ПРИМЕНЕНИЕ МОДЕЛИ MESONET ДЛЯ ОБНАРУЖЕНИЯ ДИПФЕЙКОВ

Микунов А.В., аспирант, ФГБОУ ВО ОмГУПС
Елизаров Д.А., к.т.н., доцент, ФГБОУ ВО ОмГУПС

Аннотация. В статье описывается возможность применения модели MesoNet при распознавании дипфейков.

Ключевые слова: технология дипфейк, генеративный искусственный интеллект, распознавание дипфейков, киберугроза.

Современные технологии генеративного искусственного интеллекта (ИИ) открыли новые возможности создания фотореалистичного медиаконтента. Однако вместе с креативным потенциалом появились и серьезные угрозы, связанные с распространением дипфейков – синтетических изображений и видео, созданных с помощью ИИ.

Для распознавания дипфейков на основе изображений используется специальных программ: проект с открытым исходным кодом MesoNet, платформы Reality Defender или Sentinel.

Проект MesoNet предназначен для обнаружения дипфейков, созданных с помощью методов ИИ. Основным датасетом для обучения в модели используется FaceForensics++, состоящий из 1000 оригинальных видео и их дипфейк-версии. Для оценки на более качественных дипфейках используется датасет Celeb-DF, включающий 590 оригинальных и 5639 поддельных видео высокого качества. Также в исследованиях применялся ранний датасет UADFV с 49 реальными и 49 фейковыми видео, который помог установить базовые показатели эффективности.Ключевыми особенностями этих датасетов являются сбалансированное количество реальных и поддельных изображений, разнообразие методов генерации (от простых до продвинутых), а также фокус именно на лицевых изображениях, которые чаще всего становятся объектами для дипфейк-атак.

Для улучшения обучения в MesoNet применяются различные методы аугментации данных, включая горизонтальное отражение, небольшие повороты и другие преобразования, что позволило увеличить разнообразие обучающей выборки без сбора новых данных. Также использовались техники балансировки классов и кросс-валидации на разных датасетах для обеспечения устойчивости результатов.

Далее представляем анализ кейса для демонстрации практических аспектов детектирования.В работе был развернут проект с использованием модели MesoNetдля обнаружения дипфейков. Модель анализирует входные изображения и определяет вероятность того, что они были сгенерированы или изменены алгоритмами, а не являются подлинными.

Алгоритм работы модели MesoNet:

– импорт необходимых библиотек;

– создание архитектуры MesoNet;

– загрузка веса модели (если есть);

– загрузка изображений;

– преобразование изображений;

– подача изображенийв модель;

– анализ и вывод результатов.

В начале работы импортируем необходимые библиотеки: numpy – для работы с массивами, tensorflow и keras – для нейросети, слои Conv2D, Dense, Dropout– для построения модели, Adam – оптимизатор обучения.

Архитектура MesoNet состоит из нескольких сверточных слоев, которые последовательно обрабатывают изображение, выделяя важные признаки. На вход модель принимает изображение размером 256x256 пикселей в формате RGB.

Первые слои – это сверточные блоки – последовательности слоев в нейросети, которые обрабатывают изображение, постепенно выделяя из него важные признаки, каждый из которых включает операцию двумерной свертки (Conv2D) с ядрами размером 3x3 или 5x5, за которой следует нормализация

(BatchNormalization) и операция максимизирующего пулинга (MaxPooling2D) для уменьшения размерности. Эти блоки помогают модели выявлять низкоуровневые артефакты, такие как неестественные текстуры, аномалии в цвете или искажения на границах объектов.

После сверточных слоев данные преобразуются в одномерный вектор с помощью слоя Flatten, а затем проходят через полносвязные слои – классические слои нейронной сети, где каждый нейрон связан со всеми нейронами предыдущего слоя (Dense). Первый полносвязный слой содержит 16 нейронов с активацией ReLU, за ним следует слой Dropout с вероятностью 0.5, что помогает предотвратить переобучение. Финальный слой с одним нейроном и сигмоидальной активацией выдает вероятность того, что изображение является дипфейком.

Для работы модели сначала загружаются предобученные веса –числовые параметры сети, которые она обучается подбирать в процессе тренировки, и их можно представить как коэффициенты важности связей между нейронами (если они есть), а затем изображение предварительно обрабатывается: изменяется размер до 256x256, нормализуются значения пикселей (деление на 255), и данные преобразуются в формат, подходящий для подачи в нейросеть.

После этого модель делает предсказание, и, если вероятность превышает пороговое значение (по умолчанию 0.5), изображение считается поддельным.

MesoNet определяет дипфейки, основываясь на артефактах, которые часто оставляют методы синтеза изображений, такие как нереалистичные текстуры, размытия, нарушения в освещении или неестественные границы объектов. Однако эффективность модели зависит от того, на каких данных она обучалась: если дипфейки созданы новыми методами (например, Stable Diffusion или MidJourney), модель может потребовать дообучения. Кроме того, поскольку модель работает с фиксированным размером изображения, для анализа изображений другого разрешения их необходимо масштабировать, что может повлиять на точность.

При тестировании на настоящем изображении программа выдала заключение: изображение скорее всего настоящее (рисунке 1). При тестировании на поддельном изображении – изображение скорее всего поддельное (рисунок 2).

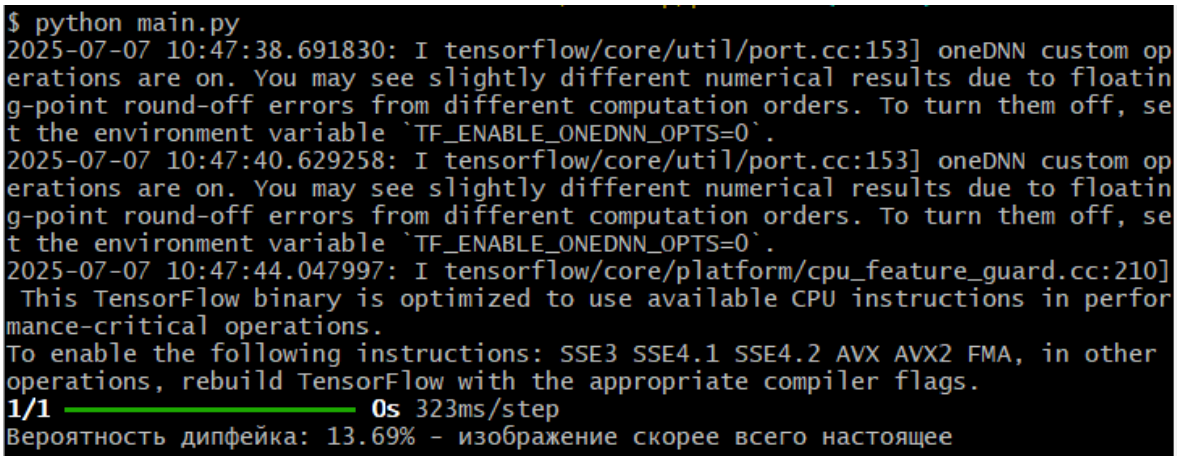

Рисунок 1 – Тестирование программы на настоящем изображении

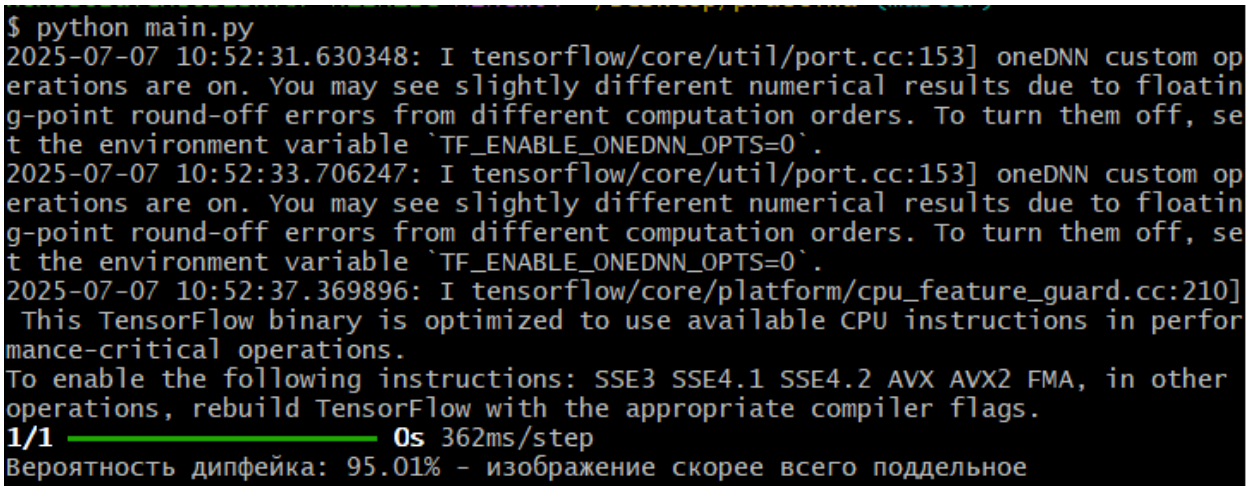

Рисунок 2 – Тестирование программы на поддельном изображении

Проект с открытым исходным кодом MesoNet можно использовать в качестве инструмента для обнаружения дипфейков. К признакам, позволяющим осуществлять детектирование, можно отнести следующие:

артефакты сжатия. DeepFake-изображения часто подвергаются повторному сжатию, что приводит к появлению характерных артефактов в текстуре кожи, волосах или фоне;

аномалии в текстуре кожи. GAN-синтезированные лица могут иметь неестественную гладкость или локальные искажения в текстуре кожи. Отсутствие микроморщин, пор или волосков, которые есть в реальных фотографиях;

несовершенства в области глаз и зубов. Глаза в DeepFake могут иметь неестественные блики, асимметрию или артефакты вокруг век. Зубы могут выглядеть размытыми или иметь неправильную геометрию;

геометрические несоответствия. Легкие искажения в пропорциях лица (например, неправильная форма носа или подбородка). Артефакты на границах лица (особенно в местах «вставки» синтезированного лица в исходное видео);

Модель MesoNet пропускает и не может детектировать:

современные генеративные модели (StyleGAN3, Diffusion Models) – из-за меньшего количества артефактов;

сложные семантические и анатомические ошибки. Модель демонстрирует слабые результаты в обнаружении:

1) логических ошибок в сцене (неправильных отражений в глазах и очках, физически невозможных теней и освещения, некорректных взаимодействий объектов в кадре);

2) анатомических аномалий (неправильного количества или формы зубов, неестественной формы ушей и других черт лица, нарушений физики волос и их взаимодействия с другими объектами).

уязвимость к профессиональной постобработке. Профессионально созданные DeepFake-изображения часто включают использование сложных масок (тщательную ручную постобработку: ретушь проблемных областей, шумоподавление и цветокоррекцию);

уязвимость к Adversarial-атакам. Современные методы обхода детекции включают: искажения, незаметные для человеческого глаза, точечные изменения ключевых пикселей и манипуляции с цветовыми каналам.

Для повышения точности обнаружения дипфейков рекомендуется проводить дополнительное обучение модели на актуальных данных и комбинировать ее с другими методами анализа.

Литература

1 Артем Опарии: Что такое дипфейк: как создать и зачем использовать в рекламе[Электронный ресурс]. – URL: https://oparinseo.ru/blog/article/cto-takoe-dipfejk-kak-sozdat-i-zacem-ispolzovat-v-reklame/(дата обращения: 16.11.2025)

2 akool: Обнаружение дипфейков[Электронный ресурс]. – URL: https://akool.com/ru/knowledge-base-article/deepfake-detection (дата обращения: 16.11.2025)

3data–light: Детекция объектов в компьютерном зрении[Электронныйресурс].–URL:https://data-light.ru/blog/detekciya-obektov-v-kompyuternom-zrenii/ (дата обращения: 16.11.2025)

4Елизаров, Д. А. К вопросу о применении генеративного искусственного интеллекта / Д. А. Елизаров, К. Д. Гуторов, Д. А. Ковальская // Медиабудущее: искусственный интеллект как вызов ноосфере: Материалы Всероссийской научно-практической конференции, Липецк, 2025. – С. 119-123. – DOI 10.18334/9785912925672.119-123. – EDN JBCTMX.

5 Гуселетова, А. Е. Инструменты обнаружения дипфейков / А. Е. Гуселетова, Д. А. Елизаров // Актуальные проблемы и тенденции развития современной экономики и информатики: Материалы Международной научно-практической конференции, Бирск, 04–06 декабря 2024 года. – Бирск: Уфимский университет науки и технологий, 2024. – С. 177-180. – EDNBWWETK.