Б

УДК 004.932.2 **DOI** 10.24412/2413-7383-94-105

М. В. Бабичева, И. А. Третьяков
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Донецкий государственный университет»
283001, Донецкая Народная Республика, г. Донецк, ул. Университетская, 24

АВТОМАТИЗАЦИЯ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ ФАЛЬШИВЫХ ИЗОБРАЖЕНИЙ ПОСРЕДСТВОМ НЕЙРОННЫХ СЕТЕЙ

M. V. Babicheva, I. A. Tretiakov
Federal State Budgetary Educational Institution of Higher Education
"Donetsk State University"
283001, Donetsk People's Republic, Donetsk, st. Universitetskaia, 24

AUTOMATION IS A PROCEDURE FOR DEEPFAKE IMAGE DETECTION USING NEURAL NETWORKS

Проблема распознавания фальшивых изображений (дипфейков) становится актуальной с развитием нейросетевых технологий. Дипфейки зачастую становятся инструментом совершения преступлений против личности и государства, поэтому необходимы технические средства для определения искусственного происхождения изображений. В ходе исследования изучены различные подходы к детектированию фальшивых изображений, включая анализ текстуры изображений, использование нейросетей и алгоритмов обнаружения аномалий. Целью работы является исследование методов создания фальшивых изображений, выявление их особенностей, а также разработка методики обнаружения таких изображений с применением глубокого обучения и нейронных сетей. Для создания дипфейков использовалась генеративная нейронная сеть (GAN), а для распознавания – сверточная (CNN). Предложенная модель распознает фальшивые изображения с точностью 89%, что не хуже большинства зарубежных аналогов.
Ключевые слова: дипфейки, генеративная нейронная сеть, сверточная нейронная сеть, Error Level Analysis

Deepfakes are often used as tools for committing crimes against individuals and states, necessitating technical means to determine the artificial origin of images. This study explores various approaches to detecting fake images, including texture analysis, neural networks, and anomaly detection algorithms. The goal of the research is to investigate methods of creating fake images, identify their distinguishing features, and develop a detection methodology using deep learning and neural networks. For generating deepfakes, a generative adversarial network (GAN) was employed, while a convolutional neural network (CNN) was used for recognition. The proposed model achieves 89% accuracy in detecting fake images, performing on par with most foreign counterparts.
Key words: deepfakes, generative adversarial network (GAN), convolutional neural network (CNN), Error Level Analysis

Введение

Сегодня средства машинного обучения, в частности нейронные сети и глубокое обучение, используются в различных сферах человеческой деятельности, в том числе в сфере информационной безопасности (в части идентификации речи и распознавания по голосу [1], [2], обнаружения сетевых вторжений и атак [3-6], обхода системы CAPTCHA [7], распознаванию спама, фейковых новостей, изображений и видео [8-14]. Одним из самых интересных явлений, порожденных эпохой глубокого обучения, являются фальшивые изображения, или же, как мы привыкли их называть — дипфейки (от английского «deep fake» - глубокая подделка). Дипфейки часто используются для создания реалистичного фото и видео контента. При этом нейронная сеть может сгенерировать искусственный образ или обработать имеющийся по запросу пользователя. Это можно использовать в развлекательных целях или для демонстрации возможностей глубокого обучения. Так, на одной из прямых линий президента России Владимира Владимировича Путина, один из вопросов задал его цифровой двойник, созданный студентом СПБГУ (в соответствии с рисунком 1) [15].

Рисунок 1 – Цифровой двойник задает вопросы президенту на прямой линии

Однако дипфейки используются и для манипулирования общественным мнением путем создания фейковых новостей, дезинформации, мошенничества, шантажа и вымогательства. Особенно это касается публичных людей. Российское законодательство не всегда успевает за развитием технологий, усложняя борьбу с такими угрозами. Сфера злоумышленного использования дипфейков нуждается как в законодательном урегулировании, так и техническом противодействии. Таким образом, создание эффективного инструмента для обнаружения фальшивых изображений является актуальной задачей в сфере информационной безопасности, и именно посредством нейронных сетей можно эффективно отличать настоящие изображения от ими же сгенерированных.

1 Визуальные особенности поддельных изображений

Иногда внимательно рассмотрев изображение можно прийти к выводу, что это фейк, потому что оно содержит различные аномалии и артефакты. Волосы на поддельных изображениях, созданных с использованием генеративных состязательных сетей (GAN), часто имеют характерные визуальные особенности, такие как чрезмерная гладкость, отсутствие детализации объема, неестественные очертания. Это связано с невозможностью искусственных алгоритмов воссоздать сложные мелкие структуры.

На рисунке 2 стрелочками показаны характерные артефакты, которые свидетельствуют об искусственном происхождении изображений. Часто нейросети генерируют нереальный, сюрреалистичный фон. Также характерным признаком является четкая граница между изображением и фоном, которая обычно размыта на настоящих фотографиях [16].

Бывает, что на сгенерированных изображениях присутствует асимметрия глаз, ушей, зубов, разная окраска глаз. Наличие темных или светлых пятен, нереальная окраска бровей. Не всегда эти мелкие неточности заметны осознанному глазу, но мозг человека, настроенный природой на выживание, на интуитивном уровне улавливает эти незначительные нестыковки и дает сигнал – «что-то странное, искусственное».

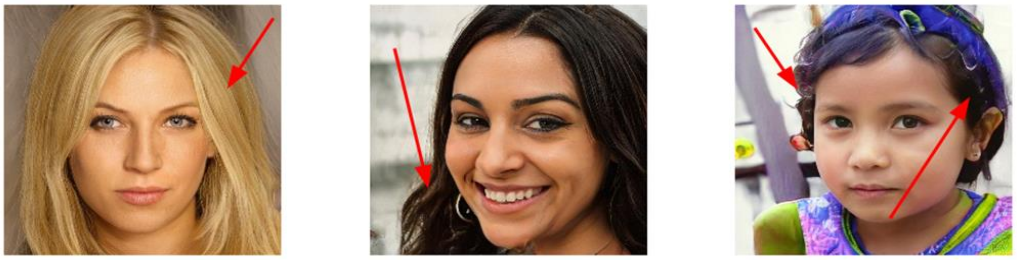

Рисунок 2 – Неестественное изображение волос на фото, сгенерированных нейросетью

Однако алгоритмы совершенствуются и все чаще сгенерированные изображения человеческий глаз и мозг не может отличить от реальной фотографии. На помощь приходит компьютер и цифровая обработка изображений.

2 Методики обнаружения изменений в изображениях

Методики обнаружения изменений в изображениях для фальшивых изображений при помощи компьютерной графики включают в себя широкий спектр подходов, включая анализ артефактов сжатия, анализ текстур и шаблонов, анализ структуры и композиции изображения, а также специализированные методы, такие как Error Level Analysis (ELA).

Анализ артефактов сжатия - основан на изучении аномалий, возникающих в процессе сжатия изображений и характерен для форматов, таких как JPEG. Этот метод анализа сравнивает уровни сжатия различных частей изображения, чтобы выявить любые несоответствия, указывающие на возможные изменения изображением. Он хорошо помогает определять вставки и редактирование.

Анализ текстур и шаблонов выявляет изменения в изображениях путем изучения характеристик их текстур, шаблонов и общей композиции. При изменении изображения текстуры и шаблоны могут быть искажены. Анализ структуры и композиции позволяет выявить неестественные изменения в расположении и форме объектов на изображении.

Однако эти методы менее эффективны, по сравнению с Error Level Analysis (ELA) по нескольким причинам. Во-первых, они требуют более сложных вычислений и алгоритмов обработки изображений и затратны по времени и ресурсам. Во-вторых, эти менее точны при обнаружении изменений. ELA основывается на анализе уровней ошибок сжатия при повторном сохранении изображения. Он сравнивает уровни яркости пикселей в исходном и повторно сжатом изображениях, чтобы выявить аномалии, связанные с манипуляциями или редактированием. Вначале изображение сжимается с использованием определенного метода сжатия, такого как JPEG, что создает базовый уровень ошибок сжатия для каждой области изображения. Затем изображение подвергается манипуляциям и сохраняется повторно с тем же методом сжатия, что создает измененные уровни ошибок сжатия. Метод ELA сравнивает эти уровни

яркости и обнаруживает области с более высокими уровнями ошибок сжатия, что указывает на возможные изменения в изображении. ELA позволяет создать большой объем размеченных данных, где измененные области изображений могут быть выделены и использованы для обучения модели [17]. Это способствует повышению эффективности обучения нейросети и улучшению ее способности распознавать фальшивые изображения. ELA предоставляет возможность автоматического анализа изображений без необходимости вручную размечать их для обучения модели. Это позволяет ускорить процесс обучения и сделать его более масштабируемым.

В работе метод ELA был реализован в программе, что позволило выявить разницу в сгенерированных и реальных изображениях. Эти изображения как создавались вручную, при помощи масок в графическом редакторе, так и генерировались моделью GAN [18]. На рисунке 3 представлены признаки ELA для реальных изображений и для фейковых.

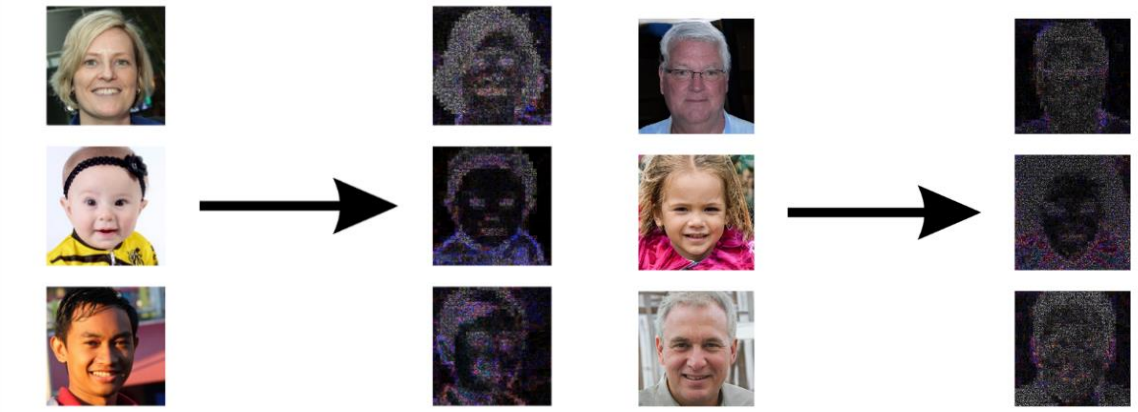

Рисунок 3 – Применение метода ELA, слева – реальные изображения, справа – изображения, сгенерированные нейронной сетью

На ELA для реальных людей присутствует отчетливый контур основных черт лица. На первом и третьем фейковых изображениях лицевые черты абсолютно размыты и не заметны. На втором изображении лицо сильно выбивается из общей картины. Оно было наложено с использованием маски. Такие же результаты можно получить, наложив лицо на фотографию в графическом редакторе и применив к нему метод ELA (в соответствии с рисунком 4).

Рисунок 4 – Метод ELA показывает место, куда было вставлено другое лицо

Можно сделать вывод, что, использование метода ELA в сверточных слоях, позволяет автоматически обнаруживать аномалии в уровнях ошибок сжатия, что является характерным признаком фейковых изображений, созданных, например, с использованием генеративных нейронных сетей (GAN). Экспериментально подтверждено, что данный метод выявляет артефакты и характерные особенности фейковых изображений, созданных как в графическом редакторе, так и сгенерированных нейронной сетью.

3 Построение сверточной нейронной сети

В работе использовалась платформа Kaggle [19]. Это сайт для проведения соревнований по машинному обучению и анализу данных. Она предоставляет участникам доступ к широкому спектру данных, инструментов и ресурсов, необходимых для решения различных задач в области машинного обучения и анализа данных. И гораздо больше мощности, чем среднестатистический компьютер.

Обучение осуществлялось на выборке из 140000 изображений. Этот датасет содержит 70 000 реальных лиц из набора данных Flickr, собранных компанией Nvidia, а также 70 000 фальшивых лиц, отобранных из 1 миллиона фальшивых лиц (сгенерированных с использованием StyleGAN).

Для создания приложения были использованы алгоритмический язык программирования Python и следующие библиотеки: OpenCV, PIL (Python Imaging Library) - для обработки изображений, numpy - для работы с тензорами, subprocess - для запуска новых процессов, подключения к ним и взаимодействия с ними через их стандартные потоки ввода, вывода, scikit-learn - для работы с данными для обучения, Keras – для создания моделей, Matplotlib – для удобного представления результатов.

Импортированный датасет для бучения прогонялся через разработанную ранее программу, которая представляла изображение, в характерном виде после выделения признаков ELA и сохранялся в таком, модифицированном виде (в соответствии с рисунком 5).

Рисунок 5 – Предобработка данных для обучения заключалась в выделении признаков алгоритмом ELA и сохранении в формате jpg

Следующим шагом был выбор архитектуры сети. После серии экспериментов была выбрана следующая архитектура:

1. Слой Conv2D с 32 фильтрами размером (3, 3) и функцией активации ReLU: Этот слой выполняет свертку входного изображения с 32 различными фильтрами, чтобы извлечь различные характеристики изображения, такие как края, текстуры и образцы.

2. Слой MaxPooling2D с размером окна (2, 2): после каждого сверточного слоя следует слой пулинга, который уменьшает размерность признаковых карт, сохраняя важные признаки и уменьшая вычислительную сложность.

3. Повторение слоев свертки и пулинга: этот процесс повторяется еще два раза, увеличивая количество фильтров в сверточных слоях до 64 и 128 соответственно. Каждый следующий слой свертки и пулинга позволяет модели извлекать более абстрактные и сложные признаки из изображений.

4. Слой Flatten: после последнего слоя пулинга применяется операция выравнивания (flatten), которая преобразует трехмерный тензор признаков в одномерный вектор, чтобы передать его на вход полносвязным слоям.

5. Полносвязные слои: затем идут два полносвязных слоя. Первый полносвязный слой имеет 128 нейронов с функцией активации ReLU, который выполняет дальнейшее извлечение признаков и создание более сложных комбинаций признаков. Dropout слой с коэффициентом 0.5 применяется после первого полносвязного слоя для уменьшения переобучения модели. Второй полносвязный слой имеет один нейрон с сигмоидной функцией активации, который выдает предсказание о наличии дипфейка на входном изображении. Для выходного слоя используется сигмоидная функция активации, которая преобразует выходное значение в диапазон от 0 до 1, интерпретируемый как вероятность принадлежности изображения к одному из классов.

Схематично архитектура представлена на рисунке 6.

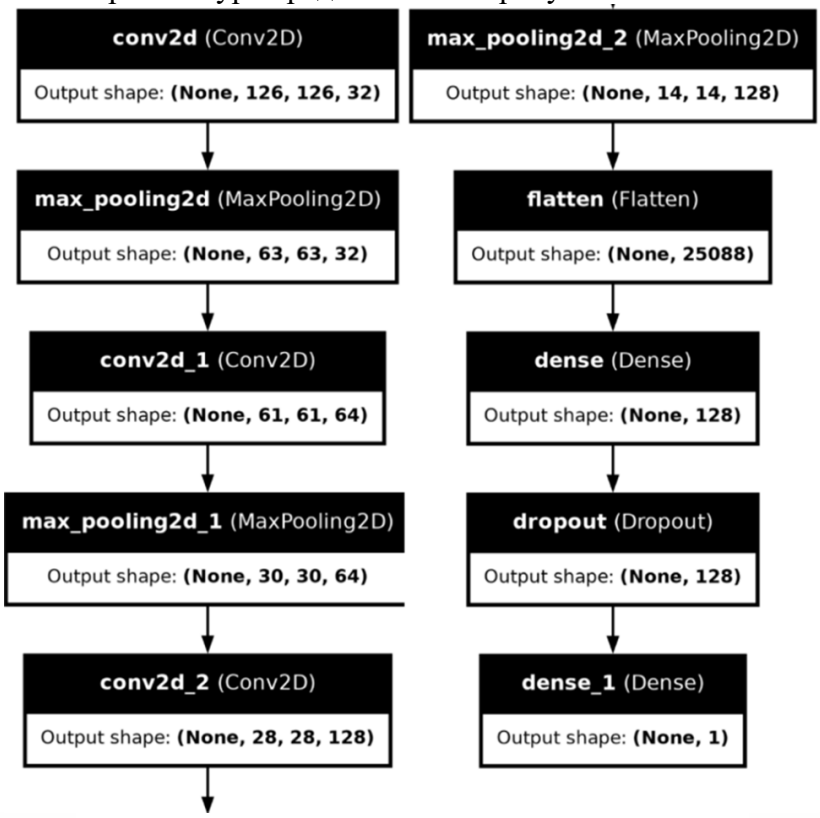

Рисунок 6 – Архитектура сверточной нейронной сети

Метод обучения – стохастический градиентный спуск с оптимизатором Adam, размер бача - 10. Функция потерь - бинарная кроссэнтропия, поскольку нужно было отнести образец к одному из двух классов – реальное изображение или дипфейк. Поскольку датасет для обучения был сбалансирован (количество дипфейков приблизительно равно количеству реальных изображений) для оценки качества обучения использовалась метрика accuracy. На рисунке 7 показано изменение accuracy в процессе обучения.

```
Epoch 31/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 171s 544ms/step - accuracy: 0.8899 - loss: 0.2495 - val_accuracy: 0.8392 - val_loss: 0.3682
Epoch 32/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 173s 551ms/step - accuracy: 0.8960 - loss: 0.2485 - val_accuracy: 0.8372 - val_loss: 0.4100
Epoch 33/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 173s 552ms/step - accuracy: 0.9028 - loss: 0.2378 - val_accuracy: 0.8241 - val_loss: 0.4135
Epoch 34/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 171s 545ms/step - accuracy: 0.9008 - loss: 0.2371 - val_accuracy: 0.8412 - val_loss: 0.4006
Epoch 35/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 171s 544ms/step - accuracy: 0.9034 - loss: 0.2275 - val_accuracy: 0.8342 - val_loss: 0.4319
Epoch 36/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 173s 552ms/step - accuracy: 0.9053 - loss: 0.2291 - val_accuracy: 0.8503 - val_loss: 0.3553
Epoch 37/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 174s 554ms/step - accuracy: 0.9060 - loss: 0.2222 - val_accuracy: 0.8523 - val_loss: 0.3652
Epoch 38/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 174s 554ms/step - accuracy: 0.9150 - loss: 0.2095 - val_accuracy: 0.8448 - val_loss: 0.4183
Epoch 39/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 174s 554ms/step - accuracy: 0.9121 - loss: 0.2090 - val_accuracy: 0.8564 - val_loss: 0.3704
Epoch 40/40
312/312 ━━━━━━━━━━━━━━━━━━━━ 172s 548ms/step - accuracy: 0.9096 - loss: 0.2225 - val_accuracy: 0.8528 - val_loss: 0.4032
```

Рисунок 7 – Изменение accuracy в процессе обучения

Визуализация изменения accuracy и loss во время обучения представлена на рисунке 8.

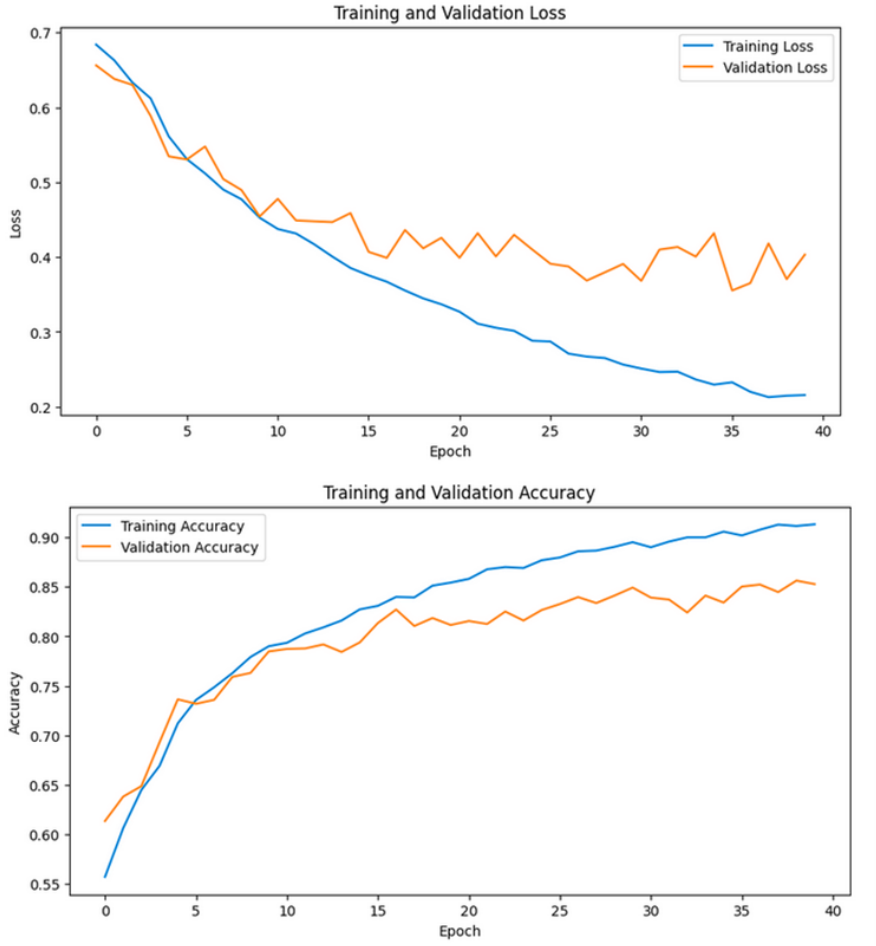

Рисунок 8 – Изменение accuracy и loss для training и validation в процессе обучения

Модель, обученная на 40 эпохах, продемонстрировала хорошую производительность. В валидационном наборе данных accuracy имел более вариативные значения. Тем не менее, модель успешно обучилась отличать фейковые изображения от настоящих с высокой точностью (на тестовой выборке accuracy = 0.89). При этом, стоит отметить, что первоначально наша модель показывала результат 87% на тренировочной и 82% на тестовой. Однако, увеличив размер обучающего датасета и количество эпох с 30 до 40, была достигнута большая эффективность.

4 Анализ результатов

Для сравнения с зарубежными аналогами модели и результаты из опубликованных исследований представлены в таблице 1.

Таблица 1 – Процент распознанных дипфейков для моделей из опубликованных исследований

№	Модель	Разработчик	% расп. дипф.	Год
1	FakeCatcher [20]	Intel	93	2022
2	GANN [21]	Sensity	90	2018
3	DeepFD [22]	National Pingtung University	96	2020
4	MesoNet [23]	National Institute of Informatics Tokyo	87	2022
5	Представленная в данной статье модель	Донецкий государственный университет	89	2024

Как видно из таблицы эффективность самых мощных нейросетей для детектирования дипфейков может быть оценена на уровне примерно 93-96% точности, что означает, что эти модели способны правильно классифицировать 9 из 10 или 19 из 20 изображений как подлинные или фальшивые. Однако эта эффективность может изменяться в зависимости от специфики данных, типов дипфейков и сложности манипуляций с контентом. Все сети, как и созданная нами, основаны на принципах глубокого обучения и сверточных сетей (CNN).

Разработанная модель может быть менее эффективной по сравнению с указанными нейросетями из-за нескольких причин: меньший объем обучающих данных, простота архитектуры, отсутствие дополнительной оптимизации гиперпараметров, отсутствие аугментации. Все эти факторы могут сказаться на точности и надежности нашей модели в обнаружении дипфейков. Однако, для сравнительно простой архитектуры сети и малой ресурсоемкости полученную точность можно считать достаточной.

Заключение

В ходе исследования были изучены различные подходы к детектированию дипфейков, включая анализ текстуры изображений, использование нейросетей и алгоритмов обнаружения аномалий. Основной целью работы было разработать и оценить эффективность модели нейронной сети в обнаружении дипфейков.

Кроме того, был проведен анализ преимуществ и недостатков каждого подхода, что позволило выявить особенности их работы и возможные области улучшений. Например, анализ текстуры изображений позволяет выявить артефакты, харак-

терные для дипфейков, такие как аномальная текстура или несоответствие освещения. Нейросети, в свою очередь, обладают высокой способностью к обучению на больших объемах данных и могут автоматически извлекать признаки из изображений для обнаружения дипфейков.

По итогам компьютерных экспериментов был выбран метод ELA (Error Level Analysis), позволяющий проанализировать разницу между оригинальными и фейковыми изображениями на основе изменения характера шума в сжатых изображениях, была написана программа для проведения тестов, которые подтвердили правильность выбранного метода.

Выбран набор оригинальных и фейковых изображений Kaggle для обучения нейронной сети, который обрабатывался генератором для перемешивания и аугментации выборки. Для тестирования создан набор изображений, которые были полностью сгенерированы сетью GAN и представляли из себя модифицированные реальные изображения. Разработана архитектура сверточной нейронной сети для распознавания дипфейков и проведено ее обучение.

По итогу работы разработан инструмент для распознавания дипфейков на основе сверточной нейронной сети с вероятностью распознавания 89% на тестовой выборке, что не хуже некоторых зарубежных аналогов [19-22].

Были выявлены возможности улучшения, которые могут быть реализованы в будущих исследованиях. Во-первых, важно увеличить объем и разнообразие обучающих данных. Дополнительные данные позволят модели получить более обширное представление о различиях между дипфейками и настоящими изображениями, что в свою очередь повысит ее обобщающую способность. Во-вторых, следует провести оптимизацию архитектуры нейросети. Это может включать в себя изменение числа слоев, размеров фильтров, использование различных функций активации и применение техник регуляризации для предотвращения переобучения. Кроме того, использование предобученных моделей для извлечения признаков может значительно ускорить процесс обучения и повысить точность модели. Это особенно полезно в случае ограниченности ресурсов для обучения нейросети на больших объемах данных.

Таким образом, представленное исследование не только расширяет научное понимание проблемы дипфейков, но и предлагает практические шаги для ее решения, что делает его важным и актуальным в контексте современных вызовов информационной безопасности и доверия к медиаинформации.

Список литературы

1. Третьяков, И. А. Исследование параметров рекуррентной нейронной сети для распознавания человека по голосу в системах безопасности / И. А. Третьяков, Е. Н. Кожекина, А. Е. Мышкин // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2021. – № 4. – С. 24-36. – EDN YUMGIV.
2. Третьяков, И. А. Текстонезависимая идентификация речи в условиях помех / И. А. Третьяков, Е. Н. Кожекина, В. И. Сыровацкий // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2022. – № 2. – С. 64-77. – EDN LUBRVH.
3. Sheluhin, O. I. Comparative analysis of informative features quantity and composition selection methods for the computer attacks classification using the unsw-nb15 dataset / O. I. Sheluhin, V. P. Ivannikova // T-Comm. – 2020. – Vol. 14. – No. 10. – P. 53-60. – DOI: 10.36724/2072-8735-2020-14-10-53-60.
4. Yang, W. Security detection of network intrusion: application of cluster analysis method / W. Yang // Computer Optics. – 2020. –Vol. 44. – No. 4. – P. 660-664. – DOI: 10.18287/2412-6179-CO-657.
5. Классификация механизмов атак и исследование методов защиты систем с использованием алгоритмов машинного обучения и искусственного интеллекта / И. В. Володин, М. М. Путято, А. С. Макарян, В. Ю. Евглевский // Прикаспийский журнал: управление и высокие технологии. – 2021. – № 2(54). – С. 91-98. – DOI 10.21672/2074-1707.2021.53.1.090-098. – EDN KWWOSM.
6. Бабичева, М. В. Применение методов машинного обучения для автоматизированного обнаружения сетевых вторжений / М. В. Бабичева, И. А. Третьяков // Вестник Дагестанского государственного технического университета. Технические науки. – 2023. – Т. 50, № 1. – С. 53-61. – DOI 10.21822/2073-6185-2023-50-1-53-61. – EDN MGBAGF.

7. Третьяков, И. А. Прохождение CAPTCHA посредством машинного обучения / И. А. Третьяков, М. В. Бабичева, К. Е. Лебедев // Искусственный интеллект: теоретические аспекты и практическое применение : материалы Донецкого международного научного круглого стола (Донецк, 30 мая 2024 г.). – Донецк: ФГБНУ «Институт проблем искусственного интеллекта», 2024. – С. 254-260.
8. Ермоленко, Т. В. Фильтрация спама методами глубокого обучения / Т. В. Ермоленко, Н. А. Шалун // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2024. – № 4. – С. 165-174. – DOI 10.5281/zenodo.14514835. – EDN ECVVBZ.
9. Довгаль, В. А. Применение глубокого обучения для создания и обнаружения поддельных изображений, синтезированных с помощью искусственного интеллекта / В. А. Довгаль // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. – 2021. – № 4(291). – С. 82-94. – DOI 10.53598/2410-3225-2021-4-291-82-94. – EDN PPECYM.
10. Джуров, А. А. Программное средство, определяющее фейковый видеоконтент с помощью технологии Deepfake алгоритма GAN / А. А. Джуров, Л. В. Черкесова, Е. А. Ревякина // Наукоемкие технологии в космических исследованиях Земли. – 2023. – Т. 15, № 4. – С. 60-67. – DOI 10.36724/2409-5419-2023-15-4-60-67. – EDN GOAPYT.
11. Алпатов, А. Н. Архитектура трёхмерной свёрточной нейронной сети для детектирования факта фальсификации видеоряда / А. Н. Алпатов, Э. З. Терлоев, В. Т. Матчин // Программные системы и вычислительные методы. – 2024. – № 3. – С. 1-11. – DOI 10.7256/2454-0714.2024.3.70849. – EDN MNOVWB.
12. Зуев, В. М. Сравнение обнаружения объектов средствами искусственного интеллекта в сравнении с классическими методами / В. М. Зуев // Проблемы искусственного интеллекта. – 2024. – № 3(34). – С. 30-35. – DOI 10.24412/2413-7383-2024-3-30-35. – EDN IIZDSH.
13. Шепель, Н. В. Использование возможностей искусственного интеллекта при создании дипфейков / Н. В. Шепель, М. О. Янгаева // Вестник Сибирского юридического института МВД России. – 2024. – № 4(57). – С. 233-239. – EDN INGWYS.
14. Криворучко, К. А. Распознавание недостоверной информации в СМИ с помощью нейронных сетей / К. А. Криворучко, И. И. Максименко // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2024. – № 4. – С. 101-109. – DOI 10.5281/zenodo.14514617. – EDN ZTGSQR.
15. Двойник Путина задал вопрос Путину о двойниках и искусственном интеллекте [Электронный ресурс] – Электрон, дан. – 2023. – Режим доступа: https://www.kp.ru/online/news/5589643/
16. Deepfake Image Detection Using Anchored Pairwise Learning Approach [Электронный ресурс] – Электрон, дан. – 2020. – Режим доступа: https://repository.tudelft.nl/islandora/object/uuid%3Aaa4c6431-6880-406c-8429-af5b04bc3b05
17. FaceForensics++: Learning to Detect Manipulated Facial Images [Электронный ресурс] – Электрон, дан. – 2020. – Режим доступа: https://arxiv.org/pdf/1901.08971.pdf
18. DeepFake Detection [Электронный ресурс] – Электрон, дан. – 2020. – Режим доступа: https://adityaanil.github.io/DeepFake-Detection/
19. A Continual Deepfake Detection Benchmark: Dataset, Methods, and Essentials [Электронный ресурс] – Электрон, дан. – 2020. – Режим доступа: https://coral79.github.io/CDDB_web/
20. Intel Newsroom Archive 2022 [Электронный ресурс] – Электрон, дан. – 2022. – Режим доступа: https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.80elpl
21. Sensity AI [Электронный ресурс] – Электрон, дан. – 2018. – Режим доступа: https://www.cbinsights.com/company/deeptrace
22. Hsu, C.-C. Deep Fake Image Detection Based on Pairwise Learning. / C.-C. Hsu, Y.-X. Zhuang, Lee C.-Y. // Applied Sciences. – 2020. V 1. – 14 p. – DOI 10.3390/app10010370.
23. Deepfake Detection [Электронный ресурс] – Электрон, дан. – 2022. – Режим доступа: https://ucladeepvision.github.io/CS188-Projects-2022Winter/2022/02/04/team18-deepfake-detection.html

References

1. Tretiakov, I. A. Issledovanie parametrov rekurrentnoi neironnoi seti dlia raspoznavaniia cheloveka po golosu v sistemakh bezopasnosti / I. A. Tretiakov. E. N. Kozhekina, A. E. Myshkin // Vestnik Donetskogo natsionalnogo universiteta. Seriia G:Tekhnicheskie nauki. – 2021. – № 4. – S. 24-36. – EDN YUMGIV.
2. Tretiakov, I. A. Tekstonezavisimaia identifikatsiia rechi v usloviiakh pomekh / I. A. Tretiakov, E. N. Kozhekina, V. I. Syrovatskii // Vestnik Donetskogo natsionalnogo universiteta. Seriia G: Tekhnicheskie nauki. – 2022. – № 2. – S. 64-77. – EDN LUBRVH.
3. Sheluhin, O. I. Comparative analysis of informative features quantity and composition selection methods for the computer attacks classification using the unsw-nb15 dataset / O. I. Sheluhin, V. P. Ivannikova // T-Comm. – 2020. – Vol. 14. – No. 10. – P. 53-60. – DOI: 10.36724/2072-8735-2020-14-10-53-60.

4. Yang, W. Security detection of network intrusion: application of cluster analysis method / W. Yang // Computer Optics. – 2020. –Vol. 44. – No. 4. – P. 660-664. – DOI: 10.18287/2412-6179-CO-657.
5. Klassifikatsiia mekhanizmov atak i issledovanie metodov zashchity sistem s ispolzovaniem algoritmov mashinnogo obucheniia i iskusstvennogo intellekta / I. V. Volodin, M. M. Putiato, A. S. Makarian, V. IU. Evglevskii // Prikaspiiskii zhurnal upravlenie i vysokie tekhnologii. – 2021. – № 2(54). – S. 91-98. – DOI 10.21672/2074-1707.2021.53.1.090-098. – EDN KWWOSM.
6. Babicheva, M.V. Primenenie metodov mashinnogo obucheniia dlia avtomatizirovannogo obnaruzheniia setevykh vtorzhenii / M. V. Babicheva, I. A. Tretiakov // Vestnik Dagestanskogo gosudarstvennogo tekhnicheskogo universiteta. Tekhnicheskie nauki. – 2023. – T. 50, № 1. – C. 53-61. – DOI 10.21822/2073-6185-2023-50-1-53-61. – EDN MGBAGF.
7. Tretiakov, I. A. Prokhozhdenie CAPTCHA posredstvom mashinnogo obucheniia / I. A. Tretiakov, M. V. Babicheva, K. E. Lebedev // Iskusstvennyi intellect: teoreticheskie aspekty i prakticheskoe primenenie: materialy Donetskogo mezhdunarodnogo nauchnogo kruglogo stola (Donetsk, 30 may 2024). – Donetsk: FGBNU «Institut problem iskusstvennogo intellekta», 2024. – S. 254-260.
8. Ermolenko, T. V. Filtratsiia spama metodami glubokogo obucheniia / T. V. Ermolenko, N. A. SHalun // Vestnik Donetskogo natsionalnogo universiteta. Seriia G: Tekhnicheskie nauki. – 2024. – № 4. – C. 165-174. – DOI 10.5281/zenodo.14514835. – EDN ECVVBZ.
9. Dovgal, V. A. Primenenie glubokogo obucheniia dlia sozdaniia i obnaruzheniia poddelnykh izobrazhenii sintezirovannykh s pomoshchiu iskusstvennogo intellekta / V. A. Dovgal // Vestnik Adygeiskogo gosudarstvennogo universiteta. Seriia 4: Estestvenno-matematicheskie i tekhnicheskie nauki. – 2021. – № 4(291). – S. 82-94. – DOI 10.53598/2410-3225-2021-4-291-82-94. – EDN PPECYM.
10. Dzhurov, A. A. Programmnoe sredstvo opredeliaiushchee feikovyi videokontent s pomoshchiu tekhnologii Deepfake algoritma GAN / A. A. Dzhurov, L. V. CHerkesova, E. A. Reviakina // Naukoemkie tekhnologii v kosmicheskikh issledovaniiakh Zemli. – 2023. – T. 15, № 4. – S. 60-67. – DOI 10.36724/2409-5419-2023-15-4-60-67. – EDN GOAPYT.
11. Alpatov, A. N. Arkhitektura trekhmernoi svertochnoi neironnoi seti dlia detektirovaniia fakta falsifikatsii videoriada / A. N. Alpatov, E. Z. Terloev, V. T. // Matchin Programmnye sistemy i vychislitelnye metody. – 2024. – № 3. – S 1-11. – DOI 10.7256/2454-0714.2024.3.70849. – EDN MNOVWB.
12. Zuev, V. M. Sravnenie obnaruzheniia obieektov sredstvami iskusstvennogo intellekta v sravnenii s klassicheskimi metodami / V. M. Zuev // Problemy iskusstvennogo intellekta. – 2024. – № 3(34). – S. 30-35. – DOI 10.24412/2413-7383-2024-3-30-35. – EDN IIZDSH.
13. SHepel, N. V. Ispolzovanie vozmozhnostei iskusstvennogo intellekta pri sozdanii dipfeikov / N. V. SHepel, M. O. IAngaeva // Vestnik Sibirskogo iuridicheskogo instituta MVD Rossii. – 2024. – № 4(57). – S. 233-239. – EDN INGWYS.
14. Krivoruchko, K. A. Raspoznavanie nedostovernoi informatsii v SMI s pomoshchiu neironnykh setei / K. A. Krivoruchko, I. I. Maksimenko // Vestnik Donetskogo natsionalnogo universiteta. Seriia G: Tekhnicheskie nauki. – 2024. – № 4. – S. 101-109. – DOI 10.5281/zenodo.14514617. – EDN ZTGSQR.
15. Putin's doppelganger asked Putin a question about doppelgangers and artificial intelligence [Electronic resource] – Electron, dan. – 2023. – Access mode: https://www.kp.ru/online/news/5589643/
16. Deepfake Image Detection Using Anchored Pairwise Learning Approach [Electronic resource] – Electron, dan. – 2020. – Access mode: https://repository.tudelft.nl/islandora/object/uuid%3Aaa4c6431-6880-406c-8429-af5b04bc3b05
17. FaceForensics++: Learning to Detect Manipulated Facial Images [Electronic resource] – Electron, dan. – 2020. – Access mode: https://arxiv.org/pdf/1901.08971.pdf
18. DeepFake Detection [Electronic resource] – Electron, dan. – 2020. – Access mode: https://adityaanil.github.io/DeepFake-Detection/
19. A Continuous Deepfake Detection Benchmark: Dataset, Methods, and Essentials [Electronic resource] – Electron, dan. – 2020. – Access mode: https://coral79 .github.io/CDDB_web/
20. Intel Newsroom Archive 2022 [Electronic resource] – Electron, dan. – 2022. – Access mode: https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.80elpl
21. Sensity AI [Electronic resource] – Electron, dan. – 2018. – Access mode: https://www.cbinsights.com/company/deeptrace
22. Hsu, C.-C. Deep Fake Image Detection Based on Pairwise Learning. / C.-C. Hsu, Y.-X. Zhuang, Lee C.-Y. // Applied Sciences. – 2020. V 1. – 14 p. – DOI 10.3390/app10010370.
23. Deepfake Detection [Electronic resource] – Electron, dan. – 2022. – Access mode: https://ucladeepvision.github.io/CS188-Projects-2022Winter/2022/02/04/team18-deepfake-detection.html

RESUME

M. V. Babicheva, I. A. Tretiakov

Automation is a procedure for deepfake image detection using neural networks

The article discusses the problems arising in connection with the expansion of opportunities for generating deepfakes, fake images that can be used by intruders for blackmail, extortion, and political discredit, which ultimately harms citizens and the state.

The study presents various approaches to detecting deepfakes, including image texture analysis, the use of neural networks and anomaly detection algorithms. A comparative analysis was carried out and, based on the results of computational experiments, it was concluded that the most promising method is the ELA method, which allows analyzing the difference between artifacts of original and fake images. A set of original and fake Kaggle images has been selected for neural network training, which has been preprocessed to identify signs of ELA.

The architecture of a convolutional neural network for deepfake recognition has been developed, which showed an 89% probability of recognizing fake images in a test sample, no worse than some foreign analogues. Steps to further improve the proposed model are proposed.

РЕЗЮМЕ

М. В. Бабичева, И. А. Третьяков

Автоматизация процедуры распознавания фальшивых изображений посредством нейронных сетей

В статье рассмотрены проблемы, которые возникают в связи с расширением возможностей генерирования дипфейков, поддельных изображений, которые могут использоваться злоумышленниками для шантажа, вымогательства, политической дискредитации, что в итоге наносит вред гражданам и государству.

В исследовании представлены различные подходы к детектированию дипфейков, включая анализ текстуры изображений, использование нейросетей и алгоритмов обнаружения аномалий. Проведен сравнительный анализ и по итогам вычислительных экспериментов сделан вывод, что наиболее перспективным является метод ELA, позволяющий проанализировать разницу между артефактами оригинальных и фейковых изображений. Выбран набор оригинальных и фейковых изображений Kaggle для обучения нейронной сети, который подвергся предварительной обработке с выявлением признаков ELA.

Разработана архитектура сверточной нейронной сети для распознавания дипфейков которая показала вероятность распознавания фальшивых изображений 89% на тестовой выборке, не хуже некоторых зарубежных аналогов. Предложены шаги по дальнейшему улучшению предложенной модели.

Бабичева М. В. – кандидат технических наук, доцент кафедры радиофизики и инфокоммуникационных технологий ФГБОУ ВО «Донецкий государственный университет», 283001, г. Донецк, ул. Университетская, 24, m.babicheva60@mail.ru. *Область научных интересов*: информационная безопасность, пентестинг, нейронные сети.

Третьяков И. А. – кандидат технических наук, доцент, доцент кафедры радиофизики и инфокоммуникационных технологий ФГБОУ ВО «Донецкий государственный университет», 283001, г. Донецк, ул. Университетская, 24, i.tretiakov@mail.ru.

Область научных интересов: автоматизация научных исследований и автоматизированные системы; оптические информационные технологии; методы и системы защиты информации, информационная безопасность.

Статья поступила в редакцию 17.01.2025.