне от 0 до 48 мА с разрешением по току 0,2 мА. При различных углах процесса и температурах максимальная погрешность для всего диапазона тока составляет менее 0,1%. Площадь ядра микросхемы составляет около 0,1 мм2. Полный прототип драйвера включает в себя микросхемы LCR, внешние силовые МОП-транзисторы, цифровой модуль и светодиодные цепи. Результаты испытаний показывают, что потери мощности линейного стабилизатора тока были значительно снижены, а КПД каждого светодиодного канала составил около 98,1%.

Ключевые слова: драйвер светодиодной подсветки; адаптивное управление; линейное регулирование тока; высокая эффективность

check for updates

Цитирование: Сюй, Х.; Чжо, К.; Цзян, К.; Чжоу, З.; Тан, Дж.; Цзоу, В. Адаптивный высокоэффективный драйвер светодиодной подсветки. Электроника 2024, 13, 3057. https://doi.org/10.3390/электроника13153057

Академический редактор: Франческо Дриусси

Поступила: 22 июня 2024 г.

Пересмотрено: 25 июля 2024 г.

Принято: 30 июля 2024 г.

Опубликовано: 2 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария.

Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Светоизлучающие диоды (LED) получили широкое распространение в качестве источников подсветки для устройств жидкокристаллических дисплеев (LCD) и других приложений. По сравнению с традиционными источниками света светодиоды обладают такими значительными преимуществами, как высокая эффективность, отличная стабильность цвета, высокая надежность, длительный срок службы, компактные размеры и экологическая безопасность [1–4]. Поскольку яркость светодиода пропорциональна его току, в приложениях светодиодной подсветки необходима эффективная и точная схема драйвера для достижения точного управления яркостью. Кроме того, для обеспечения достаточного покрытия всегда используются параллельные каналы светодиодов (цепочки светодиодов) [5,6]. Традиционный драйвер светодиодной подсветки в основном состоит из повышающего преобразователя, параллельных светодиодных каналов и соответствующих линейных регуляторов тока (LCR). Повышающий преобразователь обеспечивает достаточное напряжение для включения всех светодиодов в каждом канале, а LCR регулирует ток для каждого канала для управления яркостью [7–9]. Разработка драйверов светодиодной подсветки, являющаяся важной областью применения управления питанием , всегда была популярным направлением исследований для многочисленных университетов и исследовательских институтов по всему миру. Цель состоит в том, чтобы оптимизировать производительность системы посредством инновационных схемотехники и системной архитектуры с упором на такие показатели, как эффективность системы [10–15], коэффициент регулировки яркости [16–20] и точность тока [21–25].

Из-за различий в технологических процессах, эффектов старения и изменений температуры светодиоды одного и того же типа могут иметь разное прямое напряжение, что приводит к значительным расхождениям в напряжении на разных каналах светодиодов. Следовательно, необходимо ввести определенный запас напряжения для обеспечения правильной работы в различных услов Однако это также приводит к увеличению потерь мощности и снижению энергоэффективности [26,27]. Механизм цифрового управления на основе R-DAC был предложен в [28] для решения проблемы увеличения потерь мощности из-за чрезмерного напряжения. Сравнение напряжения

Machine Translated by Google

на LCR и подается заданное опорное напряжение, которое управляет регулировкой напряжения обратной связи повышающего преобразователя постоянного тока через цифро-аналоговый преобразователь (ЦАП). Затем выход повышающего преобразователя регулируется для поддержания напряжения LCR на фиксированном уровне. В [29] был предложен метод адаптивного отслеживания задания с использованием схемы «Минимального напряжения» и 3-разрядного регулируемого резистора в качестве резистора обратной связи для регулировки выходного сигнала повышающего преобразователя. На точность выходного напряжения влияет точность и диапазон регулировки резистора. В [30–34] с помощью «Селектора минимального напряжения» или «Селектора самого низкого канала» минимальное напряжение LCR среди каналов определяется и используется для динамической регулировки рабочего цикла сигнала переключения повышающего преобразователя и изменения выходного сигнала повышающего преобразователя. В [35,36] диоды используются для выбора минимального напряжения LCR, метод, который нелегко реализовать в стандартных КМОП-процессах. В [2] были проанализированы преимущества и недостатки различных топологий светодиодных драйверов и определены будущие направления исследований.

В целом, для большинства существующих решений основная стратегия состоит в том, чтобы отрегулировать выходное напряжение повышающего преобразователя, чтобы уменьшить избыточное напряжение на LCR, тем самым уменьшая потери и повышая эффективность. Однако в многоканальной конфигурации из-за межканальных различий более эффективно оптимизировать каждый канал отдельно, чтобы еще больше снизить потери и повысить эффективность. В этой статье мы предлагаем адаптивную высокоэффективную схему управления светодиодной подсветкой, которая решает проблему межканальных различий. Для каждого канала используются как линейное регулирование тока, так и широтно-импульсная модуляция За счет линейного регулирования тока ток канала регулируется, а напряжение на LCR в каждом канале минимизируется до заранее определенного фиксированного значения. Впоследствии, посредством широтно-импульсной модуляции, рабочий цикл тока канала регулируется, калибруя средний ток каждого канала до одного и того же уровня. Эта стратегия гарантирует, что каждый канал светодиодов работает с минимальными потерями, тем самым обеспечивая максимальную эффективность.

Структура остальных частей статьи следующая: В разделе 2 подробно описана технология адаптивного управления, разработанная в данной статье. В разделе 3 обсуждаются и анализируются результаты моделирования и результаты испытаний предлагаемого адаптивного высокоэффективного драйвера светодиодной подсветки. Раздел 4 завершает статью.

2. Методика адаптивного управления

2.1. Принцип адаптивного управления

На рис . 1 показана структура предлагаемого адаптивного высокоэффективного драйвера светодиодной подсветки, который содержит регулятор яркости, повышающий преобразователь и несколько светодиодных каналов. Каждый светодиодный канал состоит из светодиодной цепочки, LCR и высоковольтного МОП-транзистора, действующего как переключатель для управления рабочим циклом. Контроллер яркости включает в себя 9-битный R-DAC для формирования напряжения VC, который служит опорным напряжением и определяет выходной сигнал повышающего преобразователя Vboost. Сигнал SENSE указывает, находится ли LCR в правильном состоян Контроллер яркости контролирует SENSE каждого канала и генерирует сигналы VC, PWM и CM_SW, следуя определенному алгоритму и шагам. CM_SW и PWM — это многобитные цифровые сигналы, которые управляют пиковым током и рабочим циклом канала соответственно. Этот процесс гарантирует, что каждый канал светодиодов работает с высокой эффективностью.

Судя по рисунку, для канала n существует зависимость напряжения как

$$Vboost = VLED,n + VMP,n + VLCR,n \quad (1)$$

где VLED — напряжение на цепи светодиодов или сумма прямых напряжений каждого светодиода. VMP — напряжение исток-сток высоковольтного МОП-транзистора MP, а VLCR — напряжение LCR. «n» в нижнем индексе указывает канал n, что относится ко всем символам в этой статье. Если адаптивное управление отключено и канал светодиода не находится в оптимальном состоянии, то в напряжении LCR будет избыточная доля, что можно выразить следующим образом:

$$VLCR,n = VLCR_min + VLCR_exc,n \quad (2)$$

Machine Translated by Google

"=" + (2)

где VLCR_min — минимально необходимое напряжение LCR для обеспечения правильной работы, то есть где VLCR_min
и минимально необходимое напряжение LCR для обеспечения правильной работы, что
Другими словами, выходной транзистор bLCR работает правильно и обеспечивает точный выходной сигнал , то есть
выходной транзистор bLCR работает правильно и обеспечивает точный выходной сигнал.
текущий. VLCR_exc представляет собой чрезмерную часть напряжения LCR. Для одного канала ток. VLCR_exc
представляет собой чрезмерную часть напряжения LCR для одного канала.
Потери мощности можно рассчитать по формуле
Потери мощности можно рассчитать по формуле

$$P_{loss,n} = I_{ave,n} \times V_{loss,n} = \overline{I}_{ave,n} \times (V_{LCR,n} + V_{MP,n}) \quad (3)$$

где Iave,n — средний ток, который определяется пиковым током Ipeak,n и где Iave,n — средний ток, который
определяется пиковым током Ipeak,n и
текущий рабочий цикл δn. Vloss,n представляет собой сумму VLCR,n и VMP,n. Если сопротивления включения тока
скважности δn. Vloss,n представляет собой сумму VLCR,n и VMP,n. Если сопротивление включения
Высоковольтные МОП-транзисторы достаточно низкие, ВМП оказывает ограниченное влияние на Плосс. Потому что
Высоковольтные МОП-транзисторы достаточно низкие, ВМП оказывает ограниченное влияние на Плосс. Потому что
VMP и VLCR_min являются необходимыми условиями для поддержания правильной работы MOS и LCR, VMP и VLCR_min
являются необходимыми условиями для поддержания правильной работы MOS и LCR,
чрезмерная потеря мощности определяется как
чрезмерные потери мощности определяются как

$$P_{loss_exc,n} = I_{ave,n} \times V_{LCR_exc,n} \quad (4)$$

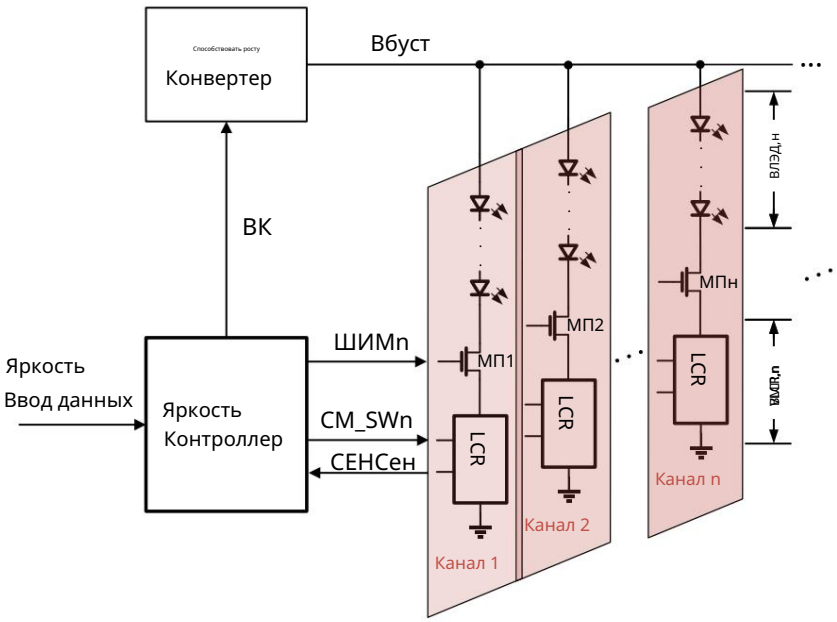


Рис. 1. Структура предлагаемого адаптивного высокоэффективного драйвера светодиодной подсветки.
Рис. 1. Структура предлагаемого адаптивного высокоэффективного драйвера светодиодной подсветки.

2.2. Процесс адаптивного управления
2.2. Процесс адаптивного управления

Процесс адаптивного управления, подходящий для многоканальных драйверов светодиодов, предназначен для
Процесс адаптивного управления, подходящий для многоканальных драйверов светодиодов, предназначен
уменьшить чрезмерные потери мощности, вызванные чрезмерным напряжением. При запуске системы появляется
для уменьшения чрезмерных потерь мощности, вызванных чрезмерным напряжением. При запуске системы появляется
яркости инициирует процесс инициализации, в ходе которого пиковый ток и
Контроллер инициирует процесс инициализации, во время которого пиковый ток и рабочий цикл каждого
светодиодного канала индивидуально устанавливаются для достижения равномерной яркости и
Цикл каждого светодиодного канала настраивается индивидуально для достижения равномерной яркости и
одновременно свести к минимуму чрезмерные потери мощности. После фазы инициализации яркость
одновременной минимизации чрезмерных потерь мощности. После фазы инициализации контроллер яркости
переходит в фазу отслеживания, где он постоянно контролирует состояние
Контроллер переходит в фазу отслеживания, где он постоянно контролирует состояние LCR. Затем настройки пикового
тока и рабочего цикла корректируются в соответствии с
ЛКР. Затем настройки пикового тока и рабочего цикла корректируются соответствующим образом с учетом колебаний
источника питания, температуры и других факторов.
умеренные колебания электропитания, температуры и других факторов.

На рисунке 2 показана блок-схема этапа инициализации. Во-первых, входные данные яркости , включая
На рисунке 2 показана блок-схема этапа инициализации. Во-первых, входные данные яркости , включая
номинальный пиковый ток Iset и рабочий цикл δset, загружаются в
данные, включая номинальный пиковый ток Iset и рабочий цикл δset, загружаются в контроллер яркости. Впоследствии
ток LCR каждого канала устанавливается на Iset через
регулятор яркости. Впоследствии ток LCR каждого канала устанавливается на Iset посредством сигнала CM_SW. Затем
контроллер яркости начинает увеличивать VC, в результате чего Vboost падает.
Сигнал CM_SW. Затем контроллер яркости начинает увеличивать VC, что приводит к соответствующему уменьшению
Vboost . Между тем, контроллер яркости одновременно контролирует
соответственно сгибайте. При этом контроллер яркости одновременно контролирует сигнал состояния SENSE с
каждого канала. Как только все сигналы SENSE изменятся с 0 на 1,
тус сигнала SENSE с каждого канала. Как только все сигналы SENSE изменятся с 0 на 1, увеличение VC прекращается .
В этот момент пиковые токи всех каналов достигли заданного значения.
сгибание упоров BK . В этот момент пиковые токи всех каналов достигают заданного значения Iset, и всегда существует
один канал (предположительно канал k), значение SENSE которого

сигнал меняется последним.

Machine Translated by Google

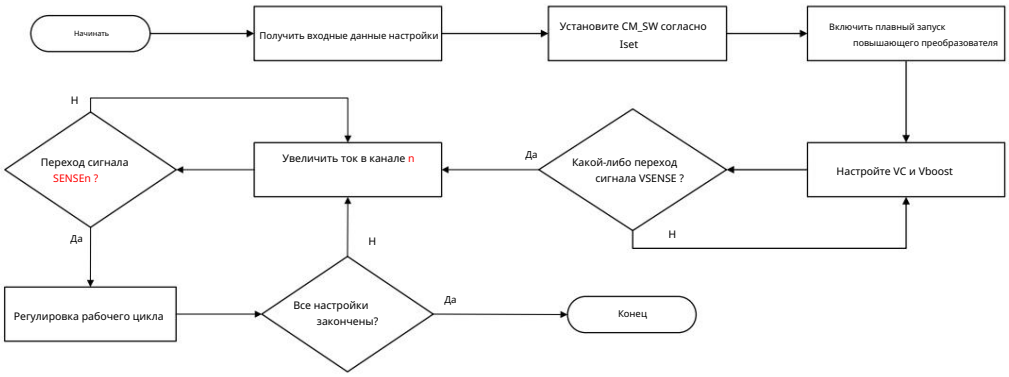


Рисунок 2. Блок-схема процесса адаптивного управления на этапе инициализации.

Далее шаг регулировки пикового тока выполняется последовательно на всех каналах, кроме канала k. Контроллер яркости пытается увеличить ток канала, предоставляя новую настройку CM_SW. Если SENSE остается на уровне 1, контроллер яркости продолжает этот процесс. Однако если SENSE изменяется с 1 на 0, контроллер яркости возвращается к предыдущей настройке CM_SW, завершая процесс настройки для этого канала. Затем соответствующий канал имеет новый пиковый ток. Регулятор яркости приступает к проведению процесса регулировки для последующего канала. После завершения всех регулировок тока напряжение на каждом LCR приближается к VLCR_min, а избыточная составляющая минимизируется.

Наконец, чтобы обеспечить одинаковую яркость, каждый канал светодиода должен иметь одинаковый средний ток. Таким образом, коэффициент заполнения канала n (δn) регулируется регулятором яркости следующим образом:

$$\delta n = \delta set \times \frac{Iset}{I_{пик,н}} \quad (5)$$

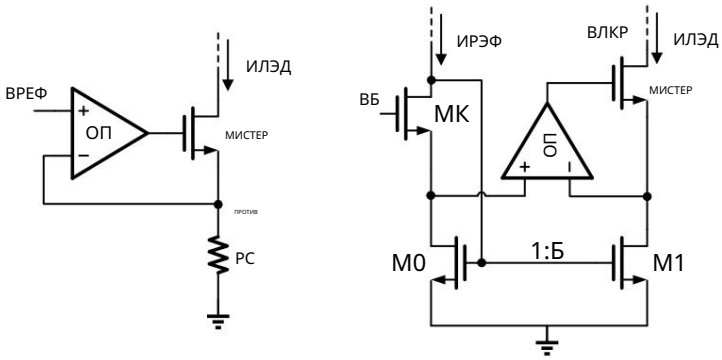
где Ipeak,n — пиковый ток n-го канала светодиода после процесса настройки, как описано выше. δset и Iset представляют начальный рабочий цикл и пиковый ток соответственно. Стоит отметить, что согласно уравнению (4) средний ток остается неизменным.

Посредством вышеупомянутого процесса пиковый ток Ipeak,n и рабочий цикл δn для каждого канала светодиода регулируются, тем самым сводя к минимуму дополнительное энергопотребление, возникающее из-за чрезмерного напряжения. Скорректированный средний ток точно соответствует первоначально установленному среднему току. Учитывая, что выходная яркость светодиодов напрямую коррелирует с их током, метод адаптивного управления гарантирует, что уровни яркости во всех каналах светодиодов остаются постоянными. При переходе на этап отслеживания основной принцип аналогичен этапу инициализации. Оба они сочетают в себе регулировку пикового тока и рабочего цикла и не будут подробно описаны в этой статье.

2.3. Линейный регулятор тока

Поскольку яркость светодиода напрямую зависит от его тока, точный контроль яркости требует точного регулирования тока. Во многих решениях драйверов светодиодной подсветки LCR обычно используется для генерации и управления током светодиодов. На рисунке 3 показаны две типичные конфигурации схемы LCR [37]. Ключевой целью проектирования схем LCR является достижение высокой точности измерения тока. На рисунке 3a операционный усилитель и транзистор M1 образуют отрицательную обратную связь, обеспечивая равенство VS VREF. Следовательно, выходной ток ILED равен VREF/RS. Очевидно, что точность RS напрямую влияет на точность ILED . Важно отметить, что токи возбуждения светодиодов подсветки обычно значительны. В этой статье максимальный ток возбуждения составляет 48 мА. Кроме того, чтобы минимизировать потери мощности, VS или VREF следует поддерживать как можно меньшими, что требует небольшого RS для достижения больших токов. Однако чрезмерно малый RS может существенно усилить негативное влияние паразитных свойств схемы на точность выходного тока.

Machine Translated by Google



(а) (б)

Рисунок 3. Типовые схемы линейного регулятора тока. а) структура регулирования источников; (б) Зеркало текущая структура.

На рисунке 3б изображен другой тип LCR. В этой конфигурации выходной ток ILED генерируется не напрямую, как на рисунке 3а, а посредством зеркального отображения тока. В этой структуре ILED определяется соотношением B × IREF, где IREF — опорный ток, а B — текущий зеркальный коэффициент, который определяется соотношением размеров транзисторов M1 и M0. Операционный усилитель (ОП) играет решающую роль, обеспечивая одинаковое VDS транзисторов M1 и M0, тем самым гарантируя точное зеркальное отображение тока.

Фактически, M0 и M1 могут работать в линейной области вместо области насыщения за счет снижения VB, тем самым уменьшая VDS, что помогает снизить VLCR и уменьшить избыточную мощность [38]. Классическое выражение тока стока для МОП-транзисторов, работающих в линейной области, имеет вид

$$\text{Транзистор теперь может быть приблизительно эквивалентен линейному резистору.} \quad \text{Как уже говорилось выше, из-за влияния ОП VDS M0 и M1 сохраняются одинаковыми. Тогда ток стока M0 и M1 имеет соотношение ff} \quad (6)$$

$$\frac{I_{D,M0}}{I_{D,M1}} = \frac{(Б/Д)_{M0}}{(Б/Д)_{M1}} = \frac{1}{Б} \quad (7)$$

Очевидно, что точное отражение тока гарантировано.

На основе структуры схемы, показанной на рисунке 3б, был разработан улучшенный LCR для адаптивного управления, изображенный на рисунке 4. Основным отличием от схемы на рисунке 3б является введение операционного усилителя OP2. OP2 вместе с транзистором M0 образует петлю отрицательной обратной связи, гарантируя, что напряжение исток-сток VDS M0, обозначенное на рисунке VA , остается равным VREF. Регулируя VREF, можно точно контролировать напряжение исток-сток VDS M0. Преимущество конструкции заключается в том, что она смягчает влияние изменений процесса и облегчает точный контроль параметров.

Когда VREF низкий, транзистор M0 работает в линейной области. Транзисторы M1, M2, . . ., MN служат токовыми зеркальными транзисторами, соответствующими M0, каждый из которых имеет одинаковый размер. Декодер определяет, активирован или деактивирован каждый зеркальный транзистор, на основе двоичного управляющего слова CM_SW, тем самым регулируя общий выходной ток. Когда затвор зеркального транзистора подключен к VCOM, он включается; при подключении к земле он выключается.

OP1 функционирует аналогично OP на рисунке 3b, гарантируя, что напряжение в узле D (обозначенном как VD) соответствует VA, тем самым поддерживая точное зеркальное отображение тока. В этой конструкции компаратор COMP работает как простой детектор состояния схемы с напряжением VCOM , установленным на 2,5 В. OP1 — это операционный усилитель с высоким коэффициентом усиления и выходной структурой повышения усиления [39], а OP2 — это операционный усилитель с выходным каскадом класса AB с полным размахом сигнала [40].

Machine Translated by Google

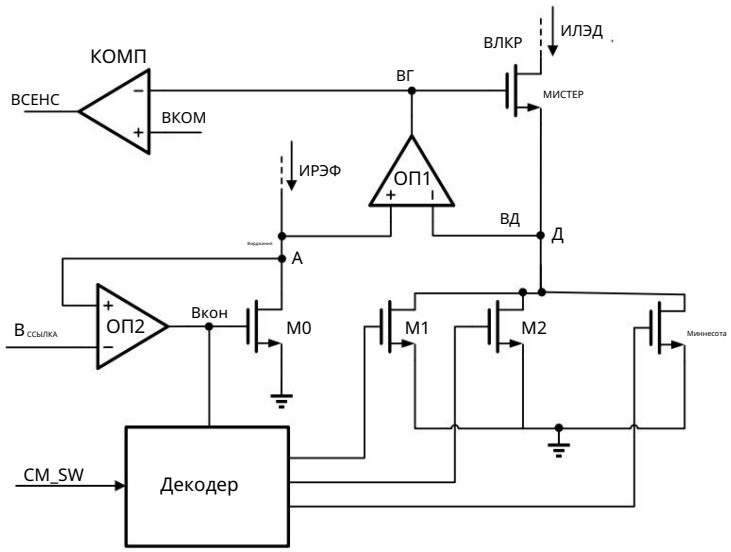


Рисунок 4. Схема LCR для адаптивного управления.

Когда Vboost достаточно высок, даже после учета падения напряжения в цепи светодиодов, VLCR остается достаточным для обеспечения правильной работы LCR, позволяя ILED достигать или поддерживать заданное значение. В этом состоянии напряжение затвора транзистора MR составляет примерно 1,86 В (VGS,MR + VREF) в соответствии с результатами моделирования в типичном углу процесса, а SENSE остается на уровне 1. Если Vboost уменьшается или если ILED увеличивается и делает прямое напряжение светодиодной цепи увеличится, VLCR соответственно уменьшится. Однако пока транзистор MR остается эффективным, LCR может продолжать нормально функционировать. Если VLCR уменьшается дальше и переводит транзистор MR в глубокую линейную область, VG необходимо увеличить, чтобы противодействовать влиянию уменьшающегося VLCR. Когда VG превышает VCOM, выход компаратора меняет состояние, и SENSE переключается на 0. Это указывает на то, что LCR больше не может поддерживать или достигать установленного ILED.

Все зеркальные транзисторы имеют одинаковый размер, а соотношение размеров каждого транзистора составляет B:1. с M0. Ток светодиода можно определить по следующему уравнению:

$$ILED = N \times B \times IREF \quad (8)$$

где N — количество активированных зеркальных транзисторов, определяемое CM_SW. IREF — опорный ток, генерируемый внешней цепью смещения. Следует отметить, что

Выходной ток ILED LCR фактически формирует вышеупомянутый пиковый ток Ipeak.

Даже на одном и том же кристалле существует определенная степень вариации процесса, которая в первую очередь влияет на размер устройства и характерные параметры, приводя к неточностям или несовпадениям текущего зеркалирования. Учитывая влияние соотношения ширины к длине (W/L) и порогового напряжения (Vth), изменение тока, вызванное изменением процесса, может быть (WL) выражено следующим образом:

$$\frac{\Delta I}{I} = \frac{\partial I}{\partial (W/L)} \frac{\Delta (W/L)}{(W/L)} + \frac{\partial I}{\partial V_{th}} \Delta V_{th} \quad (9)$$

Текущее рассогласование обычно определяется как скорость изменения тока. Предполагая корреляции между различными влияющими факторами нет, это можно выразить как:

$$\frac{\Delta I}{I} = \frac{1}{I} \frac{\partial I}{\partial V_{th}} \Delta (W/L) + \frac{\partial I}{\partial (W/L)} \Delta V_{th} \quad (10)$$

Если транзистор работает в области насыщения, по соответствующему стоку текущая формула, текущее несоответствие

$$\frac{\Delta I}{I} = \frac{\Delta (W/L)}{(W/L)} + \frac{2 \Delta V_{th}}{(V_{GS} - V_{th})} \quad (11)$$

Machine Translated by Google

В то время как, если транзистор работает в линейной области, согласно соответствующей формуле тока стока, рассогласование тока становится

$$\frac{I}{я} \text{ "=" } \frac{(Вт/Л)}{(Б/Д)} + \frac{2 \quad Vth}{2(ВГС \quad Vth) \quad VDS} \tag{12}$$

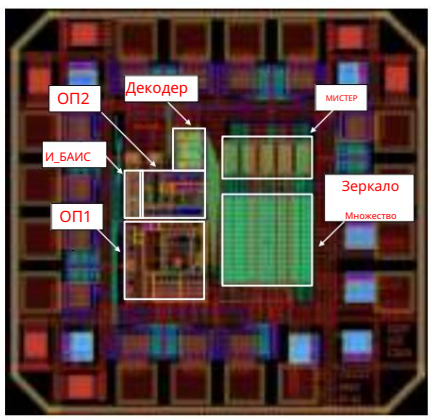
Сравнивая две приведенные выше формулы, можно видеть, что влияние изменения размера на рассогласование тока одинаково как в линейной области, так и в области насыщения. Однако влияние изменения порогового напряжения меньше в линейной области, что указывает на возможность достижения меньшего рассогласования тока в линейной области.

3. Результаты моделирования и измерений.

3.1. Проектирование системы

и микросхемы На основе рисунка 2 в данной статье создается и проверяется трехканальный адаптивный драйвер светодиодов, каждый канал которого содержит 6 светодиодов. Используется микросхема повышающего преобразователя, в частности TPS61046 от Texas Instruments. LCR спроектирован с использованием комплементарной технологии металл-оксид-полупроводник (КМОП) толщиной 0,18 мкм при напряжении 5 В. Поскольку этот процесс не поддерживает высоковольтные устройства, используются внешние МП-транзисторы. В идеале, если доступен процесс BCD, эти МР-транзисторы можно было бы интегрировать непосредственно в чип. Кроме того, предусмотрена дальнейшая интеграция контроллера яркости для получения монолитного интегрированного чипа драйвера, что является запланированной будущей работой. При моделировании схем модели PSPICE используются для повышающего преобразователя, светодиодов и транзисторов МР. Параметр напряжения включения в модели PSPICE для светодиодов намеренно настроен для внесения различий между каналами. В частности, для канала 3 напряжение включения светодиода установлено на 3,20 В, что соответствует типичному значению. Для каналов 1 и 2 параметры настроены на 2,73 В и 2,76 В соответственно. Как описано ранее, минимальное значение Vboost определяется VLED канала 3. Контроллер яркости реализован с использованием Verilog HDL и работает на частоте 100 МГц. Частота ШИМ-сигнала составляет 10 кГц. Вся схема работает в режиме моделирования смешанных сигналов для проверки процесса адаптивного управления и оценки производительности LCR.

С учетом диапазона регулирования тока и разрешения размер M0 в LCR выбран равным 1 мкм/1 мкм, а его отношение к зеркальному транзистору установлено равным 1:4. Всего имеется 240 токовых зеркальных транзисторов. Опорный ток IREF установлен на уровне 50 мкА, в результате чего выходной ток устройства составляет 200 мкА, а максимальный выходной ток — 48 мА. Размер MR выбран равным 40 мкм/2 мкм × 50. Общая схема схемы LCR показана на рисунке 5. Все зеркальные транзисторы и соответствующие им управляющие ключи расположены в виде матрицы 15 × 16. Площадь сердцевины составляет всего 325 × 310 мкм , тогда как вся площадь, включая площадки, составляет 700 × 720 мкм. Размеры ключевых транзисторов также приведены на рисунке 5.



Transistor	W/L
MR	40μm/2μm×50
M0	1μm/1μm
M1-MN	4μm/1μm

Рисунок 5. Схема LCR, изготовленная по КМОП-технологии 0,18 мкм, площадь сердцевины составляет 325 × 310 мкм2 .
Также приведены размеры ключевых транзисторов.

Рисунок 5. Схема LCR, изготовленная по КМОП-технологии 0,18 мкм, площадь сердцевины составляет 325 × 310 мкм2 .
Также приведены размеры ключевых транзисторов.

3.2. Результаты симуляции

На рисунке 6 показаны смоделированные выходные характеристики LCR при различных углах процесса и температурах. Видно, что когда выходной ток LCR находится в диапазоне от 0 мА до 48 мА, он демонстрирует хорошую линейность в различных условиях эксплуатации.

В частности, при настройке выходного тока 20 мА максимальное отклонение тока составляет 13.

Machine Translated by Google

3.2. Результаты симуляции

На рисунке 6 показаны смоделированные выходные характеристики LCR при различных углах процесса и температурах. Видно, что когда выходной ток LCR находится в диапазоне от 0 мА до 48 мА, он демонстрирует хорошую линейность в различных условиях эксплуатации. В частности, при настройке выходного тока 20 мА максимальное отклонение тока составляет 13 мкА, что приводит к ошибке тока менее 0,1%. В целом, LCR демонстрирует хорошую линейность и точность тока в указанном диапазоне тока.

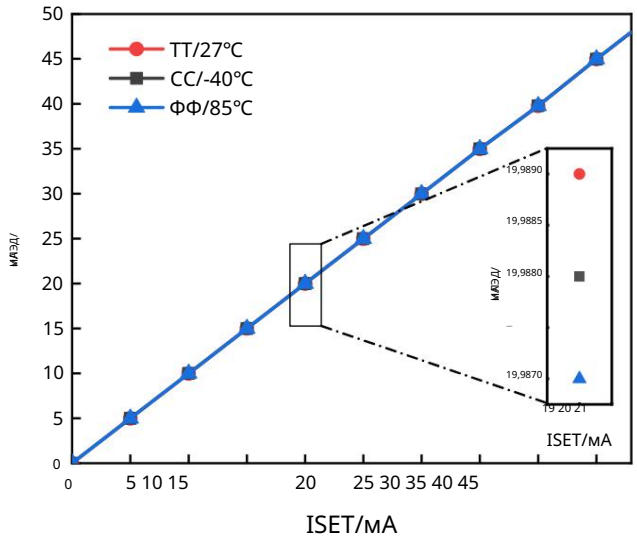


Рисунок 6. Выходной ток LCR в зависимости от настройки тока в различных рабочих условиях.

На рисунке 7 изображена форма волны напряжения в нескольких важных узлах во время адаптивного процесса. SENSE1, SENSE2 и SENSE3 представляют сигналы SENSE соответствующих каналов. T0 обозначает фазу плавного запуска повышающего преобразователя. При включении выходной сигнал повышающего преобразователя Vboost через определенный период времени постепенно достигает заданного значения. Это заданное значение представляет собой минимальное напряжение, необходимое для правильной работы схемы в различных рабочих условиях. В данной конструкции оно установлено на уровне 25,19 В. По мере увеличения Vboost сигналы SENSE каждого канала последовательно переходят в 0. T1 отмечает фазу регулировки Vboost . В течение этого времени контроллер яркости постепенно увеличивает VC, что приводит к соответствующему уменьшению Vboost. Видно, что когда Vboost достигает 21,52 В, SENSE3 переходит в 1, указывая на изменение состояния канала 3. Наконец, VC остается на уровне 0,33 В, а Vboost стабилизируется на уровне 21,62 В. T2 отмечает текущую фазу регулирования, в течение при котором ток канала 1 (канала 2) увеличивается. Когда SENSE1 (SENSE2) переходит в состояние 1, это указывает на то, что состояние канала 1 (канала 2) изменилось. Регулирование тока канала 1 (канала 2) прекращается, а рабочий цикл канала 1 (канала 2) устанавливается в соответствии с уравнением (4). T3 представляет собой нормальную рабочую фазу, которую также называют фазой непрерывного отслеживания, а сигналы SENSE динамически поддерживаются на

На рисунке 8 показана форма сигнала тока в каждом канале. Первоначально пиковое значение тока каждого канала установлено равным 20 мА с рабочим циклом по умолчанию 100%. На ранней стадии T0 Vboost не был полностью установлен, и все светодиоды выключены, что приводит к нулевому току в каждом канале. Поскольку светодиоды в канале 3 имеют наибольшее прямое напряжение, ток в канале 3 увеличивается до заданного значения последним, что соответствует времени сигналов SENSE на рисунке 7. После завершения фазы регулировки Vboost наступает фаза регулирования тока . начинается. Видно, что пиковые токи в канале 1 и канале 2 увеличены до 24,2 мА и 23,2 мА соответственно, при этом соответствующие коэффициенты заполнения отрегулированы до 82,6% и 86,2%, обеспечивая одинаковые средние токи по трем каналам. Когда начинается нормальная рабочая фаза, контроллер яркости будет постоянно контролировать состояние LCR. Если состояние изменится, то есть SENSE станет равным 1, пиковый ток и рабочий цикл будут соответственно обновлен

Machine Translated by Google

до 0. Судя по смоделированной форме сигнала тока, максимальное время нарастания и время спада составляют 173 нс и 74 нс соответственно, без заметного выброса тока.

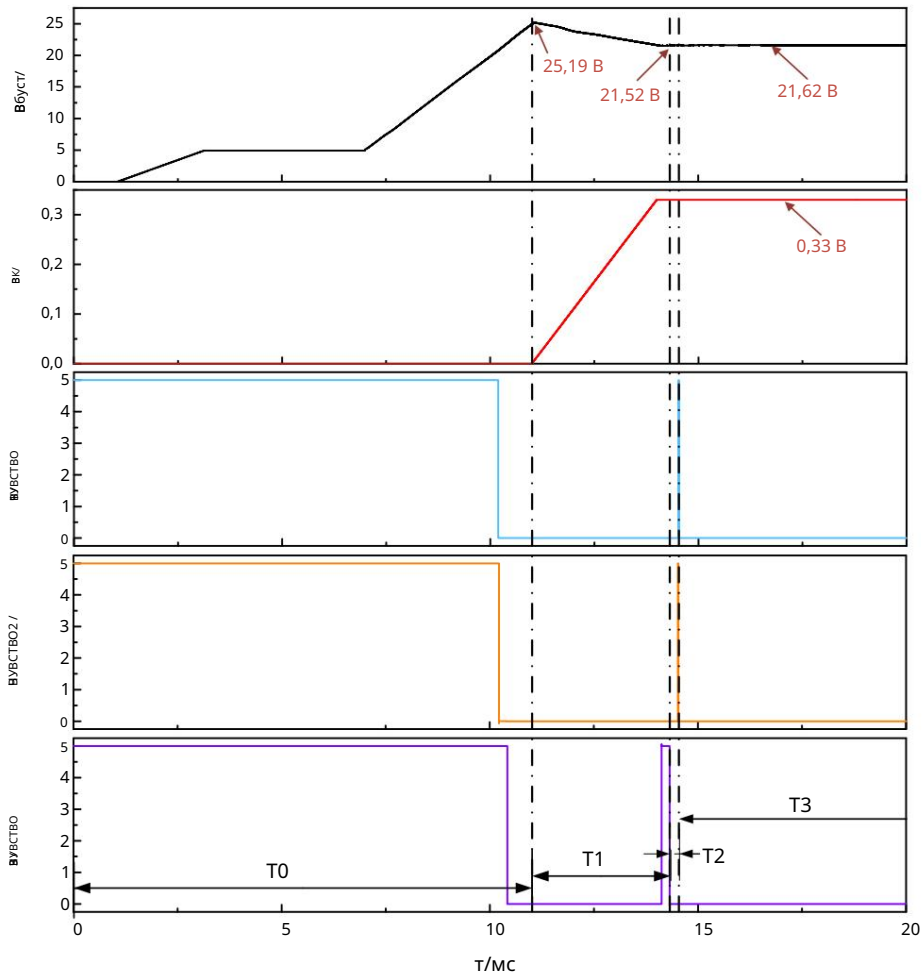


Рисунок 7. Формы напряжения в процессе адаптивного управления.

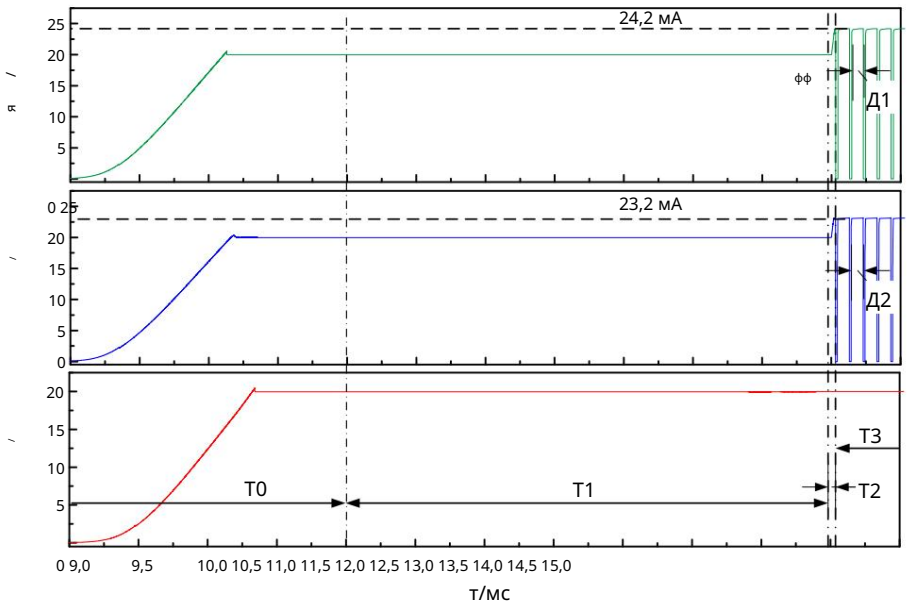


Рисунок 8. Формы тока каждого канала в процессе адаптивного управления.

Machine Translated by Google

На рисунке 9 сравниваются различия в потерях мощности между методом управления фиксированным пиковым током и методом адаптивного управления, предложенным в этой статье. В методе фиксированного пикового тока выполняется только шаг регулировки Vboost , в то время как пиковый ток и рабочий цикл остаются неизменными. Как описано ранее, при использовании метода фиксированного пикового тока канал 3 испытывает наименьшие потери мощности, тогда как каналы 1 и 2 страдают от больших потерь мощности из-за чрезмерного напряжения. На рисунке показаны потери мощности трех каналов при методе фиксированного пикового тока. Напротив, при методе адаптивного управления регулируются как пиковый ток, так и рабочий цикл, эффективно устраняя избыточное напряжение в канале 1 и канале 2, тем самым значительно снижая потери Как показано на рисунке, когда номинальный ток канала установлен на уровне 20 мА, метод адаптивного управления снижает потери мощности на 74%.

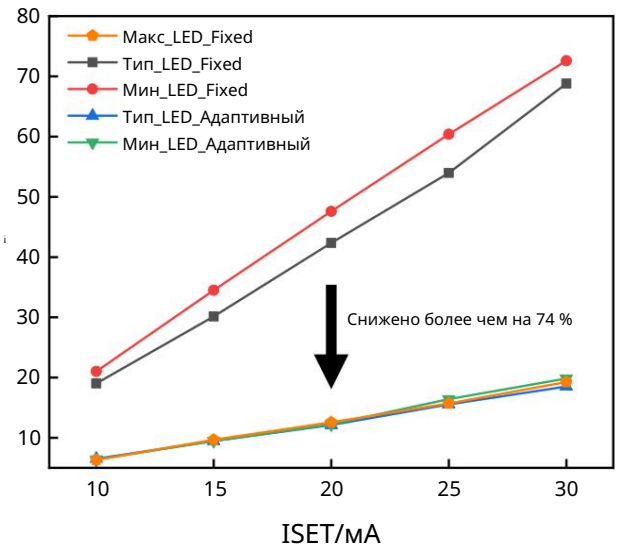


Рисунок 9. Потери мощности LCR при методе фиксированного пикового тока и предлагаемом методе адаптивного управления.

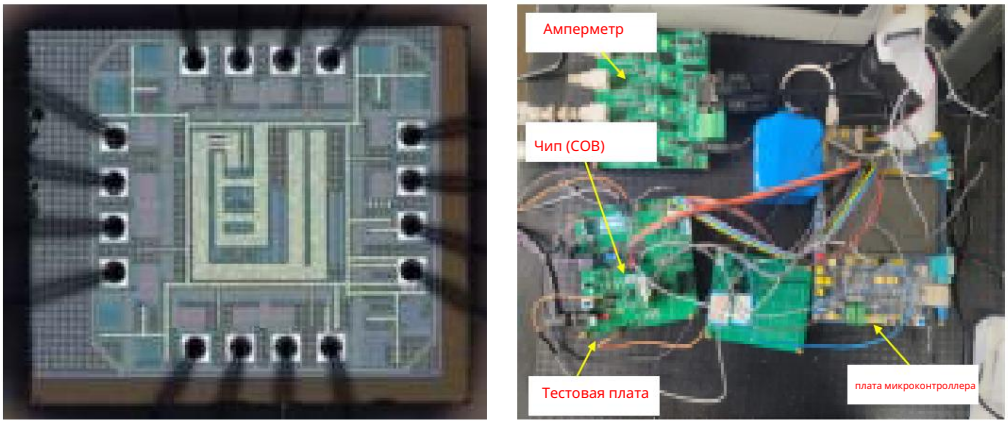
3.3. Результаты измерений

На рис . 10 представлена микрофотография изготовленного чипа с соединительными проводами, а также испытательной установки чипа. Чип помещен в корпус «чип-на-плате» (COB) и установлен на тестовой плате. Опорный ток IREF подается тестовой платой. Плата разработки микроконтроллера STM32F407 (от ST Microelectronics) служит контроллером яркости, а встроенный в микроконтроллер ЦАП используется для генерации VC. Между каждой цепочкой светодиодов и стоком высоковольтного МОП- транзистора последовательно включен многоканальный амперметр для измерения среднего тока каналов светодиода. Этот амперметр поддерживает запись данных в реальном времени через последовательный порт. Формы сигналов напряжения в узлах схемы фиксируются с помощью цифрового осциллографа, в частности RTB2004 от Rohde & Schwarz. Для простоты тестирования реализовано всего два светодиодных канала, по 6 светодиодов в каждом. Примечательно , что светодиоды намеренно выбраны так, чтобы создавать разницу в прямом напряжении на цепочке светодиодов между двумя каналами примерно в 500 мВ. Канал 2 имеет меньшее прямое напряжение. Это намеренное различие усиливает демонстрацию эффективности схемы.

Температура в помещении поддерживалась на уровне около 27 C. На рисунке 11 показаны измеренные формы напряжения VPWM и Vloss для канала 2 после начального процесса адаптивного управления. Уставка тока канала Iset составляет 20 мА. Четыре цифры соответствуют случаю, когда VREF установлен на 1 В, 0,6 В, 0,4 В и 0,3 В соответственно. Рабочий цикл VPWM в каждом случае зафиксирован как 68,57%, 68,86%, 68,43% и 68,53% соответственно. Когда VPWM высокий, канал включается, и Vloss в каждом случае измеряется на уровне 1,06 В, 0,69 В, 0,49 В и 0,36 В соответственно. Когда VPWM переходит в низкий уровень, канал закрывается, и увеличение Vloss объясняется главным образом влиянием эффектов то

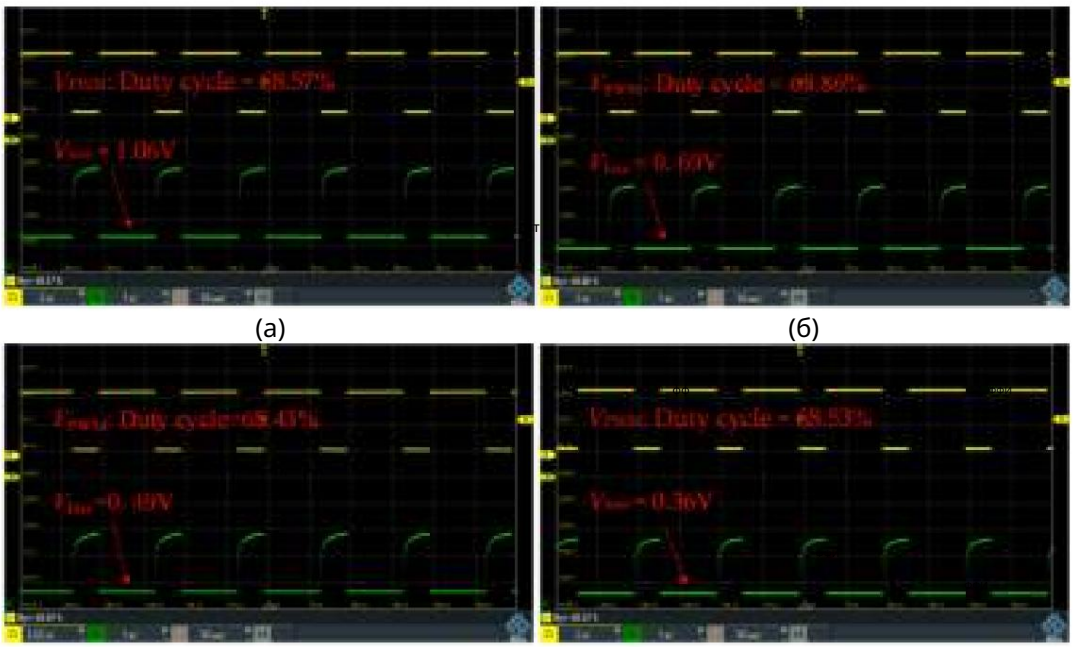
Machine Translated by Google

Из рисунков можно сделать вывод, что, несмотря на уменьшение VREF с 1 В до 0,3 В, что может привести к переходу токовых зеркальных транзисторов в LCR в линейную область, точность выходного тока все еще может сохраняться, и рабочий цикл VPWM сохраняется. согласованы во всех этих настройках.



(а) (б)

Рис. 10. (а) Микрофотография чипа LCR; (б) тест-система.

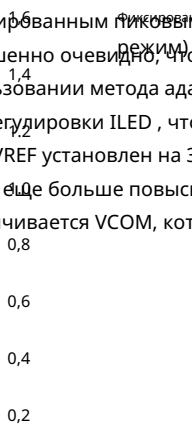


(а) (б)

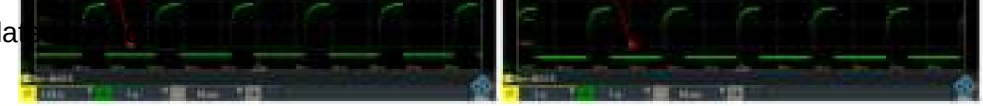
(с) (г)

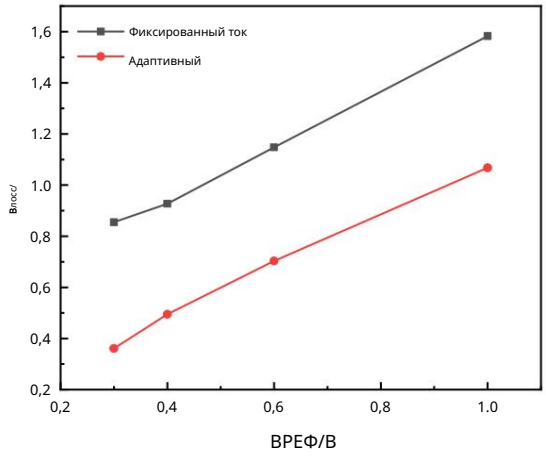
Рис. 11. Кривые напряжения VPWM и Vloss при VREF = 0,3 В (а) , VREF = 0,4 В (б) , VREF = 0,6 В (в) и VREF = 1 В (г). Iset = 20 мА.

На рисунке 12 показана разница в Vloss и энергоэффективности канала 2 как при методе управления с фиксированным пиковым током, так и при методе адаптивного управления. На рисунке 12а (Адаптивный совершенно очевидно, что в канале 2 присутствует избыточное напряжение около 500 мВ. При использовании метода адаптивного управления это избыточное напряжение эффективно снижается за счет регулировки ILED , что приводит к существенному повышению энергоэффективности. В частности, когда VREF установлен на 300 мВ, энергоэффективность может достигать 98,1% при ILED 20 мА. В принципе, можно еще больше повысить энергоэффективность, еще больше уменьшив VREF , причем верхний предел ограничивается VCOM, который зависит от источника питания.

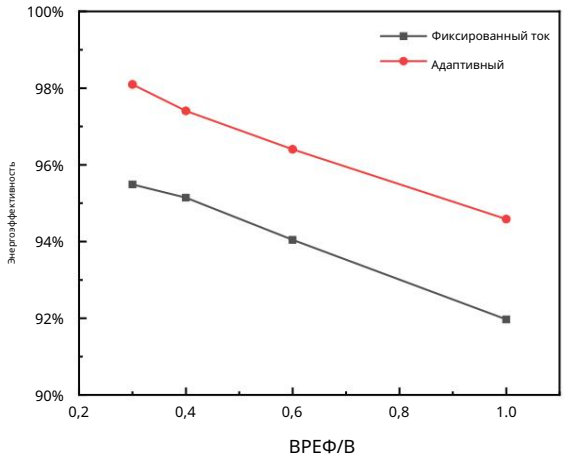


Machine Translat





(а)



(б)

Рисунок 12. Сравнение метода фиксированного пикового тока и предлагаемого адаптивного управления. ФФИ метод в отношении (а) Vloss и (b) энергоэффективности.

В таблице 1 сравниваются различные параметры и показатели производительности предлагаемых адаптивный драйвер и другие существующие ссылки. Предложенная конструкция ФФИ демонстрирует такие преимущества, как широкий диапазон выходного тока и высокая энергетическая эффективность. Фактически, Таблица 1 ФФИ показывает эффективность только одного канала. Однако, что более важно, предлагаемый Сэфф адаптивный метод направлен на повышение эффективности каждого канала в многоканальном светодиоде. Водитель. Он эффективно компенсирует чрезмерное напряжение на LCR за счет адаптивного управления. с умеренной сложностью реализации.

Таблица 1. Сравнение параметров производительности.

	[17]	[41] [42]	[43]	Эта работа
Артикул	0,35 мкм BCD 0,5 мкм CMOS	40 нм CMOS	40 нм CMOS	0,18 мкм CMOS
Процесс Площадь сердцевины (мм2)	Н/Д 6– +	Н/П/Д 3–4,2 3,2–4,2 20 25 2 6 +	Н/Д +	0,1
Источник питания (В)	27 30	5,36 22,5 ~300 ~675 94,4% 97%	3,2–4,2	5
Максимальный ILED (мА)	8	Да Да Нет Нет	25	48
Конфигурация светодиодной цепочки.				6
Вбуст (В)	26,41		6	18,96 *
Влосс (мВ)	~400		20	360 *
Энергоэффективность	98,4%		~480	98,1% *
Регулировка Vboost	Да			Да
Поканальная регулировка	Нет		97,6% Да Нет	Да

* Измерено при Iset = 20 мА, δset = 100%. + Нет в наличии.

Machine Translated by Google

4. Резюме

В этой статье адаптивный высокоэффективный драйвер светодиодной подсветки предназначен для решения проблемы чрезмерных потерь мощности, вызванных изменениями прямого напряжения светодиодов. Регулируя пиковый ток, достигается постоянство прямого напряжения светодиодной цепи внутри каждого канала. Это минимизирует напряжение на LCR в каждом канале, тем самым оптимизируя чрезмерные потери мощности в каждом канале и общую энергоэффективность. Кроме того, регулировка рабочего цикла обеспечивает постоянство яркости по каждому каналу. На основе КМОП-процесса 0,18 мкм 1P4M была спроектирована и изготовлена схема LCR, подходящая для адаптивного управления. В схеме используется структура токового зеркала и МОП-транзисторы, работающие в линейной области, что обеспечивает точность тока и энергоэффективность. Разработанная схема LCR поддерживает диапазон выходного тока 0–48 мА. Для проверки эффективности схемы адаптивного управления был построен экспериментальный двухканальный драйвер светодиодной подсветки с использованием спроектированных микросхем LCR. Каждый канал состоит из 6 светодиодов, а разница прямого напряжения между каналами намеренно увеличена примерно до 500 мВ. Результаты испытаний показывают, что представленная схема адаптивного управления устраняет эту прямую разницу напряжений и снижает чрезмерные потери мощности. Когда опорное напряжение установлено на 300 мВ, КПД мощности 98,1% достигается при токе 20 мА. Фактически экспериментальный двухканальный драйвер можно расширить до многоканальной реализации. Хотя в данном исследовании спроектирован и изготовлен только чип LCR , транзисторы MP и контроллер яркости могут быть интегрированы для реализации монолитного интегрированного чипа драйвера с использов

Вклад автора: концептуализация, WZ и QZ; методология XX, QZ, CJ и ZZ; программное обеспечение, XX; проверка, XX, QZ, CJ и ZZ; формальный анализ, XX; расследование, XX, QZ, CJ и ZZ; ресурсы, XX и QZ; курирование данных, XX и QZ; письменное – подготовка первоначального проекта, XX и QZ; написание — рецензирование и редактирование, WZ и XX; надзор, WZ и JT; администрация проекта, WZ и JT; приобретение финансирования, WZ. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Данное исследование финансировалось научно-исследовательским проектом Департамента образования провинции Хунань (23A0260) и инновационным проектом последипломной практики Университета науки и технологий Чанша (CLSJCX24079).

Заявление о доступности данных: данные, представленные в этом исследовании, доступны по запросу у соответствующего автора.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Цао, Дж. Я. Твердотельное освещение: лампы, чипы и материалы завтрашнего дня. IEEE Circuits Devices Mag. 2004, 20, 28–37. [Перекрестная ссылка]
Ядлапалли, RT; Нарасипурам, РП; Котапати А. Обзор энергоэффективных топологий драйверов твердотельных светодиодов. Межд. Дж. Энергия 2. Рез. 2020, 44, 612–630. [Перекрестная ссылка]
3. Морроу, RC светодиодное освещение в садоводстве. HortScience 2008, 43, 1947–1950. [Перекрестная ссылка]
4. Чо, Дж.; Парк, Дж. Х.; Ким, Дж. К.; Шуберт Е. Ф. Белые светодиоды: история, прогресс и будущее. Лазерная фотоника Ред. 2017, 11, 1600147. [CrossRef]
5. Сурканти, ПР; Мехротра, Д.; Furth, PM Многоканальный светодиодный драйвер с точным согласованием тока для портативных устройств. В материалах 62-го Международного симпозиума Среднего Запада по схемам и системам IEEE 2019 г. (MWSCAS), Даллас, Техас, США, 4–7 августа 2019 г.; стр. 941–944.
6. Пал, Н.; Фиш, А.; Макинтайр, В.; Гризерт, Н.; Зима, Г.; Эйххорн, Т.; Пилава-Подгурский, Р.; Ханумолу, ПК Драйвер светодиодной подсветки с гибридным повышающим SC-преобразователем на 30 В с КПД 91 % и КМОП 180 нм. В материалах конференции IEEE Custom Integrated Circuits Conference (CICC) 2020 г., Бостон, Массачусетс, США, 22–25 марта 2020 г.; стр. 1–4.
7. Прессман А. Проектирование импульсных источников питания. McGraw-Hill, Inc.: Нью-Йорк, штат Нью-Йорк, США, 1997.
8. Ван Ю.; Алонсо, Дж. М.; Руан, Х. Обзор светодиодных драйверов и связанных с ними технологий. IEEE Транс. Индийский Электрон. 2017, 64, 5754–5765. [Перекрестная ссылка]
9. Гупта, Н.; Верма, Д. Обзор и анализ топологий драйверов светодиодов (LED). В материалах 2-й Международной конференции IEEE по электроэнергетике и энергетическим системам (ICEPES) 2021 г., Бхопал, Индия, 10–11 декабря 2021 г.; стр. 1–6.
10. О, Т.-Дж.; Чо, А.; Ки, С.-Л.; Хван, И.-К. Маломощная и недорогая микросхема повышающего светодиодного драйвера с цифровым управлением для подсветки. В материалах Азиатской конференции по твердотельным схемам IEEE 2012 г. (A-SSCC), Кобе, Япония, 12–14 ноября 2012 г.; стр. 237–240.

Machine Translated by Google

11. Ли, Ю.-З.; Чоу, Х.-Дж.; Юн, Ж.-Ж. Улучшение производительности усилителя светодиодного драйвера с высоким коэффициентом усиления по напряжению для светодиодной подсветки с боковой подсветкой. IEEE Транс. Сист. цепей. II Express Briefs 2017, 65, 481–485. [Перекрестная ссылка]
12. Чен, Х.; Чжан, Ю.; Ма, Д. ИС драйвера параллельной строки SIMO для регулируемой светодиодной подсветки с оптимизацией напряжения локальной шины. и один контур регулирования с разделением времени. IEEE Транс. Силовой электрон. 2011, 27, 452–462. [Перекрестная ссылка]
13. Ким, Х.-К.; Юн, CS; Чон, Д.-К.; Ким, Дж. Одноиндукторный многоканальный светодиодный драйвер с балансировкой тока для дисплея. приложения с подсветкой. IEEE Транс. Индийское прил. 2014, 50, 4077–4081. [Перекрестная ссылка]
14. Чен, Ж.-Х.; Ван, Южная Каролина; Лю, Ю.-Х.; Ченг, Ю.-С.; Ян, З.-З. Система управления светодиодной подсветкой RGB с возможностью динамического регулирования напряжения. В материалах 10-й Международной конференции IEEE по силовой электронике и приводным системам (PEDS) 2013 г., Китакюсю, Япония, 22–25 апреля 2013 г.; стр. 349–353.
15. Лю, П.-Дж.; Хонг, З.-Ю.; Хунг, К.-К. Однокаскадный маломощный драйвер AC–DC RGB-светодиодов со схемой управления переключающим конденсатором. IEEE Транс. Индийский Электрон. 2020, 68, 3103–3112. [Перекрестная ссылка]
16. Ким, Ж.-С.; Ли, Ю.-К.; Ли, Дж.-С.; Шин, Ю.-К.; Тарк, Ж.-Х.; Рю, К.-Ц.; Ян, Б.-Д. Быстродействующий токоимпульсный драйвер для светодиодной подсветки. В материалах Международного симпозиума IEEE по схемам и системам 2009 г., Тайбэй, Тайвань, 24–27 мая 2009 г.; стр. 1775–1778.
17. Хонг, С.-И.; Хан, Дж.-В.; Ким, Д.-Х.; Квон, О.-К. Микросхема драйвера светодиодной подсветки с двойным контуром управления для ЖК-дисплеев среднего размера. В материалах Международной конференции по твердотельным схемам IEEE 2010 г. (ISSCC), Сан-Франциско, Калифорния, США, 7–11 февраля 2010 г.; стр. 116–117.
18. Хуан, Ю.-К.; Чен, Х.-В. Новый быстродействующий импульсный драйвер для светодиодной подсветки. В материалах Международной конференции IEEE по бытовой электронике-Тайвань (ICCE-TW) 2016 г., Наньтоу, Тайвань, 27–29 мая 2016 г.; стр. 1–2.
19. Ан, Х.-А.; Хонг, С.-К.; Квон, О.-К. Быстродействующий регулятор тока, использующий метод уменьшения времени нарастания для обеспечения высокого затемнения. соотношение драйверов светодиодной подсветки. IEEE Транс. Сист. цепей. II Express Briefs 2016, 63, 1014–1018. [Перекрестная ссылка]
20. Хан, Х.; Чжан, Ф.; Лю, М. Метод ШИМ-регулирования яркости для схемы разделения тока с конденсаторным ограничением в системе светодиодной подсветки. ИИЭР Дж. Экстрен. Сел. Вершина. Силовой электрон. 2018, 6, 1190–1197. [Перекрестная ссылка]
21. Хонг, С.-И.; Нам, К.-С.; Юнг, Ю.-Х.; Ан, Х.-А.; В, Х.-Ж.; Квон, О.-К. Высокоэнергетическая и надежная микросхема драйвера светодиодной подсветки для равномерного управления током жидкокристаллических дисплеев среднего размера. Ж. Инф. Дисп. 2012, 13, 73–82. [Перекрестная ссылка]
22. Лим, Б.-М.; Ко, Ю.-Х.; Джанг, Ю.-С.; Квон, О.-Х.; Хан, С.-К.; Ли, С.-Г. Понижающий светодиодный драйвер с напряжением 200 В и эффективностью 98,16 %, использующий встроенный регулятор тока для повышения точности тока для крупномасштабных однострунных светодиодных систем подсветки. IEEE Транс. Силовой электрон. 2015, 31, 6416–6427. [Перекрестная ссылка]
23. Партасарати, МК; Ганесан, Н. Новый одноиндукторный восьмиканальный драйвер светодиодов для маломощных дисплеев. приложения с подсветкой. Турок. Дж. Электр. англ. Вычислить. наук. 2018, 26, 21:30–2139. [Перекрестная ссылка]
24. Хе, Ж.-П.; Чжан, Б.; Гао, Х.; Ву, К.-Д.; Ли, З.-Дж.; Чжоу, Ж.-Ю. Высокоточный аналоговый метод затемнения светодиодов слабого напряжения. IEEE Доступ 2019, 7, 172362–172373. [Перекрестная ссылка]
25. Лин, К.-Л.; Чен, Южная Каролина; Дэн, М.-Ю.; Хо, Ю.-Х.; Лин, Калифорния; Цай, К.-Л.; Ляо, Западная Юг; Лю, К.-И.; Ву, С.-Э.; Пэн, Ж.-Т. Схема управления AM PWM для мини-светодиодной подсветки жидкокристаллических дисплеев. IEEE J. Электронные устройства Soc. 2021, 9, 365–372. [Перекрестная ссылка]
26. Шуберт, Светоизлучающие диоды EF (2018); Политехнический институт Ренсселера: Тори, Нью-Йорк, США, 2018.
27. Лю, Ч.-Х.; Се, К.-Ю.; Се, Ю.-К.; Тай, Т.-Дж.; Чен, К.-Х. Адаптивная система времени выключения с SAR-управлением без чувствительного резистора для достижения высокой эффективности и точности светодиодной системы освещения. IEEE Транс. Сист. цепей. Я Регул. Пап. 2009, 57, 1384–1394.
28. Лохаус, Л.; Россиус, А.; Дитрих, С.; Вундерлих, Р.; Хайнен, С. Светодиодный драйвер с регулируемой яркостью, использующий резистивное управление обратной связью ЦАП для адаптивного регулирования напряжения. В материалах Конгресса и выставки IEEE Energy Conversion 2013, Денвер, Колорадо, США, 15–19 сентября 2013 г.; стр. 3126–3133.
29. Се, К.-Ю.; Вэй, Ю.-С.; Чен, К.-Х.; Лин, Т.-Ц. Эффективный светодиодный драйвер с технологией адаптивного отслеживания ссылок. В материалах 16-й Международной конференции IEEE по электронике, схемам и системам 2009 г. (ICECS 2009), Ясмин Хаммамет, Тунис, 13–16 декабря 2009 г.; стр. 291–294.
30. Лю, П.-Дж.; Сюй, С.-Р.; Чанг, Ч.-В.; Ляо, К.-Ю.; Чиен, Л.-Х. Драйвер белого светодиода с регулируемой яркостью и адаптивным управлением с обратной связью по напряжению. В материалах 2-й Международной конференции IEEE по энергетической электронике будущего (IFEEC) 2015 г., Тайбэй, Тайвань, 1–4 ноября 2015 г.; стр. 1–4.
31. Юнг, Ю.Х.; Хонг, СК; Квон, ОК. Быстродействующий светодиодный драйвер с адаптивным управлением частотой в зависимости от изменения нагрузки. ЖК-подсветка большого размера. Дж. Сок. Инф. Дисп. 2017, 25, 712–724. [Перекрестная ссылка]
32. Сюй, Ю.-К.; Лин, Ж.-Ю.; Чен, СС-Р. Компактный и высокоэффективный драйвер RGB-светодиодов с адаптивным напряжением управления и энергосберегающей технологией. Энергии 2018, 11, 1422. [CrossRef]
33. Йонг, В.; Баосин, Д.; Йи, Л.; Тинг, Л.; Иньтан, Ю. Многоканальный высокоэффективный драйвер светодиодной подсветки с высоким коэффициентом затемнения. В материалах Международной конференции по электронной промышленности и автоматизации 2017 г. (EIA 2017), Сучжоу, Китай, 23–25 июня 2017 г.; стр. 110–116.
34. Се, Ю.-Т.; Лю, Б.-Д.; Ву, Ж.-Ф.; Фанг, К.-Л.; Цай, Х.-Х.; Хуанг, Ю.-З. Светодиодный драйвер с высоким коэффициентом затемнения для подсветки ЖК-дисплея. IEEE Транс. Силовой электрон. 2012, 27, 4562–4570. [Перекрестная ссылка]
35. Чиу, Х.-Дж.; Ло, Ю.-К.; Чен, Ж.-Т.; Ченг, С.-Дж.; Лин, К.-Ю.; Моу, С.-К. Высокоэффективный светодиодный драйвер с регулируемой яркостью для малой мощности. осветительные приложения. IEEE Транс. Индийский Электрон. 2009, 57, 735–743. [Перекрестная ссылка]
36. Ху, Ю.; Цзян, Л.; Вэнь, В.; Ян, В.; Дин, Р. Новый подход к адаптивному напряжению управления светодиодным драйвером. В материалах конференции и выставки прикладной силовой электроники IEEE 2019 (APEC), Анахайм, Калифорния, США, 17–21 марта 2019 г.; стр. 721–726.

Machine Translated by Google

37. By, K.-X.; Чен, K.-Л. Высокоэффективный зарядный насос с регулировкой тока для драйвера белого светодиода. IEEE Транс. Сист. цепей. II Экспресс Briefs 2009, 56, 763–767.
38. My, X.; Гэн, Л.; Лю, Дж. Высокоточный источник постоянного тока, применяемый в драйвере светодиодов. В материалах симпозиума 2011 г. по фотонике и оптоэлектронике (SOPO), Ухань, Китай, 16–18 мая 2011 г.; стр. 1–4.
39. Сансен, Основы аналогового проектирования WM; Springer Science & Business Media: Берлин, Германия, 2007 г.; Том 859.
40. Хогерворст, Р.; Теро, JP; Эшаузье, Р.Г.; Хьюйсинг, Дж.Х. Компактный энергоэффективный КМОП-порт с напряжением питания 3 В и полным входом/выходом. операционный усилитель для библиотек ячеек СБИС. IEEE J. Твердотельные схемы 1994, 29, 1505–1513. [Перекрестная ссылка]
41. Сурканти, ПР; Мехротра, Д.; Гангинени, М.; Многострунный светодиодный драйвер Furth, PM с точным согласованием тока и динамическим подавлением прямого несоответствия напряжения. В материалах 60-го Международного симпозиума Среднего Запада по схемам и системам IEEE 2017 г. (MWSCAS), Бостон, Массачусетс, США, 6–9 августа 2017 г.; стр. 229–232.
42. Шахрури, Франция; Ахмад, Х.Х.; Абуишмайс И. Аспекты проектирования многострунного драйвера WLED с одним выходом с использованием 40-нм технологии КМОП. J. Электрон малой мощности. Прил. 2022, 12, 5. [CrossRef]
43. Абуишмаис И.; Шахрури, Франция; Ахмад, Х. Методология проектирования высокоэффективного приемника тока с регулируемой яркостью для драйверов с регулируемым током. Электроника 2022, 11, 2566. [CrossRef]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Обнаружение объектов и монокулярная стабильная оценка расстояния для Дорожная среда: объединенная архитектура с использованием YOLO-RedeCa и фильтра изменения ненормальных прыжков

Хэджун Льв., Ю Ду*, Ян Ма и Ин Юань

Пекинская ключевая лаборатория разработки информационных услуг, Колледж робототехники, Пекинский союзный университет, Пекин 100101, Китай; lyuhejun@163.com (HL); jqrmayan@buu.edu.cn (ЮМ); yueeing@163.com (ГГ)
* Переписка: duyu@buu.edu.cn

Аннотация: Обеспечение быстрого и точного комплексного восприятия окружающей среды для транспортных средств представляет собой серьезную задачу. Обнаружение объектов и монокулярная оценка расстояния — две основные технологии, хотя они часто используются отдельно. Таким образом, необходимо укреплять и оптимизировать взаимодействие между ними. Движение транспортного средства или затенение объектов могут вызвать внезапные изменения положения или размеров блоков обнаружения во временных данных, что приводит к колебаниям оценок расстояния Итак, мы предлагаем способ интеграции детектора на базе YOLOv5-RedeCa, трекера Bot-Sort и фильтра скачкообразной смены аномалий. Такое сочетание позволяет более точно обнаруживать и отслеживать объекты. Фильтр скачков аномалий сглаживает изменения расстояния, вызванные внезапными изменениями размеров поля обнаружения. Наш метод повышает точность при одновременном снижении вычислительных затрат, демонстрируя выдающуюся производительность на нескольких наборах данных. Примечательно, что в наборе данных KITTI стандартное отклонение результатов непрерывного измерения дальности остается неизменно низким, особенно в сценариях с перекрытием или исчезновением нескольких объектов. Эти результаты подтверждают эффективность и точность нашего метода при решении двойных задач.

Ключевые слова: автоматическая техника вождения; обнаружение объектов; монокулярное измерение расстояния; Фильтр Калмана

check for updates

Цитирование: Льв, Х.; Ду, Ю.; Может.; Юань, Y. Обнаружение объектов и монокуляр Стабильная оценка расстояния для дороги Среды: архитектура слияния Использование YOLO-RedeCa и фильтра изменения ненормальных прыжков. Электроника 2024, 13, 3058. https://doi.org/10.3390/электроника13153058

Академический редактор: Фелипе Хименес

Поступила: 24 июня 2024 г.
Пересмотрено: 19 июля 2024 г.
Принято: 26 июля 2024 г.
Опубликовано: 2 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

Эффективное восприятие и понимание дорожной среды имеют решающее значение для технологии автономного вождения [1]. Однако достижение комплексного восприятия окружающей среды для транспортных средств вызывает сложную проблему. Эта комплексная система восприятия выполняет множество задач. Он обнаруживает различные объекты вокруг автомобиля.
Кроме того, он получает расстояния между этими объектами и транспортным средством. Поэтому мы ориентируемся на методы, решающие двойные задачи. Наши методы одновременно обнаруживают объекты вокруг автомобиля и определяют расстояния до этих объектов. Более подробную информацию можно найти на рисунке 1.

При автономном вождении точное обнаружение объектов имеет решающее значение для динамической оценки расстояния, что напрямую влияет на безопасность и эффективность системы [2]. Динамическая оценка расстояния, как часть восприятия более высокого уровня, может имитировать интуитивное восприятие водителей-людей.

Существующие методы измерения расстояний, основанные на радаре и стереозрении, хотя и эффективны, но сталкиваются с такими проблемами, как высокая стоимость, сложность системы и низкая производительность в реальном времени . Кроме того, стоит изучить, требует ли современная технология восприятия с высоты птичьего полета (BEV) более глубокой поддержки зрительных характеристик человека [3]. Существующим монокулярным методам измерения расстояния не хватает устойчивости при резких изменениях значений, что напрямую влияет на восприятие и качество принятия решений системами автономного вождения в динамичных средах.

Machine Translated by Google

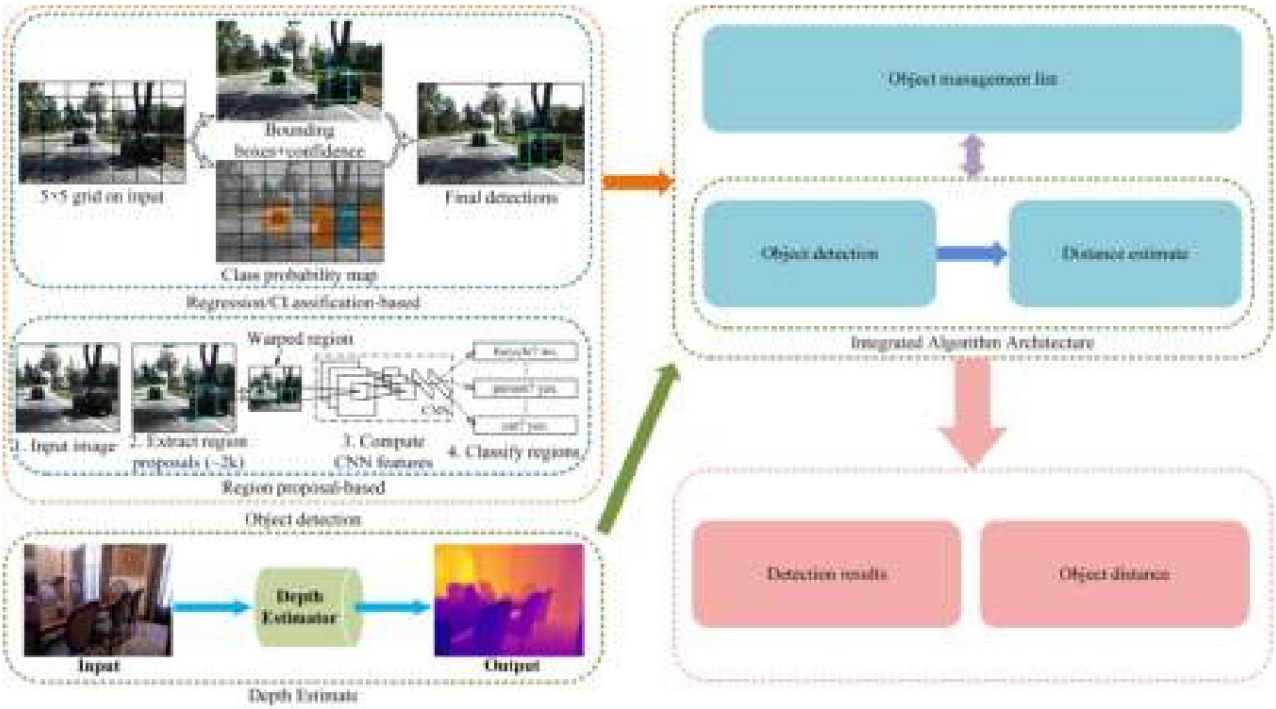


Рисунок 1. Слева показана основная схема обнаружения объектов с монокулярной оценкой глубины. Справа — архитектура алгоритма, который мы собираемся реализовать, и выходные данные.

Целью данного исследования является предложение монокулярного метода измерения расстояния. Этот Метод направлен как на обнаружение объектов, так и на определение расстояния между объектами и автомобиль. Объединив методы YOLOv5-RedeCa и фильтра скачков аномалий, подход направлен на повышение точности и надежности динамической оценки расстояния, тем самым решая проблемы восприятия и принятия решений в динамичных средах. Наш метод демонстрирует значительный потенциал в извлечении признаков и оценке расстояний . Кроме того, это предоставит более надежную и богатую информацию о восприятии для системы автономного вождения, что значительно повышает безопасность и эффективность вождения . Это делает наш метод многообещающим кандидатом на прорывные достижения в области технология автономного вождения.

Обзор предыдущей работы в области обнаружения 2D-объектов на основе предложений по регионам. методы (например, R-CNN) [4] значительно улучшили точность обнаружения благодаря двухэтапному подходу обнаружения, который работает исключительно хорошо на сложном фоне. Однако, они относительно медленны в вычислениях. Напротив, хотя одноступенчатые детекторы (например, YOLO) [5] немного идут на компромисс с точностью: они напрямую предсказывают положение и категории объектов на изображении посредством регрессии/классификации, упрощая модель. структуру и процесс вычислений, что делает их пригодными для обработки в реальном времени [6]. Поэтому мы используем их в качестве базовой линии для нашего детектора.

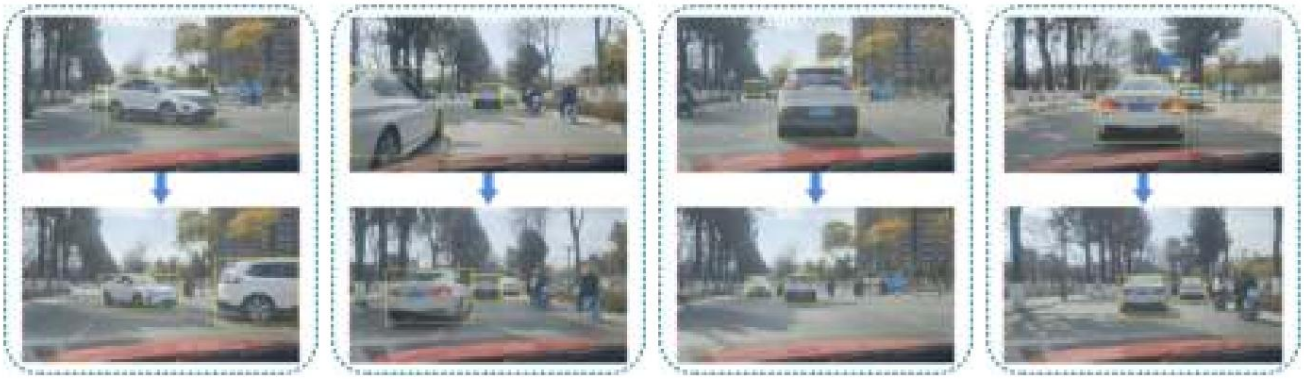
Обнаружение объектов само по себе не может полностью воспринимать окружающую среду без информации о расстоянии [7]. Традиционные методы оценки глубины используют монокулярные сигналы или разрабатываются вручную. функции [8,9], но они часто неэффективны в сценариях различной глубины.

Глубокое обучение, особенно с использованием сверточных нейронных сетей (CNN), предлагает прямые отображение пикселей в глубину [10]. Многомасштабные сети и масштабно-инвариантные функции потерь имеют повышенная точность оценки глубины. В отличие от традиционных методов, глубокое обучение автоматически генерирует карты различий и информацию о глубине без обширных аннотаций, обычно с использованием видеопоследовательностей или стереоизображений. Однако эти методы требуют значительного вычислительные ресурсы.

При автономном вождении непрерывная информация об обнаружении объектов во временной области. используется для отслеживания объектов и оценки расстояния. Однако вибрации, вызванные транспортным средством движение и окклюзии могут нарушить стабильность объектов в видеоряде. Это может

Machine Translated by Google

привести к внезапным изменениям положения или размера блоков обнаружения, как показано на рисунке 2. Поэтому необходимо сгладить поля обнаружения.



(а) (б) (с) (г)

Рисунок 2. Проиллюстрированы ограничивающие рамки для неидеальных и идеальных снимков объектов дорожной сцены. (a – d) представляют собой последовательные кадры из четырех различных видеопоследовательностей. В (a,b) размеры ограничивающей рамки резко изменяются, когда целевой объект подвергается внезапному появлению, исчезновению или окклюзии. И наоборот, (c,d) изображают сценарии, в которых цель четко видна, а ограничивающая рамка точно инкапсулирует целевой объект в нормальных условиях.

Хотя алгоритмы обнаружения 2D-объектов достигли значительного прогресса, они остаются специализированными. Аналогичным образом, значительно продвинулись алгоритмы монокулярной оценки глубины Однако исследования в этих двух областях остаются относительно отдельными. Интеграция между обнаружением объектов и оценкой глубины по-прежнему отсутствует. В существующих исследованиях при обнаружении объектов часто отсутствует интегрированная оценка глубины, а оценка глубины не включает в себя возможности распознавания объектов. Эта ситуация подчеркивает острую необходимость. Крайне важно разработать интегрированную систему, которая сочетает в себе обнаружение объектов и оценку глубины. Такая система оптимизирует взаимодействие между двумя методами. Следовательно, это повысит их эффективность в практическом применении.

Поэтому мы предлагаем и разрабатываем целостную архитектуру алгоритма, объединяющую объекты обнаружение и оценка расстояния. Наш основной вклад заключается в следующем:

1. Мы разработали модуль RedeCa, чтобы расширить возможности извлечения признаков и обеспечить более точное обнаружение объектов в сценариях дорожного движения. Чтобы обеспечить производительность в реальном времени, мы используем глубокую разделимую сверточную сетевую архитектуру.
2. Мы интегрируем механизм многокадровой ассоциации, чтобы повысить надежность алгоритма в случаях перекрытия или временной потери объекта, позволяя одновременно отслеживать и ранжировать несколько разных объектов.
3. Мы разрабатываем фильтр скачкообразных изменений аномалий для повышения точности и надежности оценки расстояния. Этот фильтр устраняет аномалии, вызванные внезапными изменениями размера поля обнаружения в кадре. Он сглаживает изменение расстояния до объекта. Следовательно, надежность определения дальности повышается.

Предложенный нами алгоритм получил золотую медаль на 3-й конференции по интеллектуальным подключенным транспортным средствам в Китае (Шэньян). Он также завоевал бронзовую медаль в группе контроля восприятия и принятия решений на конкурсе виртуального моделирования World Intelligent Driving Challenge Metaverse 2023 года .

2. Сопутствующая работа

В области автономного вождения проблема заключается в том, как заставить транспортное средство полностью воспринимать окружающую среду. Это комплексное восприятие включает в себя не только обнаружение различных окружающих объектов, но и определение расстояний между этими объектами и автомобилем. В настоящее время использование камер видимого света является распространенным и экономически эффективным подходом к решению этой проблемы. Камеры видимого света, оснащенные светочувствительными элементами, могут захватывать цветные изображения реального мира. Исследователи

Machine Translated by Google

провести углубленный анализ этих цветных изображений для выполнения ключевых задач восприятия, таких как обнаружение объектов и оценка расстояния.

Обнаружение транспортных объектов. Обнаружение объектов является одной из основных проблем в области компьютерного зрения [11]. Его цель – классифицировать и определять положение объектов на изображениях. Глубокие сверточные нейронные сети (DCNN) обеспечили мощную поддержку развитию методов обнаружения объектов благодаря своим выдающимся возможностям представления признаков [12,13]. Обычно основные архитектуры обнаружения объектов можно разделить на две категории: методы обнаружения объектов на основе предложения региона и методы обнаружения объектов на основе регрессии/классификации [6]. Каждый из них имеет свои особенности и решает различные технические задачи.

Типичные модели архитектур обнаружения объектов на основе предложений регионов включают RCNN [4] и ее последующие улучшения, такие как SPPNet [14], Fast RCNN [15], Faster R-CNN [16], Mask RCNN [17] и R-FCN. [18]. На первом этапе этих архитектур генерируются предложения регионов-кандидатов. На втором этапе выполняется детальная классификация и уточнение ограничивающей рамки этих регионов. Такая поэтапная обработка значительно повышает точность обнаружения, особенно демонстрируя превосходные характеристики на сложном фоне. Например, чтобы решить проблему медленной скорости генерации предложений, Faster R-CNN вводит сеть региональных предложений (RPN). Важно повысить качество и скорость рассмотрения предложений.
Mask RCNN дополнительно прогнозирует маски объектов, добавляя параллельную ветвь, позволяя модели одновременно выполнять как обнаружение объектов, так и сегментацию экземпляров. Это особенно ценно в приложениях, требующих детальной визуальной информации.

Чтобы упростить структуру модели и вычислительные процессы, архитектуры обнаружения объектов на основе регрессии/классификации предназначены для прогнозирования положения и категорий объектов непосредственно на изображениях, такие как OverFeat [19], серия YOLO [20–25], SSD [26], DSSD [27] и ФССР [28]. Основным преимуществом этих архитектур является высокая скорость обнаружения, что делает их пригодными для сценариев обработки в реальном времени. Среди них выделяется серия YOLO . Переосмыслив проблему обнаружения как единую задачу регрессии, можно одновременно предсказать категории и положения нескольких объектов за один прямой проход сети. Чтобы эффективно решать проблемы обнаружения объектов различных размеров и улучшать способность модели обнаруживать небольшие объекты, SSD представляет концепцию многомасштабных карт о

Монокулярная оценка глубины. Поиск определения расстояния между объектами и камерой остается ключевым направлением исследований в области компьютерного зрения с момента его создания. В первоначальных подходах оценка глубины монокуляра в основном опиралась на форму по затенению (SFS) и глубину по фокусу/дефокусировке (DFF/DFD). SFS определяет форму объекта на основе изменений освещенности на его поверхности [29]. Методы DFF/DFD оценивают глубину, используя сфокусированную и расфокусированную информацию на изображениях, а также параметры калибровки камеры [30]. Фундаментальное предположение состоит в том, что сфокусированные части изображения становятся более четкими, а расфокусированные части постепенно становятся размытыми. Хотя эти методы эффективны в конкретных сценариях, они часто основаны на дополнительных предположениях о сцене и демонстрируют низкую производительность при работе со сценами со значительной глубиной резкости.

С развитием технологий машинного обучения исследователи начали формулировать проблему оценки глубины одного пикселя как задачу вероятностного обучения, основанную на случайных полях. Например, оптимизация карт глубины с использованием марковского случайного поля (MRF) [8,31] и моделирование апостериорных вероятностей с помощью условного случайного поля (CRF) для оценки глубины [9]. Повышенная вычислительная мощность графических процессоров (GPU) и доступность крупномасштабных наборов данных изображений стимулируют глубокое обучение, открывая путь для новых направлений исследований в области оценки глубины.

Методы глубокого обучения обычно требуют реальных данных о глубине в качестве обучающих меток. Эти методы напрямую сопоставляют пиксели со значениями глубины путем построения сверточных нейронных сетей (CNN). Ссылки. [32,33] были первыми, кто применил CNN в области оценки глубины. Они использовали многомасштабную сетевую стратегию. Глобальная сеть грубой оценки предоставляет первоначальные оценки расстояния. Кроме того, локальная сеть уточнений оптимизирует детали. Они ввели масштабно-инвариантную функцию потерь для обработки ошибок. Последующие исследования

Machine Translated by Google

еще больше повысили точность оценки за счет уточнения сетевых структур [29,34–36] и интеграции вспомогательной информации [37,38]. Некоторые исследования не опираются на большое количество размеченных данных о глубине. Вместо этого они используют видеопоследовательности или стереоизображения для автоматического получения карт различий и информации о глубине с помощью алгоритмов глубокого обучения. Например, исх. [39] обрабатывают изображение слева с помощью сверточной нейронной сети. Это создает карту неравенства. Затем они реконструируют левый вид, комбинируя его с правым путем обратного отображения. Ошибка реконструкции используется в качестве управляющего сигнала для обучения. Ссылка. [40] одновременно обрабатывает левое и правое изображения для оптимизации создания карты различий, обеспечивая точную оценку глубины, даже если во время тестирования используются только монокулярные изображения. Кроме того, ссылка. [41] извлекает позу и глубину из видеопоследовательностей последовательных кадров, демонстрируя эффективность объединения нескольких сетевых моделей для реконструкции изображения. Хотя эти методы обеспечивают хорошую производительность, они требуют значительных вычислительных затрат.

Значительный прогресс был достигнут в области обнаружения 2D-объектов и монокулярной оценки глубины. Но в исследованиях эти два направления по-прежнему в значительной степени разделены, и эффективной интеграции не наблюдается. В настоящее время в системах обнаружения объектов часто отсутствует возможность оценки дальности. А система оценки глубины не включает в себя способность распознавания объектов. Такое разделение предполагает необходимость развития. Необходима целостная алгоритмическая архитектура. Такая архитектура должна интегрировать обнаружение объектов и оценку дальности. Такая интеграция усилит синергию между этими двумя методами. Следовательно, это оптимизирует их производительность в практических приложениях.

3. Метод

Чтобы обеспечить интеграцию обнаружения объектов и оценки расстояния, мы проектируем и разрабатываем алгоритмическую архитектуру. Он состоит из трёх основных компонентов: детектора на базе YOLOv5-RedeCa, многообъектного трекера и фильтра скачкообразной смены аномалий. В сценариях автономного вождения транспортные средства часто сталкиваются со сложной и динамичной средой с множеством автономно движущихся объектов, которые могут внезапно появиться, перекрыться или перекрыться [21]. Эти ситуации предъявляют жесткие требования к точности и устойчивости алгоритма. Кроме того, из-за высокой скорости движения автомобиля алгоритм должен обрабатывать данные датчиков и реагировать в течение миллисекунд, чтобы обеспечить безопасность автомобиля. Следовательно, системы автономного вождения требуют алгоритма восприятия с высокой точностью и надежностью.

Для этих целей мы разработали модуль RedeCa, который расширяет архитектуру алгоритма обнаружения объектов, повышая точность локализации объектов в дорожных сценариях. Кроме того, мы интегрируем трекер Bot-Sort. Этот трекер использует механизм многокадровой ассоциации. Это повышает надежность алгоритма при возникновении окклюзии или временной потери объекта. Таким образом, метод позволяет одновременно отслеживать несколько объектов и оценивать их расстояния. Подход также периодически оценивает непрерывность отслеживания, отслеживает существенные изменения в состоянии объекта и проверяет стабильность процесса отслеживания. Чтобы повысить точность и надежность оценки расстояния, мы также интегрируем фильтр скачкообразного изменения аномалий. Фильтр эффективно сглаживает изменение расстояния до объекта и устраняет потенциальные аномалии, вызванные внезапными изменениями размера поля обнаружения. Таким образом, это может повысить надежность измерения расстояния. Общая архитектура показана на рисунке 3.

Благодаря реализации YOLOv5-RedeCa наш метод повышает точность распознавания объектов и точность ограничивающей рамки для видеокадров. Кроме того, мы вычисляем расстояния между объектами и камерой, используя сложную модель физического изображения.

Для эффективного отслеживания и измерения нескольких объектов мы интегрируем трекер Bot-Sort. Этот механизм многокадровой ассоциации значительно повышает надежность в сценариях с перекрытиями или временной потерей объектов. Даже если объект временно исчезает из поля зрения в определенных кадрах, он продолжает последовательно отслеживать и измерять расстояние. Мы присваиваем уникальные идентификаторы вновь обнаруженным объектам, используя возрастающую последовательность. Затем он использует сортировку бота, чтобы сопоставить их с существующими объектами, различая те, которые имеют

Machine Translated by Google

вновь появившиеся и те, которые уже отслеживаются. Для объектов, которые успешно сопоставлены, информация о их статусе обновляется, а исходные идентификаторы сохраняются. И наоборот, новым объектам присваиваются новые записи отслеживания и идентификаторы. Если объект не обнаружен в текущем кадре, его запись временно сохраняется. Если объект не появится снова в течение заданного периода, он будет удален из списка отслеживания.

Сейчас большинство монокулярных методов определения дальности в первую очередь фокусируются на точности отдельных изображений, игнорируя корреляцию изменений расстояний между последовательными кадрами во временной области. Поэтому мы предлагаем стратегию многокадрового анализа данных, которая объединяет фильтрацию Калмана. Это не только повышает точность и надежность оценки расстояния, но также помогает в детальном анализе моделей движения объекта и тенденций изменения расстояния. Фильтрация Калмана сглаживает изменения расстояния, анализируя изменения одного и того же объекта в последовательных кадрах. Это также устраняет потенциальные ошибки, вызванные резкими изменениями размера блоков обнаружения одиночных кадров.

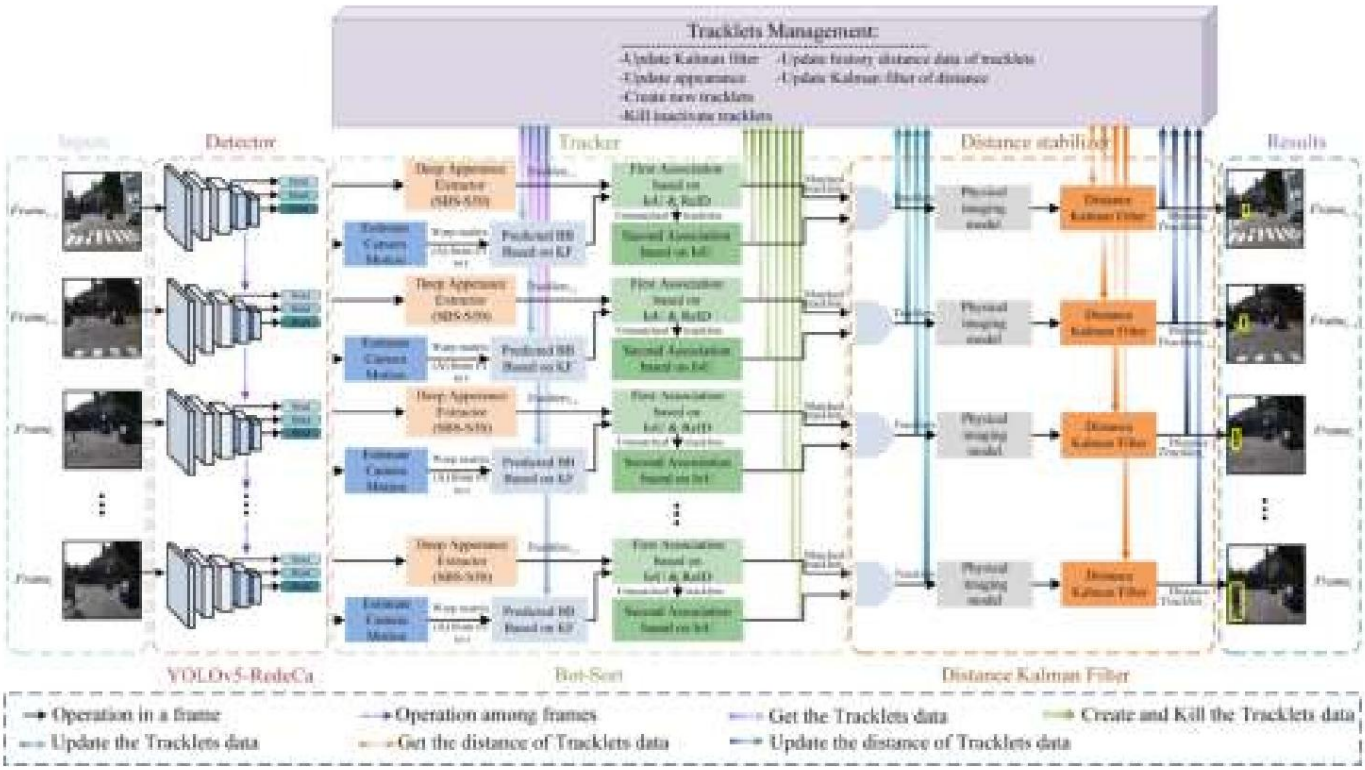


Рисунок 3. График общей архитектуры.

3.1. Обнаружение

объектов Чтобы обеспечить точное распознавание объектов и точное выделение ограничивающей рамки в каждом видеокадре, мы использовали алгоритм обнаружения объектов YOLOv5. Он служит базовой структурой для нашего модуля обнаружения, основанного на глубоких сверточных нейронных сетях . YOLOv5 прогнозирует класс и положение объектов, то есть ограничивающих рамок, непосредственно из входного изображения за один прямой проход. По сравнению с двухэтапными методами обнаружения , которые сначала извлекают области-кандидаты, а затем классифицируют их, YOLOv5 значительно повышает скорость и эффективность обнаружения. Он использует глубокую сверточную сеть для извлечения признаков из изображений, постепенно вытягивая все более сложные характер Это улучшает его способность распознавать различные объекты. За счет интеграции сети пирамиды функций (FPN) YOLOv5 объединяет карты объектов с разных уровней, тем самым улучшая свои возможности обнаружения объектов различных размеров.

Учитывая разнообразие размеров объектов в дорожных сценах, точность обнаружения объектов и особые требования дорожного движения, мы улучшили сетевую архитектуру YOLOv5. Сначала мы интегрировали модуль RedeCa в остаточные блоки. Это позволяет более точно обнаруживать более мелкие объекты в различных условиях освещения и

Machine Translated by Google

изменения погоды. Затем мы заменили традиционные свертки в остаточных блоках на отделимые по глубине свертки. Это изменение уменьшает количество параметров в сети и повышает производительность в реальном времени. Архитектура сети YOLOv5-RedeCa показана на рисунке 4.

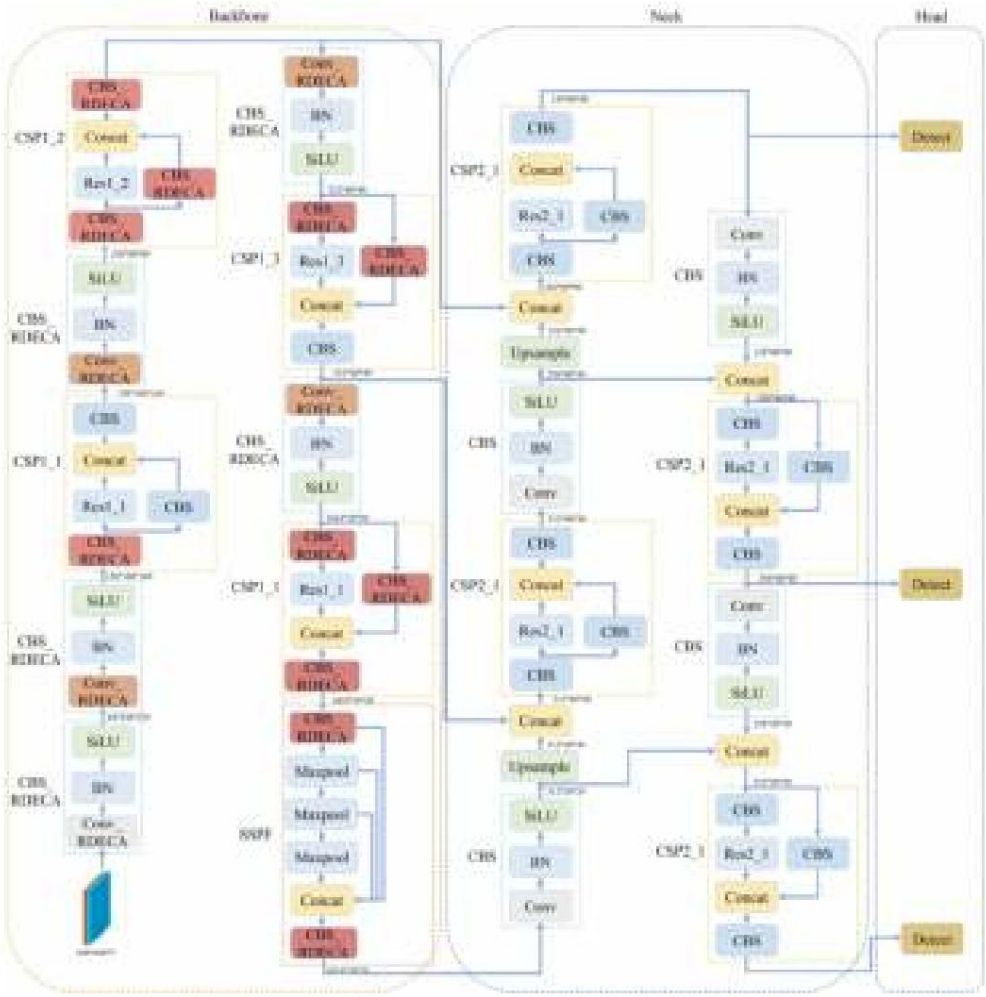


Рисунок 4. Схема сетевой архитектуры YOLOv5-RedeCa.

Модуль РедеКа

YOLOv5 продемонстрировал выдающуюся производительность в широком диапазоне сценариев обнаружения. Однако сложная дорожная среда требует универсального подхода для размещения различных типов и размеров транспортных объектов. Чтобы решить эту проблему, мы оптимизировали структуру сети YOLOv5, в частности, разработав модуль RedeCa для повышения точности распознавания транспортных объектов различного размера. Как показано на рисунке 5, модуль RedeCa основан на остаточном блоке. Он включает в себя последовательность остаточных разделяемых по глубине блоков свертки и модулей ECA для расширения возможностей извлечения признаков.

Глубоко разделяемые свертки заметно уменьшают количество параметров и вычислительную нагрузку в моделях. Это достигается путем разбиения стандартных сверток на две более простые операции: глубинные свертки и точечные свертки. Процесс подробно описан на рисунке 6. Он включает в себя сначала применение глубинной свертки независимо к каждому входному каналу. Затем поточечная свертка объединяет эти результаты с использованием ядер свертки 1 × 1. Такой подход значительно сокращает количество параметров и вычислительные требования, снижая использование памяти и энергопотребление. Следовательно, это делает модель более подходящей для развертывания на устройствах с ограниченными ресурсами.

Machine Translated by Google

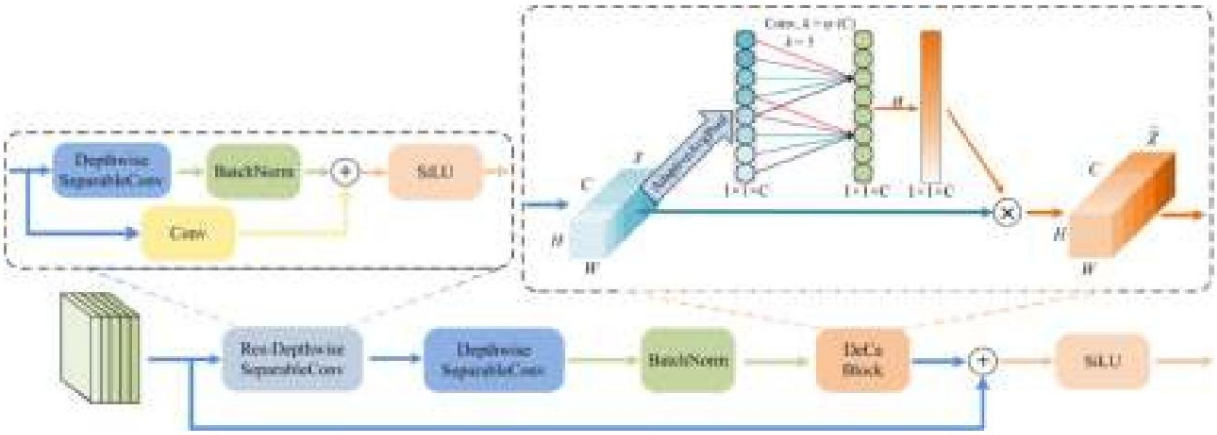


Рисунок 5. График структуры RedeCa.

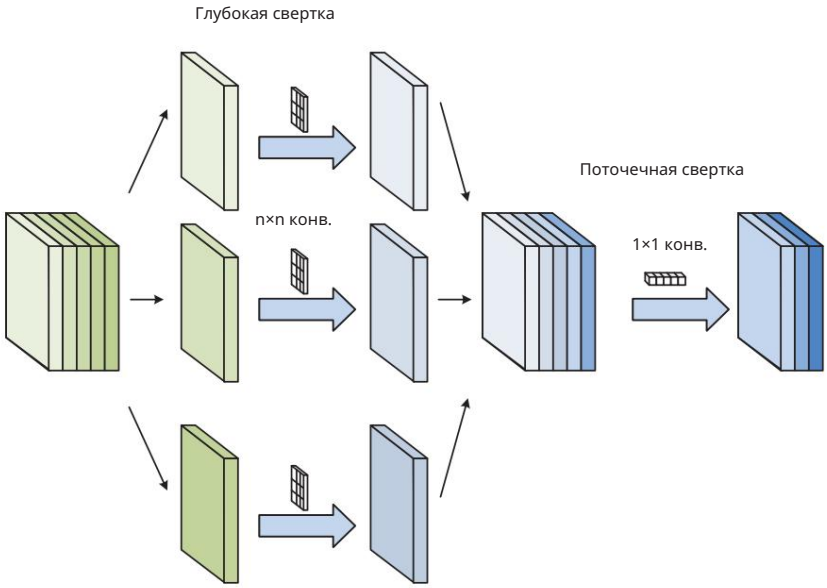


Рисунок 6. Граф глубоко отделимой сверточной структуры.

Глубоко отделимые свертки минимизируют риск переобучения, что особенно важно при обучении моделей на ограниченных данных.

За счет уменьшения количества параметров обучение становится проще и вероятность переобучения снижается. Это расширяет возможности обобщения модели в реальном мире.

Приложения. Вычислительная сложность, связанная с глубинно разделимыми свертками, подробно описана ниже.

Глубокая свертка:

$$ParamDC = K \times K \times Cin \quad (1)$$

$$FLOPSDC = Hout \times Wout \times K \times K \times Cin \quad (2)$$

Поточечная свертка:

$$ParamPC = Cin \times Cout \quad (3)$$

$$FLOPSPC = Hout \times Wout \times Cin \times Cout \quad (4)$$

Machine Translated by Google

Глубоко отделимые свертки:

$$\text{ПарамДВ} = \text{ПарамДК} + \text{ПарамПК} \quad (5)$$

$$\text{ФЛОПСДВ} = \text{ФЛОПСДК} + \text{ФЛОПСПК} \quad (6)$$

где ParamDC количественно определяет количество параметров, связанных с глубинной сверткой, а FLOPSDC определяет вычислительные требования, необходимые для глубинной свертки . Соответственно, ParamPC определяет количество параметров для поточечной свертки, тогда как FLOPSPC характеризует ее вычислительные требования. Кроме того, ParamDW указывает на количество параметров в глубоко отделимой свертке, а FLOPSDW поясняет вычислительные требования для этого типа свертки.

Модуль RedeCa запускается с помощью разделяемого по глубине блока свертки, который объединяет принципы остаточной сети. Он использует сопоставление идентификаторов для повышения эффективности обучения глубоких сетевых уровней. Он сочетает в себе высокую эффективность параметров глубинно- разделимых сверток с возможностью быстрого обучения остаточных связей, тем самым расширяя возможности модуля по извлечению признаков. Затем интегрируется стандартный блок свертки с разделением по глубине, чтобы эффективно снизить сложность модели и вычислительные требования. Наконец, модуль deca корректирует зависимости между каналами , динамически оптимизируя отклики функций для повышения точности обнаружения объекто

Математическая формулировка модуля RedeCa следующая:

$$\Phi^{\text{Рез}}_{\text{ДВКонв}} = \text{SiLU}(\text{BN}(\text{DWConv}(F)) + \text{Conv}(F)) \quad (7)$$

$$\text{FDWConv} = \text{BN}(\text{Поточечно}(\text{По глубине}(F^{\text{Рез}}_{\text{DWConv}}))) \quad (8)$$

$$\Phi_{\text{РедеКа}} = \text{SiLU}(\text{FDWConv} \times \sigma(\text{Conv}(\text{AdaptAvgPool}(\text{FDWConv}))) + F) \quad (9)$$

где F представляет входные данные карты объектов из предыдущего слоя.

3.2. Повышение устойчивости многокадровой ассоциации при оценке расстояния

При обнаружении объектов и оценке расстояния в сценах дорожного движения недостаточно работать только с одним объектом. В практических приложениях необходимо точно оценить расстояние до нескольких объектов одновременно. И эти объекты могут внезапно появляться, перекрываться, перекрываться или исчезать. Предлагаемый нами метод включает в себя трекер Bot-Sort, присваивающий уникальный идентификатор каждому объекту в каждом кадре, что способствует эффективной идентификации и отслеживанию объектов. Хотя существующие алгоритмы монокулярного определения дальности повышают точность отдельных изображений, они часто упускают из виду согласованность изменений расстояний до одного и того же объекта в последовательных кадрах. Игнорирование этой согласованности может привести к колебаниям в измерениях расстояний, особенно когда размер ограничивающей рамки обнаружения объекта резко меняется за короткий период. Наш метод повышает точность и надежность оценки расстояния за счет интеграции методов фильтрации Калмана с данными временных рядов. Такой подход сглаживает различия в расстояниях до объектов и устраняет аномалии, вызванные внезапными изменениями размеров окна обнаружения одного кадра, что приводит к превосходной производительности в сценариях дорожного движения в реальном времени.

3.2.1. Бот-Сортировка

В области визуального анализа дорожного движения простого обнаружения и оценки расстояния до отдельных объектов недостаточно. Первостепенное значение в этой области имеет возможность точно измерять расстояния между несколькими объектами в пределах сцены и назначать каждому отдельные идентификаторы для эффективного отслеживания и дифференциации. Наш метод включает в себя трекер Bot- Sort — усовершенствованный алгоритм, который умело идентифицирует и анализирует каждый объект в кадрах видео, определяя новые объекты для отслеживания. Методология Алгоритма 1 изложена в виде псевдокода [43].

Machine Translated by Google

Алгоритм 1. Рабочий процесс системы слежения за объектами на основе кадров

1: Вход: видеокадры
2: Выход: отслеживаемые траектории объекта З:
для каждого кадра в видео
4: Обнаруживайте потенциальные объекты с помощью YOLOX
5: Оцените матрицу преобразования движения камеры из предыдущего кадра
6: Извлечение характеристик обнаруженных объектов с помощью модели SBS-S50.
7: Прогнозирование местоположения объекта в текущем кадре с помощью фильтра Калмана
8: Первая ассоциация, основанная на объединенных показателях IoU и ReID.
9. Сопоставьте текущие обнаружения с существующими треклетами
10: Вторая ассоциация для непревзойденных треклетов на основе IoU
11: Обновление информации о состоянии треклетов.
12: Создавайте новые треклеты и назначайте новые уникальные идентификаторы.
13: Удалить неактивные треклеты, если они не обнаружены в течение определенного периода.
14: (Необязательно) Интерполируйте недостающие части в треклетах.
15: конец 16:
возврат Завершенные траектории для всех объектов

BoT-SORT представляет собой превосходную систему отслеживания нескольких объектов, которая значительно повышает точность отслеживания за счет синергетической интеграции данных о движении и внешнем виде. Система использует усовершенствованные фильтры Калмана наряду с методами компенсации движения камеры, чтобы умело обрабатывать динамику реального мира. Он явно объединяет пересечение через объединение (IoU) с мерами косинусного расстояния из алгоритмов повторной идентификации (ReID) для установления последовательных и надежных ассоциаций объектов, особенно в плотно динамических средах.

3.2.2. Идентификация и обслуживание объектов

Наш метод включает следующую логику для идентификации и обслуживания объектов. Во-первых, для вновь обнаруженных объектов метод присваивает каждому объекту уникальный идентификатор посредством возрастающей последовательности, чтобы различать различные объекты отслеживания. Впоследствии метод применяет алгоритм Bot-Sort. Он сравнивает вновь обнаруженные объекты с уже отслеживаемыми, чтобы определить их статус.

Для успешно сопоставленных объектов система обновляет информацию о статусе в списке отслеживания, сохраняя при этом исходные уникальные идентификаторы неизменными. Вновь появляющиеся объекты создаются как новые записи отслеживания и им присваиваются новые уникальные идентификаторы. Если отслеживаемый объект не обнаружен в текущем кадре, возможно, из-за окклюзии или выхода из поля зрения, метод временно сохраняет свою запись и идентификатор. Если объект не появится снова в течение установленного времени, он будет удален из списка отслеживания.

3.2.3. Аномалия Прыжок Изменить фильтр

При моделировании движения объекта на плоскости изображения мы обычно используем дискретный фильтр Калмана, оснащенный моделью постоянной скорости. Вектор состояния этого фильтра определяется как семерка $x = [xc, yc, s, a, \dot{x}c, \dot{y}c, \dot{s}]^T$, где (xc, yc) задает двумерные координаты центра объекта в пределах плоскость изображения. Здесь s означает масштаб (площадь) ограничивающей рамки, а a указывает на соотношение сторон. Чтобы повысить точность и устойчивость оценки расстояния, стабилизировать изменения расстояния до объекта и устранить аномалии, потенциально вызванные резкими изменениями размеров поля обнаружения в кадре, вектор состояния расширяется до восьмеричного числа, $x = [xc, yc, a, h, \dot{x}c, \dot{y}c, \dot{a}, \dot{h}]^T$. Соответственно, вектор состояния фильтра Калмана (KF) определяется, как изображено в уравнении (10), а вектор измерения очерчен в уравнении (11).

$$xk = [xc(k), yc(k), a(k), h(k), \dot{x}c(k), \dot{y}c(k), \dot{a}(k), \dot{h}(k)]T \quad (10)$$

$$zk = [zxc(k), zyc(k), za(k), zh(k)]T \quad (11)$$

Machine Translated by Google

Этот подход, который объединяет фильтрацию Калмана в анализ временных рядов, предназначен для значительного повышения точности и надежности оценки расстояния. Алгоритм фильтра Калмана методично сглаживает колебания расстояний до объектов, проводя анализ исторических данных. Он эффективно устраняет аномалии, возникающие из-за внезапных изменений размера блоков обнаружения в отдельных кадрах. Прогноз состояния алгоритма сформулирован в уравнениях (12) и (13), а механизм обновления наблюдений подробно описан в уравнениях (14)–(16).

$$^{\text{Икс}}\,k|k \quad 1 = Fkx\hat{}\,k \quad 1|k \quad 1 + Bkuk \tag{12}$$

$$Pk|k \quad 1 = FkPk \quad 1|k \quad 1F \quad {}^{T}_{\text{к}} + Kk \tag{13}$$

где $x\hat{}\,k|k$-1 представляет прогнозируемое состояние, Fk представляет собой матрицу перехода состояний, Bk обозначает модель управляющего входа, uk является вектором управления, Pk|k-1 представляет собой прогнозируемую ковариацию, а Qk означает ковариацию шума процесса.

$$K\text{к} = Pk|k \quad 1H \quad {}^{T}\,k\,(HkPk|k \quad 1H \quad {}^{T}_{\text{к}} + Pk\,)^{\,1} \tag{14}$$

$$x\hat{} = x\hat{}\,k|k \quad k|k \quad 1 + Kk\,(zk \quad Hkx\hat{}\,k|k \quad 1\,) \tag{15}$$

$$Pk|k = (I \quad KkHk\,)Pk|k \quad 1 \tag{16}$$

где Kk обозначает усиление Калмана, Hk представляет матрицу наблюдения, zk является фактическим наблюдением, Rk указывает ковариацию шума наблюдения, а I - единичная матрица.

4. Эксперименты 4.1. Наборы данных

Наборы данных COCO. Набор данных Common Objects in Context (COCO), разработанный командой Microsoft, является жизненно важным активом в исследованиях компьютерного зрения. Он содержит более 330 000 изображений с обширными аннотациями по 80 категориям объектов. Предоставляя разнообразный, крупномасштабный и сложный тест, набор данных COCO облегчил обширную исследовательскую деятельность [44]. Следовательно, предложенный нами метод был тщательно оценен на наборе данных COCO для сравнения его производительности с ведущими алгоритмами в этой области. В таблице 1 обобщаются экспериментальная установка и разделение набора данных, иллюстрирующие наш метод в наборе данных COCO.

Таблица 1. Обучение, проверка и разделение набора данных COCO.

Разделение набора данных	Количество изображений	Охватываемые категории
Обучающий набор	118 287	80 занятий
Набор проверки	5000	80 занятий
Тестовый набор	40 670	80 занятий

Наборы данных ЛОС. Набор данных классов визуальных объектов PASCAL (VOC) служит важнейшим ориентиром в области компьютерного зрения, поддерживая такие задачи, как обнаружение объектов. Он включает в себя аннотации для 20 различных категорий объектов, включая метки классов и ограничивающие рамки. Серия задач PASCAL VOC сыграла ключевую роль в продвижении технологий, основанных на зрении [45]. Кроме того, мы оцениваем улучшенный YOLOv5 на наборе данных VOC и сравниваем его с другими популярными методами. В Таблице 2 представлено подробное описание разделения набора данных VOC.

Machine Translated by Google

Таблица 2. Обучение набора данных VOC и разделение набора проверки.

Разделение набора данных	Количество изображений	Охватываемые категории
Обучающий набор	11 540	20 занятий
Набор проверки	10 991	20 занятий

Наборы данных KITTI. Набор данных KITTI является важнейшим ориентиром в области автономных систем. вождение исследований. Он охватывает множество реальных сценариев вождения, включая городские, сельской местности и шоссе. Он предоставляет подробные аннотации для таких задач, как как обнаружение, отслеживание и оценка глубины трехмерных объектов [46]. Мы используем набор данных KITTI, чтобы подтвердить точность предлагаемого нами метода оценки расстояния и продемонстрировать его применимость и эффективность в реалистичных условиях вождения.

4.2. Детали реализации

4.2.1. Настройка обучения

Чтобы обеспечить справедливость и последовательность эффекта обучения модели, детали аппаратного, программного обеспечения и экспериментальной среды представлены в Таблице 3. Обучение Настройки параметров приведены в Таблице 4.

Таблица 3. Условия эксперимента.

Среда	Версия
Операционная система	Убунту 20.04
Процессор	AMD Райзен 9 5900HX
графический процессор	NVIDIA GeForce RTX 3080 16 ГБ
Компиляция среды	Питон 3.9.0
КУДА	12.2
Структура глубокого обучения	Питорч 1.13.1
Компилятор	Пичарм 2022.3.1

Таблица 4. Параметры обучения.

Параметр	Ценить
Размер партии	16
Начальная скорость обучения	0,1
Окончательная скорость обучения	0,0001
Размер изображения	640 × 640
Оптимизатор	сингулярный доллар
Рабочие	2

4.2.2. Функция потерь

Функция потерь, используемая в модели YOLOv5-RedeCa, рассчитывается следующим образом:

$$L(tp, tgt) = \sum_{к=0}^{К} [\alpha_{к}^{баланс} \alpha box \sum_{я=0}^{С^2} \sum_{j=0}^{Б_{объект}} LC\ IoU + \alpha obj \sum_{я=0}^{С^2} \sum_{j=0}^{Б_{объект}} Лобж + \alpha cls \sum_{я=0}^{С^2} \sum_{j=0}^{Б_{объект}} Лклс] \quad (17)$$

$$_{LC\ долговое\ обязательство} = 1 \quad ИоУ + \frac{\rho^2(б, б\ гт)}{с^2} + \alpha \upsilon \quad (18)$$

где K, C ^{2}, и B представляют выходную карту объектов, ячейку и количество привязок на каждой ячейке соответственно. αx представляет вес соответствующего предмета. αbox, αobj и αcls принимают значения 0,05, 0,7 и 0,3 соответственно. Мы взяли за основу значения параметров αbox, αobj и αcls в исходном алгоритме YOLOv5. Использование стабильной и признанной базовой линии Конфигурация необходима для последовательной оценки производительности. $\alpha_{к}^{баланс}$ используется для балансировки веса выходных карт объектов в каждом масштабе со значением по умолчанию [4,0, 1,0, 0,4].

Machine Translated by Google

obj указывает, являются ли k-я выходная карта объектов, i-я ячейка и j-е поле привязки kij положительными выборками. Если это положительный образец, это 1; в противном случае это 0. tp и tgt — прогнозируемый вектор и вектор основной истины.

Функция потерь CIoU выглядит следующим образом:

$$LC_{\text{долговое обязательство}} = 1 \quad \text{ИоУ} + \frac{\rho^2 (\text{б}, \text{б гт})}{2\,\text{с}} + \alpha\upsilon \tag{19}$$

где b представляет собой центральные точки предсказанного блока, а b gt представляет блок основной истины. α — коэффициент баланса между потерями, вызванными соотношением сторон, и потерями, вызванными IoU. υ — это нормализация разницы в соотношении сторон между прогнозируемым блоком и блоком основной истины. Математически они определяются следующим образом:

$$\rho(b, b\,gt) = b \quad b\,gt \quad 2 \tag{20}$$

$$\alpha = \frac{\upsilon}{(1 \quad \text{IoU}) + \upsilon} \tag{21}$$

$$\upsilon = \frac{4(\text{арктан} \ \frac{wG}{} \frac{wP}{\pi 2} \ \text{арктан} \ \frac{hG}{} \frac{hP}{})^2}{} \tag{22}$$

Функции потерь для объекта и класса определяются следующим образом:

$$\text{Lobj}(po, pIoU) = \text{BCEsig ojb}(po, pIoU; wobj) \tag{23}$$

$$\text{Lcls}(cp, cgt) = \text{BCEsig}(cp, cgt; w_{\text{cls}}) \tag{24}$$

где (wG, hG) представляет ширину и высоту прямоугольника объекта, а (wP, hP) представляет ширину и высоту прогнозируемого блока; po — показатель достоверности объекта в прогнозируемом блоке. pIoU — это показатель достоверности объекта в прогнозируемом поле. BCEsig и BCEsig cls представляют собой двоичную кросс-энтропийную потерю. wobj и BCEcls представляют собой веса положительных образцов. клс

4.3. Индекс оценки

В следующих экспериментах мы оцениваем точность обнаружения объектов по точность и стабильность оценки расстояния с использованием следующих показателей.

- Точность. Этот показатель имеет решающее значение для оценки точности моделей обнаружения объектов. Он количественно определяет соотношение экземпляров, точно идентифицированных как объекты, ко всем случаям, помеченным моделью как объекты. Формула определяется как

$$\text{Точность} = \frac{\text{ТП}}{\text{ТП} + \text{ФП}} \tag{25}$$

где TP (True Positives) представляет собой количество правильно обнаруженных объектов, а FP (False Positives) обозначает экземпляры, ошибочно классифицированные как объекты. •

Recall: Recall оценивает долю истинных объектов, которые модель успешно обнаруживает, из всего набора реальных объектов. Более высокий уровень отзыва означает, что модель способна идентифицировать большинство истинных объектов, хотя потенциально увеличивает количество ложных срабатываний. Формула отзыва:

$$\text{Напомним} = \frac{\text{ТП}}{\text{ТП} + \text{ФН}} \tag{26}$$

с FN (ложноотрицательными результатами), указывающими количество объектов, которые не были обнаружены. • Средняя точность (AP): AP вычисляет среднее значение точности для различных уровней отзыва. Ее можно определить для каждой категории, а средняя средняя точность (MAP) представляет собой среднее значение точек доступа по всем категориям. МАП широко используется

Machine Translated by Google

для оценки комплексной производительности в задачах, связанных с обнаружением нескольких классы объектов.

$$\text{карта} = \frac{1}{\text{звонки}} \sum_{\text{я = 1}}^{\text{занятия}} \int_{0}^{1} \text{П(Р)д(Р)} \quad (27)$$

• Стандартное отклонение (SD): SD служит статистической мерой для количественной оценки степени отклонения или дисперсии от среднего значения в наборе данных. В области анализа во временной области крайне важно точно оценить близость оценок расстояния.
полученные с помощью различных алгоритмов монокулярного ранжирования до истинных значений, а также для измерения их дисперсии. Расчет стандартного отклонения определяется следующая формула:

$$\sigma = \sqrt{\frac{1}{H} \sum_{\text{я = 1}}^{H} (f(xi) \quad \mu i)^{2}} \quad (28)$$

где σ обозначает SD, f(xi) представляет собой расчетное расстояние в момент времени i, а μi соответствует фактическому значению на данный момент. Этот показатель имеет решающее значение для оценки надежность и точность этих алгоритмов в реальных приложениях.

4.4. Сравнительный эксперимент

Мы разработали улучшенный метод YOLOv5, специально оптимизированный для обнаружения мелкие объекты в дорожных сценах при различном освещении и погодных условиях. Оценить производительность нашего метода с точки зрения представления функций и обработки в реальном времени возможностей, мы провели серию экспериментов с наборами данных COCO и VOC. Наш были усовершенствованы методы, обозначенные по шкале как «Наш-с», «Наш-м» и «Наш-л». по сравнению с некоторыми современными методами обнаружения объектов.

Мы использовали точность, mAP и отзыв в качестве основных показателей производительности. Кроме того, мы учитывали сложность, оценивая количество параметров (Params) и операций с плавающей запятой в секунду (FLOPS). Мы провели сравнительный анализ различных методы на наборах данных COCO и VOC, обеспечивая надежность экспериментальных дизайн и достоверность анализа данных. Результаты представлены в сопроводительном документе. Таблица 5.

Таблица 5. Результаты обнаружения различными алгоритмами на наборах данных COCO и VOC.

Метод	COCO VOC Точность (%) mAP@.5 (%) Повторность (%) Точность (%) mAP@.5 (%) Повторность (%)						Парамс (М)	ФЛОПС (Г)
_____	50,7	45,1	38,8	31,4	54,7	57,3	26,3	65,7
Быстрее R-CNN	53,1	48,0	46,0	34,1	57,5	54,7	60,4	93,6
Маска-RCNN	51,9	48,0	45,7	33,9	57,2	69,1	63,4	147,0
YOLOv5s	63,5	54,3	52,1	69,4	60,3	52,9	7,1	16.1
YOLOv5m	70,4	61,3	55,4	82,5	74,3	67,2	21,0	48,5
YOLOv5l	72,9	64,0	59,7	82,5	75,4	69,3	46,2	108,6
YOLOv6s	64,7	52,8	46,2	65,3	58,3	51,6	18,5	45,3
YOLOv7s	65,8	47,6	49,5	66,8	58,3	52,5	6,1	13.2
YOLOv8s	68,1	56,3	53,1	68,8	56,5	53,0	11,2	28,8
YOLOv9s	67,1	55,7	50,7	66,7	61,7	54,1	9,6	38,8
YOLOв10c	66,3	57,2	51,8	68,4	60,7	56,1	8,1	24,9
Наши-с	67,2	55,0	51,3	67,0	57,5	53,5	6,7	13.2
Наш-м	71,6	61,5	57,9	81,9	77,5	73,4	17,2	34,6
Наш-л	73,6	64,4	60,1	82,5	80,0	76,0	34,3	70,4

Экспериментальные результаты на наборе данных COCO показывают, что наша оптимизированная модель Ours-m достиг точности 71,6%, mAP 61,5% и отзыва 57,9%. Эти результаты являются лучшими среди все методы используют схожие вычислительные ресурсы. В наборе данных VOC модель также отличился, зарегистрировав 77,5% mAP@.5 и 73,4% Recall. Примечательно, что, несмотря на меньшее количество параметры, чем аналогичная масштабная модель YOLOv5, Ours-s, Ours-m и Ours-l превзошли

Machine Translated by Google

другие методы Precision и mAP. Это демонстрирует наш успех в повышении производительности при оптимизации эффективности вычислений.

По сравнению с SSD и Faster R-CNN Ours-m продемонстрировал значительные преимущества как на наборах данных COCO, так и на VOC. Это было особенно эффективно в Recall для обнаружения небольших объектов. Например, в наборе данных COCO Recall of Ours-l был на 21,3% выше, чем у SSD, и на 14,1% выше, чем у Faster R-CNN, что подчеркивает эффективность модуля RedeCa в расширении возможностей отзыва.

Хотя Mask-RCNN предназначен для сегментации экземпляров, модель Ours-l преуспела в задачах обнаружения объектов. В наборе данных VOC он достиг mAP@.5 80,0% по сравнению с 57,2% у Mask-RCNN. Это подчеркивает, что наша улучшенная модель обеспечивает выдающуюся производительность обнаружения даже в областях, специализирующихся на Mask-RCNN.

Мы также контролировали сложность модели, стремясь к более высокой производительности. По сравнению с оригинальным YOLOv5, серия Ours продемонстрировала превосходное обнаружение мелких объектов в сложных средах с меньшим количеством параметров. В наборе данных COCO Ours-l улучшил mAP и Recall на 0,4% каждый по сравнению с YOLOv5l. Наши методы превосходят существующие методы по ключевым показателям производительности, сохраняя при этом вычислительную эффективность, закладывая прочную основу для практического применения.

Мы сравнили наш метод с новейшими методами обнаружения объектов от YOLOv6 до YOLOv10 [47–51]. Наш метод соответствует YOLOv8 и YOLOv9 с точки зрения точности и превосходит YOLOv6, v8, v9 и v10 по облегчению модели и вычислительной нагрузке. Данные , представленные в таблице 5, показали, что, хотя наша модель достигает точности, сравнимой с последними версиями YOLO, она по-прежнему имеет преимущества в облегчении модели и вычислительной нагрузке. Эти преимущества способствуют внедрению и эффективности нашего метода в практических приложениях.

Модуль RedeCa расширяет возможности нашего метода по представлению особенностей, особенно в точной обработке информации о краях и деталях, что имеет решающее значение для обнаружения небольших объектов . Это также повышает адаптивность к изменяющемуся освещению и погодным условиям, тем самым повышая надежность в сложных условиях.

В сценариях дорожного движения решающее значение имеет производительность в режиме реального времени. Мы сравнивали наш метод с популярными методами с точки зрения FPS и количества параметров, как подробно описано в таблице 5. Количество параметров и вычислительная сложность нашего метода заметно ниже, чем у Faster R-CNN и Mask-RCNN. По сравнению с базовым YOLOv5 наш метод обеспечивает сопоставимую производительность при уменьшении количества параметров и сложности вычислений

Результаты экспериментов показывают, что, добавив модуль RedeCa в YOLOv5, мы заметно повышаем точность обнаружения мелких объектов, сохраняя при этом высокую скорость отзыва.

Эти результаты подтверждают дальнейшее изучение методов глубокого обучения для обнаружения мелких объектов в сложных средах, подтверждая эффективность и потенциал наших улучшений.

Чтобы подтвердить превосходную точность и надежность нашего метода, мы провели комплексный количественный и качественный сравнительный анализ. Мы сравнили наш метод с популярными монокулярными методами измерения расстояния в сценариях дорожного движения.

Результаты показаны на рисунке 7.

Как показано на рисунке 7, наш метод монокулярного измерения расстояния эффективно снижает потенциальные ошибки, вызванные внезапными изменениями размера ограничивающей рамки обнаружения.

Это приводит к уменьшению разброса результатов измерений расстояний, гарантируя, что они будут более согласовываться с фактическими расстояниями с течением времени.

Мы также провели качественный анализ, чтобы сравнить наш метод с основными методами. В отличие от методов, которые создают карты глубины для всего изображения для определения расстояний до объектов, наш метод одновременно обнаруживает объекты и измеряет их расстояния, что повышает эффективность. Результаты измерения расстояния показаны на рисунке 8.

Machine Translated by Google

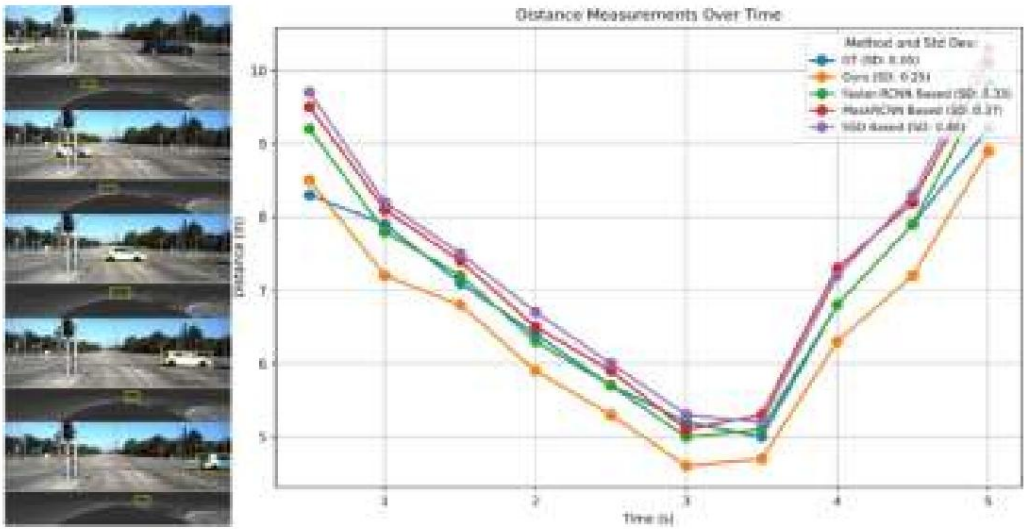


Рисунок 7. Статистические результаты и расчеты стандартных отклонений различных монокулярных расстояний. методы измерения. Стандартное отклонение нашего метода составляет 0,25. Стандартное отклонение Faster-RCNN составляет 0,33. Стандартное отклонение Маска R-CNN составляет 0,37. Стандартное отклонение SSD составляет 0,40.

(а) (б) (с) (г)

Рисунок 8. Результаты визуализации различных монокулярных методов измерения расстояния. (а) является основная истина. (б) — результат, основанный на Faster R-CNN. (с) — результат на основе YOLOv5. (г) является результат на основе YOLO-RedeCa.

Machine Translated by Google

4.5. Эксперимент по абляции

Мы использовали метод эксперимента по абляции, чтобы оценить влияние различных модулей на эффективность обнаружения модели YOLOv5m в наборе данных ЛОС. Экспериментальный результаты показаны в таблице 6. Мы исследуем эффект отделимой по глубине свертки (DW) о производительности модели. Кроме того, мы изучаем влияние эффективных каналов внимание (ЭКА). Мы также исследуем прямое сочетание DW и ECA. Наконец, мы оцениваем производительность модуля RedeCa.

Таблица 6. Результаты обнаружения с использованием различных модулей в наборе данных ЛОС.

Метод	DW ECA RedeCa (DW + ECA) Точность (%) mAP (%) IOU (%) Params (M) FLOPS (G)				
YOLOv5m	82,5	74,3	67,2	21,0	48,5
YOLOv5m	70,5	62,9	38,4	9,1	18,3
YOLOv5m	82,7	76,3	74,1	21,8	48,6
YOLOv5m	79,3	74,6	71,3	20,8	47,6
YOLOv5m + RedeCa	81,9	77,5	73,4	17,2	34,6

Как показано в Таблице 6, базовая модель YOLOv5m достигла точности 82,5 %. mAP 74,3% и IOU 67,2% в наборе данных VOC. Он имел 21,0 миллиона параметров и 48,5 ГФЛОПС.

После добавления Depthwise Separable Convolutions (DW) к базовой модели точность снизилась до 70,5%, mAP упала до 62,9%, а IOU значительно снизилась до 38,4%. Однако количество параметров и FLOPS сократились до 9,1 миллиона и 18,3 GFLOPS соответственно. Это указывает на то, что, хотя модуль DW эффективно снижает сложность модели, это отрицательно влияет на производительность обнаружения.

Включение модуля Efficient Channel Attention (ECA) привело к небольшому улучшение точности до 82,7%, увеличение mAP до 76,3% и значительный рост долговая расписка до 74,1%. Количество параметров и FLOPS немного увеличились, демонстрируя что модуль ECA повышает эффективность обнаружения без значительного увеличения вычислительная нагрузка.

Использование глубоко отделимых сверток (DW) и эффективного внимания канала (ECA), модель продемонстрировала небольшие улучшения в средней средней точности (mAP) и Пересечение через объединение (IoU). Но прирост был минимальным и не было никаких существенных снижение параметров сети и вычислительной нагрузки. Однако, эффективно при интеграции DW и ECA в модуль RedeCa точность модели немного снизилась до 81,9%, но MAP увеличился до 77,5%, а IoU вырос до 73,4%, при этом параметры и FLOPS сократились до 17,2 M и 34,6 Г соответственно. Это указывает на то, что использование модуля RedeCa может сбалансировать производительность обнаружения и сложность модели до некоторых степени, обеспечивая эффективное решение для приложений обнаружения объектов в реальном времени.

Проверить, что использование фильтра скачков аномалий помогает сгладить изменения расстояний и устранить потенциальные ошибки, вызванные внезапными изменениями размера ограничивающей рамки обнаружения, мы провели сравнительный анализ. Мы сравнили архитектуру алгоритма с фильтр скачка аномалии на фильтр без него.

Как показано на рисунке 9, эффективное использование фильтра Калмана, интегрированного с временным рядом. устраняет потенциальные ошибки, вызванные внезапными изменениями размера ограничивающей рамки обнаружения. Этот уменьшает дисперсию измерений расстояний, делая их более согласованными с фактические расстояния с течением времени.

Machine Translated by Google

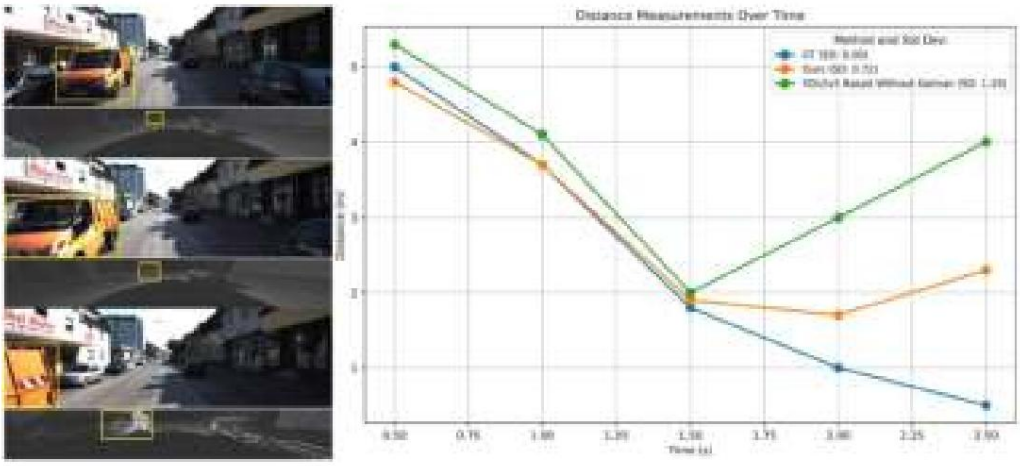


Рисунок 9. Влияние использования фильтрации Калмана на колебания оценки расстояния. Стандартное отклонение нашего метода составляет 0,72. Стандартное отклонение для YOLOV5-RedeCa.

метод без использования фильтрации Калмана составляет 1,29.

5. Выводы

Наша улучшенная сеть обнаружения объектов превосходно работает с набором данных COCO. Он достигает точности 71,6%, средней точности 61,5% (mAP@.5) и полноты 57,9% соответственно. Эти результаты лучше, чем модели сравнения при тех же вычислительных ресурсах. Кроме того, в наборе данных VOC наша улучшенная сеть обнаружения объектов также показывает отличную производительность в mAP@.5 и отзыве. Результаты составляют 77,5% и 73,4% соответственно. За счет интеграции алгоритма Bot-Sort надежность предлагаемого метода значительно повышается при работе с перекрытием или временной потерей объектов, так что система может отслеживать и удалять несколько объектов одновременно. В то же время фильтр аномальных скачков может эффективно сглаживать внезапное изменение дальности объекта, устранять потенциальную аномалию, вызванную изменением размера окна обнаружения, и тем самым повышать надежность определения дальности.

Комплексная архитектура алгоритма, которую мы разработали и внедрили, успешно преодолевает традиционный исследовательский разрыв между обнаружением 2D-объектов и монокулярной оценкой глубины. Мы объединили эти два метода, разработав новую архитектуру алгоритма. Это значительно повышает производительность обнаружения объектов и оценки расстояния в динамичных и сложных средах. Интеграция этих технологий не только оптимизирует соответствующие функции, но и повышает эффективность системы в практическом применении. В частности, он показал большой потенциал применения в области мониторинга дорожного движения и автоматического вождения.

Вклад автора: концептуализация, HL, YM, YD и YY; методология, HL и YM; программное обеспечение, ХЛ; проверка, HL и YY; формальный анализ, HL, YY и YM; расследование, ХЛ; ресурсы, HL, YD и YM; курирование данных, HL; письмо – подготовка оригинального черновика, HL; написание — просмотр и редактирование, HL, YM и YY; визуализация, ХЛ; надзор, YD и YM; администрация проекта, HL; приобретение финансирования, YD. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Эта работа в основном поддерживается проектом совместного управления автономным вождением транспортных средств и дорог и спонсируется академическими исследовательскими проектами Пекинского союзного университета (номера ZK80202003, ZK90202105), Программой науки и технологий Пекинской муниципальной комиссии по образованию (№№ KM202111417007, KM202211417006).

Заявление о доступности данных. Используемый набор данных COCO можно получить по адресу https://cocodataset. opr/#дом (по состоянию на 24 июня 2024 г.). Набор данных VOC можно получить по адресу http://host. robots.ox.ac.uk/pascal/VOC/ (по состоянию на 24 июня 2024 г.). Набор данных KITTI можно получить по адресу https://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=3d . (по состоянию на 24 июня 2024 г.). Доступна архитектура слияния с использованием YOLO-RedeCa и фильтра ненормальных скачкообразных изменений.

Machine Translated by Google

на GitHub по адресу https://github.com/Mr-Lv-BUU/The-Fusion-Architecture-Using-YOLO-RedeCa- and-Abnormal-Jumping-Change-Filter.git. (по состоянию на 24 июня 2024 г.).

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Тан, К.; Ву, Дж.; Чжоу, Х.; Ван, Ю.; Чен Дж. Интеграция передового компьютерного зрения и алгоритмов искусственного интеллекта для автономного вождения Системы. Дж. Теоретическая практика. англ. наук. 2024, 4, 41–48.
2. Харис, М.; Гловач, А. Обнаружение дорожных объектов: сравнительное исследование алгоритмов, основанных на глубоком обучении. Электроника 2021, 10, 1932. [Перекрестная ссылка]
3. Ли, Ю.; Хуанг, Б.; Чен, З.; Цюи, Ю.; Лян, Ф.; Шен, М.; Лю, Ф.; Се, Э.; Шэн, Л.; Оуян, В.; Шао, Дж. Fast-BEV: базовый уровень быстрого и четкого восприятия с высоты птичьего полета. IEEE Транс. Паттерн Анал. Мах. Интел. 2024, 1–14.
4. Гиршик, Р.; Донахью, Дж.; Даррелл, Т.; Малик, Дж. Рич использует иерархии для точного обнаружения объектов и семантической сегментации. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Колумбус, Огайо, США, 23–28 июня 2014 г.; стр. 580–587.
5. Диван, Т.; Анируд, Г.; Тембхерн, СП. Обнаружение объектов с помощью YOLO: проблемы, архитектурные преемники, наборы данных и Приложения. Мультимед. Инструменты Прил. 2023, 82, 9243–9275. [Перекрестная ссылка] [ПабМед]
6. Чжао, ZQ; Чжэн, П.; Сюй, СТ; Ву, Х. Обнаружение объектов с помощью глубокого обучения: обзор. IEEE Транс. Нейронная сеть. Учиться. Сист. 2019, 30, 3212–3232. [Перекрестная ссылка] [ПабМед]
7. Поджи, М.; Тоси, Ф.; Батсос, К.; Мордохай, П.; Матточча, С. О синергии машинного обучения и бинокулярного стерео для оценки глубины по изображениям: обзор. IEEE Транс. Паттерн Анал. Мах. Интел. 2021, 44, 5314–5334. [Перекрестная ссылка] [ПабМед]
8. Саксена А.; Сан, М.; Нг, А.Ю. Make3d: изучение структуры 3D-сцены по одному неподвижному изображению. IEEE Транс. Паттерн Анал. Мах. Интел. 2008, 31, 824–840. [Перекрестная ссылка]
9. Карш, К.; Лю, К.; Канг, С.Б. Передача глубины: извлечение глубины из видео с использованием непараметрической выборки. IEEE Транс. Шаблон Анальный. Мах. Интел. 2014, 36, 2144–2158. [Перекрестная ссылка]
10. Фу, К.; Юань, Х.; Сюй, Х.; Чжан, Х.; Шен, Л. TMSO-Net: Многомасштабное наблюдение с адаптацией к текстурам для определения глубины изображения светового поля. оценка. Дж. Вис. Коммун. Изображение представляет. 2023, 90, 103731. [CrossRef]
11. Цзоу, З.; Чен, К.; Ши, З.; Го, Ю.; Йе, Дж. Обнаружение объектов за 20 лет: обзор. Учеб. IEEE 2023, 111, 257–276. [Перекрестная ссылка]
12. Крижевский А.; Суцкевер И.; Хинтон, Классификация GE Imagenet с глубокими сверточными нейронными сетями. Адв. Нейронная инф. Процесс. Сист. 2012, 25, 1097–1105. [Перекрестная ссылка]
13. Сяо, Ю.; Тиан, З.; Ю, Дж.; Чжан, Ю.; Лю, С.; Ду, С.; Лан, Х. Обзор обнаружения объектов на основе глубокого обучения. Мультимед. Инструменты Прил. 2020, 79, 23729–23791. [Перекрестная ссылка]
14. Он, К.; Чжан, Х.; Рен, С.; Сан, Дж. Объединение пространственных пирамид в глубокие сверточные сети для визуального распознавания. IEEE Транс. Паттерн Анал. Мах. Интел. 2015, 37, 1904–1916. [Перекрестная ссылка] [ПабМед]
15. Гиршик, Р. Фаст р-кнн. В материалах Международной конференции IEEE по компьютерному зрению, Сантьяго, Чили, 7–13 декабря. 2015 г.; стр. 1440–1448.
16. Рен, С.; Он, К.; Гиршик, Р.; Сан, Дж. Фастер r-cnn: На пути к обнаружению объектов в реальном времени с помощью сетей предложений регионов. Адв. Нейронная инф. Процесс. Сист. 2015, 28, 91–99. [Перекрестная ссылка] [ПабМед]
17. Мартин, Д.; Фаулкс, К.; Таль, Д.; Малик, Дж. База данных сегментированных естественных изображений человека и ее применение для оценки алгоритмов сегментации и измерения экологической статистики. В материалах восьмой Международной конференции IEEE по компьютерному зрению, ICCV 2001, Ванкувер, Британская Колумбия, Кандада, 7–14 июля 2001 г.
18. Дай, Дж.; Ли, Ю.; Он, К.; Сан, Дж. Р-фкн: Обнаружение объектов с помощью полностью сверточных сетей на основе регионов. Адв. Нейронная инф. Процесс. Сист. 2016, 29, 379–387.
19. Серманет, П.; Эйген, Д.; Чжан, Х.; Матье, М.; Фергюс, Р.; ЛеКун, Ю. Overfeat: Интегрированное распознавание, локализация и обнаружение с использованием сверточных сетей. arXiv 2013, arXiv:1312.6229.
20. Редмон, Дж.; Диввала, С.; Гиршик, Р.; Фархади, А. Вы смотрите только один раз: унифицированное обнаружение объектов в реальном времени. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Лас-Вегас, Невада, США, 26 июня – 1 июля 2016 г.; стр. 779–788.
21. Хан, Х.; Чанг, Дж.; Ван, К. Обнаружение объектов в реальном времени на основе YOLO-v2 для крошечных транспортных средств. Процессия Компьютер. наук. 2021, 183, 61–72. [Перекрестная ссылка]
22. Редмон, Дж.; Фархади, А. YOLO9000: Лучше, быстрее, сильнее. В материалах конференции IEEE по компьютерному зрению и Распознавание образов, Гонолулу, Гавайи, США, 21–26 июля 2017 г.; стр. 7263–7271.
23. Редмон, Дж.; Фархади, А. Йолов3: Постепенное улучшение. arXiv 2018, arXiv:1804.02767.
24. Бочковский А.; Ван, СУ; Ляо, ХИМ Йолов4: Оптимальная скорость и точность обнаружения объектов. arXiv 2020, arXiv:2004.10934.
25. Джохер, Г.; Чаурасия, А.; Стокен, А.; Боровец, Дж.; Квон, Ю.; Майкл, К.; Фанг, Дж.; Вонг, К.; Ифу, З.; Монтес, Д.; и другие. ultralytics /yolov5: v6.2 — модели классификации YOLOv5, Apple M1, воспроизводимость, интеграция ClearML и Deci.ai, v6.2; Зенодо: Женева, Швейцария, 2022 г.
26. Лю, В.; Ангелов Д.; Эрхан, Д.; Сегеди, К.; Рид, С.; Фу, СУ; Берг, АС Ssd: Мультибоксовый детектор одиночного действия. В материалах Computer Vision –ECCV 2016: 14-я Европейская конференция, Амстердам, Нидерланды, 11–14 октября 2016 г.; Слушания, часть I 14; Springer: Берлин/Гейдельберг, Германия, 2016 г.; стр. 21–37.

Machine Translated by Google

27. Фу, Калифорния; Лю, В.; Ранга, А.; Тьяги, А.; Берг, АС Dssd: Деконволюционный детектор одиночных выстрелов. arXiv 2017, arXiv:1701.06659.
28. Ли, З.; Ян, Л.; Чжоу, Ф. FSSD: Функциональный однозарядный многокамерный детектор Fusion. arXiv 2017, arXiv:1712.00960.
29. Рен, Д.; Ян, М.; Ву, Дж.; Чжэн, Н. Нормальные к поверхности и весовые ограничения по Гауссу для завершения внутренней глубинной конструкции. Распознавание образов. 2023, 138, 109362. [CrossRef]
30. Фаваро, П.; Соатто, С. Геометрический подход к формированию путем расфокусировки. IEEE Транс. Паттерн Анал. Мах. Интел. 2005, 27, 406–417. [Перекрестная ссылка] [ПабМед]
31. Лю, Б.; Гулд, С.; Коллер, Д. Оценка глубины одного изображения на основе предсказанных семантических меток. В материалах конференции IEEE Computer Society 2010 г. по компьютерному зрению и распознаванию образов, Сан-Франциско, Калифорния, США, 13–18 июня 2010 г.; стр. 1253–1260.
32. Эйген, Д.; Пурш, К.; Фергус, Р. Прогнозирование карты глубины по одному изображению с использованием многомасштабной глубокой сети. Адв. Нейронная инф. Процесс. Сист. 2014, 27, 2366–2374.
33. Кузнецов Ю.; Стаклер, Дж.; Лейбе, Б. Полуконтролируемое глубокое обучение для монокулярного прогнозирования карты глубины. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Гонолулу, Гавайи, США, 21–26 июля 2017 г.; стр. 6647–6655.
34. Эйген, Д.; Фергюс, Р. Прогнозирование глубины, нормалей поверхности и семантических меток с помощью общей многомасштабной сверточной архитектуры. В материалах Международной конференции IEEE по компьютерному зрению, Сантьяго, Чили, 7–13 декабря 2015 г.; стр. 2650–2658.
35. Лю, К.; Ян, Г.; Цзо, В.; Занг, Т. DPDFormer: модель от грубого к точному для оценки глубины монокуляра. АКМ Транс. Мультимед. Вычислить. Коммун. Прил. 2024, 20, 1–21. [Перекрестная ссылка]
36. Ли, З.; Чен, З.; Лю, Х.; Цзян, Дж. Депффомер: Использование корреляции на большом расстоянии и локальной информации для точного монокуляра оценка глубины. Мах. Интел. Рез. 2023, 20, 837–854. [Перекрестная ссылка]
37. Цзяо Дж.; Цао, Ю.; Сонг, Ю.; Лау, Р. Взгляните глубже в глубину: монокулярная оценка глубины с семантическим усилением и потерей внимания. В материалах Европейской конференции по компьютерному зрению (ECCV), Мюнхен, Германия, 8–14 сентября 2018 г.; стр. 53–69.
38. Миклеа, В.К.; Недевски, С. Монокулярная оценка глубины с повышенной точностью на больших расстояниях для восприятия окружающей среды БПЛА. IEEE Транс. Геосци. Дистанционная сенсорика 2021, 60, 1–15. [Перекрестная ссылка]
39. Гарг, Р.; Бг, ВК; Карнейро, Г.; Рид, И. Неконтролируемая CNN для оценки глубины одним видом: геометрия на помощь. В материалах Computer Vision –ECCV 2016: 14-я Европейская конференция, Амстердам, Нидерланды, 11–14 октября 2016 г.; Слушания, часть VIII 14; Springer: Берлин/Гейдельберг, Германия, 2016 г.; стр. 740–756.
40. Годар, К.; Мак Аода, О.; Бростоу, Г.Дж. Неконтролируемая монокулярная оценка глубины с согласованностью слева и справа. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Гонолулу, Гавайи, США, 21–26 июля 2017 г.; стр. 270–279.
41. Чжоу, Т.; Браун, М.; Снавли, Н.; Лоу, Д.Г. Неконтролируемое обучение глубине и движениям эго с помощью видео. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, Гонолулу, Гавайи, США, 21–26 июля 2017 г.; стр. 1851–1858.
42. Махаур, Б.; Мишра, К. Обнаружение мелких объектов на основе YOLOv5 в системах автономного вождения. Распознавание образов. Летт. 2023, 168, 115–122. [Перекрестная ссылка]
43. Аарон, Н.; Орфейг, Р.; Бобровский, БЗ BoT-SORT: Надежные объединения многопешеходного слежения. arXiv 2022, arXiv:2206.14651.
44. Лин, Тайвань; Мэр, М.; Белонги, С.; Хейс, Дж.; Перона, П.; Раманан, Д.; Доллар, П.; Зитник, К. Л. Microsoft Coco: Общие объекты в контексте. В материалах Computer Vision –ECCV 2014: 13-я Европейская конференция, Цюрих, Швейцария, 6–12 сентября 2014 г.; Слушания, часть V 13; Springer: Берлин/Гейдельберг, Германия, 2014 г.; стр. 740–755.
45. Эверингем, М.; Эслами, SA; Ван Гул, Л.; Уильямс, КК; Винн, Дж.; Зиссерман, А. Задача классов визуальных объектов Паскаля: ретроспектива. Межд. Дж. Компьютер. Вис. 2015, 111, 98–136. [Перекрестная ссылка]
46. Гейгер А.; Ленц, П.; Стиллер, К.; Уртасун, Р. Vision встречает робототехнику: набор данных Kitti. Межд. Джей Робот. Рез. 2013, 32, 1231–1237. [Перекрестная ссылка]
47. Ли, К.; Ли, Л.; Цзян, Х.; Венг, К.; Гэн, Ю.; Ли, Л.; Ке, З.; Ли, К.; Ченг, М.; Не, В.; и другие. YOLOv6: одноэтапное обнаружение объектов рамки для промышленного применения. arXiv 2022, arXiv:2209.02976.
48. Ван, К.-Ю.; Бочковский А.; Ляо, Х.-ЮМ YOLOv7: Обучаемый набор подарков устанавливает новый уровень развития детекторов объектов в реальном времени. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Ванкувер, Британская Колумбия, Канада, 17–24 июня 2023 г.; стр. 7464–7475.
49. Лу, Х.; Дуань, Х.; Го, Дж.; Лю, Х.; Гу, Дж.; Би, Л.; Чен, Х. DC-YOLOv8: Алгоритм обнаружения объектов малого размера на основе камеры датчик. Электроника 2023, 12, 2323. [CrossRef]
50. Ван, К.-Ю.; Ага, И.-Х.; Ляо, Х.-ЮМ YOLOv9: Изучение того, чему вы хотите научиться, с использованием программируемой градиентной информации. arXiv 2024, arXiv:2402.13616.
51. Ван, А.; Чен, Х.; Лю, Л.; Чен, К.; Лин, З.; Хан, Дж.; Дин, Г. YOLOv10: Сквозное обнаружение объектов в режиме реального времени. apXiv 2024, arXiv: 2405.14458.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Подход классификации зашифрованного трафика на основе пути Фирменные функции и LSTM

Йихе Мэй [1], Нурбол Луктархан 2,*, Годун Чжао [1] и Сяотун Ян [2]

[1] Школа программного обеспечения Синьцзянского университета, Урумчи 830091, Китай; 107552204789@stu.xju.edu.cn (ЮМ); 107552204851@stu.xju.edu.cn (GZ)
[2] Колледж информатики и инженерии Синьцзянского университета, Урумчи 830046, Китай; 107552204103@stu.xju.edu.cn *
Переписка: nurbol@xju.edu.cn

Аннотация: Классификация зашифрованного трафика является важнейшим аспектом сетевой безопасности. Однако популярные методы сталкиваются с рядом ограничений, таких как зависимость от проектирования признаков и необходимость создания сложной архитектуры моделей для обеспечения эффективной классификации. Чтобы решить эти проблемы, мы предлагаем метод, который сочетает в себе функции подписи пути с моделями долгосрочной краткосрочной памяти (LSTM) для классификации типов услуг в зашифрованном трафике. Наш подход строит пути трафика, используя размер пакета и время прибытия. Мы генерируем признаки пути в различных масштабах, используя инновационную многомасштабную технику извлечения кумулятивных признаков. Эти функции служат входными данными для сетей LSTM для выполнения классификации. Примечательно, что, используя только 24 функции последовательного пакетирования в сочетании с моделями LSTM, наш метод достиг значительных успехов в классификации типов у Результаты экспериментов подчеркивают превосходство предложенного нами метода по сравнению с ведущими подходами в этой области.

Ключевые слова: классификация зашифрованного трафика; функции подписи пути; длительная кратковременная память; глубокое обучение

check for updates

Цитирование: Мэй, Ю.; Луктархан, Н.; Чжао, Г.; Ян, Х. Подход к классификации зашифрованного трафика, основанный на функциях сигнатуры пути и LSTM. Электроника 2024, 13, 3060. https://doi.org/10.3390/electronics13153060 .

Академический редактор: Мартин Рейсляйн

Поступила: 27 июня 2024 г.
Пересмотрено: 29 июля 2024 г.
Принято: 31 июля 2024 г.
Опубликовано: 2 августа 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

В эпоху повсеместного подключения к Интернету классификация сетевого трафика стала краеугольным камнем кибербезопасности. Важность этой дисциплины подчеркивается ее критически важными приложениями для обнаружения вторжений, анализа тенденций трафика, управления качеством обслуживания и повышения прозрачности сети [1]. С широким распространением технологий шифрования ситуация с сетевым трафиком резко изменилась, поскольку все больше пользователей и поставщиков услуг предпочитают передавать информацию через зашифрованный трафик [2]

Зашифрованный трафик отличается от незашифрованного трафика тем, что для кодирования данных, передаваемых между устройствами, используются протоколы шифрования, что предотвращает доступ к ним неавторизованных сторон. Такое шифрование обеспечивает значительные преимущества с точки зрения конфиденциальности и безопасности, защищая конфиденциальную информацию от перехвата и использования злоумышленниками . Зашифрованный трафик обеспечивает конфиденциальность и целостность пользовательских данных, сообщений и транзакций, делая их защищенными от несанкционированного доступа. Напротив, незашифрованный трафик передает данные в виде обычного текста, который может быть легко перехвачен и прочитан любым, имеющим доступ к сети. Хотя незашифрованный трафик упрощает мониторинг и анализ сетевыми администраторами, он подвергает пользователей более высокому риску утечки данных и нарушения конфиденциальности. Основными преимуществами незашифрованного трафика являются его простота и меньшие вычислительные затраты, что может привести к более быстрой передаче данных. Однако преимущества безопасности зашифрованного трафика намного перевешивают эти преимущества, что делает шифрование необходимым в современной цифровой среде.

Хотя зашифрованный трафик защищает конфиденциальность и анонимность большинства пользователей, он одновременно представляет собой серьезную проблему для кибербезопасности [3]. Преступники все чаще используют сложные методы шифрования для ведения незаконной деятельности.

Machine Translated by Google

подрывают стабильность киберпространства и создают угрозы национальной безопасности. Исследователям срочно нужны новые методы классификации, чтобы решить эту сложную задачу.

Появление шифрования создало существенную проблему для традиционных методов классификации трафика [4]. Методы, которые когда-то полагались на проверку полезной нагрузки и классификацию на основе портов, теперь оказались неэффективными против завесы шифрования [5,6].

Исторически эта область опиралась на статистические методы, а в последнее время и на алгоритмы машинного обучения для решения сложных задач классификации зашифрованного трафика [7–10].
Однако эти методы имеют свои собственные проблемы. Они часто требуют значительных усилий по проектированию функций и сильно зависят от качества и количества доступных данных. Более того, несмотря на практичность, сложные конструкции моделей глубокого обучения требуют огромных вычислительных ресурсов и больших наборов данных для достижения удовлетворительной точности [4].

В настоящее время к реальным приложениям для классификации зашифрованного трафика относятся Cisco Encrypted Traffic Analytics (ETA), Darktrace, Symantec Encrypted Traffic Management (ETM) и Vectra AI. Cisco ETA идентифицирует вредоносное ПО в зашифрованном трафике с помощью сетевой телеметрии и машинного обучения. Darktrace использует обнаружение аномалий на основе искусственного интеллекта для выявления угроз в зашифрованных данных. Symantec ETM расшифровывает, проверяет и повторно шифрует трафик, используя глубокую проверку пакетов для классификации. Платформа Cognito от Vectra AI использует машинное обучение и поведенческий анализ для обнаружения угроз в зашифрованном трафике. Эти решения подчеркивают зависимость отрасли от машинного обучения, искусственного интеллекта и расширенного сетевого анализа для безопасного и эффективного управления зашифрованным трафиком.

Функция подписи пути, предложенная Ченом [11], в последние годы была успешно применена в таких областях, как машинное обучение, распознавание образов и анализ путей. Эта функция также недавно была введена в область классификации зашифрованного трафика, где она продемонстрировала свою осуществимость на нескольких наборах данных [12]. Однако присущая сложность характеристик сигнатур пути означает, что результаты классификации не всегда являются удовлетворительными. При классификации зашифрованного трафика сохраняется проблема полного использования функций сигнатур пути для достижения удовлетворительных результатов, что требует дальнейших исследований и инноваций.

В этой статье мы представляем многомасштабный метод извлечения признаков кумулятивных сигнатур пути, чтобы максимизировать информационное содержание функций сигнатур пути. Чтобы полностью раскрыть потенциал этих функций, мы изучили их интеграцию с глубокими нейронными сетями.
Эта интеграция привела к удовлетворительным результатам классификации, продемонстрировав эффективность сочетания функций сигнатуры пути с глубоким обучением при классификации зашифрованного трафика

Основные положения данной статьи заключаются в следующем:

1. Мы предлагаем классификатор зашифрованного трафика, который сочетает в себе функции подписи пути с моделью LSTM. Этот классификатор извлекает сложную динамику и геометрические структуры зашифрованного трафика, используя функции сигнатуры пути, и использует модель LSTM для выполнения задачи классификации, что значительно сокращает вычислительные ресурсы, необходимые для классификации трафика.

2. Мы представляем многомасштабный метод извлечения кумулятивных признаков для создания признаков сигнатуры пути, которые оптимально подходят для моделей LSTM. Используя этот метод, точность классификации модели LSTM можно повысить на 5–30 %, что подтверждает эффективность нашего нового подхода к выделению признаков.

3. Предлагаемый нами метод позволяет достичь результатов конкурентной классификации, используя только 24 последовательных признака длины пакета и временных интервалов. Он демонстрирует точность классификации 94,74%, 90,53%, 93,86% и 95,03% для наборов данных ISCX-VPN, ISCX-nonVPN, ISCX-Tor и ISCX-nonTor соответственно, что доказывает эффективность нашего подхода.

2. Сопутствующая работа

2.1. Классификация зашифрованного трафика

Классификацию интернет-трафика можно разделить на три основные области:

1. Различение шифрованного и незашифрованного трафика [3]; 2. Определение конкретных приложений, связанных с трафиком [13];

Machine Translated by Google

3. Классификация трафика по разным типам услуг. Типы услуг относятся к цели трафика, т. е. трафику, создаваемому для удовлетворения конкретных потребностей пользователя. Например, трафик передачи файлов генерируется, когда пользователи выполняют операции загрузки или скачивания в сети, а потоковый трафик генерируется, когда пользователи используют службы потокового мультимедиа, такие как прослушивание музыки или просмотр видео [14].

В этом разделе обобщены основные методы классификации сетевого трафика, разделенные на категории в зависимости от используемых методов: на основе портов, на основе полезной нагрузки, построение отпечатков пальцев, статистические методы и модели глубокого обучения.

На основе портов: этот метод, один из старейших в классификации сетевого трафика, основывался на фиксированных номерах портов, назначенных Управлением по присвоению номеров Интернета (IANA) на раннем этапе развития Интернета. Этот подход облегчил ассоциацию сетевых приложений с соответствующим им трафиком посредством сканирования портов, что позволило идентифицировать трафик [15]. Однако с экспоненциальным ростом сетевых приложений и методов динамического портирования сканирование портов больше не применимо к текущей сетевой среде [5]. Наш метод, в частности, не зависит от номеров портов.

На основе полезной нагрузки. Чтобы устранить ограничения сканирования портов, исследователи предложили методы глубокой проверки пакетов (DPI) [6] для идентификации трафика. DPI проверяет часть данных пакета (полезную нагрузку) и использует известные сигнатуры для определения типа трафика. Основной принцип DPI заключается в том, что пакеты данных различных протоколов связи имеют определенные форматы пакетов, а определенные символы или строки со значимой информацией можно найти путем извлечения полезной нагрузки пакетов. Это конкретное содержимое полезной нагрузки служит индикаторами распознавания типа трафика [6]. Однако этот метод создает риски нарушения конфиденциальности пользователей, больших затрат на вычисления и становится неэффективным при работе с зашифрованным трафиком [16]. Наш подход позволяет избежать зависимости от информации о полезной нагрузке.

Создание отпечатков пальцев. Чтобы устранить ограничения методов классификации трафика на основе полезной нагрузки, в нескольких исследованиях для классификации трафика использовалась незашифрованная информация поля протокола [3]. Например, Т. Ван Эде и др. [17] использует такие функции, как тип устройства, сведения о сертификате, размер пакета и информацию о времени для представления каждого потока, создавая библиотеку отпечатков пальцев с помощью методов кластеризации и взаимной корреляции для эффективной классификации трафика. Однако эти методы в значительной степени полагаются на информацию в виде открытого текста, которую можно легко подделать во время передачи, что приводит к потере точного смысла. И наоборот, наша модель работает независимо от текстовой информации в пакетах данных.

Статистические методы: эта категория предполагает ручное извлечение таких характеристик, как размер и время трафика, с использованием алгоритмов машинного обучения для классификации трафика. Классификатор случайного леса — это классический алгоритм машинного обучения, который строит несколько деревьев решений во время обучения и выводит шаблон класса для задачи классификации. Этот ансамблевый метод повышает надежность и точность модели, устраняя распространенную проблему переобучения в отдельных деревьях решений. Например, В.Ф. Тейлор и др. [9] использует классификатор случайного леса с функциями размера пакета для задач классификации. С другой стороны, Винай Джордж Биджу [18] использовал алгоритм машины опорных векторов (SVM) для классификации зашифрованного трафика. SVM также является классическим алгоритмом машинного обучения, который работает путем поиска оптимальной гиперплоскости, которая максимизирует разницу между различными классами в пространстве признаков и исключительно хорошо работает в многомерных пространствах. Хотя эти методы не требуют проверки содержимого пакетов и подходят для классификации зашифрованного трафика, они также сталкиваются с ограничениями. Простые статистические признаки часто не соответствуют точной классификации, а получение сложных признаков требует значительных усилий и ресурсов. Наша методология опирается на простые статистические характеристики для получения характеристик сигнатур пути, повышая информативность признаков без сложного проектирования.

Модели глубокого обучения. Появление глубокого обучения существенно повлияло на использование этих моделей при классификации зашифрованного трафика. DF [19] впервые использует сверточную нейронную сеть (CNN) для извлечения признаков из зашифрованного трафика, а FS-NET [20] использует рекурсивную нейронную сеть для классификации зашифрованного трафика. DeepPacket [6] и TSCRNN [21] строят модели, сочетая извлечение признаков и построение классификаторов.

Machine Translated by Google

автоматическая обработка извлечения и классификации признаков при вводе в модель полезных данных исходных пакетов. Чжан и др. [22] использовали графовые нейронные сети для извлечения функций уровня пакетов для классификации трафика. MT-FlowFormer [23] использует полуконтролируемый преобразователь потока для завершения классификации зашифрованного потока. Лин и др. [16] фиксировали надежные и общие ассоциации из необработанных пакетов трафика с использованием больших предварительно обученных моделей, таких как двунаправленное представление кодировщиков от трансформаторов (BERT), добившись отличных результатов классификации. Тем не менее, для оптимальной производительности модели глубокого обучения обычно требуют сложной сетевой архитектуры и обширных наборов данных. Напротив, наша модель спроектирована так, чтобы быть эффективной и требует меньше ресурсов.

2.2. Особенности сигнатуры

пути Сигнатуры пути были первоначально предложены Ченом [11] в качестве основного понятия теории неровных путей, где «грубые» относятся к непрерывным путям, которые демонстрируют резкие колебания в неровных путях. Эти пути непрерывны, но не везде дифференцируемы, а непрерывные последовательности размеров пакетов в зашифрованном трафике можно рассматривать как грубые пути. В последние годы сигнатуры путей вновь привлекли внимание исследователей и успешно применяются в областях машинного обучения, распознавания образов и анализа данных. Грэм [24] использовал функции подписи пути на основе скользящего окна с CNN для распознавания рукописных символов, добившись отличных результатов. Л. Г. Гюрко [25] использовал сигнатуры путей для извлечения признаков из потоков финансовых данных для классификации финансовых данных. И. П. Аррибас [26] применил сигнатуры путей в области медицины для дифференциации психических расстройств. В [12] была представлена новая методология описания сигнатур с помощью сигнатур пути, нормализованных по длине. В исследовании также были изучены особенности сигнатур путей, которые использовались в сочетании с алгоритмом случайного леса для классификации зашифрованного трафика. В целом, эти приложения демонстрируют эффективность извлечения признаков с использованием функций сигнатуры пути и ценность представления векторов, насыщенных информацией. В следующем разделе представлены определения сигнатур пути, их геометрическая интерпретация низкого порядка и соответствующие свойства. Дополнительные математические доказательства можно найти в [11].

2.2.1. Определение сигнатур пути

Сигнатура пути — это преобразующая функция отображения, которая преобразует необработанные данные пути в последовательность действительных чисел [12]. Каждое число в этой последовательности получается уникальным образом через точки данных вдоль исходного пути, отражая отдельные геометрические характеристики этого пути. Для пути X в d-мерном евклидовом пространстве Rd, определяемом как непрерывное отображение, интервала [a, b] → Rd, параметр, t [a, b] определяет траекторию внутри этого d-мерного евклидова пространства:

$$X_t = X\left(X^1_t, X^2_t, \ldots, X^d_t\right), \quad (1)$$

, где X_t представляет собой произвольный непрерывный путь в d-м измерении. Подпись пути , состоящая из подписи нулевого порядка (1) представляет собой бесконечную последовательность пути X: [a, b] R и различных членов сигнатуры порядка (также называемых повторными интегралами):

$$S(X)_{a,b} = \left(1, S(X)^1_{a,b}, \ldots, S(X)^d_{a,b}, S(X)^{1,1}_{a,b}, S(X)^{1,2}_{a,b}, \ldots, S(X)^{d,d}_{a,b} \ldots\right) \quad (2)$$

Для d-мерного пути X его k-кратный повторный интеграл можно выразить следующим образом:

$$S(X)^{i_1,\ldots,i_k}_{a,b} = \int_{a<t_k<t} \cdots \int_{a<t_1<t_2} dX^{i_1}_{t_1} \cdots dX^{i_k}_{t_k}. \quad (3)$$

В формуле (3) t1, t2,. . ., tk относятся к переменным интегрирования, отличным от t. Это различие сделано для того, чтобы избежать путаницы с параметром пути t во время многократного интегрирования . Если взять в качестве примера двумерный путь $X_t = X^1_t, X^2_t$, показанный на рисунке 1 , его характеристики сигнатуры пути третьего порядка будут следующими: {57.17, 54.67, 23.67, 30.33, 44.17, 34.33, 22.83, 20.83}.

Machine Translated by Google

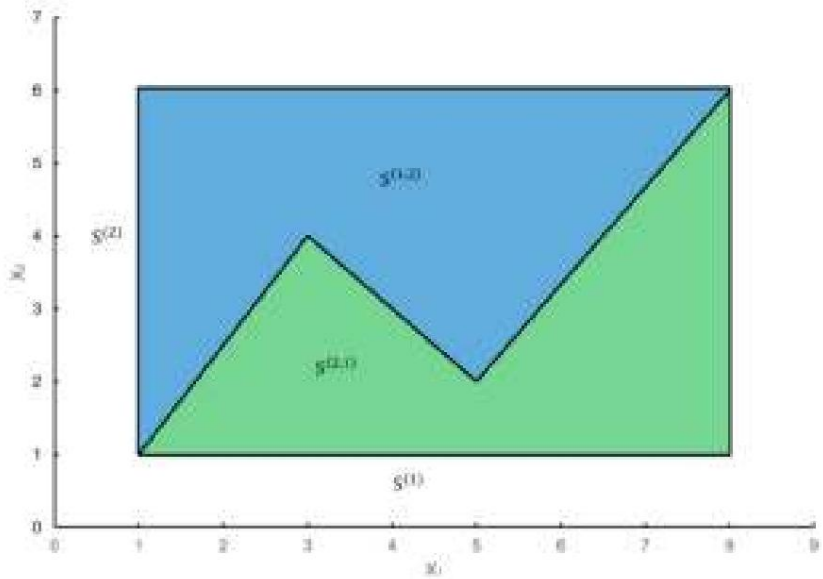


Рисунок 1. Геометрическая интерпретация двумерного пути.

2.2.2. Геометрическая интерпретация сигнатур

пути. Для дальнейшего объяснения геометрической интерпретации особенностей сигнатуры пути мы используем в качестве примера двумерный путь Xt = X , где t [0,4], (a = 0,b = 4). Путь определяется X = {1, 4, 2, 6}. Важно отметить, что t — непрерывный параметр, для удобства мы выбрали четыре дискретные точки для построения пути. Форма пути показана на рисунке 1. Вычисленные сигнатуры пути нулевого, первого и второго порядка следующие:

$$C = 1, C(1) = S(), C(2) = 1, (), (), (),

Machine Translated by Google

конкретные типы приложений остаются стабильными, на них не влияет разнообразие параметров. Использование этого атрибута позволяет устранить несоответствия, вызванные различными изменениями параметров трафика в одной и той же категории, подчеркивая решающее преимущество использования функций сигнатуры пути.

• Инвариантность размерности: Размерность особенностей сигнатуры пути определяется исключительно выбранной глубиной усечения, независимо от фактической длины пути [27]. Например , рассматривая ранее обсуждавшийся двумерный путь Xt = X 1 t [0,4], (a = 0,b = 4) с X = {1, 3, 5, 8} и X = {1, 4, 2, 6}, усечение до одного измерения приводит к длине сигнатуры пути, равной 3, а усечение до двух измерений дает длину X_T^1, X_T^2, 7. Такое извлечение признаков фиксированной длины из путей различной длины значительно упрощает процесс извлечения признаков. , особенно для моделей, требующих функций ввода фиксированной длины.

2.3. Модель ЛСТМ

Сети LSTM, подкласс рекуррентных нейронных сетей (RNN), отличаются своей способностью поддерживать долгосрочные зависимости [28]. Шен и др. использовал LSTM для анализа коэффициента аэродинамического сопротивления кузовов автомобилей [29]. Ссылка. [30] применили LSTM для прогнозирования уровня грунтовых вод, продемонстрировав способность модели обрабатывать изменчивые данные об окружающей среде и обеспечивать точные прогнозы. Ссылка. [31] исследовали перспективы применения LSTM в области автономного вождения, подчеркнув его потенциал для обработки и обучения на огромных объемах последовательных данных. Эти разнообразные приложения подчеркивают универсальность LSTM в различных областях и его надежные возможности обработки данных временных рядов. Как показано на рисунке 2, блоки LSTM отличаются от традиционных RNN за счет интеграции трех отдельных «ворот», которые в совокупности расширяют возможности управления памятью модели [32]: • Входной шлюз: контролирует поток входящих данных в ячейку, принимая решения. какой объем новой информации следует хранить;

• Выходной вентиль: определяет степень, в которой текущее состояние ячейки влияет на другие части сети, контролирующие выходной поток;

• «Ворота забыть»: регулирует самоповторяющиеся соединения ячейки с помощью сигмовидных функций для масштабирования значений от 0 до 1, определяя, какая информация отбрасывается, а какая сохраняется.

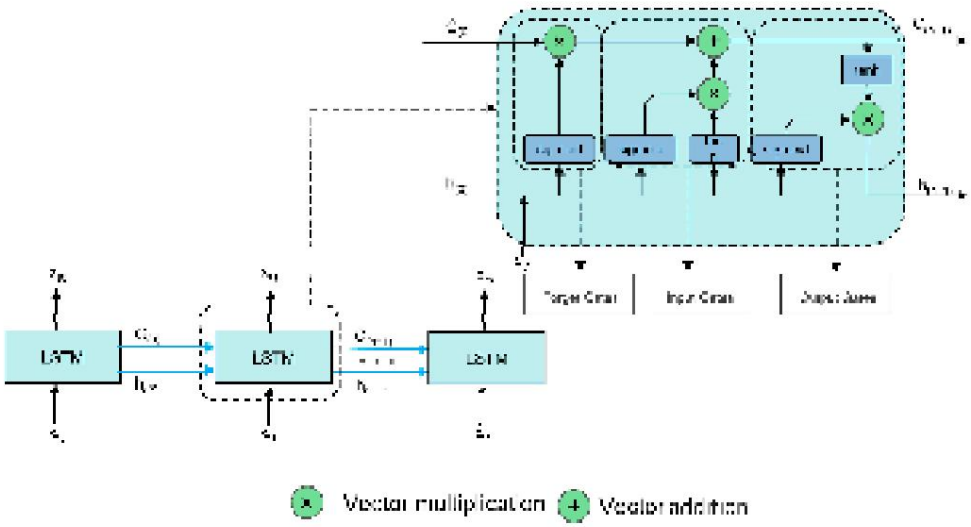


Рисунок 2. Архитектура модели LSTM.

Как типичная модель временных рядов, LSTM эффективно обрабатывает функции сигнатур пути внутри последовательностей. Известный своей способностью фиксировать долгосрочные зависимости в последовательных данных [29–31], LSTM особенно хорошо подходит для этой задачи из-за присущей данным сетевого трафика последовательной природы . По сравнению с другими известными моделями глубокого обучения, такими как CNN или Transformers, LSTM имеет следующие преимущества в этой задаче классификации зашифрованного трафика:

ффи

ффи

тт

Machine Translated by Google

• Лучшая обработка временных зависимостей: хотя CNN превосходно фиксируют пространственные иерархии и закономерности, они обычно более эффективны при обработке данных с пространственными корреляциями (например, изображений) и не могут фиксировать временные зависимости
так же эффективно, как LSTM [33]. • Более простая и непосредственная обработка данных. Применение CNN к данным последовательности требует преобразования временных рядов в формат, подходящий для операций свертки, что требует более сложной обработки данных. Напротив, LSTM позволяет напрямую вводить характеристики временных рядов в модель.
• Снижение требований к вычислительным ресурсам. Модели-трансформеры в последнее время завоевали популярность благодаря своему успеху в задачах последовательного преобразования, особенно при обработке естественного языка. Трансформаторы используют механизмы самообслуживания для фиксации зависимостей между различными частями входной последовательности, независимо от их расстояния [3
Хотя это полезно для учета глобальных зависимостей, для практического обучения Transformers обычно требуются значительные вычислительные ресурсы и большие наборы данных.
Сети LSTM, использующие собственные механизмы вентилирования для обработки последовательных данных, требуют меньше вычислительных ресурсов для обучения и развертывания.

3. Методика 3.1.

Обзор подхода

Технологическая схема нашего метода классификации зашифрованного трафика показана на рисунке 3. Первоначально, на этапе предварительной обработки данных, мы отделяем файлы pcap от общедоступных наборов данных, используя критерии пяти кортежей, впоследствии отбрасывая пакеты, в которых отсутствует допустимая полезная нагрузка. Этот шаг обеспечивает создание потоков данных одинаковой длины. Переходя к этапу извлечения признаков, мы получаем атрибуты размера, направления и времени прибытия из этих однородных потоков данных для установления путей трафика. К этим путям трафика применяется процесс преобразования , чтобы обогатить их дополнительной информацией. Следующий этап включает в себя использование методов многомасштабного накопления для выделения функций сигнатур путей из этих расширенных путей трафика. После извлечения этих функций мы применяем методы балансировки данных для получения хорошо сбалансированного обучающего набора. Эти оптимизированные функции сигнатур пути вводятся в модель LSTM и полностью связанные слои для выполнения окончательной задачи классификации. В следующих подразделах каждый этап этой методологии будет рассмотрен более подробно.

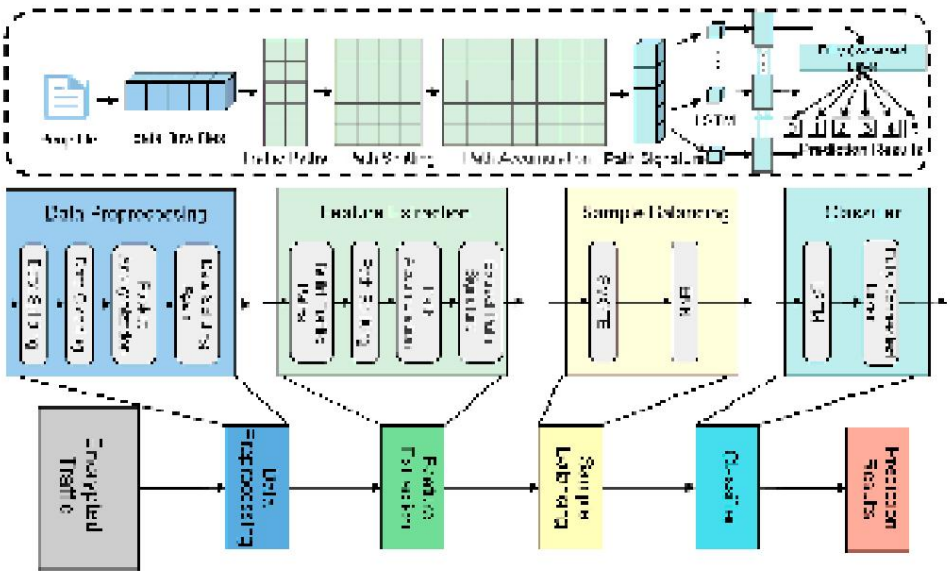


Рисунок 3. Общая структура нашего подхода.

Machine Translated by Google

3.2. Предварительная

обработка данных. При обработке необработанных файлов pcap из набора данных мы используем SplitCap для разделения их на отдельные файлы сеанса, используя пять критериев: IP-адрес источника, порт источника, IP-адрес назначения, порт назначения и протокол транспортного уровня [3]. . Следующие шаги включают удаление сеансов, лишенных данных и дубликатов. Чтобы снизить риск переоснащения модели этим набором данных, мы анонимизируем MAC- и IP-адреса в пакетных данных, заменяя их нулями [6]. После этого мы дополнительно уточняем файлы сеансов, чтобы гарантировать, что каждый содержит ровно 24 пакета данных — параметр, значение которого будет подробно рассмотрено в следующем экспериментальном разделе. Сеансы, не соответствующие критерию 24 пакетов данных, исключаются, в результате чего создаются образцы файлов, соответствующие нашим предварительным условиям эксперимента.

3.3. Построение путей

трафика. Приведя файлы сеансов к единообразной длине, мы классифицируем пакеты данных , передаваемые от клиента к серверу, с положительными значениями, а пакеты, движущиеся в обратном направлении, обозначаем отрицательными значениями. Результатом этого процесса является создание последовательности длин пакетов данных L = (l1, l2, ..., ln), где l представляет длину каждого пакета данных [26]. Кроме того, мы фиксируем время прибытия каждого пакета данных. Чтобы синхронизировать эти временные метки, мы нормализуем данные для составления временной последовательности для пакетов, обозначенной как T = (t1, t2, ..., tn), где t представляет время прибытия для каждого пакета данных.

3.4. Преобразование путей

трафика После извлечения двумерных характеристик сигнатуры пути, которые инкапсулируют время и длину пакета данных, мы понимаем, что представление межсеансовой информации остается неполным. Чтобы улучшить использование межсессионных характеристик, мы применяем методологию, вдохновленную работой Сюй и др. [12]. Этот подход предполагает преобразование исходных путей трафика для раскрытия дополнительных слоев информации, тем самым облегчая более подробный и информативный анализ.

3.4.1. Разделение путей

В нашем методе термин «клиент» относится к пользователям, которым требуются интернет-услуги. Использование клиентами различных сервисов в Интернете приводит к генерации различных типов зашифрованного трафика. Например, когда пользователь загружает или скачивает файлы в Интернете, создается трафик передачи файлов. Аналогично, когда пользователь использует потоковые сервисы, например, слушает музыку или смотрит видео, генерируется потоковый трафик. Чтобы более точно отобразить динамику потока трафика между клиентом и сервером, мы разделили последовательность длин пакетов данных на две отдельные последовательности. Одна последовательность фиксирует передачу данных от клиента к серверу, а другая детализирует передачу от сервера к клиенту [12]. Ниже это разделение формализуется следующим образом.

Для передачи клиент-сервер:

$$C = (c1, c2, \ldots, cn),\ ci = \begin{matrix} 0, \text{ли} < 0 \\ \text{ли} , \text{ли} > 0.. \end{matrix} \quad (5)$$

И наоборот, для передачи от сервера к клиенту:

$$S = (s1,s2, \ldots ,sn),si = \begin{matrix} \text{ли} , \text{ли} < 0 \\ 0, \text{ли} > 0.. \end{matrix} \quad (6)$$

Присваивая размеры пакетов данных в противоположном направлении передачи равными нулю, мы подчеркиваем отличительные свойства потока, присущие каждому направлению, закладывая фундаментальную основу для дальнейших преобразований пути.

3.4.2. Накопление пути

Функции накопления играют решающую роль в классификации зашифрованного трафика [35], прежде всего потому, что размер полезной нагрузки имеет тенденцию оставаться постоянным для одного и того же типа трафика.

Machine Translated by Google

услуг, что приводит к аналогичным закономерностям накопления. Чтобы использовать эту согласованность для раскрытия информации в зашифрованном трафике, мы определяем функции накопления как следует ниже.

Прямое накопление длин пакетов:

$$л' = л \quad {}'_{1,2,\cdots,} \quad {}'_{л_н,} \quad {}'_{л_я} \quad "=" \quad \overset{к}{\underset{я=1}{}} \quad {}_{ли,} \tag{7}$$

Для передач, направляемых клиентом:

$$С' = с \quad {}'_{1,2,} \quad {}'_{с_в,} \quad \cdots, с' \quad с' \quad "=" \quad \overset{к}{\underset{я=1}{}} \quad {}_{ци,} \tag{8}$$

Для передач, направляемых сервером:

$$С' = с \quad {}'_{1,} \quad {}'_{с2,} \quad \cdots, с' \quad {}_{н,} \quad с' \quad {}_{я} \quad "=" \quad \overset{к}{\underset{я=1}{}} \quad {}_{си.} \tag{9}$$

Таким образом, мы достигаем комплексного представления пути движения через следующее:

$$X \; {}'_{т,} с, s, {}'_{,} с' \; {}'_{,} с' \; {}'_{,} Т \quad = Л, С, С, Л \quad ', С', С', Т \quad , t \quad [0, n]. \tag{10}$$

Это представление объединяет последовательности длин исходных пакетов данных и их накопленные версии вместе с последовательностью времени прибытия, предлагая многогранное представление о транспортный поток для адекватной классификации. Выбор конкретного пути трафика L = [194, 83, 32,53,86] и T = [0,2,5,8,10] в качестве примера соответствующий преобразованный путь трафика показан на рисунке. Таблица 1.

Таблица 1. Примеры семи различных путей трафика.

			$Икс_{5}^{л,к,с,л',с,с',т}$			
л	С	С	л'	С'	С'	Т
194	194	0	194	194	0	0
83	0	83	111	194	83	2
32	0	32	79	194	115	5
53	53	0	132	247	115	8
86	86	0	218	333	115	10

3.5. Извлечение подписи пути

После преобразования путей трафика наш следующий шаг включает в себя извлечение сигналов пути. особенности. Мы определяем признак пути f для каждого преобразованного пути следующим образом:

$$ft = \varphi(X \quad {}^{л,к,с,л',с',с',т}_{т}), t \quad [0, n]. \tag{11}$$

Здесь φ(X) обозначает функцию, ответственную за извлечение сигнатуры пути, а ft представляет собой сигнатуру пути первых t пакетов данных. Важно отметить, что размерность особенностей сигнатуры пути зависит исключительно от выбранного пути глубина усечения, независимая от фактической длины каждого пути. В рамках «многие к одному» В модели LSTM принято использовать характеристики последнего временного шага в качестве входных данных для окончательного результата. классификация.

Использование уникальных свойств сигнатуры пути и модели LSTM, как изображенный на рисунке 4, мы представляем новый метод извлечения признаков многомасштабного накопления. подход. Этот метод повышает информационную ценность, извлекаемую на последнем временном шаге. вход в модель. Подробно описаны формулировка и структура входной последовательности признаков. следующее:

$$F = (f2, f3, \ldots, fn). \tag{12}$$

Machine Translated by Google

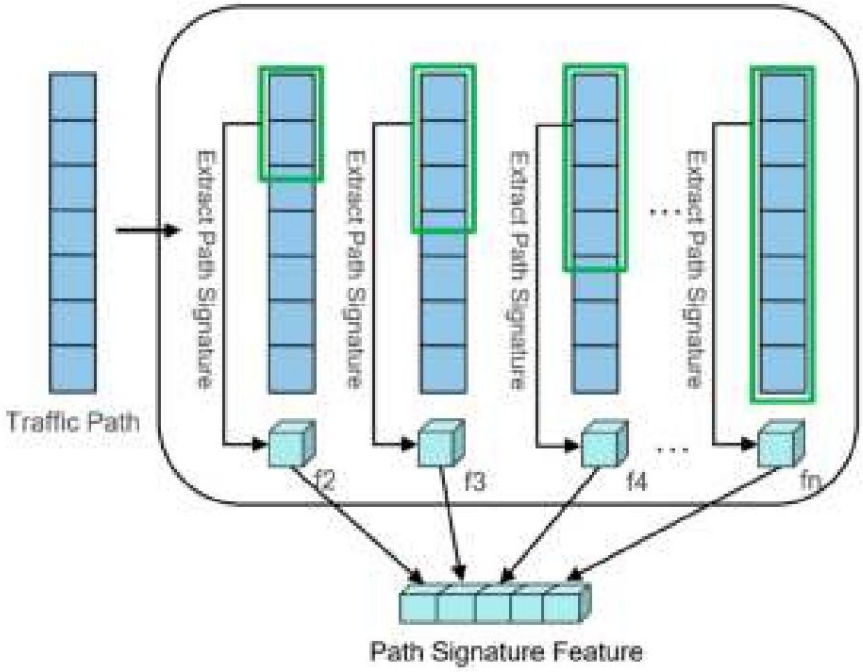


Рисунок 4. Метод выделения признаков многомасштабного накопления.

Требование n 2 обусловлено тем, что сигнатура пути, основанная на менее чем двух пакетах данных, не имеет существенной значимости. Когда n = 1, сигнатура пути сводится к простому отражению длины и времени прибытия одного пакета данных, не удовлетворяя критериям нашей экспериментальной установки. Следовательно, мы компилируем последовательность сигнатур пути F, где каждый элемент содержит характеристики сигнатур пути первых n пакетов данных, и все сигнатуры пути имеют одинаковую размерность.

3.6. Балансировка

данных Встреча несбалансированных данных в разных классификациях является распространенной проблемой при сборе экспериментальных данных из наборов данных. Внедрение методов балансировки данных имеет решающее значение для повышения эффективности учебных задач. В этом разделе описываются стратегии, которые мы приняли для достижения баланса данных.

3.6.1. SMOTE (метод синтетической избыточной выборки меньшинства)

Представленный Чавлой [36], SMOTE является ведущей методологией передискретизации для устранения проблем дисбаланса данных. Он дополняет набор данных, создавая новые синтетические экземпляры класса меньшинства, полученные на основе существующих наблюдений. В пространстве функций SMOTE создает синтетические образцы, напоминающие класс меньшинства. В частности, для каждой выборки класса меньшинства SMOTE идентифицирует n ближайших соседей, используя евклидово расстояние. После этого случайным образом выбирается один сосед, генерируя вектор, связывающий этого соседа и выборку. Затем формируется синтетический экземпляр путем добавления случайной части этого вектора к исходной выборке. Эта процедура повторяется до тех пор, пока не будет достигнута целевая частота передискретизации [36]. Формула алгоритма SMOTE имеет следующий вид:

$$n = s + d * s - s^{p}, \quad 0 \quad d \quad 1. \tag{13}$$

где s символизирует выборки класса меньшинства, s обозначает соседей, определяемых k ближайших соседей, d представляет собой случайно выбранное значение в диапазоне [0,1], а n представляет собой сгенерированную синтетическую выборку. Как показано на рисунке 5, две точки разного цвета представляют два разных типа образцов. Три ближайших соседа образца s — это s1 R, s2 R и s3 Используя метод SMOTE, для выборки s случайным образом выбирается сосед s1 , и путем расчета получается новая выборка n.

Machine Translated by Google

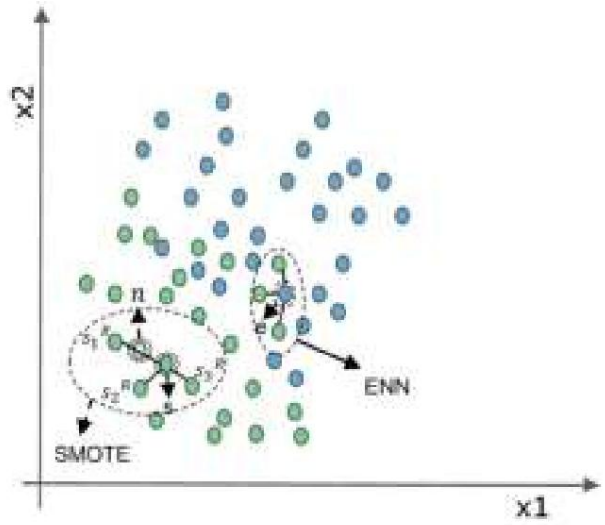


Рисунок 5. Пример SMOTE и ENN.

3.6.2. ENN (отредактированные ближайшие соседи)

ENN использует алгоритм k ближайших соседей в качестве стратегии вторичной выборки [37]. Он направлен на сокращение набора данных путем удаления экземпляров, которые могут быть неправильно классифицированы. ENN вычисляет евклидово расстояние для каждого экземпляра, чтобы точно определить его k ближайших соседей (где k обычно устанавливается равным 3). Если большинство из этих соседей не имеют общего класса экземпляра, экземпляр и его соседи исключаются из набора данных [37]. Таким образом, ENN может эффективно уменьшить присутствие выборок большинства классов в обучающем наборе. Если взять в качестве примера образец e на рисунке 5 , то все его k соседей (установленных равными 3) принадлежат к другим классам. При использовании метода ENN все четыре образца будут удалены.

3.6.3. СМОТ-ЭНН

SMOTE-ENN объединяет сильные стороны SMOTE и ENN в единый подход к балансировке данных, как это было предложено Батистой и др. [38]. Первоначально набор данных подвергается усовершенствованию с помощью алгоритма SMOTE, который усиливает представление класса меньшинства за счет создания синтетической выборки. Как только будет создано достаточное количество выборок классов меньшинства , алгоритм ENN применяется для дальнейшего уточнения набора данных путем удаления экземпляров, близких к границе классификации. Такое последовательное применение SMOTE и ENN способствует более равномерному распределению набора выборок по классам [38].

3.7. Входные данные в модель LSTM

Последовательность сигнатур путей, обозначенная как F, служит входными данными для модели LSTM, создавая основу для процесса классификации. Конфигурация временных шагов внутри LSTM напрямую коррелирует с длиной исходного пути трафика, уменьшенной на единицу. Следовательно, признак fn+1 на n-м временном шаге инкапсулирует сигнатуру пути трафика, содержащего n + 1 пакет данных. Такое расположение гарантирует, что с каждым дополнительным временным шагом будет включено все больше информации, обогащая представление данных . Примечательно, что функция на последнем временном этапе объединяет сигнатуру всего пути трафика, воплощая наиболее полную доступную информацию.

Присущая модели LSTM способность памяти [28] играет ключевую роль, позволяя ffi на каждом временном этапе использовать и передавать информацию из предыдущих шагов. В результате заключительный временной шаг эффективно агрегирует информацию по всей последовательности, обходя потенциальные недостатки использования исключительно глобальных функций. Такая комплексная интеграция информации значительно повышает точность классификации модели.

Machine Translated by Google

3.8. Полностью связный слой

После обработки входной последовательности LSTM выходные данные последнего временного шага подвергается экстракции и затем направляется к полностью связанному слою, где попытка классификации реализована. Сущность этого слоя заключается в его способности к переходу данные, полученные с помощью LSTM, помещаются в область пространства меток. Операция в рамках полносвязный слой математически изображается следующим образом:

$$Y = Ш \times X + B. \quad (14)$$

где Y — выходной вектор, W — вес полносвязного слоя, X — входной вектор. вектор, полученный из последнего временного шага LSTM, а B — смещение.

3.9. Полный рабочий процесс

В этом разделе мы подробно объясним рабочий процесс предлагаемого нами метода. После захватывая сегмент трафика, мы сначала извлекаем две основные статистические характеристики длины пакета. и время прибытия пакета для формирования пути трафика. Затем мы разделяем последовательность длин пакетов на два пути, представляющие направление передачи пакета. Далее мы выполняем процесс накопления для получения трех путей, представляющих совокупную длину пакета передача инфекции. В результате получается семимерный путь движения.

Используя функцию признака сигнатуры пути, мы преобразуем семимерную путь трафика в многомерные функции сигнатур пути, которые затем передаются в LSTM. модель. Модель LSTM обрабатывает входную последовательность постепенно, шаг за шагом. На каждом временной шаг, входные функции взаимодействуют со скрытым состоянием и состоянием ячейки памяти из предыдущий временной шаг. Через ряд механизмов шлюзования скрытое состояние и память состояние ячейки обновляется на каждом временном шаге, что позволяет модели эффективно фиксировать временные изменения. зависимости в данных о трафике.

После обработки всей последовательности скрытое состояние модели LSTM на финальном этапе временной шаг выбирается в качестве выходных данных функции. Это скрытое состояние имеет интегрированную информацию из всей входной последовательности и представляет глобальные характеристики данных последовательности. Выбранное скрытое состояние затем вводится в полностью связный слой. Полностью подключенный слой с помощью ряда весов и функций активации дополнительно обрабатывает скрытые состояние для генерации распределения вероятностей для результата классификации. Таблица 2 иллюстрирует изменения формы на каждом этапе обработки зашифрованной последовательности трафика, содержащей 24 пакета.

Таблица 2. Изменение формы данных на каждом этапе обработки.

Этап	Описание	Форма
Необработанные данные	Захват зашифрованной последовательности трафика	—
Постройте пути трафика	Извлеките длину пакета и время прибытия	(24, 2)
Разделение пути	Разделить последовательность длин на два пути	(24, 4)
Накопление пути	Накопить последовательности трех длин	(24, 7)
Извлечение пути Подпись	Используйте функцию подписи пути для извлечения функции подписи пути	(24, 56)
LSTM-вход	Используйте функции подписи пути в качестве входных данных	(1, 24, 56)
Скрытые модули LSTM	для скрытых единиц LSTM 64 входных данных процесса уровня LSTM. последовательность, возвращающая скрытые состояния для все шаги времени	(1, 24, 64)
Последнее скрытое состояние	Выберите скрытое состояние последнего временного шага в модель LSTM и введите полностью связный слой	(1, 64)
Выход полностью связанного слоя	Полносвязный слой выводит распределение вероятностей результат классификации	(1, 6)

Machine Translated by Google

4. Дискуссия

В этом разделе мы подробно остановимся на различных аспектах нашей экспериментальной структуры, проливая свет на различные аспекты, такие как экспериментальная среда, метрики оценки, используемые наборы данных, методы обработки данных и выбор гиперпараметров. Кроме того, чтобы установить эффективность и конкурентные преимущества нашей методологии, мы провели сравнительные эксперименты с преобладающими основными методами. Также были проведены исследования абляции, чтобы подчеркнуть вклад и влияние отдельных компонентов в рамках предложенной нами структуры.

4.1. Экспериментальная среда

Эксперименты проводились в операционной системе Windows 10 с Python версии 3.9. Мы реализовали модель LSTM, используя PyTorch версии 1.13.1. Мы установили начальную скорость обучения для обучения модели на 0,0001, а периоды обучения — на 500. Модель LSTM, используемая для этого эксперимента, имеет 64 скрытых измерения, а выходное измерение последующего полносвязного слоя равно 6, что соответствует классификации из шести классов. задача типов обслуживания зашифрованного трафика.

4.2. Метрики оценки

Для оценки эффективности метода мы использовали четыре оценочных показателя: точность (AC), прецизионность (PR), полноту (RC) и показатель F1 (F1). Истинно положительные (TP) — это правильно идентифицированные положительные образцы, ложноположительные (FP) — это отрицательные образцы, ошибочно классифицированные как положительные, истинно отрицательные (TN) — это правильно идентифицированные отрицательные образцы, а ложноотрицательные (FN) — это положительные образцы, ошибочно классифицированные как отрицательные.

$$\text{Точность} = \frac{\text{ТП} + \text{ТН}}{\text{ТП} + \text{ТН} + \text{ФП} + \text{ФН}}, \tag{15}$$

$$\text{Точность} = \frac{\text{ТП}}{\text{ТП} + \text{ФП},} \tag{16}$$

$$\text{Напомним} = \frac{\text{ТП}}{\text{ТП} + \text{ФН}}, \tag{17}$$

$$\text{F1-оценка} = \frac{2 \quad \text{Точность} \quad \text{Напомним}}{\text{Точность} + \text{отзыв}}. \tag{18}$$

Эти показатели дают комплексное представление об эффективности модели, отражая ее суть. точность, точность и чувствительность при классификации данных по соответствующим категориям.

4.3. Наборы данных

Чтобы всесторонне проверить производительность предложенного нами метода, мы использовали два общедоступных набора данных: набор данных ISCX VPN-nonVPN [39] и набор данных ISCX Tor-nonTor [40]. Эти наборы данных были опубликованы Университетом Нью-Брансуика.

Набор данных ISCX VPN-nonVPN состоит из двух частей: ISCX-VPN и ISCX-nonVPN. Набор данных ISCX-VPN содержит трафик, перехваченный в среде VPN для шести типов услуг: чат, электронная почта, передача файлов, потоковая передача, P2P и VoIP. Набор данных ISCX-nonVPN включает те же типы трафика, но захватывается вне среды VPN.

Аналогичным образом набор данных ISCX Tor-nonTor делится на ISCX-Tor и ISCX-nonTor. Набор данных ISCX-Tor собирает трафик через сеть Tor, что обеспечивает высокий уровень анонимности. Набор данных ISCX-nonTor содержит трафик, собранный вне среды Tor, с типами трафика, идентичными типам в наборе данных ISCX-VPN. В Таблице 3 представлены основные приложения, соответствующие каждому типу услуг, и дано их краткое описание.

Machine Translated by Google

Таблица 3. Подробности набора данных.

Услуга	Приложение	Описание
Чат	АИМ, ICQ, Скайп, Фейсбук, Hangouts	Трафик, генерируемый во время онлайн-чата коммуникация
Электронная почта	Электронная почта, Gmail	Трафик, генерируемый во время электронной почты передача инфекции
Передача файла	Skype SFTP, FTPS, SCP	Трафик, генерируемый во время загрузки файлов и загрузки
Потоковое вещание	Вимео, Ютуб, Нетфликс, Спотифай	Трафик, генерируемый при использовании потоковые приложения
P2P	uTorrent, БитТоррент	Трафик, генерируемый при совместном использовании торрента ресурсы с использованием P2P-программ
VoIP	Фейсбук, Скайп, Hangouts, VoipBuster	Трафик, генерируемый во время голосовых онлайн-звонков

После получения набора данных мы использовали длину пакета и время прибытия для построения наши особенности подписи пути. Длина пакета относится к размеру каждого пакета в байтах. поток данных. Поскольку передача трафика в сети обычно двунаправленная, мы добавили положительные и отрицательные знаки, обозначающие направление трафика: трафик от клиента к сервер был настроен как положительный, а трафик от сервера к клиенту был установлен как отрицательный. Это создало последовательность длин пакетов, которая представляет как направление передачи, так и размер загрузки пакета.

Время прибытия пакета относится ко времени получения каждого пакета, начиная с прием первого пакета и запись времени прибытия каждого последующего пакета. Это приводит к увеличению временного ряда, который представляет частоту передачи пакеты. Путем сегментации и накопления последовательности длин пакетов и нормализации время прибытия пакетов, мы получили окончательный путь трафика. Этот путь затем был введен в функция функции подписи пути для получения функций подписи пути трафика, необходимых для наш метод.

4.4. Балансировка данных

Чтобы устранить дисбаланс выборки, присущий набору данных ISCX VPN-nonVPN для При классификации зашифрованного трафика в нашей методологии использовался многогранный подход. обеспечить справедливое представительство различных типов транспортных услуг. Первоначально порог было установлено 5000 выборок, при этом категории, превышающие это число, имели случайный выбрано подмножество из 5000 образцов, тогда как используются категории, находящиеся ниже этого порога. все доступные образцы для максимального использования данных. После этого выбора категория с было выявлено наименьшее количество образцов, что стало эталоном единообразия. Порция соответствующее одной пятой этого минимального количества, было извлечено из каждой категории для собрать комплексный набор тестов, обеспечив сбалансированное представление. SMOTE-ENN был применен для устранения остаточного дисбаланса в наборе обучающих данных, синтезируя новые выборки для классов меньшинств и удаление неоднозначных выборок для улучшения качества набора данных. На рисунке 6 сравнивается количество образцов для каждой категории до и после балансировки. данные обучающего набора в наборе данных ISCX-VPN.

Machine Translated by Google

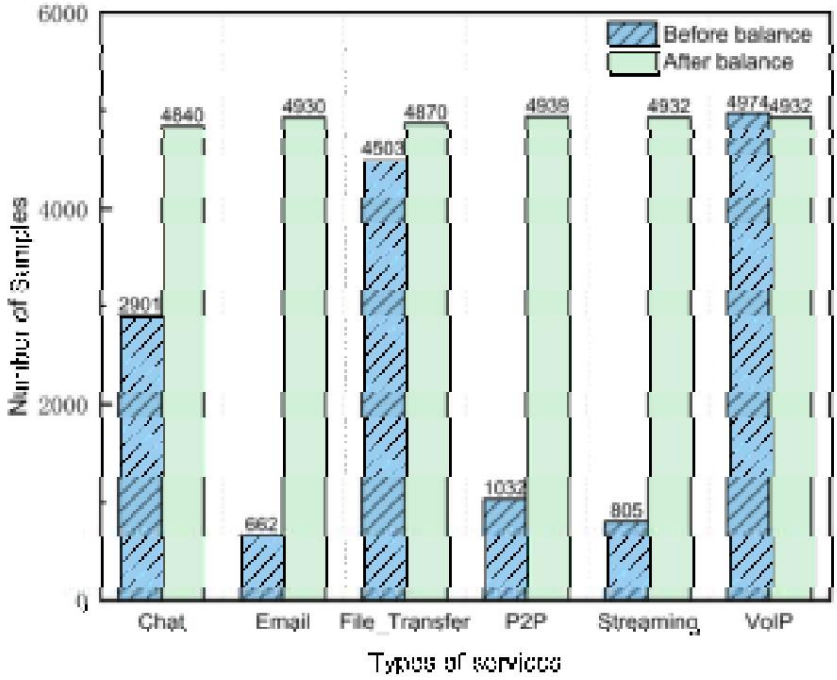


Рисунок 6. Количество образцов.

4.5. Выбор параметра

В наших экспериментах значения двух важнейших гиперпараметров, а именно длины путей трафика и размеров сигнатур путей, играли значительную роль и были тесно связаны с производительностью модели. Они напрямую сформировали способность модели учиться на данных и интерпретировать их, влияя на глубину и широту анализа.

Длина пути трафика определяет входные данные для каждого экземпляра анализа, влияя на понимание модели непрерывности сеанса. Между тем, размерность сигнатур путей влияет на сложность и детализацию функций, извлекаемых из путей трафика, балансируя между информационным богатством и вычислительной осуществимостью. В этом разделе мы подробно опишем процесс получения этих гиперпараметров.

Оптимальная длина пути трафика продиктована необходимостью включения достаточного количества последовательных пакетов данных с допустимой полезной нагрузкой в файл сеанса, который составляет основу исходного пути трафика. Слишком краткая длина пути может привести к тому, что файлы ffi сеанса будут лишены адекватных данных для эффективных задач классификации трафика. И наоборот, слишком длинные пути трафика могут неоправданно увеличить требования к сбору данных без пропорционального повышения точности классификации [41]. Определение оптимальной длины путей движения транспорта было жизненно важным в качестве основополагающего элемента в нашем экспериментальном проекте.

Классифицируя типы сервисов для зашифрованного трафика в условиях VPN, мы стремились определить оптимальную длину пути. Как показано на рисунке 7, когда длина пути была относительно короткой, точность классификации колебалась и постепенно улучшалась с увеличением длины пути для любого заданного измерения. Когда длина пути достигла 20, модели различных размерностей достигли относительно хороших характеристик классификации. Среди них двумерная модель пути достигла наибольшей точности классификации (94,74%), когда длина пути составляла 24. В дальнейшем, по мере увеличения длины пути, точность классификации существенно не улучшалась, а в некоторых измерениях даже снижалась. Следовательно, мы выбрали 24 в качестве оптимального параметра длины пути.

Более того, выбор размеров сигнатуры пути не менее важен. Сигнатуры путей более высокого порядка по своей сути инкапсулируют больше информации, чем их аналоги более низкого порядка. Тем не менее, увеличение размеров сигнатуры пути ускоряет экспоненциальный рост размеров объектов [11]. Например, сигнатура второго порядка в семимерном маршруте трафика будет иметь размерность 56, увеличивающуюся до 399 для третьего порядка и 2800 для четвертого порядка. Эта тенденция подчеркивает необходимость баланса

Machine Translated by Google

между информационным выигрышем от увеличения размеров сигнатуры пути и результирующим вычислительная сложность.

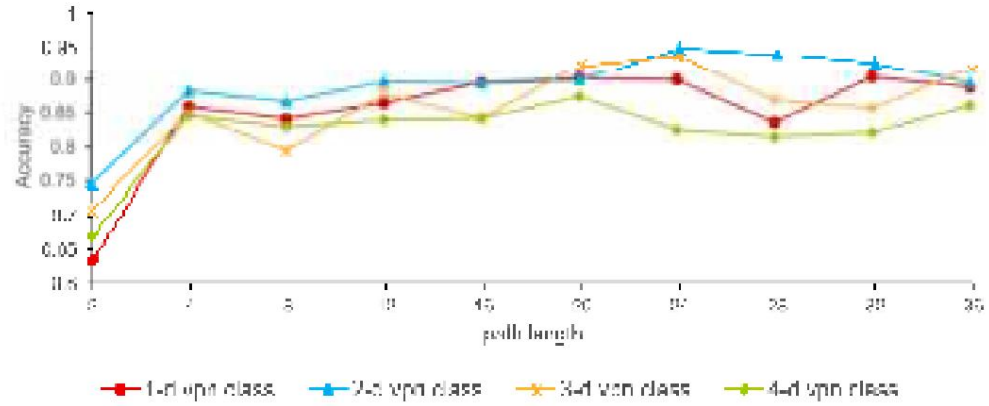


Рисунок 7. Влияние длины и размера пакета (точность).

Как показано на рисунке 7, согласно нашим экспериментам, увеличение размеров не привело к увеличению размеров. значительно улучшить производительность классификации. Произошло небольшое снижение производительности при увеличении размеров до четырех. Мы выбрали двумерный путь сигнатура как оптимальный параметр с учетом производительности классификации и эффективности вычислений.

4.6. Анализ чувствительности

Чтобы оценить надежность нашего метода, мы провели анализ чувствительности. Анализ чувствительности — это метод, используемый для оценки реакции модели на изменения параметров. Мы наблюдали влияние на производительность модели, изменяя скорость обучения и параметры регуляризации во время обучения модели. В частности, при оценке воздействия скорость обучения мы установили параметр регуляризации L2 равным 0,0001, и, оценивая Влияние параметра регуляризации, мы установили скорость обучения 0,0001. Экспериментальный результаты показаны в таблицах 4 и 5.

Таблица 4. Точность классификации при разных скоростях обучения.

Скорость обучения	0,1	0,01	0,001	0,0001	0,00001
Точность	0,8596	0,8918	0,9152	0,9474	0,9298

Таблица 5. Точность классификации при различных параметрах регуляризации.

Параметры регуляризации	0,01	0,001	0,0001	0,00005	0,00001
Точность	0,8099	0,9006	0,9474	0,921	0,9152

Можно заметить, что по мере изменения скорости обучения и параметров регуляризации, Точность классификации предложенного нами метода оставалась стабильно высокой, несмотря на некоторые колебания. Это указывает на высокую надежность нашего метода. Способность поддерживать высокий уровень точности в различных условиях демонстрирует надежность и стабильность нашего подхода.

4.7. Исследование абляции

Чтобы подчеркнуть вклад каждого компонента в предлагаемом нами методе классификации зашифрованного трафика мы провели серию исследований по абляции. Эти исследования были предназначенный для индивидуальной оценки влияния функции сигнатуры пути, многомасштабный метод извлечения совокупных признаков и модель LSTM для определения общей производительности.

Machine Translated by Google

Удаление функций сигнатуры пути. В этой части исследования мы больше не извлекаем функции сигнатуры пути из путей трафика. Вместо этого необработанные семимерные образцы объектов были переданы непосредственно в модель LSTM, минуя этап обработки сигнатуры пути. Все остальные условия эксперимента поддерживались с использованием установленного метода.

Абляция метода извлечения признаков: при оценке роли нашего многомасштабного метода извлечения совокупных признаков мы отказались от этого продвинутого подхода в пользу классического метода, на который ссылается Сюй и др. [12]. Характеристики, полученные с помощью этого традиционного процесса извлечения, затем были введены в модель LSTM для классификации. Опять же, эта корректировка была сделана при сохранении всех остальных экспериментальных параметров неизменными.

Абляция модели LSTM: Чтобы изучить конкретный вклад модели LSTM, мы заменили ее алгоритмом случайного леса для задачи классификации, используя настройки параметров, соответствующие тем, которые были предложены Сюй [12]. Эта замена позволила нам напрямую сравнить эффективность LSTM с альтернативным, хорошо зарекомендовавшим себя методом классификации.

Результаты этих экспериментов по абляции, как показано на рисунке 8, демонстрируют значительные различия ff при сопоставлении с результатами, достигнутыми с использованием всего нашего метода. Такие различия подчеркивают критическую роль каждого компонента в нашей методологии, подтверждая их коллективную необходимость и эффективность в повышении точности классификации.

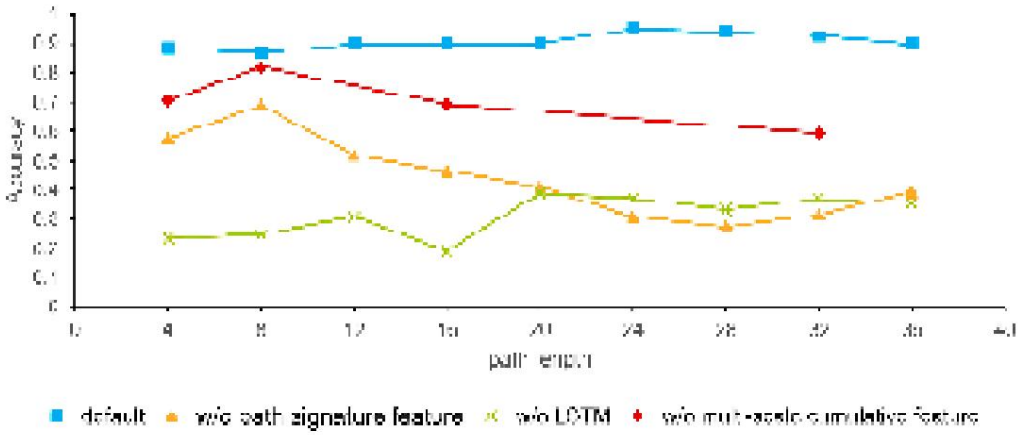


Рисунок 8. Результаты исследований абляции.

Благодаря этим исследованиям абляции были продемонстрированы интегральная ценность и синергетический эффект функций сигнатуры пути, нового метода извлечения признаков и модели LSTM , что подтвердило их незаменимый вклад в превосходную производительность нашего метода.

4.8. Сравнительные эксперименты

4.8.1. Сравнительные методы

Чтобы убедиться в эффективности предложенного нами метода классификации зашифрованного трафика, он был сопоставлен с 11 современными основными подходами. Эти методологии охватывают следующее: 1. Создание отпечатков пальцев: FlowPrint [17]; ff Статистические методы: AppScanner [9], CUMUL [35], K-FP (K-Fingerprinting) [42], 2. GRAIN [43] и ETC-PS [12];

3. Модели глубокого обучения: FS-net [20], DF [19], MT-FlowFormer [23], GraphDApp [44] и ET-BERT (поток) [16].

Примечательно, что предлагаемый нами метод использует исключительно функции зашифрованного трафика на уровне потока. Справедливости ради мы решили сравнить его с ET-BERT (поток), который также полагается исключительно на функции уровня потока, а не с ET-BERT (пакет), который обеспечивает лучшую производительность классификации, но требует функций уровня пакета.

Machine Translated by Google

4.8.2. Результаты эксперимента

Мы оценили эффективность различных методов, в том числе и нашей, на ISCX. Наборы данных VPN-nonVPN и ISCX Tor-nonTor, как показано в таблицах 6 и 7. В таблице 6 представлены эффективность классификации типов обслуживания трафика в средах VPN и без VPN, а в Таблице 7 показана эффективность классификации в средах Tor и без Tor. Наш метод продемонстрировал превосходную и конкурентоспособную эффективность по большинству показателей. по сравнению с большинством эталонных методов.

Таблица 6. Эффективность классификации различных методов для набора данных ISCX VPN-nonVPN.

Набор данных	ISCX-VPN				ISCX-без VPN			
Метод	Точность	Точность	Отзывать	F1-Оценка	Точность	Точность	Отзывать	F1-Оценка
ФлоуПринт [17]	0,8538	0,7451	0,7917	0,7566	0,6944	0,7073	0,7310	0,7131
AppScanner [9]	0,8889	0,8679	0,8851	0,8722	0,7576	0,7594	0,7465	0,7486
КУМУЛ [35]	0,7661	0,7531	0,7852	0,7644	0,6187	0,5941	0,5971	0,5897
К-ФП [42]	0,8713	0,8750	0,8748	0,8747	0,7551	0,7478	0,7354	0,7387
ЗЕРНО [44]	0,8129	0,8077	0,8109	0,8027	0,6667	0,6532	0,6664	0,6567
ЭТЦ-ПС [12]	0,8889	0,8803	0,8937	0,8851	0,7273	0,7414	0,7133	0,7208
ФС-сеть [20]	0,9298	0,9263	0,9211	0,9234	0,7626	0,7685	0,7534	0,7355
ДФ [19]	0,8012	0,7799	0,8152	0,7921	0,6742	0,6857	0,6717	0,6701
МТ-ФлоуФормер [23]	0,9327	0,9152	0,9243	0,9193	0,8549	0,8473	0,8268	0,8344
ГрафДапп [44]	0,6491	0,5668	0,6103	0,5740	0,4495	0,4230	0,3647	0,3614
ET-BERT (поток) [16]	0,9532	0,9436	0,9507	0,9463	0,9167	0,9245	0,9229	0,9235
Предложенный	0,9474	0,9480	0,9474	0,9472	0,9053	0,9064	0,9053	0,9050

Таблица 7. Эффективность классификации различных методов для набора данных ISCX Tor-nonTor.

Набор данных	ISCX-Top				ISCX-nonTor			
Метод	Точность	Точность	Отзывать	F1-Оценка	Точность	Точность	Отзывать	F1-Оценка
ФлоуПринт [17]	0,2400	0,0300	0,1250	0,0484	0,5243	0,7590	0,6074	0,6153
AppScanner [9]	0,7543	0,6629	0,6042	0,6163	0,9153	0,8435	0,814	0,8273
КУМУЛ [35]	0,6686	0,5349	0,4899	0,4997	0,8605	0,8143	0,7393	0,7627
К-ФП [42]	0,7771	0,7417	0,6209	0,6313	0,8741	0,8653	0,7792	0,8167
ЗЕРНО [44]	0,6914	0,5253	0,5346	0,5234	0,7895	0,6714	0,6615	0,6613
ЭТЦ-ПС [12]	0,7486	0,6811	0,5929	0,6033	0,9155	0,8710	0,8311	0,8486
ФС-сеть [20]	0,8286	0,7487	0,7197	0,7242	0,9278	0,8368	0,8254	0,8285
ДФ [19]	0,6514	0,4803	0,4767	0,4719	0,8568	0,8003	0,7415	0,7590
МТ-ФлоуФормер [23]	0,8750	0,8252	0,8217	0,8220	0,8941	0,8742	0,8651	0,8670
ГрафДапп [44]	0,4286	0,2557	0,2509	0,2281	0,6936	0,5447	0,5398	0,5352
ET-BERT(поток) [16]	0,9543	0,9242	0,9606	0,9397	0,9029	0,8560	0,8217	0,8332
Предложенный	0,9386	0,9400	0,9386	0,9385	0,9503	0,9510	0,9503	0,9502

В этих задачах классификации наш метод продемонстрировал значительные улучшения в точности , воспроизводимости и показателе F1 по сравнению с ETC-PS, который также использует сигнатуру пути. функции. По сравнению с самой эффективной моделью ET-BERT (поток) наш метод точность была всего на 0,58% ниже в среде VPN и на 1,57% ниже в задачах классификации среды Tor. Однако важно признать, что ET-BERT (поток) – это сложная, крупномасштабная предварительно обученная модель, резко контрастирующая с нашим более рациональным подходом. используя классическую мелкомасштабную модель LSTM.

4.8.3. Анализ производительности

Для более полной оценки сложности модели и вычислительных затрат В качестве показателей оценки мы использовали параметры (Params), операции с плавающей запятой (FLOP), время вывода и время задержки. Параметры обозначают накопительную сумму всех обучаемых весов и смещений в нейронной сети, в то время как FLOP в первую очередь относятся к количество арифметических операций с плавающей запятой, необходимых во время работы нейронной сети процессы прямого и обратного распространения.

В нашем эксперименте мы использовали партию из 32 образцов. Время, затраченное на ввод модель получения результатов прогнозирования называется временем вывода, тогда как Время задержки относится к продолжительности от загрузки 32 образцов в память до получения результаты прогнозирования. Важно отметить, что количество параметров и FLOP может использоваться только для оценки моделей глубокого обучения. Поэтому мы оценивали только те методы, которые использовать модели глубокого обучения.

Machine Translated by Google

В Таблице 8 представлены параметры, FLOP, время вывода и время задержки для различных методы. Данные таблицы наглядно показывают, что наш метод имеет существенные преимущества. в нескольких аспектах для одной и той же задачи классификации зашифрованного трафика.

Таблица 8. Сложность разных методов.

Метод	Парамс (М)	Флопы (G)	Время вывода (мс)	Время задержки (мс)
ФС-сеть [20]	2,17	24,86	39,94	45,72
ДФ [19]	1,83	3,06	29,57	33,56
МТ-ФлоуФормер [23]	0,26	1,07	57,79	63.08
ГрафДапп [44]	0,22	0,59	78,43	84,72
ET-BERT (поток) [16]	85,70	10,87	104,04	110,94
Предложенный	0,03	7,32 × 10–4	4,02	5,56

Во-первых, наш метод использует значительно меньше параметров (Params), чем другие методы. по сложности модели. Это означает, что структура нашей модели более краткая и эффективная, что помогает снизить риск переобучения и одновременно улучшить интерпретируемость модели. и ремонтопригодность.

Во-вторых, что касается требований к вычислительным ресурсам, наш метод требует меньшего количества операции с плавающей запятой (FLOP). Это указывает на то, что наша модель требует меньше вычислений. во время прямого и обратного распространения, что позволяет ускорить процессы обучения и вывода. Это особенно важно для приложений с ограниченными ресурсами, таких как встроенные системы или мобильные устройства.

Кроме того, наш метод демонстрирует явные преимущества как по времени вывода, так и по времени. время задержки. Время вывода короче, а время задержки меньше. Это указывает что наша модель не только эффективна в вычислительном отношении, но и очень отзывчива, что делает ее подходит для приложений реального времени или почти реального времени. Учитывая все эти факторы, мы считаем небольшое расхождение в точности классификации считается приемлемым.

Мы объясняем короткое время вывода и время задержки нашего метода следующими причинами: две основные причины:

- Наша модельная архитектура имеет меньшую сложность и требует меньшего количества вычислений. Ресурсы. В результате расчеты выполняются быстро после отбора образцов. входные данные в модель, тем самым сокращая время вывода.
- Размер нашей выборки относительно небольшой: входной размер равен (1,23,56). Кроме того, вычислительная нагрузка на ЦП минимальна, что позволяет быстро передавать данные из процессора к графическому процессору, тем самым сокращая время задержки.

4.8.4. Статистический анализ

Для дальнейшей проверки эффективности нашего метода мы использовали критерий знаковых рангов Уилкоксона, чтобы определить, существуют ли статистически значимые различия между результатами нашей классификации и результатами других моделей. Знако-ранговый критерий Уилкоксона широко распространен. использовали статистический метод для сравнения двух наборов данных [45]. Наша нулевая гипотеза утверждает следующее: «Два метода обнаружения имеют одинаковую эффективность классификации». Если значение p больше 0,05, мы принимаем нулевую гипотезу. Если значение p меньше 0,05, мы отвергаем нулевую гипотезу с уровнем достоверности 95%. P -значения, сравнивающие наши предлагаемый метод для каждой базовой модели показан в таблицах 9 и 10.

Таблица 9. Значения p при сравнении нашего метода с построением отпечатков пальцев и статистическими методами.

Метод	FlowPrint [17]	AppScanner [9]	CUMUL [35]	К-ФП [42]	ЗЕРНО [44]	ЭТЦ-ПС [12]
р-значение	0,006956	0,036267	0,000186	0,005447	0,000104	0,041655

Таблица 10. Значения p для сравнения нашего метода с моделями глубокого обучения.

Метод	ФС-Нет [20]	ДФ [19]	МТ-ФлоуФормер [23]	ГрафДапп [44]	ET-BERT (поток) [16]
р-значение	0,015764	0,000183	0,016351	0,000322	0,167007

Machine Translated by Google

По результатам можно сделать вывод, что значения p для всех методов по сравнению с нашим методом, за исключением ET-BERT, составляют менее 0,05. Это указывает на то, что, за исключением ET-BERT, существуют значительные различия в эффективности классификации между нашим методом и другими методами. В частности, значение p меньше 0,05 означает, что мы можем отвергнуть нулевую гипотезу с уровнем достоверности 95%, подразумевая, что предположение о том, что два метода имеют одинаковую эффективность классификации, не выполняется.

Таким образом, наш метод значительно превосходит другие базовые модели с точки зрения эффективности классификации, за исключением ET-BERT. Для ET-BERT значение p превышает 0,05, что указывает на отсутствие статистически значимой разницы в эффективности классификации между нашим методом и ET-BERT. Это еще раз подтверждает эффективность нашего метода, демонстрируя, что преимущества нашей модели в сложности и вычислительных ресурсах не достигаются за счет производительности классификации.

5. Выводы и будущие работы.

В этом исследовании мы предложили метод, который сочетает в себе функции подписи пути с моделью LSTM для классификации типов услуг зашифрованного трафика, используя только 24 последовательных функции межтрафика. Суть нашего метода заключается в извлечении фундаментальных характеристик (таких как размер пакета и время прибытия) из последовательности зашифрованных пакетов трафика для построения подробных путей трафика. Эти пути подвергаются процессу трансформации и улучшаются с использованием многомасштабных кумулятивных методов извлечения признаков, в конечном итоге генерирующих многомасштабные признаки сигнатур пути. Эти функции затем эффективно используются в качестве входных данных для модели LSTM, что приводит к точным результатам классификации.

Мы оценили наш метод на общедоступных наборах данных ISCX VPN-nonVPN и ISCX Tor-nonTor. Результаты эксперимента показали точность классификации 94,74%, 90,53%, 93,86% и 95,03% в средах VPN, без VPN, Tor и без Tor соответственно . Эти результаты сопоставимы с современными методами. Кроме того, наши эксперименты по анализу производительности и статистическому анализу продемонстрировали преимущества нашего метода перед крупномасштабными моделями нейронных сетей.

Однако, как метод классификации, основанный на длине пакета, наш подход, как и другие методы [12] , основанные на характеристиках длины пакета, чрезвычайно ограничен по эффективности против зашифрованного трафика, который использует методы обфускации длины пакета, такие как заполнение пакетов [46]. Заполнение пакетов является эффективной защитой от анализа трафика, делая все пакеты одинаковой длины. Изучение более отличительных особенностей зашифрованного трафика может решить эту проблему. Кроме того, наша модель использует относительно простую структуру LSTM; Таким образом, он может работать плохо при решении более сложных реальных задач классификации, что указывает на необходимость в более надежных классификаторах.

В будущей работе мы планируем расширить наш подход, используя больше наборов данных, чтобы применить его к более сложным сетевым средам. Более того, изучение применимости нашей методологии в области обнаружения аномального трафика остается центральным элементом наших будущих исследовательских целей.

Вклад авторов: концептуализация, YM и NL; методология, ЮМ; программное обеспечение, ЮМ; проверка , YM, GZ и XY; формальный анализ, ЮМ; расследование, ГЗ; ресурсы, ЮМ; курирование данных, XY; письмо – подготовка первоначального черновика, ЮМ; написание — обзор и редактирование, ЮМ; визуализация, ГЗ; надзор, Нидерланды; администрация проекта, Нидерланды; приобретение финансирования, Нидерланды. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Эта работа частично финансировалась Национальным фондом социальных наук Китая в рамках гранта 20&ZD293.

Заявление о доступности данных: В этом исследовании были проанализированы общедоступные наборы данных. Эти данные можно найти здесь: https://www.unb.ca/cic/datasets/vpn.html. (по состоянию на 1 июня 2024 г.).

Благодарности: Авторы хотели бы поблагодарить анонимных рецензентов за их вклад в эту статью.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Machine Translated by Google

Рекомендации

1. Бадер О.; Личи, А.; Хаджадж, К.; Дубин Р.; Двир, А. MalDIST: от классификации зашифрованного трафика к обнаружению и классификации трафика вредоносных программ. В материалах 19-й ежегодной конференции потребительских коммуникаций и сетей IEEE 2022 г. (CCNC), Лас-Вегас, Невада, США, 8–11 января 2022 г.; стр. 527–533.

2. Ван Ю.; Он, Х.; Лай, Ю.; Лю, А.С. Двухэтапный подход к быстрой и точной классификации зашифрованного трафика. IEEE/ACM Пер. Сеть. 2023, 31, 1071–1086. [Перекрестная ссылка]

3. Ван, В.; Чжу, М.; Ван, Дж.; Цзэн, Х.; Ян, З. Классификация сквозного зашифрованного трафика с помощью одномерных сверточных нейронных сетей. В материалах Международной конференции IEEE по разведке и информатике безопасности (ISI) 2017 г., Пекин, Китай, 22–24 июля 2017 г.; стр. 43–48.

4. Шен, М.; Лю, Ю.; Чжу, Л.; Сюй, К.; Ду, Х.; Гуизани, Н. Оптимизация выбора функций для эффективной классификации зашифрованного трафика: систематический подход. Сеть IEEE. 2020, 34, 20–27. [Перекрестная ссылка]

5. Караяннис Т.; Папаганнаки, К.; Фалуцсос, М. БЛИНК: Многоуровневая классификация дорожного движения в темноте. ACM SIGCOMM Comput. Коммун. Ред. 2005, 35, 229–240. [Перекрестная ссылка]

6. Лотфоллахи, М.; Джафари Сиавошани, М.; Ширали Хосейн Заде, Р.; Сабериан, М. Deep package: новый подход к классификации зашифрованного трафика с использованием глубокого обучения. Мягкий компьютер. 2020, 24, 1999–2012. [Перекрестная ссылка]

7. Аль-Наами, К.; Чандра, С.; Мустафа, А.; Хан, Л.; Турайсингем, Б.М. Адаптивное определение отпечатков зашифрованного трафика с двунаправленной зависимостью. В материалах конференции по приложениям компьютерной безопасности, Лос-Анджелес, Калифорния, США, 5–9 декабря 2016 г.

8. Тейлор, В.Ф.; Сполаор, Р.; Конти, М.; Мартинович, И. Надежная идентификация приложений для смартфонов посредством анализа зашифрованного сетевого трафика. IEEE Транс. Инф. Криминалистическая безопасность. 2018, 13, 63–78. [Перекрестная ссылка]

9. Тейлор, В.Ф.; Сполаор, Р.; Конти, М.; Мартинович, И. AppScanner: автоматическое распознавание отпечатков приложений для смартфонов на основе зашифрованного сетевого трафика. В материалах Европейского симпозиума IEEE по безопасности и конфиденциальности 2016 г. (EuroS&P), Саарбрюкен, Германия, 21–24 марта 2016 г.; стр. 439–454.

10. Се, Г.; Ли, К.; Цзян, Ю. Метод глубокого обучения для классификации онлайн-трафика и его интерпретируемость. Вычислить. Сеть. 2021, 196, 108267. [CrossRef]

11. Чен К.Т. Интегрирование путей — достоверное представление путей некоммутативным формальным степенным рядом. Пер. Являюсь. Математика. Соц. 1958, 89, 395–407. [Перекрестная ссылка]

12. Сюй, С.-Дж.; Гэн, Г.-Г.; Джин, Х.-Б.; Лю, Д.-Дж.; Венг, Дж. Просмотр путей трафика: классификация зашифрованного трафика с функциями подписи пути. IEEE Транс. Инф. Криминалистическая безопасность. 2022, 17, 2166–2181. [Перекрестная ссылка]

13. Ямансаваскилар, Б.; Гувенсан, Массачусетс; Явуз, АГ; Карслигил, М.Э. Идентификация приложений посредством классификации сетевого трафика. В материалах Международной конференции по вычислительной технике, сетям и коммуникациям (ICNC) 2017 г., Силиконовая долина, Калифорния, США, 26–29 января 2017 г.; стр. 843–848.

14. Чен, RTQ; Рубанова Ю.; Бетанкур, Ж.; Дювено, Д.К. Нейронные обыкновенные дифференциальные уравнения. Адв. Нейронная инф. Процесс. Сист. 2018, 31, 07366.

15. Дайнотти, А.; Пескейп, А.; Клаффи, К.К. Проблемы и будущие направления в классификации дорожного движения. Сеть IEEE. 2012, 26, 35–40. [Перекрестная ссылка]

16. Лин, Х.; Сюн, Г.; Гоу, Г.; Ли, З.; Ши, Дж.; Ю, Дж. Эт-берт: Контекстуализированное представление дейтаграмм с преобразователями предварительного обучения для классификации зашифрованного трафика. В материалах веб-конференции ACM 2022, Лион, Франция, 25–29 апреля 2022 г.; стр. 633–642.

17. Ван Эде, Т.; Бортоламеотти, Р.; Континелла, А.; Рен, Дж.; Дюбуа, диджей; Линдорфер, М.; Чоффнес, Д.; ван Стин, М.; Питер, А. Flowprint: полуконтролируемый сбор отпечатков пальцев мобильных приложений по зашифрованному сетевому трафику. В материалах симпозиума по безопасности сетей и распределенных систем (NDSS), Сан-Диего, Калифорния, США, 23–26 февраля 2020 г.

18. Биджу, В.Г.; Оценки Прашанта, К. Фридмана и Уилкоксона, сравнивающие SVM, пакетирование, повышение, K-NN и дерево решений классификаторы. Дж. Прил. Вычислить. наук. Методы 2017, 9, 23–47. [Перекрестная ссылка]

19. Сиринам, П.; Имани, М.; Хуарес, М.; Райт, М. Глубокий сбор отпечатков пальцев: подрыв защиты веб-сайтов от отпечатков пальцев с помощью глубокого обучения. В материалах конференции ACM SIGSAC 2018 г. по компьютерной и коммуникационной безопасности, Торонто, Онтарио, Канада, 15–19 октября 2018 г.; стр. 1928–1943.

20. Лю, К.; Он, Л.; Сюн, Г.; Цао, З.; Ли, З. Fs-net: сеть последовательности потоков для классификации зашифрованного трафика. В материалах конференции IEEE INFOCOM 2019-IEEE по компьютерным коммуникациям, Париж, Франция, 29 апреля – 2 мая 2019 г.; стр. 1171–1179.

21. Лин, К.; Сюй, Х.; Гао, Х. TSCRNN: Новая схема классификации зашифрованного трафика, основанная на пространственно-временных характеристиках потока для эффективное управление IIoT. Вычислить. Сеть. 2021, 190, 107974. [CrossRef]

22. Чжан Х.; Ю, Л.; Сяо, Х.; Ли, К.; Меркальдо, Ф.; Луо, Х.; Лю, К. TFE-GNN: кодировщик временного слияния, использующий графовые нейронные сети для детальной классификации зашифрованного трафика. В материалах веб-конференции ACM 2023, Остин, Техас, США, 30 апреля – 1 мая 2023 г.; стр. 2066–2075.

23. Чжао Р.; Дэн, Х.; Ян, З.; Ма, Дж.; Сюэ, З.; Ван, Ю. Mt-flowformer: полуконтролируемый преобразователь потока для классификации зашифрованного трафика. В материалах 28-й конференции ACM SIGKDD по обнаружению знаний и интеллектуальному анализу данных, Вашингтон, округ Колумбия, США, 14–18 августа 2022 г.; стр. 2576–2584.

24. Грэм, Б. Разреженные массивы подписей для онлайн-распознавания символов. arXiv 2013, arXiv:1308.0371.

25. Гюрко, Л.Г.; Лайонс, Т.; Контковский, М.; Филд, Дж. Извлечение информации из подписи потока финансовых данных. apXiv 2013, arXiv: 1307.7244.

Machine Translated by Google

26. Перес Аррибас, И.; Гудвин, генеральный менеджер; Геддес, младший; Лайонс, Т.; Сондерс, К.Э. Модель машинного обучения на основе сигнатур для различения биполярного расстройства и пограничного расстройства личности. Перевод Психиатрия 2018, 8, 274. [CrossRef] [ПабМед]
27. Хэмбли, Б.; Лайонс, Т. Единственность сигнатуры пути ограниченной вариации и сокращенной группы путей. Анна. Математика. 2010, 171, 109–167. [Перекрестная ссылка]
28. Хохрайтер, С.; Шмидхубер, Дж. Длинная кратковременная память. Нейронный компьютер. 1997, 9, 1735–1780. [Перекрестная ссылка]
29. Шен, С.; Хан, Т.; Панг, Дж. Прогнозирование коэффициента лобового сопротивления автомобиля с использованием нейронной сети долговременной памяти и LASSO. Размер 2024, 225, 113982. [CrossRef]
30. Йегане, А.; Ахмади, Ф.; Вонг, Ю.Дж.; Шадман, А.; Барати, Р.; Саиди, Р. Меллоу и модели глубокого обучения для прогнозирования уровня подземных вод: подход к интеграции данных с нескольких пьезометров. Вода Воздух Почва. Загрязнение. 2024, 235, 441. [CrossRef]
31. Чжао, Л.; Фархи, Н.; Валеро, Ю.; Христофору, З. Нейронные сети с длинной кратковременной памятью для моделирования поведения человека при вождении. Трансп. Рез. Procedia 2023, 72, 2589–2596. [Перекрестная ссылка]
32. Эсварсай. Изучение различных типов LSTM. Доступно онлайн: https://medium.com/analytics-vidhya/exploring- Different-types-of-lstms-6109bcb037c4. (по состоянию на 28 июля 2024 г.).
33. Сунь, Ю.; Ван, Х.; Тан, Х. Глубоко изученные представления о лицах редки, избирательны и надежны. В материалах конференции IEEE по компьютерному зрению и распознаванию образов 2015 г., Бостон, Массачусетс, США, 7–12 июня 2015 г.; стр. 2892–2900.
34. Васвани А.; Шазир, Н.; Пармар, Н.; Ушкорейт, Дж.; Джонс, Л.; Гомес, АН; Кайзер, Л.; Полосухин И. Внимание – это все, что вам нужно. In Proceedings of the Advances in Neural Information Processing Systems 30, Лонг-Бич, Калифорния, США, 4–9 декабря 2017 г.; стр. 5998–6008.
35. Панченко А.; Ланце, Ф.; Пеннекамп, Дж.; Энгель, Т.; Зиннен, А.; Хенце, М.; Верле, К. Отпечатки веб-сайтов в масштабах Интернета. В Материалы NDSS 2016, Сан-Диего, Калифорния, США, 21–24 февраля 2016 г.
36. Чавла, Невада; Бойер, КВ; Холл, Лоу; Кегельмейер, WP SMOTE: Техника чрезмерной выборки синтетического меньшинства. Дж. Артиф. Интел. Рез. 2002, 16, 321–357. [Перекрестная ссылка]
37. Уилсон Д.Л. Асимптотические свойства правил ближайшего соседа с использованием отредактированных данных. IEEE Транс. Сист. Человек Киберн. 1972, 2, 408–421. [Перекрестная ссылка]
38. Батиста, Дж. Е.; Прати, РЦ; Монард, МС Исследование поведения нескольких методов балансировки данных обучения машинного обучения. ACM SIGKDD Эксплор. Новости 2004, 6, 20–29. [Перекрестная ссылка]
39. Дрейпер-Гил, Г.; Лашкари, АХ; Мамун, MSI; Горбани А.А. Характеристика зашифрованного и VPN-трафика с использованием временной привязки. В материалах 2-й Международной конференции по безопасности и конфиденциальности информационных систем (ICISSP), Рим, Италия, 19–21 февраля 2016 г.; стр. 407–414.
40. Лашкари, А.Х.; Гил, Грузия; Мамун, MSI; Горбани А.А. Характеристика Tor-трафика с использованием временных характеристик. В материалах Международной конференции по безопасности и конфиденциальности информационных систем, Рим, Италия, 29–31 мая 2017 г.; стр. 253–262.
41. Шапира Т.; Шавитт, Ю. FlowPic: общее представление для классификации зашифрованного трафика и идентификации приложений. IEEE Транс. Сеть. Серв. Менеджер. 2021, 18, 12:18–12:32. [Перекрестная ссылка]
42. Хейс, Дж.; Данезис, Г. k-fingerprinting: надежная масштабируемая техника снятия отпечатков веб-сайтов. В материалах 25-й конференции USENIX Симпозиум по безопасности (USENIX Security 16), Остин, Техас, США, 10–12 августа 2016 г.; стр. 1187–1203.
43. Заки, Ф.; Афифи, Ф.; Абд Разак, С.; Гани, А.; Ануар, NB GRAIN: Детальная классификация зашифрованного трафика с несколькими метками с использованием классификатора цепь. Вычислить. Сеть. 2022, 213, 109084. [CrossRef]
44. Шен, М.; Чжан, Дж.; Чжу, Л.; Сюй, К.; Ду, Х. Точная идентификация децентрализованных приложений посредством анализа зашифрованного трафика с использованием графические нейронные сети. IEEE Транс. Инф. Криминалистическая безопасность. 2021, 16, 2367–2380. [Перекрестная ссылка]
45. Вс, П.; Ли, С.; Се, Дж.; Сюй, Х.; Ченг, З.; Ян, Р. ГПМТ: Генерация практического вредоносного трафика на основе состязательных атак с мало предварительных знаний. Вычислить. Безопасность. 2023, 130, 103257. [CrossRef]
46. Ю, С.; Чжао, Г.; Доу, В.; Джеймс, С. Прогнозирование заполнения пакетов для анонимного просмотра веб-страниц от атак с анализом трафика. IEEE Транс. Инф. Криминалистическая безопасность. 2012, 7, 1381–1393. [Перекрестная ссылка]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Machine Translated by Google

электроника

MDPI

Статья

Обнаружение сетевых вторжений на основе сети Deep Belief Network Широкая система обучения по выравниванию

Миаолей Дэн [1,2,*], Чуанчуань Сунь [1,2], Юпэй Кан [1,2], Хайхан Сюй [1,2], Синь Чжоу [1,2] и Шаоцзюнь Фань [1,2]

[1] Колледж информатики и инженерии Хэнаньского технологического университета, Чжэнчжоу 450001, Китай; scc@stu.haut.edu.cn (CS); 2022930951@stu.haut.edu.cn (ЮК); 2022930927@stu.haut.edu.cn (HX); 2023930889@stu.haut.edu.cn (XZ); 2023930813@stu.haut.edu.cn (Сан-Франциско)

[2] Хэнаньская международная совместная лаборатория обработки информации о зерне, Чжэнчжоу 450001, Китай

* Переписка: dengmiaolei@haut.edu.cn

Аннотация: Системы обнаружения сетевых вторжений являются важной защитной технологией, гарантирующей информационную безопасность и защищающей сеть от атак. В последние годы система широкого обучения привлекла большое внимание и с некоторым успехом была внедрена в системы обнаружения вторжений. Однако, поскольку традиционная система широкого обучения представляет собой простую линейную структуру, при работе с несбалансированными наборами данных она часто игнорирует изучение особенностей выборок классов меньшинств, что приводит к более низкому уровню распознавания выборок классов меньшинств. Во-вторых, высокая размерность и избыточность функций в наборах данных для обнаружения вторжений также серьезно влияют на время обучения и производительность обнаружения традиционной системы широкого обучения. Для решения вышеуказанных проблем мы предлагаем систему обучения с широким выравниванием сети глубоких убеждений. Модель полностью изучает крупномасштабный набор многомерных данных через сеть глубоких убеждений и представляет его как оптимальный набор данных низкой размерности, а затем вводит идею повторного взвешивания потерь v2 для выравнивания в широкую систему обучения и учится классифицировать низкоразмерные данные. набор данных через широкую систему обучения с выравниванием. Модель была экспериментально протестирована с использованием набора данных CICIDS2017 и полностью проверена с использованием набора данных CICIDS2018. По сравнению с другими алгоритмами в той же области, модель сокращает время обучения и имеет высокий уровень обнаружения и низкий уровень ложных тревог.

Ключевые слова: система обнаружения сетевых вторжений; широкая система обучения; несбалансированные наборы данных; потери выравнивания v2; CICIDS2017; CICIDS2018

check for updates

Цитирование: Дэн, М.; Сан, К.; Кан, Ю.; Сюй, Х.; Чжоу, Х.; Фан, С. Сеть Обнаружение вторжений на основе системы обучения с широким выравниванием сети Deep Belief. Электроника 2024, 13, 3014. https://doi.org/10.3390/электроника13153014

Академический редактор: Раджендра В. Боппана

Поступила: 17 июня 2024 г.
Пересмотрено: 26 июля 2024 г.
Принято: 28 июля 2024 г.
Опубликовано: 31 июля 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария. Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons Лицензия с указанием авторства (CC BY) (https://creativecommons.org/licenses/by/4.0/).

1. Введение

С быстрым развитием современных компьютерных технологий и сетевых информационных технологий сети влияют на все аспекты нашей жизни, сетевые атаки становятся все более сложными, а сетевая безопасность никогда не была более важной. Хотя Интернет принес людям много удобств, он также сталкивается со многими новыми проблема Проблемы сетевой безопасности становятся все более серьезными, например, фишинговые веб-сайты, распределенные атаки типа «отказ в обслуживании» и сетевые черви, угрожающие сетевой безопасности. Системы обнаружения сетевых вторжений (NIDS) были предложены Андерсоном [1] в 1980 году. Принцип работы системы обнаружения вторжений заключается в основном в мониторинге и фильтрации поведения сети путем анализа данных аудита хоста, данных сетевого трафика и других характеристик , выявления ненормального доступа в сетевых коммуникациях и оперативного уведомления сетевых администраторов для достижения цели обеспечение сетевой информационной безопасности.

В последние годы, благодаря бурному развитию искусственного интеллекта, алгоритмы машинного обучения могут лучше обрабатывать сложные данные для автоматизированного обучения по сравнению с традиционными системами обнаружения вторжений, а также обладают лучшей адаптируемостью и обобщением. Поэтому все больше исследователей внедряют в систему обнаружения вторжений технологию машинного обучения [2] и добиваются успехов [3]. Однако традиционные модели машинного обучения плохо работают при работе со сложными и многомерными задачами.

Machine Translated by Google

данные [4], а надежность модели машинного обучения ограничивает производительность модели [5]. Глубокое обучение позволяет автоматически извлекать и полностью изучать сложные характеристики данных для достижения высокой точности и надежности, поэтому методы глубокого обучения широко используются в системах обнаружения вторжений [6]. Однако процесс обучения модели глубокого обучения требует больших вычислительных ресурсов, длительного времени обучения и высокого энергопотребления [7]. Кроме того, модели глубокого обучения обычно представляют собой модели «черного ящика», которым не хватает интерпретируемости, что затрудняет понимание и объяснение их внутреннего процесса принятия решений, что снижает производительность систем обнаружения вторжений. Более того, надежность моделей глубокого обучения также ограничивает безопасность систем обнаружения вторжений. Аббас и др. [8] провели углубленное исследование надежности моделей и добились определенных успехов. Все больше и больше ученых [9] начинают применять новые методы в области обнаружения вторжений.

Недавно Чен и др. [10] предложили новую рандомизированную нейронную сеть, названную системой широкого обучения (BLS), которая основана на плоской сетевой архитектуре. Эта модель отличается от структуры глубокого обучения с расширением до глубины тем, что для обновления весов не требуется использовать градиентный спуск, а сетевая структура имеет небольшой размер и, следовательно, вычислительно быстрее [11]. Кроме того, обновление веса BLS основано на псевдообратных операциях с лучшей интерпретируемостью и масштабируемостью модели. Поэтому некоторые люди внедрили BLS в системы обнаружения вторжений с хорошими результатами [12]. Однако, с одной стороны, наборы данных NIDS имеют тенденцию быть сильно несбалансированными, а традиционные модели BLS не могут хорошо обрабатывать несбалансированные наборы данных, что приводит к очень низкой эффективности обнаружения атак меньшинства, что фатально для систем обнаружения вторжений, обнаруживающих атаки. С другой стороны, наборы данных NIDS, как правило, имеют высокую размерность и содержат больше избыточной информации, что может привести к увеличению затрат времени и снижению эффективности обнаружения модели BLS. Поэтому для решения этих проблем необходимо принять определенные меры, чтобы улучшить эффективность обнаружения модели.

Традиционное решение проблемы дисбаланса набора данных при обнаружении вторжений состоит в изменении количества выборок класса меньшинства или выборки класса большинства путем избыточной или недостаточной выборки для выравнивания выборок. Например, метод синтетической избыточной выборки меньшинства (SMOTE) [13] увеличивает количество выборок классов меньшинств за счет синтеза новых выборок классов меньшинств, чтобы сбалансировать количество выборок разных классов в наборе данных. Однако выборки, созданные SMOTE, могут расширить исходное пространство выборки и изменить распределение выборочных данных, что приводит к плохой классификации или даже путанице модели в граничной области и часто приводит к увеличению затрат времени из-за введения новых выборок. . Для решения проблемы высокой размерности наборов данных при обнаружении вторжений традиционный метод основан на уменьшении размерности анализа главных компонентов (PCA) [14]. Хотя он может уменьшить размерность набора данных, в качестве метода уменьшения линейной размерности он может найти только структуру линейной корреляции в данных, а для нелинейной структуры наборов данных обнаружения вторжений PCA не может эффективно собирать эту информацию, и ее уменьшение размерности достигается путем проецирования данных в новое низкоразмерное пространство, что часто приводит к серьезной потере информации в исходных данных.

Для решения вышеуказанных проблем мы предлагаем систему обучения с широким выравниванием сети глубоких убеждений (DBELS), в которой сеть глубоких убеждений (DBN) выполняет уменьшение размерности данных в наборе данных NIDS для устранения избыточной информации. Идея выравнивания потерь v2 (EQL v2) введена в BLS, и модель уравновешивает различные классы выборок путем корректировки положительных и отрицательных факторов градиента для повышения производительности при обнаружении выборок атак классов меньшинства и классифицирует низкоразмерные наборы данных для достичь цели классификации и обнаружения атак. Основные положения данной

статьи заключаются в следующем. • Мы вводим идею EQL v2 в BLS и предлагаем новую модель, добавляя положительные и отрицательные факторы градиента и пересчитывая ее веса, что улучшает плохую обучаемость модели BLS для выборок классов меньшинств путем корректировки положительных и отрицательных факторов градиента и смягчает дефект модели BLS в

Machine Translated by Google

что нецелесообразно иметь дело с несбалансированным набором данных, чтобы повысить производительность обнаружения выборок
класса меньшинства. • Мы оценили два типа моделей уменьшения размерности на основе DBN — традиционную модель DBN и модель DeBN, которая вводит идею EQL v2 — и экспериментально сравнили влияние двух моделей уменьшения размерности на эффект классификации.

• Мы провели множество экспериментов с моделью DBELS на основе общедоступного эталонного набора данных CICIDS2017, предоставив подробные экспериментальные установки, включая бинарную и мультиклассификацию, и оценивая модель с точки зрения точности, полноты, частоты ложных срабатываний, времени и работы приемника. характеристические кривые, обнаружив явные преимущества перед другими моделями. Мы также проверили влияние гиперпараметров на модель посредством множества экспериментов и провели сравнительный анализ с другими современными моделями. • Мы
дополнительно проверили пригодность и масштабируемость предложенной модели с помощью набора данных CI-CIDS2018. Производительность и полезность предложенной модели были оценены путем сравнительного анализа с другими моделями.

Остальная часть статьи организована следующим образом: В разделе 2 рассматриваются текущие исследования систем обнаружения вторжений, основанных на BLS, уменьшении размерности данных и дисбалансе данных. В разделе 3 мы описываем детали предлагаемого DBELS. Подробные экспериментальные результаты представлены и проанализированы в разделах 4 и 5. Раздел 6 завершает эту статью.

2. Сопутствующие работы

В этом разделе основное внимание уделяется текущему состоянию исследований и недостаткам системы обнаружения сетевых вторжений (NIDS) на основе системы широкого обучения (BLS). Мы даем традиционные решения ограничений традиционного BLS и анализируем преимущества и недостатки этих методов.

2.1. NIDS на основе BLS

Модель BLS — это модель мелкой сети, которая преобразует и изучает функции путем сопоставления узлов и узлов улучшения, а также быстро рассчитывает веса модели с помощью псевдообратных операций, что упрощает и ускоряет обучение. Когда это было предложено, широкое обучение вызвало широкий интерес, и его применение исследуется во многих областях исследований, таких как компьютерное зрение, обработка изображений, анализ медицинских данных и обработка естественного языка. Также имеются работы в области обнаружения сетевых вторжений. Ли и др. [15] предложили гибридную модель обнаружения вторжений на основе рекуррентной нейронной сети (RNN) и BLS, и результаты экспериментов показывают, что модель на основе BLS имеет лучший обучающий эффект и более короткое время обучения. Лаура и др. [16] реализовали каскад узлов сопоставления признаков, каскад узлов улучшения, а также каскад узлов сопоставления признаков и узлов улучшения.

Авторы пришли к выводу, что каскад узлов улучшения требует более длительного времени обучения, чем другие варианты BLS. Впоследствии они предложили систему обнаружения DDoS-атак, основанную на широком обучении сетей связи [17]. Авторы пришли к выводу, что наилучшей производительности обнаружения обычно можно достичь при использовании каскадного BLS. Ли и др. [18] предложили систему тройного обучения (TBLS), основанную на модели BLS, которая изучает особенности по трем измерениям временной детализации, содержания данных и пространственной детализации набора данных . Экспериментально было доказано, что модель TBLS может обеспечить лучшую производительность обнаружения за счет изучения трехмерных функций.

Хотя модель BLS добилась определенных успехов в области обнаружения вторжений, будучи мелкой нейронной сетью, она слишком сильно полагается на топологию, количество выборок и информацию о классах обучающих выборок и не способна справиться с несбалансированными наборами данных.
Это приводит к тому, что он уделяет слишком много внимания выборкам классов большинства и почти игнорирует изучение выборок классов меньшинства при изучении классификации, что приводит к плохой производительности обнаружения на несбалансированных наборах данных, особенно для меньшинства.

Machine Translated by Google

образцы класса, которые хуже или даже необнаружимы. Во-вторых, наборы данных большой размерности увеличивают сложность моделей BLS и увеличивают затраты на вычисления и хранение. Кроме того, чем выше размерность данных, тем больше влияние шума и избыточных функций, которые могут маскировать полезную информацию и снижать точность и стабильность модели BLS. Следовательно, необходимы определенные методы для адаптации BLS к несбалансированным и многомерным наборам данных для обнаружения вторжений.

2.2. NIDS на основе дисбаланса данных

Дисбаланс данных означает тот факт, что в реальном мире разные классы данных имеют разное распределение, при этом определенные типы данных значительно недопредставлены, что оказывает серьезное негативное влияние на классификацию моделей. Наборы данных для обнаружения вторжений часто сильно несбалансированы, что губительно для обучения и производительности обнаружения в модели; например, в случае CICIDS2017 нормальное поведение Доброкачественного поведения составляет 82,248%, а ненормальное поведение Атаки составляет только 17,752%. Примеры атак DoS/DDoS составляют 14,571%, сканирование портов — 2,607%, брутфорс — 0,389%, веб-атака — 0,096%, ботнет — 0,088%; таким образом, очевидно, что нормальный класс образцов намного больше, чем класс образцов атаки, что приведет к тому, что модель будет слишком много изучать образцы большинства класса. В ответ на несбалансированность наборов данных по обнаружению вторжений некоторые учёные уже провели у By T и др. [13] рассмотрели несбалансированные данные при обнаружении сетевых вторжений с использованием кластеризации k-средних и метода синтетической избыточной выборки меньшинства (SMOTE). Они кластеризовали данные, определили кластеры меньшинств и применили SMOTE для создания синтетических образцов, продемонстрировав эффективность в экспериментах. Ахмад Т. и др. [19] разработали гибридную модель , сочетающую выбор признаков и анализ шаблонов. Они использовали анализ на основе правил для уменьшения размерности, SMOTE для балансировки данных и адаптивное повышение (AdaBoost) для обучения, эффективно выявляя выборки класса меньшинства. Благодаря достижениям в области глубокого обучения Hao X et al. [20] использовали GAN для создания синтетических наборов данных, повышая производительность классификатора для классов меньшинств за счет реалистичной генерации данных.

Однако большинство методов, используемых для решения проблемы дисбаланса данных при обнаружении вторжений, основаны на генерации данных, будь то создание выборок класса меньшинства путем избыточной выборки или сокращение выборок класса большинства путем недостаточной выборки, что уравновешивает различные выборки. путем изменения номера набора данных; однако это меняет распределение исходных данных, что может повлиять на эффект обучения модели классификации набора данных. Кроме того, метод передискретизации увеличит количество наборов данных, что увеличит дополнительные вычислительные затраты модели и приведет к дополнительным затратам времени. В последние годы Джин и др. [21] предложили выравнивание потерь v2 (EQL v2), которое руководствуется положительными и отрицательными градиентами и перенастраивает веса, чтобы сбалансировать обучение каждого типа задач и улучшить обучение образцов класса меньшинства, и добилось определенных результатов. . Поскольку он не меняет распределение и количество исходных выборок и корректирует веса так, чтобы модель изучала каждый класс выборок более сбалансированным образом, это имеет большое значение для решения проблемы дисбаланса систем обнаружения вторжений и введения этого Идея в BLS способствует повышению эффективности обнаружения модели в условиях несбалансированных наборов данных.

2.3. NIDS на основе уменьшения размерности данных

Исследователи предложили некоторые решения, позволяющие справиться с проблемой большой размерности в области обнаружения вторжений. Чжан Б и др. [14] предложил метод обнаружения вторжений с использованием расширенного анализа главных компонентов (PCA) в сочетании с простым байесовским алгоритмом Гаусса. Взвешивая основные векторы признаков в PCA, этот метод снижает загрязнение данных. Этот подход, сопровождаемый байесовским алгоритмом обнаружения, значительно сокращает время обнаружения по сравнению с использованием только классификаторов. Шен З и др. [22] представили улучшенный алгоритм наивной байесовской классификации, который объединяет анализ главных компонент с линейным дискриминантным анализом для уменьшения размерности пространства выборки и уточняет байесовские вы

Machine Translated by Google

корреляция. Этот подход позволяет классификатору учитывать как частоту атрибутов, так и проблемы корреляции, решая проблемы в реальном времени и точности при обнаружении вторжений с многочисленными функциями. Сало Ф и др. [23] разработали метод интеграции обработки признаков , который сочетает в себе получение информации и анализ главных компонентов для извлечения низкоразмерного оптимального подпространства. Затем они объединяют несколько классификаторов, используя машины опорных векторов , алгоритмы обучения экземпляров и стратегии принятия решений многоуровневого перцептрона, применяя правило комбинации средних вероятностей для голосования. Тестирование на различных наборах данных показало, что этот метод не только обеспечивает высокую точность, но и значительно снижает вычислительные затраты, что делает его более эффективным для обнаружения к

Однако большинство текущих решений проблемы высокой размерности наборов данных NIDS представляют собой алгоритмы машинного обучения, которые в определенной степени уменьшают размерность набора данных, но все же имеют большие ограничения перед лицом нелинейных функций высокой размерности. , что приводит к более низкой скорости обнаружения модели. Поскольку алгоритмы глубокого обучения эффективны при обучении высокой размерности, а нелинейные функции могут автономно завершать обучение и извлекать признаки из необработанных данных без слишком большого вмешательства человека, алгоритмы глубокого обучения были внедрены в крупномасштабные системы обнаружения вторжений, которые могут лучше завершить извлечение и уменьшение размерности объектов данных. Сеть глубоких убеждений (DBN) как репрезентативная модель глубокого обучения была предложена Джеффри Хинтоном и др. [24] в 2006 году. Из необработанных данных он может изучать многоуровневые абстрактные представления функций, которые эффективно отражают сложную структуру и изменяющиеся шаблоны данных. Этот иерархический процесс извлечения признаков помогает сопоставить входные данные высокой размерности с пространством представления низкой размерности, сохраняя при этом важную информацию данных для уменьшения размерности.

По сравнению с традиционными методами уменьшения размерности, уменьшение размерности DBN не только лучше сохраняет структуру и актуальность данных, но также адаптивно изучает нелинейные отношения в данных, что улучшает выразительные способности и эффективность классификации уменьшенных данных. Таким образом, ДБН превосходно справляется с задачей уменьшения размерности моделей обнаружения вторжений.

3. Методология В

этом разделе подробно строится система обучения с широким выравниванием сети глубоких убеждений (DBELS) на основе BLS, EQL v2 и DBN в три этапа: предварительная обработка данных, уменьшение размерности данных и классификация модели.

3.1. ДБЭЛС Архитектура

Структура DBELS показана на рисунке 1 и разделена на три этапа.

Первый этап — предварительная обработка данных, которая в основном преобразует исходные данные в форму данных, подходящую для уменьшения размерности ДБН. Набор данных CICIDS2017 содержит 79 функций, содержащих множество избыточных функций, которые затрудняют обучение и классификацию модели. В этой статье предварительная обработка данных включает две части: очистку данных и стандартизацию данных. Цель очистки данных — уменьшить избыточные признаки путем расчета вклада признаков в классификацию с помощью коэффициентов корреляции, исключить неверные значения данных и избежать их негативного влияния на модель.

Цель стандартизации данных состоит в том, чтобы модель сбалансированно фокусировалась на различных функциях данных, игнорируя при этом влияние функций с разными весами на модель из-за размера значения. Из-за серьезного дисбаланса данных в наборе данных CICIDS2017 метод избыточной выборки SMOTE можно использовать для увеличения выборок атак меньшинства, когда DBN уменьшает размерность, чтобы облегчить проблему слишком малого количества выборок для класса меньшинства.

Второй этап — уменьшение размерности ДБН, цель которого — представить исходные многомерные данные как оптимальные низкоразмерные данные. Модель уменьшения размерности DBN, которую мы используем, использует классическую структуру, которая максимально сохраняет основные функции при максимально возможном уменьшении размеров функций. Модель полностью обучает каждую ограниченную машину Больцмана (RBM) посредством предварительного обучения, настраивает всю DBN

Machine Translated by Google

тс
структуру посредством точной настройки веса и полностью изучает особенности стандартизированного набора данных, который представлен как оптимальный малоразмерный набор данных. На третьем этапе используется система обучения с широким выравниванием (BELS) для классификации оптимальных низкоразмерных данных. Модель BELS уравнивает обучение модели выборкам класса меньшинства путем корректировки положительных и отрицательных факторов градиента и увеличения ее весов для улучшения класса обнаружения и распознавания образцов класса меньшинства.

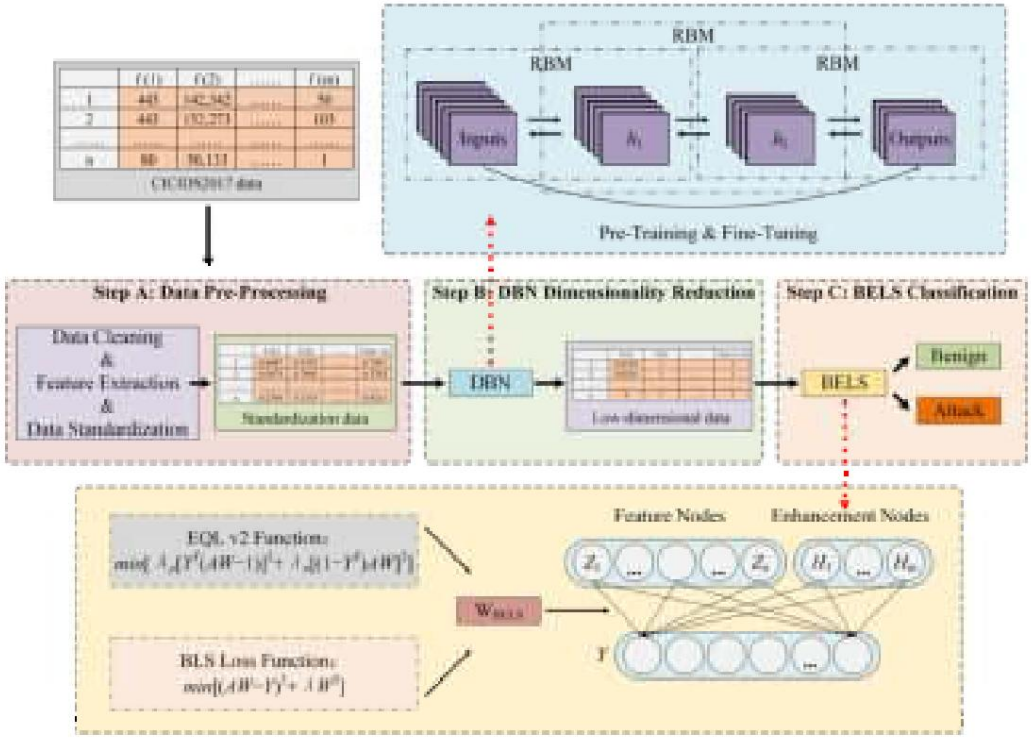


Рисунок 1. Структура DBELS.

3.2. Сеть глубоких
убеждений 3.2.1. Ограниченная машина Больцмана

тс
Ограниченная машина Больцмана [24] представляет собой базовую структуру ДБН, которая представляет собой мелкую нейронную сеть и состоит из двухслойной структуры видимых и скрытых слоев. Связи в RBM ненаправлены, и между нейронами одного слоя нет связей ; его структуру можно увидеть на рисунке 2. Обычно используемые RBM обычно являются двоичными. Независимо от того, находятся ли они в скрытом или видимом слое, их нейрон принимает только значение 0 или 1. Принцип работы RBM основан на энергетической функции, а энергетическая функция видимого и скрытого слоев определяется выражением , E(v, h) = bivi cjhj vihjwij

(1)

где i принадлежит видимому слою, а j принадлежит скрытому слою, b и c — смещения модели , w — вес модели, первая часть представляет вклад узлов видимого слоя в энергию системы, вторая часть представляет собой вклад узлов скрытого слоя в энергию системы, а последняя часть представляет собой вклад энергии системы за счет взаимодействия видимого слоя и скрытого слоя.

Machine Translated by Google

(|h)

h

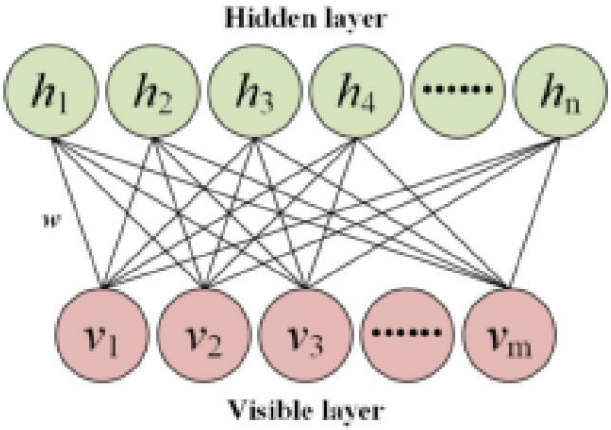


Рисунок 2. Структура RBM.

RBM определяет совместную плотность вероятности видимого и скрытого слоев с точки зрения энергетической функции. Совместная вероятность видимого слоя и скрытого слоя определяется как

$$p(v, h) = \frac{1}{3} e^{E(v,h)} \quad (2)$$

где Z — статистическая сумма, которая используется для нормализации распределения вероятностей и обеспечения равенства суммы распределений вероятностей 1:

$$Z = \sum_{в,ч} e^{E(v,h)} \quad тт \quad (3)$$

При совместной плотности вероятности видимого и скрытого слоев, суммируя по h дает предельное распределение плотности видимого слоя:

$$p(v) = \frac{1}{Z} \quad e^{E(v,h)} \quad (4)$$

В процессе обучения модели RBM переменная скрытого слоя h1 выбирается и вычисляется с использованием апостериорной вероятности p(h1 |v), а новая переменная видимого слоя v1 выбирается с помощью апостериорной вероятности p(v1 |h1); впоследствии новая переменная скрытого слоя h2 снова выбирается повторяющимся образом, чтобы получить видимый слой v2. Вышеуказанные шаги повторяются несколько раз до тех пор, пока параметры не сойдутся или не достигнут заданного количества итераций, совместное распределение вероятностей не станет близким к гладкому распределению и обучение RBM не будет завершено.

3.2.2. Обучение сети Deep Belief

Сеть глубоких убеждений [24] представляет собой глубокую нейронную сеть, состоящую из нескольких RBM. Структура БДД показана на рисунке 3. В процессе обучения БДД сначала использует неконтролируемый жадный алгоритм послойного обучения для каждого уровня УОР, за исключением первого и последнего слоев структуры ДБД, где каждый уровень Скрытый слой RBM служит видимым слоем следующего уровня RBM для передачи функциональной информации и информации обучения (этот этап также называется этапом предварительного обучения DBN). После завершения предварительного обучения небольшое количество размеченных данных прикрепляется к последнему слою модели DBN, и контролируемое обучение выполняется на каждом уровне RBM. Алгоритм обратного распространения распространяет информацию об ошибках на каждый уровень RBM и использует функцию максимального правдоподобия для целевой функции, которая может оптимизировать всю DBN, получая таким образом оптимальное низкоразмерное представление исходных да

Machine Translated by Google

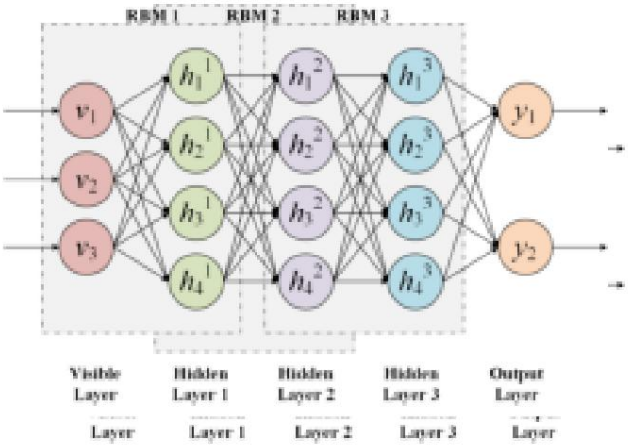


Рисунок 3. Структура ДБН.

3.3. Широкая система обучения

Носителем системы широкого обучения [10] является случайная векторная функция, связанная с нейронными связями. сеть. По сравнению с методом глубокого обучения, BLS имеет только два слоя нейронной сети. сеть, а ее структура имеет форму расширения по ширине. Показана структура его модели. на рисунке 4. Во-первых, BLS отображает входные данные в матрицу узлов объектов через функция отображения. Во-вторых, матрица узлов признаков формирует матрицу узлов улучшения путем с помощью функции улучшения. Затем узлы сопоставления объектов и улучшения узлы используются вместе в качестве входных данных скрытого слоя. Наконец, псевдообратное операция используется для расчета весовой матрицы между скрытым слоем и выходным уровень для достижения цели обучения модели BLS.

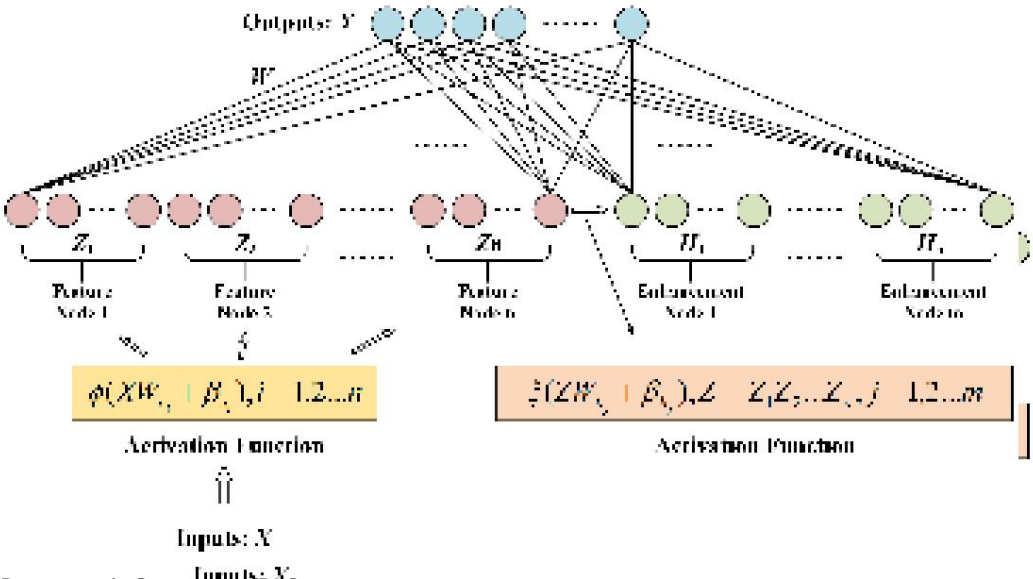


Рисунок 4. Структура BLS.

Предположим, что входные данные X = [x1, x2, . . . , xn] принадлежат R(N M) и выход матрица, Y = [y1, y2, . . . , yn], принадлежит R(N C), где N — общая выборка данных, M — размерность выборки C — количество категорий выборки данных Z1, Z2, . . . , Zn признаков, H1, H2, . . . , Hn — матрица узла улучшения, а W — выходная весовая матрица между скрытым слоем и выходным слоем.

Конкретная процедура BLS заключается в следующем. Сначала данные выборки X сопоставляются по признакам группы i , чтобы получить матрицу узлов признаков группы i Zi , где i (1, 2, .., n):

$$Z_i = \varphi_i(XW_{ei} + \beta_{ei}) \quad R^{N \times q} \; i = 1, 2, \ldots, n \tag{1}$$

Machine Translated by Google

где φi — функция активации, q — количество узлов признаков, соответствующих каждому отображению признаков, Wei — весовая матрица i-го сопоставления признаков, а βei — матрица смещения i-го сопоставления признаков. И Wei , и βei генерируются случайным образом. Затем все матрицы узлов признаков интегрируются для получения общей матрицы узлов признаков:

$$З^{Н} = [Z1, Z2, \cdots , Zn] \quad R\ N\times nq \tag{6}$$

Во-вторых, общая матрица узлов признаков Zn подвергается преобразованию улучшения группы j для получения матрицы узлов улучшения группы j Hj , где j (1, 2,..., m), следующим образом:

$$Hj = \xi j\ Z\ iWhj + \beta hj \quad R^{Н\times р} \quad j = 1, 2, \cdots , м \tag{7}$$

где ξj — функция активации, r представляет собой количество узлов улучшения, соответствующих каждой группе преобразований улучшения, Whj — весовая матрица отображения улучшения группы j , а βhj — матрица смещения отображения улучшения группы j . Общая матрица узлов улучшения получается путем интегрирования всех матриц узлов улучшения:

$$^{ЧАС}{}^{М} = [H1, H2, \cdots , Hm] \quad R^{Н\times мр} \tag{8}$$

Затем узлы сопоставления объектов объединяются с узлами расширения в качестве входных данных для BLS, определяемых как A:

$$A = [Z^{Н} |H\ m] \quad R\ N\times(nq+mr) \tag{9}$$

Наконец, прогнозируемая ценность алгоритма широкого обучения может быть выражена как

$$Yˆ = [Z\ n\ |H\ м]W = AW \tag{10}$$

где W — весовая матрица узлов признаков и узлов улучшения выходного слоя.

Чтобы минимизировать ошибку между прогнозируемым значением Yˆ и истинным значением Y и найти подходящее значение W, модель BLS оптимизируется с помощью следующей функции:

$$\arg\min_{Вт} FBLS = Y \quad Yˆ \quad {}^{2}_{2} + \lambda \quad Вт \quad {}^{2}_{2} = Y \quad AW \quad {}^{2} + \lambda \quad Wm \quad {}^{2} \tag{11}$$

где первый член используется для управления минимизацией ошибки обучения, второй термин используется для предотвращения переобучения модели, а значение λ обозначает дополнительные ограничения на сумму квадратов весов; мы берем производную указанной выше функции и позволяем ей принимать значение нуля, получая W как

$$W = (A\ TA + \lambda I) \quad {}^{1}ТИ \tag{12}$$

где A — матрица отображения и расширения, λ — штрафной коэффициент, I — единичная матрица, а Y — истинная метка.

3.4. Потери при

выравнивании v2 Основная идея потери при выравнивании v2 [21] заключается в выравнивании степени обучения модели для разных классов выборок путем взвешивания положительных и отрицательных градиентов каждого классификатора в соответствии с совокупными коэффициентами градиентов положительных и отрицательных градиентов. классификатора соответственно. При решении проблемы несбалансированных наборов данных классическая стратегия является фокусной стратегией, и Eqlv2 улучшен на этой основе. Тем не менее, EQL v2 предлагает явные преимущества по сравнению с потерей фокуса, особенно в устранении дисбаланса классов без внесения предвзятости в сторону классов меньшинств. В отличие от очаговой потери, которая корректирует вклад потерь в зависимости от сложности прогнозирования, EQL v2 динамически балансирует градиенты во время обучения, уменьшая переоснащение редких категорий, сохраняя при этом фокус на общих классах. Этот подход повышает стабильность и производительно

Machine Translated by Google

обеспечивая более детальный ответ на проблемы дисбаланса. Положительные и отрицательные градиенты выхода Z каждого классификатора относительно потерь L равны

$$\text{поз}_{zj}(Л) = \frac{1}{|Я|}\sum_{y\,j\,i\ I}^{я}(п_{дж}^{я} \quad 1) \tag{13}$$

$$\text{отрицательный}_{j}(Л) = \frac{1}{|Я|}\sum_{iєI}(1 \quad y_{дж}^{я})п_{дж}^{я} \tag{14}$$

где I — общее количество выборок, p j ^{я} вероятность того, что i-й образец будет предсказан принадлежать к классу j, а ^{я} — значение матрицы правдивой разметки One-Hot, в которой i-й образец y j относится к классу j.

Веса EQL v2 обновляются путем определения gj как отношения совокупного положительного градиента к отрицательному градиенту задачи j до итерации t. В этой итерации веса положительного градиента qt и отрицательного градиента rt можно вычислить следующим образом:

$$q_{jj}^{(t)}(t) \text{ знак равно } 1 + \alpha\ 1 \quad f(g) \tag{15}$$

$$p_{дж}^{(t)} = f(g\ j^{т}) \tag{16}$$

где f — отображающая функция как

$$f(x) = \frac{1}{\gamma(x \quad \mu)\ 1 + e} \tag{17}$$

Получаются положительный градиент q и отрицательный градиент r , а затем они jj применяются к положительному и отрицательному градиентам текущей партии соответственно, и перевзвешенный градиент равен

$$\text{pos}'_{zj}(L(t)) = q_{дж}^{(t)}\,\text{pos}_{zj}(Л(т)) \tag{18}$$

$$zj^{\text{отрицательный}}(L(t)) = p^{(t)}\,\text{neg}_{zj}\ j(Л(т)) \tag{19}$$

Наконец, вычисляется $^{T+1}$ накопленного положительного градиента к отрицательному градиенту отношение g j к обновлению итерации t + 1 и обозначается как

$$г_{ж}^{т+1} \text{ "=" } \frac{\sum_{т=0}^{T}\text{пос}'_{zj}(L(t))}{\sum_{т=0}^{T}\text{neg}'_{zj}(Л(т))} \tag{20}$$

EQL v2 выбирает статистику градиента как меру того, находится ли задача в сбалансированном состоянии обучения, балансируя модель на основе того, насколько легко или сложно каждую классификацию обучать модель, а не просто учитывая количество ее положительных и отрицательных выборок. . Введение идеи градиентного взвешивания в DBN и BLS помогает модели обучаться с большим вниманием к особенностям меньшинственных классов выборок, что делает модель более полной и сбалансированной при оценке каждого класса выборок.

Модель корректирует оптимальный весовой вклад итерации на основе влияния различных категорий выборок на производительность классификации.

3.5. Широкая система обучения по выравниванию

W системы обучения с широким выравниванием решается следующим образом; на основе Положительные и отрицательные градиенты EQL v2, положительные и отрицательные градиенты выборки могут быть заданы следующим образом:

$$p(W) = Y^{T}(AB - 1) \tag{21}$$

$$n(W) = (1 \quad Y^{T})\ ОУ \tag{22}$$

Machine Translated by Google

где A — матрица признаков сопоставленных и расширенных узлов, Y — истинная метка, а W — матрица весов.

Функция градиента может быть построена из положительных и отрицательных градиентов как

$$G(W) = \lambda p[Y^{T}(AW - 1)]2 + \lambda n[(1 - Y^{T})AY]^{2} \quad (23)$$

где λp — коэффициент положительного градиента, а λn — коэффициент отрицательного градиента. Чем меньше G(W) , тем больше вероятность положительных выборок и меньше вероятность отрицательных выборок. Корректировка параметров λp и λn позволяет регулировать эффект обучения модели на разных выборках, чтобы добиться выравнивания степени вклада каждого класса выборок в классификацию обучения модели.

На основе функции потерь BLS:

$$L(W) = (AW - Y)^{2} + \lambda W2 \quad (24)$$

где λ – штрафной коэффициент; тогда целью оптимизации BELS является функция градиента EQL v2 и функция потерь BLS:

$$\text{argmin} : F(W) = (AW - Y)^{2} + \lambda W2 + \lambda p[Y^{T}(AW - 1)]2 + \lambda n[(1 - Y^{T})AY]^{2} \quad (25)$$

Среди них первая часть — это функция потерь, которая используется для контроля ошибки классификации модели. Вторая часть — это штрафной член, который в основном предотвращает переобучение модели. Третья часть — это градиентный термин, который используется для балансировки степени обучения различных выборок и снижения низкой производительности модели при обнаружении выборок классов меньшинств. $\frac{\partial F(W)}{\partial W}$

Существующее F(W) выражается как частная производная от W:

$$W = [\lambda pA\,TYYTA + \lambda nA^{T}(J1 - Y)(J2 - Y^{T})A + A\,TA + \lambda I]^{1}(\lambda pA\,TYYT\,J3 + A\,TY) \quad (26)$$

где J — матрица, все элементы которой равны 1, λp — положительный коэффициент градиента, λn — отрицательный фактор градиента, положительные и отрицательные факторы градиента используются для контроля степени обучения модели на разных выборках, а λ — это штрафной коэффициент, чтобы избежать переобучения модели, вызванного слишком большим значением W.

4. Эксперименты

В этом разделе представлены подробные эксперименты. Подробно анализируем влияние гиперпараметров на модель. Мы используем поиск по сетке, чтобы зафиксировать положительные и отрицательные факторы градиента в соответствующих диапазонах, ища оптимальную комбинацию параметров, которая максимизирует эффективность обнаружения модели. Чтобы оценить эффективность предлагаемого метода, мы проводим полные эксперименты по абляции. В заключение мы даем сравнительный анализ предлагаемой модели с другими современными методами.

4.1. Наборы данных CICIDS2017

Набор данных CICIDS2017 [25] , предложенный Канадским институтом кибербезопасности, является одним из важных эталонных наборов данных для оценки моделей обнаружения вторжений. Авторы использовали систему профилей поведения для анализа абстрактного поведения человеческих взаимодействий на основе различных протоколов, создавая 25 абстрактных моделей поведения пользователей и генерируя естественный дружественный фоновый трафик. Набор данных содержит безопасные и новые распространенные атаки и включает в себя семь типов сетевых атак, а именно атаки типа «отказ в обслуживании», атаки методом перебора защищенной оболочки, атаки ботнетов, распределенные атаки типа «отказ в обслуживании», атаки веб-приложений, эксплойты Heartbleed и проникновение. тестирование атак. Распределение выборки и особенности набора данных показаны в таблице 1.

Machine Translated by Google

Таблица 1. Распределение классов набора данных CICIDS2017.

Категория	Этикетка	Образец	Особенность
Доброкачественный	Доброкачественный	2 273 097	Порт назначения, продолжительность потока, Tot Fwd |Bwd Pkts, Tot Лен Фвд |Бвд Пкц, Фвд |Бвд Пкт Лен Макс |Мин |Среднее |Станд., Мин |Макс Pkt Len, Pkt Len Среднее |Стандарт |Вар, Средний размер участка, Среднее прямое |Bwd сегмент Размер, поток Байты |Pkts, Fwd |Bwd Pkts, Fwd |Bwd Средняя объемная скорость, Fwd |Bwd Avg Bytes |Pkts, Flow IAT Среднее |Стандарт |Макс |Мин, Прямое |Bwd IAT Итого |Среднее |Стандарт |Макс |Мин, коэффициент понижения-повышения, Активный |Средний холостой ход |Стандарт |Макс |Мин, вперед |Bwd PSH Флаги, флаги Fwd |Bwd URG, байты инициализации Win Fwd |Bwd, Fwd |Bwd Заголовок Len, Act Data Pkt Вперед, FIN |SYN |RST |PSH |ACK |URG |CWE |ECE Флаги, минимальный размер сегмента вперед, подпоток вперед |Bwd Pkts |Байты, Метка
DoS/DDoS	DoS Халк, DDoS, DoS GoldenEye, DoS Слоулорис, Медленный DoShttptest, Сердцекровие	380 699	
ПортСкан	ПортСкан	158 930	
Грубая сила	FTP-Пататор, SSH-Пататор	13 835	
Веб-атака	Веб-атака — грубая сила, Веб-атака — XSS, Веб-атака: SQL-инъекция	2180	
Ботнет	Бот	1966 год	
Общий	-	2 830 707	79

4.2. Детали реализации

Параметры модели DBELS были установлены, как указано в таблице 2, включая параметры обучения. модели снижения размерности ДБН, функции активации БЭЛС, количества узлов в каждой из групп отображения, количество узлов в каждом из расширений группы и штрафной коэффициент λ. Модель уменьшения размерности ДБН состояла из стопка из трех УКР: (49,64), (64,32) и (32,16). Все эксперименты проводились с использованием 64-разрядный процессор Intel(R) Core (TM) i7-11700 с 32 ГБ оперативной памяти в среде Windows 11. Модели были реализованы на Python v3.9.16 с использованием библиотеки PyTorch v2.1.0.

Таблица 2. Параметры ДБЭЛС.

Параметр	Предварительное обучение	Тонкая настройка	Описание
Эпохи	30	100	-
Скорость обучения	0,0001	0,00001	-
Размер партии	64	128	-
Оптимизатор	сингапурский доллар	Адам	-
Шаг Гиббса	-	-	5
Сопоставление |Количество узлов группы улучшений	-	-	16
Сопоставление |Функция активации улучшения	-	-	Релу
λ	-	-	0,001

4.3. Показатели эффективности

Метрики оценки обнаружения вторжений в основном включают точность, отзыв и уровень ложноположительных результатов (FPR). В формуле истинно положительный результат (TP) представляет собой количество образцы правильно идентифицированы как положительные, истинно отрицательный результат (TN) — это количество образцов правильно идентифицировано как отрицательное, ложноположительный результат (FP) — это количество отрицательных образцов неправильно идентифицированы как положительные, а ложноотрицательный (ЛН) – это количество положительных образцы ошибочно идентифицированы как отрицательные. Точность обозначает соотношение количества выборки, правильно классифицированные по модели, к общему числу выборок:

$$\text{Точность} = \frac{\text{ТП} + \text{ТН}}{\text{ТП} + \text{ТН} + \text{ФП} + \text{ФН}} \quad (27)$$

Machine Translated by Google

Отзыв обозначает долю положительных случаев в выборке, которые правильно прогнозируемый, также известный как уровень обнаружения:

$$\text{Напомним} = \frac{\text{ТП}}{\text{ТП} + \text{ФН}} \quad (28)$$

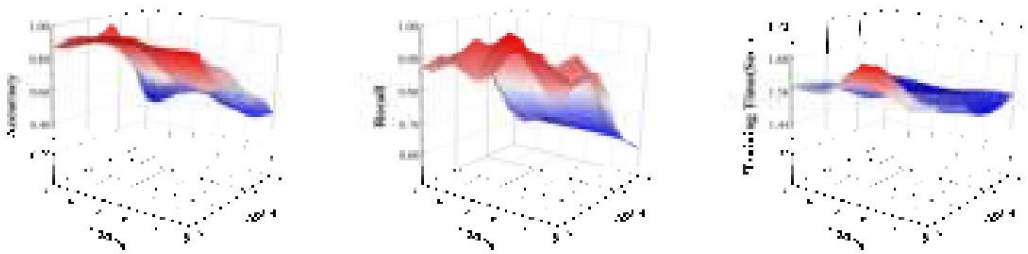
Доля ложноположительных результатов — это доля действительно отрицательных образцов, которые модель ошибочно относит к положительным:

$$\text{ФПР} = \frac{\text{ФП}}{\text{ФП} + \text{ТН}} \quad (29)$$

4.4. Анализ гиперпараметров

4.4.1. Влияние λ_p и λ_n на бинарную классификацию

Для двух параметров λ_p и λ_n в модели DBELS их влияние на производительность модели в различных задачах классификации. Их влияние на производительность точности, полноты и времени в задаче двоичной классификации показаны на рисунке 5. Для Точность и полнота, производительность имеет тенденцию увеличиваться с уменьшением λ_p и λ_n. Другой комбинации параметров приводят к большим различиям, поскольку модель более чувствительна к два параметра в задаче двоичной классификации, и по мере улучшения порядка величины производительность модели кардинально меняется. Эксперименты показывают, что при около 10^{-5} порядков, а λ_n составляет около 10^{-4} порядков, модель может поддерживать более высокую точность и полноту — около 0,99 и 0,98 соответственно. Для время обучения модели при изменении λ_p и λ_n время обучения незначительно колеблется выше и ниже 1,6 с, что указывает на то, что время обучения модели в бинарной классификации Задача не чувствительна к положительным и отрицательным факторам.



(а) (б) (с)

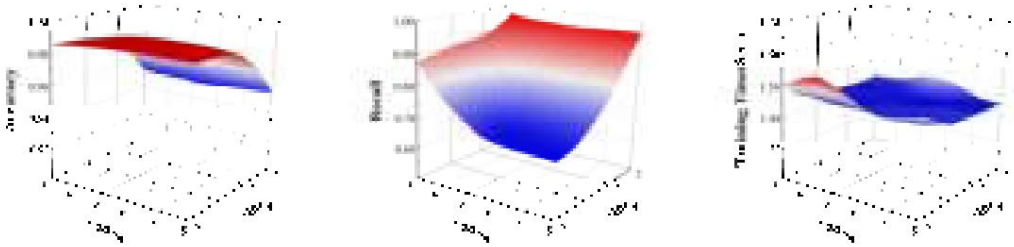
Рисунок 5. Влияние λ_p и λ_n на бинарную классификацию. Влияние на производительность: а) влияние на точность; (б) влияние на отзыв; (с) влияние на время обучения.

далее 4.4.2. Влияние λ_p и λ_n на мультиклассификацию

В задаче мультиклассификации ее влияние на точность, полноту и производительность по времени. показаны на рисунке 6. По точности модель практически нечувствительна к λ_n и почти стремится быть стабильным при изменении λ_n ; когда $10^{-3} < \lambda_p < 10^{-2}$, точность увеличивается с увеличением λ_p . Напомним, что по мере увеличения λ_p и λ_n запоминаемость модели имеет тенденцию к увеличению. Когда λ_p порядок величины составляет 10^{-2} , а порядок величины значения λ_n составляет около 10^{-3}. , отзыв вырастает максимум примерно до 0,95. Для времени обучения модели с изменением λ_p и λ_n время модели колеблется от 1,4 с до 1,6 с и практически стабилизируется, что показывает, что время обучения не чувствительно к положительным и отрицательным факторным параметрам.

Для достижения лучшей общей производительности порядок величины λ_p составляет около 10^{-4}., и порядок величины λ_n составляет около 10^{-3}, что может стабилизировать точность до 0,96. и отзыв стабилизируется до 0,95.

Machine Translated by Google

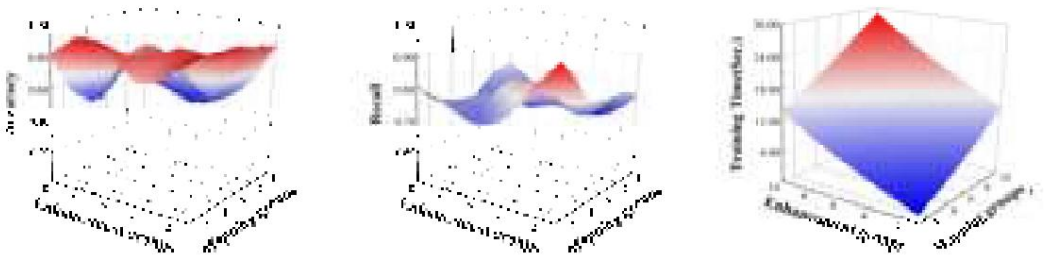


(а) (б) (с)

Рисунок 6. Влияние λp и λn на мультиклассификацию. Влияние на производительность: (а) эффект по точности; (б) влияние на отзыв; (с) влияние на время обучения.

4.4.3. Влияние групп отображения и расширения на двоичную классификацию

В ходе экспериментов изучалось влияние групп улучшения модели DBELS и сопоставление групп производительности модели для различных задач классификации. Для задача бинарной классификации, влияние групп расширения и групп отображения на точность, отзыв и время обучения показаны на рисунке 7. Для точности и отзыва модель больше плавает при изменении параметров, поскольку модель задачи двоичной классификации более чувствителен к этим двум параметрам. С увеличением матрицы отображения Z и матрица улучшения H, ее входная матрица A увеличивается, а производительность обнаружения модели сильно зависит от матрицы A. Для времени обучения оно увеличивается с увеличением группы расширения и группы отображения, поскольку входная матрица A увеличивается с увеличением увеличение параметров. Из результатов экспериментов известно, что при усилении группы и группы сопоставления выбраны равными 1, это может обеспечить как меньшее время обучения, так и более высокая точность и отзыв.



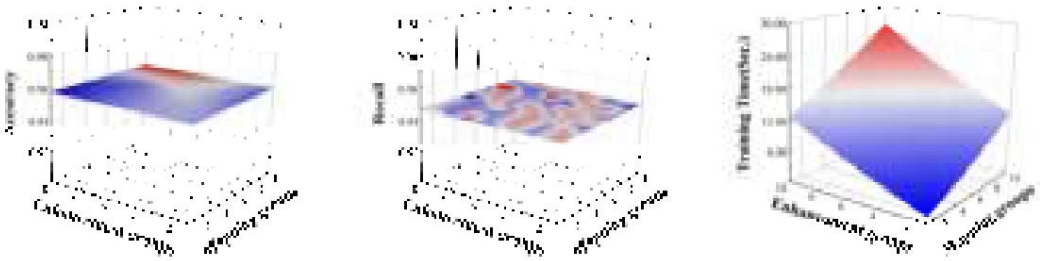
(а) (б) (с)

Рисунок 7. Влияние групп расширения и групп сопоставления на бинарную классификацию. Эффекты на производительность: (а) влияние на точность; (б) влияние на отзыв; (с) влияние на время обучения.

4.4.4. Влияние групп сопоставления и улучшения на мультиклассификацию

Влияние задачи мультиклассификации на точность и полноту показано на рисунке 8. С увеличением групп расширения и групп отображения точность и полнота почти имеют тенденцию к стабилизации, что указывает на то, что для задачи мультиклассификации эти две метрики модель DBELS не чувствительна к группам сопоставления и расширения. Тренировка время показывает тенденцию к увеличению затрат времени, как с увеличением картографирования группы и группы расширения. Причина в том, что увеличение картографических групп и группы расширения увеличивают размер матрицы отображения Z и матрица H, что приводит к увеличению матрицы A, увеличению нагрузки и потребления модели. Можно сделать вывод, что, поскольку увеличение групп усиления и группы сопоставления оказывают меньшее влияние на производительность обнаружения модели и увеличивают потребление затрат времени при выборе групп расширения и групп сопоставления как 1, точность и полнота обеспечивают более высокую производительность и затраты времени. сведен к минимуму.

Machine Translated by Google



(а) (б) (с)

Рисунок 8. Влияние групп расширения и групп сопоставления на мультиклассификацию. Эффекты на производительность: (а) влияние на точность; (б) влияние на отзыв; (с) влияние на время обучения.

4.5. Исследования абляции

Чтобы оценить эффективность предложенного метода, мы провели эксперименты по абляции на основе трех моделей уменьшения размерности и двух моделей классификации. Трехмерные модели уменьшения размерности: традиционная PCA, традиционная DBN и DeBN с введением EQL v2. Методы классификации: традиционные BLS и наши улучшенные БЭЛС соответственно. Мы оценили эффективность обнаружения в общей сложности восемь методов в рамках задачи бинарной классификации и задачи мультиклассификации на основе Набор данных CICIDS2017. Все эксперименты проводились в одной и той же среде с идентичные формы обработки данных, а характеристики обнаружения различных моделей показаны в таблицах 3 и 4. Уровень распознавания каждого образца по разным классификациям задачи представлены в таблицах 5 и 6.

Таблица 3. Эффективность различных методов бинарной классификации.

Метод	Алгоритм					Метрика оценки		
	БЛС	БЕЛС	СПС	ДБН	ДеБН	Точность	Отзывать	Время(а)
БЛС						0,88105	0,63062	1,83126
ПБЛС						0,86780	0,59044	0,75575
ДБЛС						0,91531	0,73711	0,60949
ДеБЛС						0,29939	0,58103	0,60531
БЕЛС						0,96670	0,92483	4,86849
ПБЕЛС						0,92279	0,85250	1,64884
ДБЭЛС						0,99240	0,98579	1,63914
ДЕБЕЛС						0,98896	0,97516	1,57186

Знак в таблице указывает на то, что метод содержит алгоритм.

Таблица 4. Эффективность различных методов мультиклассификации.

Метод	Алгоритм					Метрика оценки		
	БЛС	БЕЛС	СПС	ДБН	ДеБН	Точность	Отзывать	Время(а)
БЛС						0,97450	0,47382	1,74388
ПБЛС						0,93755	0,29756	0,51006
ДБЛС						0,99286	0,65797	0,51341
ДеБЛС						0,98936	0,67712	0,50913
БЕЛС						0,97060	0,52527	5.00825
ПБЕЛС						0,93848	0,47024	1,77999
ДБЭЛС						0,98719	0,80346	1,49983
ДЕБЕЛС						0,95790	0,94741	1,63619

Знак в таблице указывает на то, что метод содержит алгоритм.

Machine Translated by Google

Таблица 5. Отзыв каждого образца по бинарной классификации.

Категория	Метод							
	БЛС	ПБЛС	ДБЛС	ДеБЛС	БЕЛС	ПБЕЛС	ДБЭЛС	ДЕБЕЛС
Доброкачественный	0,99983	0,99934	0,99983	0,16582	0,98655	0,95613	0,99554	0,99551
Атака	0,26141	0,18154	0,47439	0,99625	0,86312	0,74888	0,97604	0,95480

Таблица 6. Отзыв каждого образца по мультиклассификации.

Категория	Метод							
	БЛС	ПБЛС	ДБЛС	ДеБЛС	БЕЛС	ПБЕЛС	ДБЭЛС	ДЕБЕЛС
Доброкачественный	0,99094	0,99252	0,99787	0,99665	0,99057	0,95042	0,98886	0,95903
Ботнет	0	0	0	0,13660	0	0	0	0,94845
АРЕС								
Грубая сила	0	0	0,98480	0,98655	1	0	0,99766	0,98830
DoS/DDoS	0,91515	0,79286	0,97416	0,95507	0,87920	0,89234	0,98244	0,94544
Сканирование портов	0,93683	0	0,99101	0,98787	0,85313	0,97871	0,99093	0,98709
Веб-атака	0	0	0	0	0	0	0,86085	0,85613

4.5.1. Анализ производительности бинарной классификации

В задаче двоичной классификации традиционная структура BLS менее эффективна при распознавании набора данных CICIDS2017, главным образом из-за многомерной избыточной информации. в наборе данных, что влияет на эффективность классификации, что приводит к снижению точности и Напомним, особенно отсутствие внимания к выборкам класса меньшинства в несбалансированном набор данных. После внедрения EQL v2 точность и полнота модели BELS значительно улучшились, что указывает на то, что EQL v2 расширяет возможности распознавания представителей класса меньшинства. образцы решения проблемы дисбаланса. Модель системы широкого обучения сети глубоких убеждений (DBLS) уменьшает избыточность данных за счет размерности DBN. сокращение, улучшающее точность и полноту, которые были улучшены на 0,034 и 0,106 соответственно. Модель DeBLS чрезмерно фокусируется на выборках из класса меньшинства во время Уменьшение размерности DeBN, что приводит к снижению эффективности обнаружения для образцы большинства класса.

DBELS имеет самую высокую производительность обнаружения в задаче двоичной классификации и введение EQL v2 поверх DBLS повышает точность и отзыв за счет 0,077 и 0,249 соответственно, что указывает на то, что сочетание размерности ДБН сокращение и EQL v2 значительно улучшают производительность модели. Хотя ДеБЕЛС модель представляет EQL v2 в наборе данных уменьшения размерности DeBN и улучшает производительность модели путем настройки параметров положительного и отрицательного градиента факторов точность и полнота несколько снизились на 0,004 и 0,01 соответственно по сравнению с с DBELS, что отражает негативное влияние чрезмерного обучения аномальных образцов на классификационная производительность. Для анализа главных компонентов широкая система обучения (PBLS), после уменьшения размерности PCA точность и полнота уменьшились на 0,014 и 0,040 соответственно по сравнению с BLS, в основном из-за большего количества информации. потери при уменьшении размерности PCA и невозможность справиться с несбалансированными эффективно работать с наборами данных. Система обучения с широким выравниванием анализа главных компонентов (PBELS) имела точность 0,923 и полноту 0,853 по сравнению с PBLS, при этом улучшение на 0,055 и 0,262, что указывает на то, что EQL v2 все еще может значительно улучшиться. производительность обнаружения образцов меньшинства даже при размерности PCA снижение неэффективно. Однако PBELS снизил точность и полноту на 0,070 и 0,106 соответственно по сравнению с DBELS, что еще раз демонстрирует, что размерность DBN сокращение превосходит уменьшение размерности PCA в задачах двоичной классификации.

Machine Translated by Google

4.5.2. Анализ производительности при

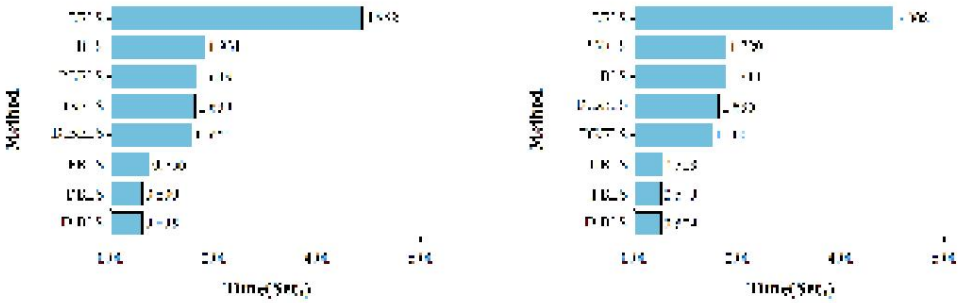
мультиклассификации. В задаче мультиклассификации традиционный BLS показал плохие результаты с точностью 0,975 и полнотой всего 0,474, главным образом из-за серьезного дисбаланса набора данных, который влияет на производительность классификации. Модель BELS после внедрения EQL v2 имела улучшенную полноту — 0,051, главным образом потому, что EQL v2 улучшает проблему дисбаланса данных, что позволяет модели увеличить вес выборок атак для классов меньшинства и, таким образом, повышает уровень обнаружения. Модель DBLS после уменьшения размерности с помощью DBN имела улучшенную точность 0,018 и улучшенную полноту 0,184, что связано с тем, что уменьшение размерности DBN уменьшает избыточность данных, смягчает негативное влияние необработанных данных и улучшает производительность обнаружения модели. Модель DeBLS улучшила запоминаемость на 0,02 по сравнению с DBLS, что объясняется тем, что DeBN узнает больше о выборках класса меньшинства, устраняет избыточность между различными выборками и в то же время сохраняет ключевые особенности образцов класса меньшинства. что улучшает эффективность классификации и обнаружения. Модель DBELS улучшилась на 0,018 в уменьшении размерности DBN, а с введением EQL v2 точность улучшилась на 0,017, а полнота улучшилась на 0,278; по сравнению с DBLS, запоминание улучшилось с 0,658 до 0,803, что связано с тем, что EQL v2 уравновешивает обучающий вес выборок из класса меньшинства и увеличивает внимание к образцам

Точность модели DeBELS 0,957 и полнота 0,947 были лучшими показателями в задаче мультиклассификации, а полнота была значительно улучшена на 0,27 по сравнению с DeBLS, поскольку улучшенная модель BLS лучше справляется с несбалансированными задачами. наборы данных и расширяет возможности распознавания атак меньшинства. Отзыв DeBELS улучшился на 0,144 по сравнению с DBELS, а общая эффективность классификации значительно улучшилась. Точность модели PBLS снизилась на 0,037 , а полнота уменьшилась на 0,176 после уменьшения размерности PCA, поскольку уменьшение размерности PCA приводит к большей потере информации, что влияет на эффект мультиклассификации . Точность модели PBELS составила 0,938, полнота — 0,470; по сравнению с PBLS точность и полнота были улучшены на 0,001 и 0,173 соответственно, что указывает на то, что введение EQL v2 повышает уровень обнаружения образцов класса меньшинства. Однако по сравнению с DeBELS точность и полнота PBELS были снижены на 0,019 и 0,477 соответственно, что указывает на то, что снижение размерности DBN намного лучше, чем модель уменьшения размерности PCA с точки зрения производительности обнаружения, и еще раз доказывает превосходство модели уменьшения размерности PCA. предложенная модель в задачах мультиклассификации.

4.5.3. Анализ затрат

времени Сравнение времени обучения для различных моделей показано на рисунке 9. Для бинарной классификации и мультиклассификации время обучения традиционного BLS составляло около 1,8 с, что отражает преимущество неглубоких сетей с меньшими временными затратами при классификации моделей. Время обучения BELS составило около 5 с, а увеличение времени связано с тем, что внедрение EQL v2 требует дополнительного выполнения множества матричных операций. Время обучения модели DBLS составило около 0,5 с, что является значительным сокращением по сравнению с моделью BLS, что указывает на то, что многомерные данные оказывают существенное негативное влияние на время обучения модели, а уменьшенная размерность набора данных снижает время обучения до 25% от первоначального. Время обучения DeBLS также составило около 0,5 с, что указывает на то, что введение EQL v2 после уменьшения размерности DBN практически не влияет на время обучения. Время обучения DBELS составило около 1,5 с, что значительно меньше по сравнению с BELS благодаря уменьшению размерности модели DBN для многомерных данных. Однако по сравнению с DBLS произошло увеличение из-за некоторых дополнительных вычислений, необходимых для улучшенной модели BELS. Время обучения DeBELS составило около 1,6 с, что почти такое же, как и у DBELS, что указывает на то, что введение низкоразмерного набора данных с EQL v2 не оказывает существенного влияния на время обучения, и наблюдается небольшое увеличение. в затратах времени за счет небольшого количества дополнительных вычислений матриц положительного и отрицательного градиента по сра

Machine Translated by Google



(а) (б)

Рисунок 9. Сравнение времени обучения для разных методов: (а) бинарная классификация; (б) мультиклассификация.

Результаты показывают, что гибридные модели DBELS и DeBELS на основе DBN и BLS сохраняют низкие затраты времени по сравнению с традиционными моделями BLS, следуя преимуществам неглубоких сетей BLS, которые являются быстрыми в вычислительном отношении и легко обучаемыми. Хотя модель уменьшения размерности PCA также сокращает время обучения , ее производительность обнаружения значительно ухудшается, что еще раз подтверждает преимущества модели уменьшения размерности на основе DBN. Предложенная модель сравнима с традиционной моделью глубокого обучения, имеющей существенное преимущество в затратах времени , что способствует экономии определенного количества вычислительных ресурсов в практических приложениях.

4.5.4. Анализ результатов для отзыва каждого образца при бинарной классификации

Уровень обнаружения выборок доброкачественного типа и атаки при выполнении задачи двоичной классификации показан в таблице 5. Традиционный BLS показал плохие результаты при обнаружении выборок атаки с отзывом всего 0,261, что связано с переобучением образцов мажоритарного класса при работе с несбалансированные наборы данных, что приводит к низкому уровню обнаружения выборок класса меньшинства. С внедрением модели BELS с EQL v2 уровень обнаружения атак значительно улучшился до 0,863, что значительно улучшает идентификацию образцов класса меньшинства. Модель DBLS с уменьшением размерности DBN улучшила уровень обнаружения атаки на 0,213, что говорит о том, что уменьшение размерности DBN снижает избыточность данных и положительно влияет на последующую классификацию модели. В модели DeBLS уровень обнаружения атаки достигал 0,996, но уровень обнаружения доброкачественного состояния составлял всего 0,165, что показывает, что снижение размерности DeBN чрезмерно фокусируется на меньшинствах классов образцов, что приводит к снижению эффективности распознавания большинство классов образцов.

Модель DBELS после уменьшения размерности за счет введения EQL v2 имела коэффициент обнаружения 0,996 для доброкачественных заболеваний и 0,976 для атак, что улучшило уровень обнаружения двух классов образцов на 0,009 и 0,113 соответственно по сравнению с моделью BELS. . Поскольку уменьшение размерности DBN уменьшает избыточные функции и улучшает эффективность классификации модели по сравнению с DBLS, DBELS значительно улучшил уровень обнаружения атаки с 0,474 до 0,955, что объясняется тем фактом, что EQL v2 уравновешивает веса выборки, так что Модель фокусируется на каждом классе образцов более сбалансированным образом. Коэффициенты обнаружения доброкачественных явлений и атак для модели DeBELS составляли 0,995 и 0,955 соответственно, и по сравнению с DeBLS уровень обнаружения доброкачественных явлений был значительно выше, а уровень обнаружения атак был немного ниже, но EQL v2 улучшил производительность обнаружения, сделав модель более сбалансированный подход к каждому классу образцов. По сравнению с DBELS, уровень обнаружения доброкачественных заболеваний для DeBELS практически не изменился, а уровень обнаружения атак немного снизился, что объясняется переобучением образцов класса меньшинства, представленных DeBN, но классификация по-прежнему работала хорошо за счет корректировки положительных результатов. и факторы

Machine Translated by Google

Уровень обнаружения атаки для модели PBLS составил всего 0,182 после уменьшения размерности PCA, что меньше, чем у модели BLS, равной 0,080. Уменьшение размерности PCA приводит к потере информации, что отрицательно влияет на эффект обнаружения классификации выборок класса меньшинства. Коэффициент обнаружения атак модели PBELS составил 0,749, что улучшилось на 0,567 по сравнению с PBLS. Внедрение EQL v2 позволяет модели более сбалансированно концентрироваться на каждом классе образцов, что повышает эффективность обнаружения образцов меньшинства. Несмотря на улучшение модели PBELS, ее уровень обнаружения атак все еще был на 0,227 ниже, чем у DBELS, что еще раз доказывает превосходство модели уменьшения размерности DBN в производительности обнаружения.

4.5.5. Анализ результатов для отзыва каждого образца при мультиклассификации

Коэффициенты обнаружения различных методов для каждого типа выборки в рамках задачи мультиклассификации показаны в Таблице 6. Традиционный BLS не работал хорошо при работе с несбалансированным набором данных CICIDS2017, и хотя он имел лучшую производительность обнаружения для большего количества (например, Beign, DoS/DDoS, сканирование портов), он плохо обнаруживал образцы некоторых классов (например, ботнет ARES, Brute Force, Web Attack). Это связано с тем, что модель BLS не способна сбалансировать вклады различных классов выборок при работе с несбалансированными наборами данных, что приводит к очень низкому уровню обнаружения выборок меньшинства. Внедрение модели BELS с EQL v2 значительно улучшило уровень обнаружения атак грубой силы, что, в свою очередь, улучшило идентификацию атак.
Однако ботнет ARES и Web Attack по-прежнему оставались незамеченными, поскольку характеристики этих образцов меньшинства очень похожи на другие образцы, которые модель не может дифференцировать, что приводит к ложным срабатываниям. Модель DBLS устраняет определенную избыточность данных после уменьшения размерности с помощью DBN, что повышает эффективность обнаружения . Коэффициент обнаружения грубой силы был улучшен до 0,985, однако ботнет ARES и Web Attack по-прежнему не обнаруживались. Это указывает на то, что само по себе снижение размерности не может решить проблему несбалансированных наборов данных, особенно для классов с очень небольшим количеством выборок.

Модель DeBLS еще больше улучшила эффективность обнаружения, и большинство типов атак были обнаружены, включая ботнет ARES. Это связано с тем, что снижение размерности DeBN позволяет больше узнать об образцах класса меньшинства и сохранить их ключевые особенности, но Web Attack по-прежнему не обнаруживается, что требует дальнейшего внимания к образцам класса меньшинства . Модель DBELS показала значительное улучшение уровня обнаружения большинства типов атак по сравнению с BELS и DBLS, особенно Web Attack, с 0 до 0,856.
Это связано с тем, что уменьшение размерности DBN уменьшает избыточность данных, а EQL v2 позволяет модели лучше ориентироваться на выборки класса меньшинства. Однако ботнет ARES так и не был обнаружен, что говорит о том, что уменьшения размерности DBN недостаточно для решения проблемы.
Модель DeBELS поддерживала высокий уровень обнаружения всех типов атак. По сравнению с DeBELS, уровень обнаружения ботнета ARES улучшился с 0,137 до 0,948, а веб-атаки — с 0 до 0,856. Это связано с тем, что внедрение EQL v2 улучшает способность лучше распознавать и изучать образцы меньшинства, что значительно повышает уровень обнаружения. DeBELS еще больше улучшил уровень обнаружения ботнета ARES по сравнению с DBELS, поскольку уменьшение размерности DeBN сохраняет ключевые особенности выборок класса меньшинства с более тонкой детализацией и смягчает проблему дисбаланса данных.

Модель PBLS, использующая уменьшение размерности PCA, обнаружила только наибольшее количество DoS/DDoS-атак, тогда как сканирование портов, ботнет ARES, брутфорс, веб-атака и другие атаки не были обнаружены. По сравнению с моделью BLS, сканирование портов не может быть обнаружено из-за потери информации, вызванной уменьшением размерности PCA, что влияет на эффект мультиклассификации. После внедрения EQL v2 в модель PBELS уровень обнаружения сканирования портов увеличился до 0,978, но несколько других классов атак (например, ботнет ARES, брутфорс и веб-атака) по-прежнему оставались незамеченными. Это говорит о том, что уменьшение размерности PCA приводит к потере ключевых функций. По сравнению с

Machine Translated by Google

DeBELS, производительность обнаружения PBELS была хуже, что указывает на то, что снижение размерности на основе DBN превосходит традиционное снижение размерности PCA с точки зрения производительности обнаружения.

4.6. Анализ ROC-AUC На

рисунке 10 показаны кривые рабочих характеристик приемника (ROC) для двоичной классификации DBELS . Общая средняя кривая ROC имеет площадь под кривой (AUC) 0,9949, что указывает на то, что классификатор DBELS превосходно работает со всем набором данных и близок к идеальному классификатору. Во-вторых, можно наблюдать кривые ROC для двух категорий (Атака и Доброкачественная), каждая из которых имеет AUC 0,9916, что указывает на то, что классификатор также имел высокую производительность при рассмотрении каждой категории в отдельности, особенно при работе с меньшинством образцы. Значение AUC, близкое к 1, означает , что классификатор имеет очень высокий уровень истинно положительных результатов (TPR) в своих прогнозах для категории, сохраняя при этом относительно низкий FPR. Это означает выявлять как можно больше истинно положительных образцов, сохраняя при этом уровень ошибочного диагноза как можно ниже. В совокупности классификатор DBELS показал отличную производительность на всем наборе данных о дисбалансе как для выборок класса меньшинства, так и для выборок большинства класса

Кривая ROC мультиклассификации DeBELS показана на рисунке 11. Общая средняя кривая ROC имеет AUC 0,9760, что указывает на то, что классификатор хорошо работает на всем наборе данных. Обнаружения Benign, Botnet ARES, Brute Force, DoS/DDoS и сканирования портов показали хорошие результаты; все их области AUC были больше 0,9700, хотя у Web Attack была самая низкая область AUC 0,9100, но это уже лучше, чем у других моделей того же уровня. В целом, после введения положительных и отрицательных градиентных факторов EQL v2, корректируя веса параметров, модель может уделять больше внимания меньшинствам классов выборок перед лицом несбалансированных наборов данных, тем самым улучшая производительность обнаружения.

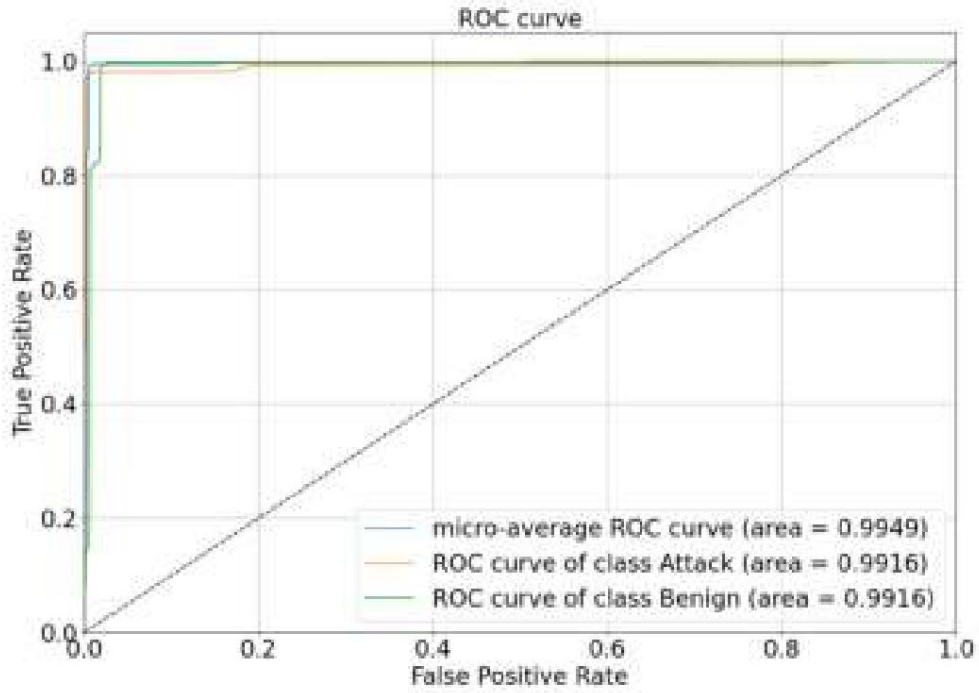


Рисунок 10. ROC DBELS по бинарной классификации.

Machine Translated by Google

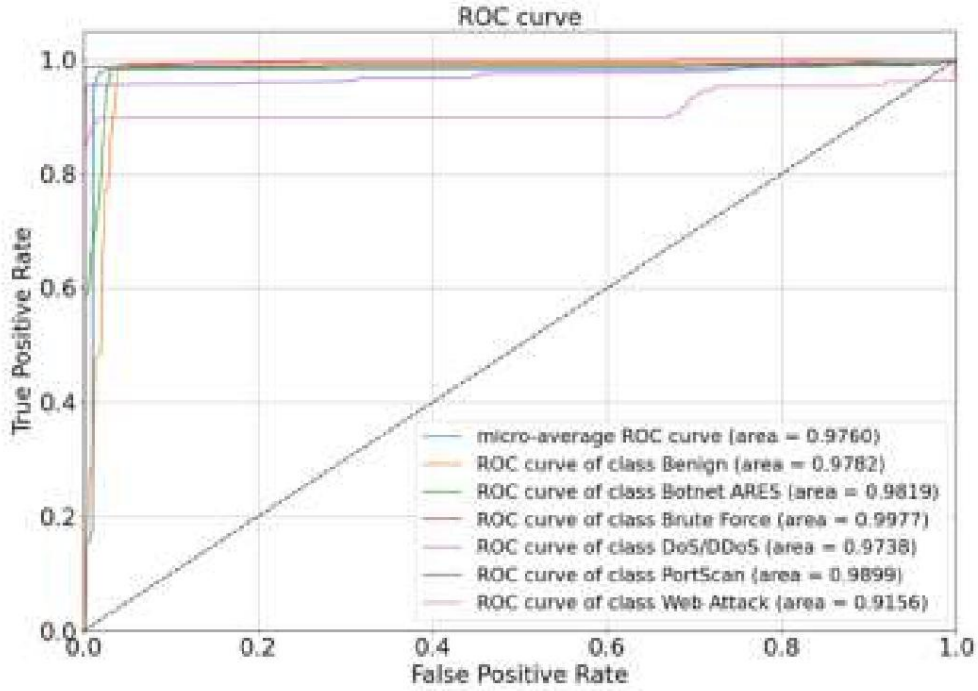


Рисунок 11. ROC DeBELS по мультиклассификации.

4.7. Сравнение с современными методами 4.7.1.

Бинарная классификация.

Частота ложных тревог и временные затраты всех алгоритмов сравнивались при выполнении задачи двоичной классификации, как показано на рисунке 12. Для FPR предложенная модель имела более низкую частоту ложных тревог — 0,014, что было лучше, чем у других моделей. Это свидетельствует о том, что предложенная модель реже неправильно маркирует нормальные ситуации как ненормальные и чаще правильно идентифицирует и классифицирует нормальные ситуации, по сравнению с другими методами. Для времени обучения модели предложенная модель потребовала меньше времени обучения и времени обнаружения в задаче двоичной классификации, при этом затраты времени составили около 1 секунды. Временные затраты модели намного ниже, чем у модели глубокого обучения, и имеют преимущество перед другими алгоритмами машинного обучения, что позволяет значительно сократить временные затраты и обеспечить более быстрое обучение модели на крупномасштабных наборах данных.

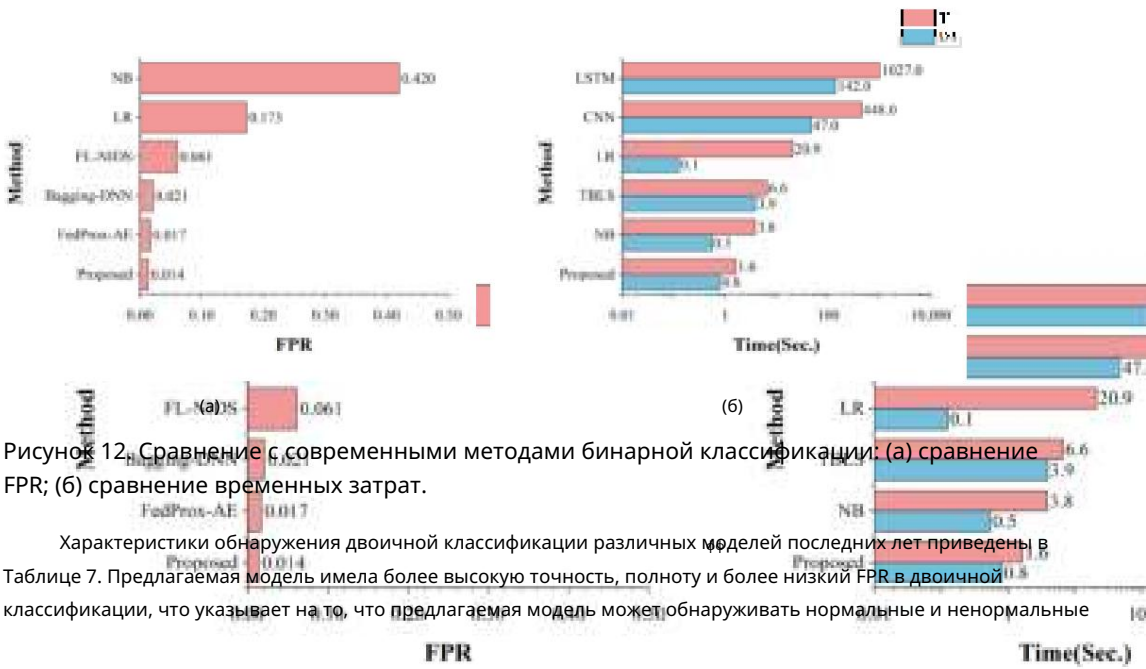


Рисунок 12. Сравнение с современными методами бинарной классификации: (а) сравнение FPR; (б) сравнение временных затрат.

Характеристики обнаружения двоичной классификации различных моделей последних лет приведены в Таблице 7. Предлагаемая модель имела более высокую точность, полноту и более низкий FPR в двоичной классификации, что указывает на то, что предлагаемая модель может обнаруживать нормальные и ненормальные

Machine Translated by Google

поведение хорошо. По сравнению с традиционными моделями машинного обучения предлагаемая модель превосходит логистическую регрессию (LR) и наивный Байес (NB) по точности, отзыву и ФПР. По сравнению с другими моделями на основе BLS точность и полнота лучше, чем у TBLS. По сравнению с моделями глубокого обучения предлагаемая модель превосходит большинство моделей. а производительность слияния статистической глубокой нейронной сети (FS-DNN) и многоцелевой эволюционной сверточной нейронной сети (MECNN) немного выше, чем предлагаемая модель; однако, поскольку обе модели основаны на моделях глубокого обучения, их размер модели и затраты времени на обучение намного выше, чем у модели на основе BLS. Учитывая модель глубокой нейронной сети обучения ансамбля пакетирования (Baging-DNN), ее точность ниже, чем у предлагаемой модели, а ее отзыв немного выше, чем у предлагаемой модели, но в дополнение к более высокой пространственной и временной сложности, вызванной моделью глубокого обучения, уровень ложных тревог также выше, чем у предлагаемой модели. Подводя итог, предлагаемый модель имеет лучшую производительность обнаружения и более низкий уровень ложных срабатываний при двоичной классификации. и имеет очевидные преимущества перед другими моделями.

Таблица 7. Сравнение с современными методами бинарной классификации.

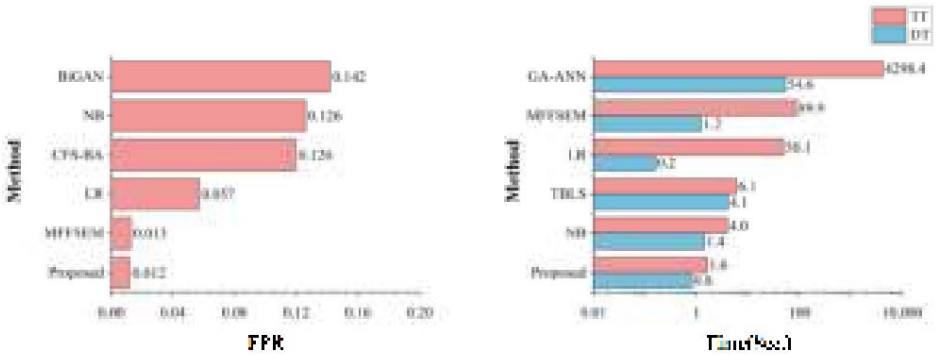
Метод	Точность	Отзывать	ФПР
Наше исследование	0,992	0,986	0,014
ЛР	0,934	0,827	0,173
Примечание:	0,307	0,580	0,420
ДКАЭ	0,925	0,925	-
ТБЛС(Ж) [18]	0,982	0,975	-
ДНН [18]	0,868	-	-
Си-Эн -Эн [18]	0,844	-	-
ЛСТМ [18]	0,365	-	-
2D-CNN [26]	0,980	-	-
CNN-LSTM [27]	0,930	0,768	-
ФС-ДНН [28]	0,998	0,999	0,012
МЕКНН [29]	0,998	0,998	-
Бэггинг-ДНН [30]	0,987	0,999	0,021
ФЛ-НИДС [31]	0,943	0,947	0,061
ФедПрокс-АЭ [32]	0,935	-	0,017

4.7.2. Мультиклассификация

Частота ложных срабатываний и временные затраты всех остальных алгоритмов мультиклассификации задачи показаны на рисунке 13. Графики FPR показывают, что предложенная модель имела только Уровень ложноположительных результатов 0,012, что ниже, чем у других моделей. Это означает, что модель может лучше распознавать различные категории атак и реже ошибочно классифицировать их во время задача мультиклассификации и выполнение задачи обнаружения и идентификации образцов атак лучше. Модель потребовала меньше времени на обучение и время обнаружения по сравнению с другими модели, предлагающие преимущества при обработке крупномасштабных данных и повышающие их пригодность для реальные среды набора данных.

Производительность обнаружения мультиклассификации различных моделей в последние годы приведены в таблице 8. Предлагаемая модель имеет более высокую точность и полноту и меньший FPR при задача мультиклассификации, которая указывает на то, что предлагаемая модель может обнаруживать как нормальные и ненормальное поведение и лучше обнаруживают каждый конкретный класс атак. Сравнивая традиционные алгоритмы машинного обучения, мы видим, что предлагаемая модель превосходит по производительности LR и NB с точки зрения точности, отзыва и FPR. По сравнению с другими моделями BLS, Модель TBLS(W) имеет немного более высокую точность и полноту, чем предлагаемая модель, поскольку для TBLS авторы выбрали для модели набор данных только одного дня, среды. обучение и тестирование. На его долю приходится только 30% от общего набора данных, а типы атак все атаки типа DoS/DDoS, которые лучше сбалансированы по сравнению с общим набором данных, и эффективность обнаружения модели будет в некоторой степени затронута.

Machine Translated by Google



(а) (б)

Рисунок 13. Сравнение с современными методами мультиклассификации: (а) сравнение FPR; (б) сравнение временных затрат.

Таблица 8. Сравнение с современными методами мультиклассификации.

Метод	Точность	Отзывать	ФПР
Наше исследование	0,958	0,947	0,012
ЛР	0,934	0,301	0,057
ГА-АНН	0,286	0,746	0,126
ГСПСО-АНН	0,815	-	-
ТБЛС(Ж) [18]	0,840	-	-
БиГАН [33]	0,978	0,977	-
МФФСЭМ(Вт) [2]	0,823	0,763	0,142
ДНН [34]	0,999	0,999	0,013
МЕКНН [29]	0,946	0,846	-
АдаБуст [35]	0,997	0,791	-
СДАЭ-СВМ [36]	0,889	0,234	-
ГРУ [37]	0,954	0,444	-
ЛСТМ [38]	0,985	0,742	-
КФС-БА(Ж) [39]	0,989	0,748	-
	0,999	0,999	0,120

По сравнению с другими алгоритмами машинного обучения, для многомерных
Механизм объединения и укладки ансамблей функций (MFFSEM) и функция на основе корреляции модели алгоритма выбора летучих мышей (CFS-BA), эффективность обнаружения лучше, чем эта предлагаемой модели, но их FPR выше, чем у предлагаемой модели, что приведет они слишком ошибочно воспринимают нормальное поведение как нападение. Обе модели также обучены. с тем же набором данных только за один день, среду. По сравнению с долгосрочными краткосрочными памяти (LSTM), ее точность немного выше, чем у предлагаемой модели, но ее отзыв намного ниже, чем в предлагаемой модели, а это означает, что модель LSTM не может хорошо распознавать атаки. Таким образом, по сравнению с другими моделями, предлагаемая модель имеет такие преимущества, как лучшая производительность обнаружения и более низкий уровень ложных тревог, когда выполнение задачи мультиклассификации.

5. Проверка модели

Сетевой трафик в реальных средах является большим и сложным; оценить DBELS лучше моделировать и имитировать эффект реальной среды, мы выбрали CICIDS2018 набор данных для тестирования и проверки производительности модели в различных средах данных. Набор данных CICIDS2018 — это широко используемый набор данных в исследованиях кибербезопасности, созданный Канадским институтом исследований кибербезопасности, который моделирует реальную сеть трафик, включая обычный трафик и многие типы атак. По сравнению с CICIDS2017 датасета, его размер увеличился примерно в четыре раза и содержит более восьми миллионов записи дорожного движения; однако здесь также наблюдается серьезный дисбаланс данных. Набор данных содержит более

Machine Translated by Google

восемь миллионов записей трафика с до 80 функциями в каждой выборке. Эти данные охватывают оба обычный трафик и широкий спектр атакующего трафика, причем примеры типов в основном включают Доброкачественные атаки, DoS/DDoS, брутфорс, ботнеты и проникновение. Разнообразие и сложность Типы выборок в наборе данных CICIDS 2018 позволяют лучше отражать реальные киберугрозы, что позволяет всесторонне оценить эффективность и возможности обобщения систем обнаружения вторжений.

Эксперименты по проверке модели с использованием набора данных CICIDS2018 включали предварительную обработку данных и уменьшение размерности данных, при этом форма операций осталась неизменно соответствует CICIDS2017. Модели для их сравнительных экспериментов включают BLS, DT, NB, MLP и CNN. Для обеспечения достоверности экспериментов предварительная обработка данных проводится то же самое, и все эксперименты проводятся в условиях согласия. Оценка Метрики сравнительного эксперимента включали точность, отзыв, частоту ложных срабатываний и Тренировочное время. Производительность различных моделей в задаче бинарной классификации показано в таблице 9, а производительность при выполнении задачи множественной классификации показана. в таблице 10.

Таблица 9. Эффективность различных методов двоичной классификации для набора данных CICIDS2018.

Метод	Метрика оценки			
	Точность	Отзывать	ФПР	Время(а)
Наше исследование	0,981	0,956	0,044	5.295
БЛС	0,958	0,931	0,069	18,826
ДТ	0,917	0,862	0,138	16.770
Примечание:	0,791	0,742	0,258	2.556
МЛП	0,963	0,933	0,067	861,788
Си-Эн-Эн	0,957	0,926	0,074	949,476

Таблица 10. Эффективность различных методов мультиклассификации для набора данных CICIDS2018.

Метод	Метрика оценки			
	Точность	Отзывать	ФПР	Время(а)
Наше исследование	0,946	0,923	0,016	4.731
БЛС	0,953	0,702	0,031	12,729
ДТ	0,891	0,499	0,060	17.266
Примечание:	0,828	0,860	0,047	2,694
МЛП	0,934	0,485	0,046	1025.025
Си-Эн-Эн	0,966	0,785	0,026	916.550

Результаты показывают, что предложенная нами модель превосходит большинство других моделей по точности. Напомним, FPR и затраты времени при различных задачах классификации. Для бинарной классификации задачи, по сравнению с традиционной структурой BLS, точность и полнота улучшены на 0,023. и 0,025 соответственно, а частота ложных тревог снижается на 0,025. По сравнению с традиционные алгоритмы машинного обучения DT и NB, оба они значительно улучшают производительность обнаружения, в то время как уровень ложных тревог значительно снижается. Для алгоритмов глубокого обучения MLP и CNN улучшена производительность обнаружения. и время обучения модели значительно сокращается. Для задачи мультиклассификации предлагаемая модель также обеспечивает высокую производительность обнаружения и малое время обучения модели, и по сравнению с моделью NB она имеет значительное преимущество в производительности обнаружения, хотя время обучения модели немного улучшено. По сравнению с моделью CNN, хотя точность немного снижается на 0,02, она лучше модели CNN с точки зрения времени отзыва, FPR и обучения модели.

Machine Translated by Google

Из приведенных выше экспериментальных результатов можно сделать вывод, что предлагаемая модель по-прежнему может демонстрировать лучшую производительность обнаружения при работе с набором данных CICIDS2018 с более крупным и сложным масштабом, что также указывает на то, что модель имеет определенную степень масштабируемости и адаптируемости. Масштаб этого набора данных значительно увеличен по сравнению с набором данных CICIDS2017; однако предложенная модель по-прежнему может поддерживать низкие затраты времени , что еще раз подтверждает, что модель обладает определенной степенью адаптивности и легкости для реальных сложных сред. Предлагаемая нами модель неизменно демонстрирует высокую производительность обнаружения и низкие затраты времени на различные наборы данных, будь то бинарная или многоклассовая классификация. Эта эффективность имеет решающее значение для реальных приложений, где важно поддерживать высокую точность и скорость. Кроме того, адаптируемость модели к различным средам данных и устойчивость к различным угрозам повышают ее практическую ценность. Его масштабируемость гарантирует, что он сможет обрабатывать растущие объемы данных, что делает его универсальным решением в динамичных средах кибербезопасности.

6. Выводы

В этой статье, основываясь на модели BLS и представляя идею повторного взвешивания EQL v2, мы предлагаем модель DBELS, целью которой является решение проблемы традиционного BLS, сталкивающейся с низкой скоростью обнаружения многомерных и несбалансированных наборов данных в области NIDS. а также низкое признание нападений со стороны меньшинств. Во-первых, модель полностью изучает крупномасштабный набор многомерных данных через DBN, полностью обучает каждую структуру RBM, фиксирует оптимальную структуру и ключевую информацию набора многомерных данных и представляет его как оптимальный набор низкоразмерных данных. Впоследствии модель BELS применяется для изучения классификации набора данных низкой размерности, а степень обучения модели на выборках из класса меньшинства улучшается за счет корректировки положительных и отрицательных факторов градиента для увеличения классификационного веса модели и повысить признание нападений со стороны меньшинства. DBELS полностью протестирован экспериментально с использованием набора данных CICIDS2017, и предлагаемая нами модель превосходит другие алгоритмы по точности, полноте, FPR и времени. Наконец, полезность предлагаемой модели дополнительно тестируется и подтверждается с помощью набора данных CICIDS2018. Результаты показывают, что предложенная нами модель DBELS имеет значительные преимущества для решения проблемы многомерных и несбалансированных данных обнаружения вторжений в сеть, а также смягчает проблемы длительного времени обучения систем обнаружения вторжений и низкого распознавания атак класса меньшинства, что делает ее практический метод обнаружения вторжений. В будущем мы планируем внедрить стратегии инициализации и методы потокового обучения, чтобы улучшить стабильность и производительность обнаружения BLS. Кроме того, мы планируем использовать передовые а

Вклад авторов: Концептуализация, доктор медицинских наук; методология, MD и CS; программное обеспечение, MD, CS и YK; валидация, MD; расследование, CS, MD и SF; ресурсы, XZ; курирование данных, CS и SF; письменное - подготовка первоначального проекта, MD и HX; написание — обзор и редактирование, MD и HX; визуализация, МД и ЮК; надзор, XZ и HX. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Это исследование финансировалось Национальным фондом естественных наук Китая (62276091).

Заявление о доступности данных: данные будут предоставлены по запросу.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов.

Рекомендации

1. Андерсон, Дж. П. Мониторинг и наблюдение угроз компьютерной безопасности; Технический отчет; Компания Джеймса П. Андерсона: Вашингтон, Округ Колумбия, США, 1980 г.

2. Чжан Х.; Ли, Ж.-Л.; Лю, Х.-М.; Донг, К. Механизм многомерного объединения и укладки ансамблей для сети обнаружения вторжений. Будущий генерал. Вычислить. Сист. 2021, 122, 130–143. [Перекрестная ссылка]

3. Сайн, Х.; Пурнами, С.В. Объединение векторной машины поддержки выборки для классификации несбалансированных данных. Процессия Компьютер. наук. 2015, 72, 59–66. [Перекрестная ссылка]

Machine Translated by Google

4. Тапа, Н.; Лю, З.; Кс, ДБ; Гокараджу, Б.; Рой, К. Сравнение моделей машинного обучения и глубокого обучения для сети системы обнаружения вторжений. Будущий Интернет 2020, 12, 167. [CrossRef]
5. Намакшенас, Д.; Яздинежад, А.; Дегантанха, А.; Шривастава, Г. Федеративная квантовая модель обнаружения угроз для потребительского Интернета вещей, сохраняющая конфиденциальность. IEEE Транс. Потребление. Электрон. 2024. [Перекрестная ссылка]
6. Альмиани, М.; Абу Газле, А.; Аль-Рахайфе, А.; Атьеви, С.; Разак, А. Глубокая рекуррентная нейронная сеть для обнаружения вторжений в Интернет вещей система. Симул. Модель. Практика. Теория 2020, 101, 102031. [CrossRef]
7. Ву П. Глубокое обучение для обнаружения сетевых вторжений: распознавание атак с помощью вычислительного интеллекта; UNSW: Сидней, Австралия, 2020 г.
8. Яздинежад, А.; Дегантанха, А.; Каримипур, Х.; Шривастава, Г.; Паризи, Р.М. Надежная модель федеративного обучения, сохраняющая конфиденциальность и защищающая от атак, вызывающих отравление моделей. IEEE Транс. Инф. Криминалистическая безопасность. 2024, 19, 6693–6708. [Перекрестная ссылка]
9. Яздинежад, А.; Дегантанха, А.; Шривастава, Г.; Каримипур, Х.; Паризи, Р.М. Гибридная конфиденциальность, сохраняющая федеративное обучение от нерегулярных пользователей в Интернете вещей следующего поколения. Дж. Сист. Архит. 2024, 148, 103088. [CrossRef]
10. Чен, КП; Лю, З. Широкая система обучения: эффективная и действенная система поэтапного обучения без необходимости глубокого обучения. архитектура. IEEE Транс. Нейронная сеть. Учиться. Сист. 2017, 29, 10–24. [Перекрестная ссылка] [ПабМед]
11. Гонг, Х.; Чжан, Т.; Чен, КП; Лю, З. Обзор исследований широкой системы обучения: алгоритмы, теория и приложения. IEEE Транс. Киберн. 2021, 52, 8922–8950. [Перекрестная ссылка]
12. Ли, З.; Риос, Алжир; Сюй, Г.; Трайкович, Л. Методы машинного обучения для классификации сетевых аномалий и вторжений. В материалах Международного симпозиума IEEE по схемам и системам (ISCAS) 2019 г., Саппоро, Япония, 26–29 мая 2019 г.; стр. 1–5.
13. Ву, Т.; Фан, Х.; Чжу, Х.; Ты, К.; Чжоу, Х.; Хуанг, Х. Система обнаружения вторжений, сочетающая расширенный случайный лес с SMOTE алгоритм. EURASIP J. Adv. Сигнальный процесс. 2022, 2022, 39. [CrossRef]
14. Чжан Б.; Лю, З.; Цзя, Ю.; Рен, Дж.; Чжао, Х. Метод обнаружения сетевых вторжений на основе алгоритма PCA и Байеса. Безопасность. Коммун. Сеть. 2018, 2018, 1914980. [CrossRef]
15. Ли, З.; Батта, П.; Трайкович, Л. Сравнение алгоритмов машинного обучения для обнаружения сетевых вторжений. В материалах Международной конференции IEEE по системам, человеку и кибернетике (SMC) 2018 г., Миядзаки, Япония, 7–10 октября 2018 г.; стр. 4248–4253.
16. Риос, Алма-Ата; Ли, З.; Сюй, Г.; Алонсо, AD; Трайкович Л. Обнаружение сетевых аномалий и вторжений в сетях связи. В материалах 23-й Международной конференции IEEE по интеллектуальным инженерным системам (INES) 2019 г., Гёдёллё, Венгрия, 25–27 апреля 2019 г.; стр. 000029–000034.
17. Риос, Алма-Ата; Ли, З.; Бекшентаева К.; Трайкович, Л. Обнаружение атак типа «отказ в обслуживании» в сетях связи. В материалах Международного симпозиума IEEE по схемам и системам 2020 г. (ISCAS), Севилья, Испания, 12–14 октября 2020 г.; стр. 1–5.
18. Ли, Дж.; Чжан, Х.; Лю, З.; Лю, Ю. Обнаружение сетевых вторжений с помощью системы тройного обучения, основанной на пространственно-временной детализации. Дж. Суперкомпьютер. 2023, 79, 9180–9205. [Перекрестная ссылка]
19. Ахмад, Т.; Азиз, М.Н. Предварительная обработка данных и выбор функций для систем обнаружения вторжений с машинным обучением. ICIC Express Lett. 2019, 13, 93–101.
20. Хао, Х.; Цзян, З.; Сяо, К.; Ван, К.; Яо, Ю.; Лю, Б.; Лю, Дж. Производство большего с меньшими затратами: подход к обнаружению сетевых атак на основе GAN для несбалансированных данных. В материалах 24-й Международной конференции IEEE 2021 года по совместной компьютерной работе в дизайне (CSCWD), Далянь, Китай, 5–7 мая 2021 г.; стр. 384–390.
21. Тан, Дж.; Лу, Х.; Чжан, Г.; Инь, К.; Ли, К. Потери при выравнивании v2: новый подход градиентного баланса для обнаружения объектов с длинным хвостом. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, Нэшвилл, Теннесси, США, 20–25 июня 2021 г.; стр. 1685–1694.
22. Шен З.; Чжан, Ю.; Чен, В. Метод обнаружения вторжений с байесовской классификацией, основанный на слиянии PCA и LDA. Безопасность. Коммун. Сеть. 2019, 2019, 6346708. [CrossRef]
23. Сало, Ф.; Нассиф, АБ; Эссекс, А. Уменьшение размерности с помощью IG-PCA и ансамблевого классификатора для обнаружения сетевых вторжений. Вычислить. Сеть. 2019, 148, 164–175. [Перекрестная ссылка]
24. Хинтон, Джорджия; Осиндеро, С.; Тех, Ю.-В. Алгоритм быстрого обучения для глубоких сетей доверия. Нейронный компьютер. 2006, 18, 1527–1554. [Перекрестная ссылка] [ПабМед]
25. Беларби, О.; Хан, А.; Карнелли, П.; Спиридопулос, Т. Система обнаружения вторжений на основе сетей глубокого доверия. В материалах Международной конференции по науке о кибербезопасности, Симанэ, Япония, 10–12 августа 2022 г.; стр. 377–392.
26. Андрезини, Г.; Аппис, А.; Малерба, Д. Обнаружение вторжений на основе ближайшего кластера с помощью сверточных нейронных сетей. Знать. На основе сист. 2021, 216, 106798. [CrossRef]
27. Ким А.; Парк, М.; Ли, Д.Х. AI-IDS: Применение глубокого обучения для обнаружения веб-вторжений в режиме реального времени. IEEE Доступ 2020, 8, 70245–70261. [Перекрестная ссылка]
28. Таккар, А.; Лохия, Р. Объединение статистической значимости для выбора функций в обнаружении вторжений на основе глубоких нейронных сетей Система. Инф. Fusion 2023, 90, 353–363. [Перекрестная ссылка]
29. Чен Ю.; Лин, К.; Вэй, В.; Джи, Дж.; Вонг, К.-К.; Коэлло, САС Обнаружение вторжений с использованием многоцелевой эволюционной сверточной нейронной сети для вычислений Интернета вещей в тумане. Знать. На основе сист. 2022, 244, 108505. [CrossRef]
30. Таккар, А.; Лохия, Р. Классификация атак на несбалансированные данные о вторжениях в сеть Интернета вещей с использованием глубокого ансамблевого обучения. нейронная сеть. IEEE Internet Things J. 2023, 10, 11888–11895. [Перекрестная ссылка]

Machine Translated by Google

31. Мульянто, М.; Фейсал, М.; Пракоса, Юго-Запад; Леу, Ж.-С. Эффективность потери фокуса для классификации меньшинства при вторжении в сеть системы обнаружения. Симметрия 2020, 13, 4. [CrossRef]
32. Идрисси, МЮ; Алами, Х.; Эль Махдауи, А.; Эль Мекки, А.; Уалил, С.; Яртауи, З.; Беррада, И. Феданиды: Федеративное обучение для системы обнаружения сетевых вторжений на основе аномалий. Экспертная система. Прил. 2023, 234, 121000. [CrossRef]
33. Яо, В.; Ши, Х.; Чжао, Х. Масштабируемое обнаружение вторжений на основе аномалий для безопасного Интернета вещей с использованием генеративно-состязательных сетей в туманной среде. Дж. Сеть. Вычислить. Прил. 2023, 214, 103622. [CrossRef]
34. Хосе Дж.; Хосе Д.В. Алгоритмы глубокого обучения для систем обнаружения вторжений в Интернете вещей с использованием набора данных CIC-IDS 2017. Межд. Дж. Электр. Вычислить. англ. (IJECE) 2023, 13, 1134–1141. [Перекрестная ссылка]
35. Рохас, Р. AdaBoost и суперкубок классификаторов: учебное введение в адаптивное повышение. Свободный университет. Берл. Тех. Респ. 2009, 1, 1–6.
36. Льв, З.; Цяо, Л.; Ли, Дж.; Сонг, Х. Проблемы безопасности в Интернете вещей, основанные на глубоком обучении. IEEE Internet Things J. 2020, 8, 9531–9538. [Перекрестная ссылка]
37. Ассис, М.В.; Карвалью, Л.Ф.; Льорет, Дж.; Проэнса-младший, ML Система глубокого обучения ГРУ против атак в программно-конфигурируемых сетях. Дж. Сеть. Вычислить. Прил. 2021, 177, 102942. [CrossRef]
38. Имрана, Ю.; Сян, Ю.; Али, Л.; Абдул-Рауф, З. Двунаправленный подход глубокого обучения LSTM для обнаружения вторжений. Экспертная система. Прил. 2021, 185, 115524. [CrossRef]
39. Чжоу, Ю.; Ченг, Г.; Цзян, С.; Дай, М. Построение эффективной системы обнаружения вторжений на основе выбора функций и ансамблевого классификатора. Вычислить. Сеть. 2020, 174, 107247. [CrossRef]

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и участникам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.