Программные системы и вычислительные методы
Правильная ссылка на статью:
Алпатов А.Н., Терлоев Э.З., Матчин В.Т. Архитектура трёхмерной свёрточной нейронной сети для детектирования факта фальсификации видеоряда // Программные системы и вычислительные методы. 2024. № 3. DOI: 10.7256/2454-0714.2024.3.70849 EDN: MNOVWB URL: https://nbpublish.com/library_read_article.php?id=70849

Архитектура трёхмерной свёрточной нейронной сети для детектирования факта фальсификации видеоряда

Алпатов Алексей Николаевич

ORCID: 0000-0001-8624-1662

доцент; кафедра ИиППО; МИРЭА - Российский технологический университет

119454, Россия, г. Москва, пр-т Вернадского, 78

aleksej01-91@mail.ru

Терлоев Эмиль Зияудинович

аспирант; кафедра инструментального и прикладного программного обеспечения; МИРЭА — Российский технологический университет

119454, Россия, г. Москва, пр-т Вернадского, 78

emil199@yandex.ru

Матчин Василий Тимофеевич

старший преподаватель; институт информационных технологий; МИРЭА — Российский технологический университет

119454, Россия, г. Москва, пр-т Вернадского, 78

matchin@mirea.ru

Статья из рубрики "Базы знаний, интеллектуальные системы, экспертные системы, системы поддержки принятия решений"

DOI:

10.7256/2454-0714.2024.3.70849

EDN:

MNOVWB

Дата направления статьи в редакцию:

26-05-2024

Дата публикации:

10-06-2024

Аннотация: В статье отражено использование нейросетевых технологий для определения фактов фальсификации содержимого видеорядов. В современном мире новые технологии стали неотъемлемой частью мультимедийной среды, однако их распространение также создало новую угрозу - возможность неправомерного использования для фальсификации содержимого видеорядов. Это приводит к возникновению серьезных проблем, таких как распространение фейковых новостей, дезинформация общества. В научной статье рассматривается данная проблема и определяется необходимость использования нейронных сетей для ее решения. В сравнении с другими существующими моделями и подходами, нейронные сети обладают высокой эффективностью и точностью в обнаружении фальсификации видеоданных благодаря своей способности к извлечению сложных признаков и обучению на больших объемах исходных данных, что особо важно при снижении разрешения анализируемого видеоряда. В рамках данной работы представлена математическая модель идентификации фальсификации аудио и видеоряда в видеозаписи, а также модель на основе трехмерной свёрточной нейронной сети для определения факта фальсификации видеоряда, путём анализа содержимого отдельных кадров. В рамках данной работы было предложено рассмотреть задачу идентификации фальсификатов в видеозаписи, как совместное решение двух задач: идентификации фальсификации аудио- и видеоряда, а сама результирующая задача, была преобразована в классическую задачу классификации. Любая видеозапись может быть отнесена к одной из четырёх групп, описанных в работе. Только видеозаписи, относящиеся к первой группе, считаются аутентичными, а все остальные - сфабрикованными. Для повышения гибкости модели, были добавлены вероятностные классификаторы, что позволяет учитывать степень уверенности в предсказаниях. Особенность полученного решения состоит в возможности настройки пороговых значений, что позволяет адаптировать модель к различным уровням строгости в зависимости от задачи. Для определения сфабрикованных фоторядов предложена архитектура трёхмерной свёрточной нейронной сети, включающей слой предобработки и нейросетевой слой. Полученная модель обладает достаточной степенью точности определения фальсифицированных видеорядов, с учетом значительного понижения разрешения кадров. Апробация модели на тренировочном наборе данных показала долю корректного определения фальсификации видеорядов выше 70%, что заметно лучше угадывания. Несмотря на достаточную точность модель может быть доработана для более существенного увеличения доли корректных предсказаний.

Ключевые слова:

машинное обучение, нейронные сети, свёрточные нейронные сети, фальсификация видео, дипфейки, детектирование дипфейков, фальсификация аудио, предобработка данных, обнаружение аномалий, пакетная нормализация

Введение

В настоящее время большую популярность стали набирать нейронные сети, основной целью которых является генерация изображений и голосовых аудиозаписей. Высокая степень доступности для обыкновенного обывателя делает их более популярными.

Самыми популярными сервисами являются DALL-E от openAI, midjourney, stable diffusion, FaceApp, FaceSwap и подобные им [1] [2]. Для генерации голосовых аудиозаписей используются такие популярные сервисы, как elevenlabs, Microsoft custom neural voice и speechify [3].

В большинстве случаев данные утилиты используются в безобидных целях, для представления получившихся изображений друзьям и знакомым, для публикации на своей странице в социальной сети, для ускорения рабочего процесса в области дизайна или для ускорения процесса создания аудио книг. Несложно представить значительное упрощение рабочих процессов в художественных сферах, в том числе и киноиндустрии. Помимо этого, возможно «воскрешение» умерших актеров при помощи инструментов генерации голосовых аудиозаписей, а также инструментов переноса лица [4].

С другой стороны, данные технологии ставят под вопрос необходимость актеров и художников в кинематографе, а большая доступность делает их более привлекательными инструментами для злоумышленников [5]. Среди сценариев использования инструментов нейросетевой генерации фото и аудио рядов, возможно создание видеозаписи, в котором популярная политическая или медийная личность делает спорное заявление, способное нанести большой репутационный урон. Также возможна кража личности и дальнейшие преступные действия [6].

Примером кражи личности при помощи нейросетевых технологий является случай, произошедший весной 2022-го года, когда на видеохостинговом сервисе YouTube начали появляться трансляции, с участием нейросетевой копии Илона Маска, предлагающей зрителям передать ему свои криптовалютные вложения для получения их обратно с процентами. Пример трансляций представлен на рисунке 1. [7]

Рисунок 1 – Трансляции с участием нейросетевой копией Илона Маска на платформе YouTube[7]

С одной стороны, мошенническая схема достаточно очевидна. С другой стороны, незнающий пользователь может не придать этому значения, так как трансляцию ведет достаточно популярная личность, что повышает доверие пользователя к получаемой информации.

Описание модели идентификации факта фальсификации видеозаписи

Задачу идентификации фальсификации аудио и видеоряда, в рамках данной работы, сведена к классической задачи классификации.

Любая видеозапись будет определена одной из 4-х групп:

1. Фоторяды и аудио ряд видеозаписи аутентичны;

2. Фоторяды видеозаписи сфабрикован, аудио ряд видеозаписи аутентичен;

3. Фоторяды видеозаписи аутентичны, аудио ряд видеозаписи сфабрикован;

4. Фоторяды и аудио ряд видеозаписи сфабрикованы.

Видеозаписи, относящиеся к первой группе, считаются аутентичными, а не относящиеся к ней видеозаписи считаются сфабрикованными.

Обозначим через X фоторяды видеозаписи, а через A - аудиоряд видеозаписи. Выделим два классифкатора C_X и C_A. Первый классификатор определяет подлинность фоторяда (аутентичен или сфабрикован), а классификатор C_A, определяет подлинность аудиоряда (аутентичен или сфабрикован). Тогда $C_X(X) = 1$, если фоторяды аутентичны, иначе $C_X(X) = 0$, то есть фоторяды сфабрикованы. В случае аудиорядов $C_A(A) = 1$, если аудиоряды аутентичны, иначе $C_A(A) = 0$, то есть аудиоряды сфабрикованы. Тогда аутентичность видеозаписи может быть определена как $\Phi_{authenticity}(X,A) = \begin{cases} 1, \text{если } C_X(X) = 1 \text{ и } C_A(A) = 1 \\ 0 \text{ иначе} \end{cases}$.

Тогда процедуру определения группы видеозаписи можно обобщить. Для этого, вначале, определим значения $C_X(X)$ и $C_A(A)$ с помощью классификаторов. Далее сравним результаты с возможными комбинациями для соотнесения с каждой группой. Так, если $(C_X(X), C_A(A)) = (1,1)$, то видеозапись относится к первой группе, а значит она аутентична. Если $(C_X(X), C_A(A)) = (0,1)$, то видеозапись не аутентична, так как сфабрикован видеоряд (относится ко второй группе). Если $(C_X(X), C_A(A)) = (1,0)$, то видеозапись не аутентична, так как сфабрикована аудиодорожка(относится к третьей группе). Иначе, если $(C_X(X), C_A(A)) = (0,0)$, то видеозапись не аутентична, так как сфабрикована аудиодорожка и видеоряд(относится к четвёртой группе). Однако такой «жёсткий» порог для классификации, требует уверенности в том, что модель верна в своих предсказаниях. Для повышения гибкости модифицируем предложенную модель, добавив вероятностные классификаторы.

Пусть $P(C_X(X)) = 1|X$ обозначает вероятность того, что фоторяды аутентичны, и $P(C_A(A)) = 1|A$. Далее Используем байесовский подход для определения совместной вероятности подлинности фоторядов и аудиоряда. Пусть Θ_X и Θ_A являются пороговыми значениями для определения аутентичности фоторядов и аудиоряда соответственно. Тогда подлинность фоторядов и аудиоряда на основе пороговых значений определяется через $\hat{C}_X = 1$, если $P(C_X(X) = 1|X) \geq \Theta_X$, иначе $\hat{C}_X =$ 0. Аналогично, $\hat{C}_A = 1$, если $P(C_A(A) = 1|A) \geq \Theta_A$, иначе $\hat{C}_A =$ 0 Тогда аутентичность видеозаписи может быть определена $\Phi_{authenticity}(X,A) = \begin{cases} 1, \text{если } P(C_X(X) = 1|X) \geq \Theta_X \text{ и } P(C_A(A) = 1|A) \geq \Theta_A \\ 0 \text{ иначе} \end{cases}$.

Возможность настройки пороговых значений Θ_X и Θ_A позволяет адаптировать модель к различным уровням строгости в зависимости от задачи. Например, в критически важных случаях, можно установить высокие пороги для минимизации ложных положительных результатов, что делает предложенную модель более настраиваемой. Это особенно полезно в ситуациях, где данные могут быть неоднозначными или шумными. В свою

очередь, такой подход позволяет улучшить надёжность системы, так как решения принимаются на основе распределения вероятностей, а не на основе одиночного детерминированного результата.

Технически, выявление сфабрикованных видеорядов возможно при помощи анализа кадров и поиска аномалий, при помощи анализа аудио ряда на предмет аномалий или при комбинированном анализе. В данной работе будет рассматриваться только анализ фоторядов.

Для определения сфальфицированных фоторядов можно воспользоваться трёхмерной свёрточной нейронной сетью. Свёрточные слои в нейронной сети позволяют уменьшить размерность входа, тем самым ускоряя процесс обучения. Трёхмерный свёрточный слой имеет размерность N x M x K,

где:

N — количество кадров во временной оси,

M и K — пространственные размерности (высота и ширина кадра).

Одиночный трёхмерный слой в данном случае будет декомпозирован на слой с размерностью 1 x M x K, называемым пространственной свёрткой, и слой с размерностью N x 1 x 1, называемым временной свёрткой. Таким образом достигается уменьшение количества обучаемых параметров, по сравнению с использованием обычного трёхмерного слоя с размерностью N x M x K, а также показывает лучший результат при определении действий на видео [9].

Обозначим входное видеоданные как $X \in \mathbb{R}^{N \times M \times K \times C}$, где C — количество каналов (например, 3 для RGB-видео). Пространственная свёртка, в данном случае, применяется для обработки пространственных признаков каждого кадра, то есть двумерные свёртки применяются к каждому кадру независимо. Тогда, пусть f_{spat} — операция пространственной свёртки с ядром $f_{spat}: \mathbb{R}^{N \times M \times K \times C} \longrightarrow \mathbb{R}^{N \times M' \times K' \times C'}$

Новые высота и ширина каждого кадра после свёртки будут зависеть от размера ядра $H \times W$, шага свёртки (англ.stride) и паддинга (англ.padding). Конкретные значения M' и K' можно вычислить следующим образом (если stride = 1 и padding = 0)

$$K' = K - W + 1$$

Временная же свёртка применяется для обработки временных признаков, то есть одномерные свёртки применяются вдоль временной оси. Пусть f_{temp} — операция временной свёртки с ядром $N' = N - T + 1$

Визуализация такой свёртки показана на рисунке 2.

Рисунок 2 – Визуализация трехмерной свертки с декомпозицией на пространственную и временную

Архитектура искусственной нейронной сети

В качестве набора данных был взят набор ZF DeepFake Dataset[10]. Данный набор состоит из коротких видеозаписей, из которых 199 видеозаписей фальсифицированы и 176 аутентичны (на момент написания статьи).

В представленное технологическое решение состоит из слоя предобработки видеозаписи и нейросетевого слоя. В слое предобработки видео разбивается на 10 кадров и размер каждого кадра уменьшается или увеличивается до размера 224 на 224 пикселей. Входная размерность каждого вектора видеозаписи получается 10 x 224 x 224 x 3, где последняя размерность – цветовые каналы: красный, зеленый и синий.

После предобработки видео векторы передаются в нейросетевую модель. Модель машинного обучения состоит входного слоя; слоя трехмерной свертки с декомпозицией сверток на пространственную и временную с набивкой до выходного размера, 16-ю фильтрами и размером ядра 3 x 7 x 7; слоя пакетной нормализации, слоя кусочно-линейной функции активации (ReLU)[11]; слоя изменения размерности кадров до 112 x 112; остаточного слоя с 32-я фильтрами и ядром размером 3 x 3 x 3; слоя снижения размерности кадров до 64 x 64; остаточного слоя с 64-я фильтрами и ядром размером 3 x 3 x 3; трехмерного слоя субдискретизации на основе среднего значения [12], слоя выравнивания (flatten) и полносвязного слоя с 10-ю выходами. Функция ошибки модели – категориальная перекрестная энтропия с оптимизатором Adam и скоростью обучения 0.0001.

Набор данных для обучения модели состоит из 100 видеозаписей, из которых 50 фальсифицированы и 50 аутентичны. Наборы тестирования и валидации состоят из 40 записей, в каждый набор входят по 20 аутентичных и 20 фальсифицированных. Каждая конкретная видеозапись в наборах не используется более чем в одном из наборов одновременно. Обучение проводится в течение 10 эпох. Структура нейронной сети представлена на рисунке 3.

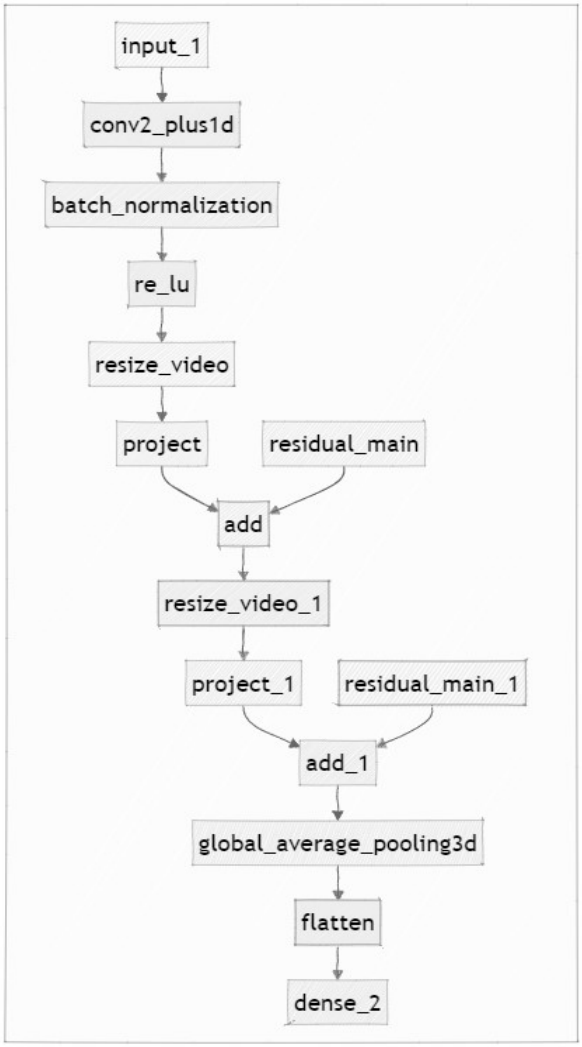

Рисунок 3 – Структура нейронной сети в виде блок-схемы

Тестирование решения

Доля корректных предсказаний (аккуратность) тренировочного набора в последнюю эпоху обучения составляла 75%. Изменение значения функции ошибки с течением обучения представлено на рисунке 4. Изменение значения аккуратности с течением обучения представлено на рисунке 5.

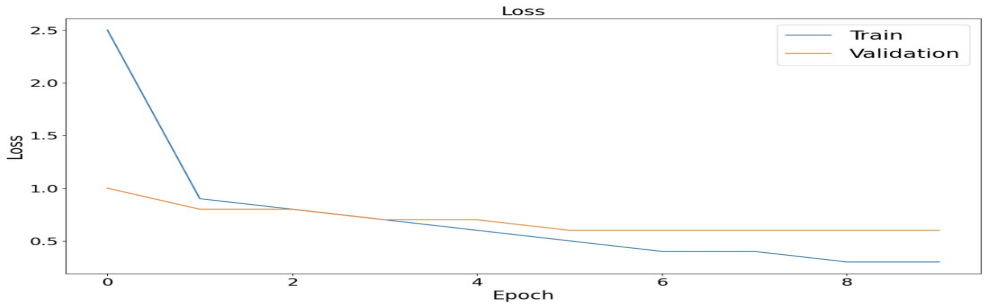

Рисунок 4 – График изменения значения функции ошибки с течением обучения для тренировочного и валидационного набора данных

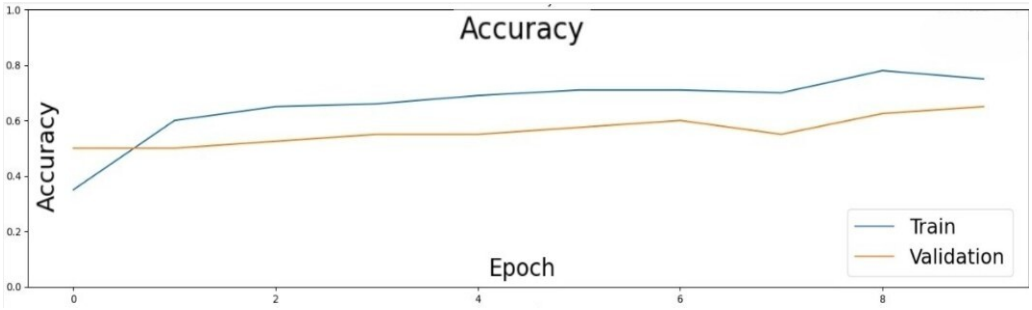

Рисунок 5 – График изменения значения аккуратности с течением обучения для тренировочного и валидационного набора данных

Несмотря на заметное увеличение аккуратности предсказаний на обучающем наборе, на валидационном наборе данных видно только незначительное улучшение. Матрица несоответствий обучающего набора данных, представлена на рисунке 6.

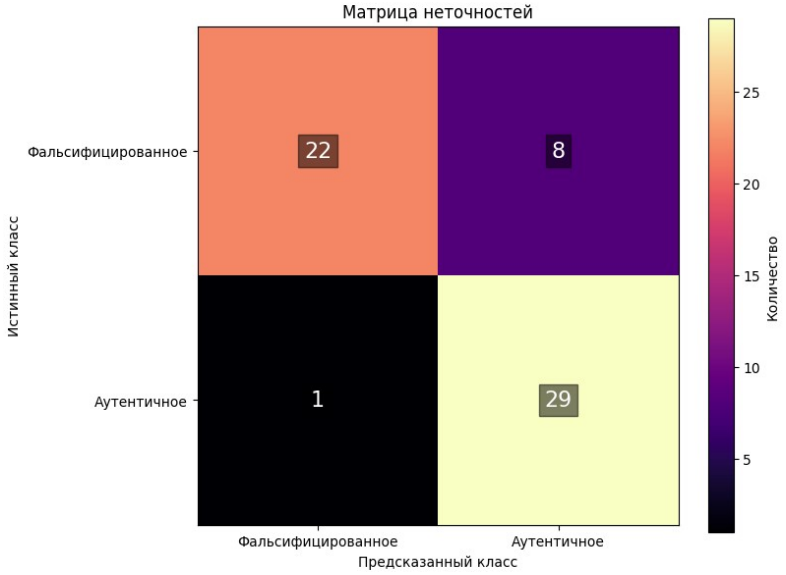

Рисунок 6 – Матрица несоответствий обучающего набора

Исходя из матрицы несоответствий видно, что модель чаще определяет видео, как аутентичное, из-за чего возникает много ложноотрицательных предсказаний. Матрица несоответствий для тестового набора данных представлена на рисунке 7.

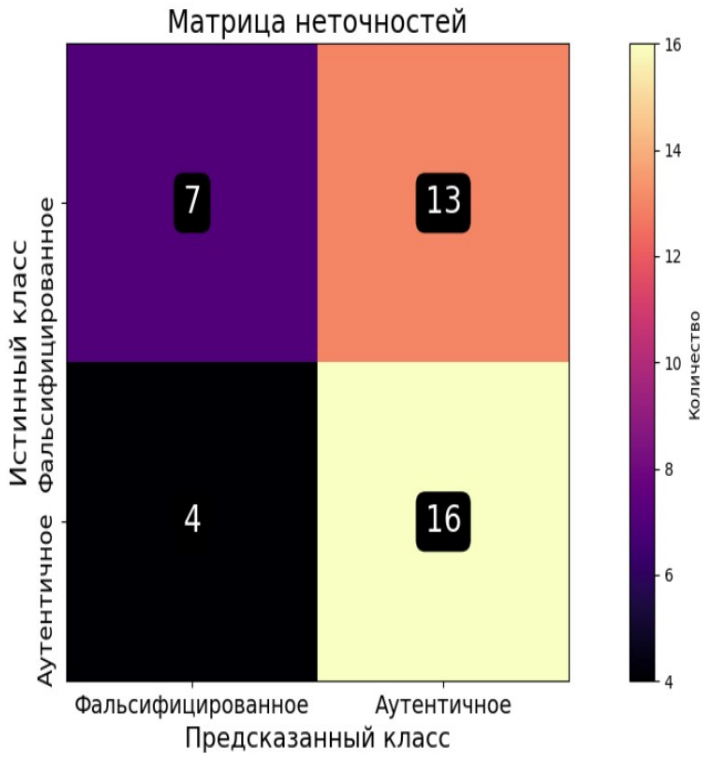

Рисунок 7 – Матрица несоответствий тестового набора

Значения точности, полноты, и F1-меры для возможных классов представлены в таблице 1.

Таблица 1 – Значения точности, полноты, и F1-меры для предсказываемых классов

Класс/метрика	Аутентичное	Фальсифицированное
Точность	0.552	0.6364
Полнота	0.8	0.35
F1-мера	0.653	0.451

Заключение

В настоящей работе представлена нейронная сеть для определения факта фальсификации видеорядов с существенной долей корректного определения. Несмотря на это, модель может быть значительно улучшена путем дополнительного наполнения обучающего набора данных и последующего увеличения доли тренировочного набора; сокращения рабочей области, путем выделения и последующего анализа конкретных зон возможной фальсификации; изменения структуры нейронной сети.

Дальнейшая работа над проблемой может также быть направлена на разработку метода определения факта фальсификации без использования моделей машинного обучения, с целью снижения риска возможных проблем с переобучением и снижения доли корректных предсказаний в случае изменений в технологии фальсификации видеорядов с использованием нейронных сетей.

Библиография

1. Beyan E.V. P., Rossy A.G.C. A review of AI image generator: influences, challenges, and future prospects for architectural field // Journal of Artificial Intelligence in Architecture. 2023. V. 2. №. 1. Pp. 53-65.
2. Huang Y. F., Lv S., Tseng K.K., Tseng P.J., Xie, X., Lin, R.F.Y. Recent advances in artificial intelligence for video production system // Enterprise Information Systems. 2023.

V. 17. №. 11. Pp. 2246188.
3. Albert V. D., Schmidt H. J. AI-based B-to-B brand redesign: A case study // Transfer. 2023. P. 47.
4. Алиев Э. В. Проблемы использования цифровых технологий в киноиндустрии //European Journal of Arts. 2023. No1. С. 33–37. DOI: https://doi.org/10.29013/EJA-23-1-33-37
5. Chow, P. S. Ghost in the (Hollywood) machine: Emergent applications of artificial intelligence in the film industry // NECSUS_European Journal of Media Studies. 2020. V. 9. №. 1. Pp. 193-214.
6. Лемайкина С. В. Проблемы противодействия использования дипфейков в преступных целях // Юристъ-Правоведъ. 2023. No 2(105). С. 143–148.
7. Vakilinia I. Cryptocurrency giveaway scam with youtube live stream // 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). 2022. Pp. 0195-0200.
8. Tran D., Wang H., Torresani L., Ray J., LeCunY., Paluri M. A closer look at spatiotemporal convolutions for action recognition // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018. Pp. 6450-6459.
9. Naik K. J., Soni A. Video classification using 3D convolutional neural network // Advancements in Security and Privacy Initiatives for Multimedia Images. IGI Global. 2021. Pp. 1-18.
10. ZF DeepFake Dataset [Электронный ресурс] URL: https://www.kaggle.com/datasets/zfturbo/zf-deepfake-dataset (дата обращения: 20.01.2024).
11. Garbin C., Zhu X., Marques O. Dropout vs. batch normalization: an empirical study of their impact to deep learning // Multimedia tools and applications. 2020. V. 79. №. 19. Pp. 12777-12815.
12. Zhou D. X. Theory of deep convolutional neural networks: Downsampling // Neural Networks. 2020. V. 124. Pp. 319-327.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

В статье рассматривается разработка и тестирование модели трёхмерной свёрточной нейронной сети (3D CNN) для детектирования факта фальсификации видеоряда. Целью исследования является создание системы, способной эффективно распознавать аутентичные и сфабрикованные видеофайлы.
Методология включает использование 3D CNN, где свёрточные слои декомпозируются на пространственные и временные, что позволяет уменьшить количество обучаемых параметров и улучшить результаты при анализе видео. Датасет ZF DeepFake был использован для обучения и тестирования модели, что обеспечивает достаточную достоверность результатов. Модель обучалась и тестировалась на различных наборах данных, включающих аутентичные и фальсифицированные видеозаписи.
С развитием технологий нейронных сетей и их доступностью для широких масс увеличивается риск использования этих технологий в мошеннических целях, таких как создание дипфейков. Актуальность исследования подчеркивается необходимостью разработки надёжных методов распознавания фальсификаций, что может помочь предотвратить преступления и сохранить репутацию публичных лиц.
Научная новизна работы заключается в предложении усовершенствованной архитектуры

3D CNN для детектирования фальсификаций видеоряда, а также в использовании вероятностного подхода для повышения точности классификации. Предложенная модель позволяет гибко настраивать пороговые значения для различных задач, что делает её универсальной и более точной.

Статья написана в научном стиле с чёткой структурой и логичным изложением материала. Введение подробно описывает текущие проблемы и цели исследования. Описание методологии и архитектуры модели дано детально, что позволяет понять ключевые аспекты работы. Тестирование модели и обсуждение результатов выполнены с использованием графиков и таблиц, что делает выводы прозрачными и понятными.

В заключении авторы подчеркивают эффективность предложенной модели и необходимость дальнейшего её совершенствования. Указывается на возможность улучшения модели за счёт увеличения объёма данных для обучения и изменения архитектуры нейросети. Дальнейшее исследование также может быть направлено на разработку методов детектирования фальсификаций без использования машинного обучения, что может снизить риск переобучения.

Статья будет интересна исследователям в области искусственного интеллекта, компьютерного зрения и информационной безопасности. Представленные результаты могут найти применение в различных областях, включая медиаиндустрию, правовую сферу и кибербезопасность.

Для дальнейшего развития работы предлагаю увеличить объем данных для обучения. Это включает расширение датасета за счет использования большего объема данных для обучения и тестирования модели. Важно рассмотреть использование различных источников данных, включая публичные датасеты и собственные сборы видеозаписей. Также следует диверсифицировать данные, включив различные типы фальсификаций, что позволит более полно представить все возможные сценарии.

Статья представляет собой важный вклад в область детектирования фальсификаций видеоряда и рекомендуется к публикации. Представленные результаты демонстрируют высокий потенциал предложенной модели и её применимость в реальных условиях.

Маленькое замечание: в предложении «Значения точности, полноты, и F1-меры для возможных классов представлены …» перед «и» запятая не нужна.