ИНСТРУМЕНТЫ ОБНАРУЖЕНИЯ ДИПФЕЙКОВ

Гуселетова А.Е., студент III курса, ФГБОУ ВО ОмГУПС
Елизаров Д.А., к.т.н., доцент, ФГБОУ ВО ОмГУПС

Аннотация. Статья посвящена анализу методов и инструментов обнаружения дипфейков.

Ключевые слова:технология дипфейк, генеративный искусственный интеллект, дезинформация.

Технология дипфейк (deepfake) упрощает процесс синтезирования изображения и создания звуковых дорожек с заданными параметрами за счет использования нейронных сетей. Они обучаются на сотнях или даже тысячах примеров лиц и голосов, с ними связанными.

В большинстве случаев в основе метода лежат генеративно-состязательные нейросети (GAN). При использовании данного метода работают две нейросети. Первая из них (генеративная, Generator, G) генерирует изображения, а вторая (дискриминативная, Discriminator, D) отвечает за поиск отличий между ними и настоящими образцами.

Используя набор переменных латентного пространства, генеративная сеть пытается «слепить» новый образец, смешав несколько исходных образцов. Дискриминативная сеть обучается различать подлинные и поддельные образцы, а результаты различения подаются на вход генеративной сети так, чтобы она смогла подобрать лучший набор латентных параметров, и дискриминативная сеть уже не смогла бы отличить подлинные образцы от поддельных. Таким образом, целью сети G является повысить процент ошибок сети D, а целью сети D является, наоборот, улучшение точности распознавания.

Дискриминативная сеть D, анализируя образцы из оригинальных данных и из подделанных генератором, достигает некоторой точности различения. Генератор при этом начинает со случайных комбинаций параметров латентного

пространства, а после оценки полученных образцов сетью D, применяется метод обратного распространения ошибки, который позволяет улучшить качество генерации, подправив входной набор латентных параметров. Постепенно искусственные изображения на выходе генеративной сети становятся всё более качественными. Сеть D реализуется как свёрточная нейронная сеть, в то время как сеть G наоборот разворачивает изображение на базе скрытых параметров.

Далее рассмотрим особенности работы инструментов и методов обнаружения дипфейков. Каждый инструмент и метод предлагает уникальный подход к обнаружению дипфейков: от анализа тонких элементов видео в градациях серого до отслеживания выражений лица и движений субъектов.

Платформа Sentinel позволяет противостоять угрозе дипфейков на основе искусственного интеллекта (ИИ) [1]. Пользователь может загрузитьаудио- и видеофайлы через веб-сайт или API. Технология обнаружения дипфейков Sentinel предназначена для защиты целостности цифровых носителей. Она использует алгоритмы искусственного интеллекта для анализа загруженного мультимедиа и определения наличия процесса синтезирования. Система предоставляет подробный отчет о своих выводах, включая визуализацию областей носителя, которые были изменены. Это позволяет пользователям точно видеть, где и как манипулировали медиа.

Intel представила Детектор дипфейков FakeCatcher от Intel в реальном времени может распозновать поддельные видео с точностью 96%, возвращая результаты за миллисекунды[2]. FakeCatcher ищет подлинные подсказки в реальных видео, оценивая «кровяной поток» в пикселях видео. Когда сердце перекачивает кровь, вены меняют цвет. Эти сигналы кровотока собираются со всего лица, и алгоритмы переводят эти сигналы в пространственно-временные карты. Затем, используя алгоритмы глубокого обучения, можно определить, является ли видео настоящим или фальшивым.

Проект WeVerify направлен на разработку интеллектуальных методов и инструментов для проверки контента и анализа дезинформации в социальных сетях и веб-контенте [3]. Это достигается за счет кросс-модальной проверки

контента, анализа социальных сетей, микроцелевого разоблачения и общедоступной базы данных известных подделок на основе блокчейна.

Инструмент Microsoft Video Authenticator Tool анализирует неподвижное фото или видео, чтобы предоставить оценку достоверности, указывающую, были ли манипуляции с мультимедиа [4]. Он обнаруживает границу смешивания дипфейковых и тонких элементов в градациях серого, которые не видны человеческому глазу. Инструмент обеспечивает оценку достоверности в реальном времени, позволяя пользователям быстро определить, является ли исследуемый файл подлинным или нет.

Метод обнаружения дипфейков с использованием несоответствий фонемы и виземы, разработанный исследователями из Стэнфордского университета и Калифорнийского университета, использует тот факт, что виземы, которые обозначают динамику формы рта, иногда отличаются или несовместимы с произносимой фонемой [4]. Это несоответствие является распространенным недостатком дипфейков, поскольку ИИ часто изо всех сил пытается идеально сопоставить движения рта с произнесенными словами.Техника несоответствия фонемы и виземы использует алгоритмы искусственного интеллекта для анализа видео и обнаружения этих несоответствий. Он сравнивает движения рта (виземы) с произносимыми словами (фонемами) и ищет любые несоответствия. Если обнаружено несоответствие, это явный признак того, что видео является дипфейком.

Основной задачей технологии на основе открытого исходного кода Deepware – выявлять видео, созданные нейросетью[5]. На веб-сайте есть сканер, в который можно загружать видео, чтобы понять, были ли они обработаны или сгенерированы ИИ. Модели Deepware ищут признаки обработки в изображениях человеческих лиц. Главное ограничение этого инструмента – в его неспособности выявлять техники изменения голоса, а манипуляции с голосом представляют гораздо более серьезную угрозу, чем подмена лиц.

На практике рассмотрим работу с открытым сервисом проверки дипфеков Deepware. Главная страница сайта предлагает проверить видео на достоверность по ссылке, а также предоставляет собственное API для внедрения сервиса в свои продукты (рисунок 1). Данный сервис запрашивает проверку у четырех моделей для большей надёжности.

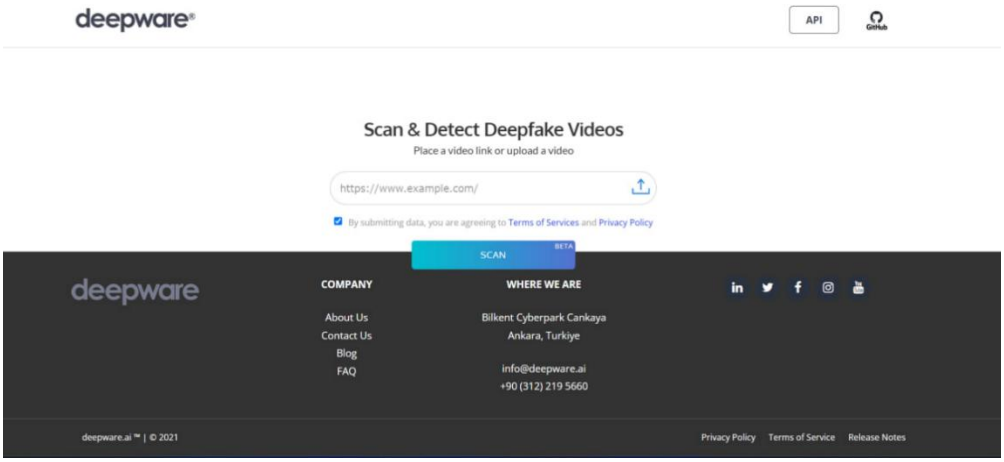

Рисунок 1 – Главная страница сервиса

Для первой проверки было загружено оригинальное видео без обработки. Сервис выдал результаты четырёх моделей, где три из них показали, что видео не дипфейк. Результаты представлены на рисунке 2.

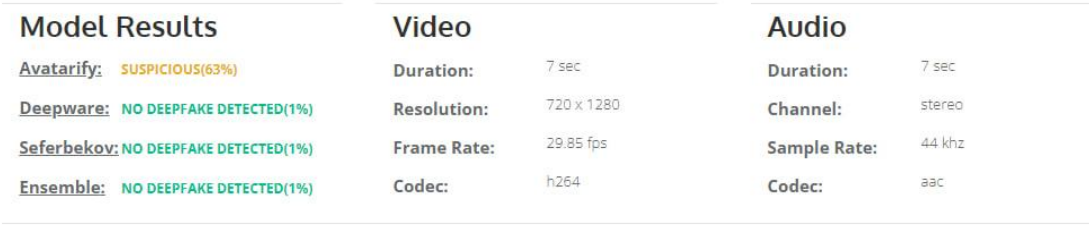

Рисунок 2 – Результаты обычного видео

Дальше было загружено обработанное видео. На этот раз две модели подтвердили признаки дипфейка, одна модель была не уверена, и одна модель дала ошибочный результат. Результаты представлены на рисунке 3.

Рисунок 3 – Результаты обработанного видео

Технология создания дипфейков уже не исчезнет, её нельзя запретить или технически ограничить. Поэтому лучшая защита – соблюдение правил безопасности, а также информирование пользователей о подобных типах атак и готовность ко встрече с ними.

Литература

1 SentinelAI [Электронный ресурс]. – URL: https://thesentinel.ai/(дата обращения: 08.10.2024)

2 Intel FakeCatcher [Электронный ресурс]. – URL: https:/ /www .intel . com/content/www/us/en/support/ru-banner-inside.html (дата обращения: 08.10.2024)

3Weverify[Электронныйресурс].–URL:https://weverify.eu/tools/deepfake-detector/ (дата обращения: 08.10.2024)

4Программы для обнаружения дипфейков [Электронный ресурс]. – URL: https://www.unite.ai/ru/best-deepfake-detector-tools-and-techniques/ (дата обращения: 08.10.2024)

5 Deepware [Электронный ресурс]. – URL: https://deepware.ai/ (дата обращения: 08.10.2024)