УДК 004.03

РАЗРАБОТКА МЕТОДА ОПРЕДЕЛЕНИЯ СИНТЕТИЧЕСКИХ ЛИЦ, СГЕНЕРИРОВАННЫХ ДИФФУЗИОННЫМИ МОДЕЛЯМИ

Роговой В., Герасимов В.М., Кишеев В.В.
Научный руководитель – ассистент Есипов Д.А.
Университет ИТМО

Аннотация. В данной работе была исследована проблема, которая появилась в связи с развитием технологий машинного обучения и их применением для подделки лиц на фото и видео контенте. Поскольку метода, удовлетворяющего необходимым условиям, по выявлению манипуляций с цифровым контентом, созданного диффузионными моделями, не представлено, была продемонстрирована собственная архитектура.

Введение. Обнаружение манипулируемого визуального контента, такого как фото и видеозаписи, является предметом пристального внимания научного общества [1]. Уже сейчас манипулируемый контент создается без согласия человека, участвующего в видео. Все эти видео считаются мошенническими, и они кажутся реалистичными, в связи чем миллионы пользователей не могут их идентифицировать как ложные.

Потенциальные угрозы варьируются от создания порнографических видео с лицом какой-нибудь знаменитости [2], до провокационных синтезированных видеороликов политических лидеров, манипулирующих выборами. Таким образом, контент, над которым произвели такие манипуляции, может нарушать конфиденциальность, создавая фальшивые новости и злонамеренные мистификации, а обилие медиаконтента в том числе в социальных сетях обеспечивает применение указанной технологии, формирования «дипфейков», необходимыми данными, тем самым обуславливая актуальность рассматриваемой угрозы.

Основная часть. Ограничением существующих подходов по выявлению синтезированных видеороликов является то, что большинство предлагаемых методов делают прогнозы по кадрам в видео и усредняют прогнозы, чтобы получить окончательную оценку прогноза для всего видео. Таким образом, они не учитывают отношения между кадрами. Чтобы это преодолеть мы предлагаем новый видео-трансформер для извлечения пространственных признаков с временной информацией [3, 4, 5]. Чтобы извлечь больше обучающих признаков, мы на вход нейронной сети подаем 3 составляющие: исходное изображение, а также спектральную и шумовую компоненту этого же изображения.

В своей работе мы будем использовать стратегию пошагового обучения, чтобы точно настроить предлагаемую модель на новых наборах данных, не жертвуя ее производительностью на предыдущих наборах данных. Функция потерь в пошаговом обучении состоит из двух частей: одна часть измеряет сходство между весами из нового набора данных со старыми весами из предыдущего набора данных, а другая часть функции потерь нужна для измерения точности модели обучения на новом наборе данных [6]. Таким образом первая часть заставляет веса быть как можно ближе к старым, поэтому модель хорошо работает с предыдущим набором данных, а вторая часть гарантирует, что модель хорошо работает с новым набором данных.

Выводы. В результате проделанной работы мы предлагаем видео-трансформер с пошаговым обучением для обнаружения дипфейков. Данная архитектура модели позволяет получить информативное обучение признакам, тем самым улучшая производительность в обнаружении дипфейков, а также стратегия постепенного обучения повышает способность к обобщению данной модели.

Список используемых источников

1. Sharma M., Kaur M. A review of Deepfake technology: an emerging AI threat //Soft Computing for Security Applications. – 2022. – С. 605-619.
2. S. Lyu, Deepfake detection: Current challenges and next steps. IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, United Kingdom., pp. 1–6 (2020).
3. Oscar de Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George. 2020. Deepfake Detection using Spatiotemporal Convolutional Networks. arXiv: Computer Vision and Pattern Recognition abs/2006.14749 (2020).
4. David Güera and Edward J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE. https://ieeexplore.ieee.org/document/ 8639163
5. Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. IEEE, 80–87.
6. Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. 2018. End-to-end incremental learning. In Proceedings of the European conference on computer vision (ECCV). 233–248.

Герасимов В. М. (автор) _____ (подпись)

Кишеев В. В. (автор) _____ (подпись)

Роговой В. (автор) _____ (подпись)

Есипов Д. А. (научный руководитель) _____ (подпись)