УДК 004.891.3

ОПРЕДЕЛЕНИЕ ДИПФЕЙКОВ С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

Тимощук-Бондарь А.И. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Кугаевских А.В. (ИТМО)

Введение. Дипфейк представляет собой методику синтеза изображения, основанную на искусственном интеллекте и используется для соединения и наложения существующих изображений и видео на исходные изображения или видеоролики. Цифровые изображения и видеофайлы в наше время стали неотъемлемой частью информационной среды. Однако, с развитием технологий, возникает проблема дополненной реальности в виде генерации дипфейков — искусственно созданных медиаданных, в которых лица заменяются с использованием технологий глубокого обучения. Это создает угрозу для безопасности и доверия в обществе, так как дипфейки могут использоваться для распространения дезинформации, манипуляции и киберпреступлений. Существующие методы обнаружения дипфейков часто оказываются неэффективными [1, 2, 3] в условиях постоянного развития технологий генерации, что порождает необходимость разработки новых методов и алгоритмов. Предполагается, что в результате комбинации принципов работы вариационного автоэнкодера и генеративно-состязательной нейросети мы получим дискриминатор, который условно назовём DeepDetect, способный отличать искусственно сгенерированные кадры от реально существующих. Добавив к DeepDetect несколько рекурентных слоёв и дообучив его на последовательности кадров полученных из реальных и искусственно сгенерированных видео, мы сможем получать оценку принадлежности видео к классу дипфейку или реальному.

Основная часть. Используя созданную и обученную нейронную сеть, решаются следующие задачи:

1) Определение дипфейков в ранее созданных, существующих видео с точностью большей, чем у современных моделей.
2) Определение дипфейков в реальном времени, с задержкой от десятой доли секунды, до нескольких секунд.

В основе работы нейросети использован подход перевода исходного видео в латентное пространство, последующего его декодирования с помощью вариационного автокодировшика, и, в последствии, классификации получившегося видео с помощью дискриминатора. Процесс обучения базируется на интеграции генеративно-состязательной нейронной сети с вариационным автоэнкодером. Классификация осуществляется при помощи дискриминатора, в то время как вариационный автоэнкодер выявляет различия между входным изображением и его восстановленным эквивалентом, определяя тем самым аномалии (дипфейки).

Выводы. Проведён анализ существующих методов выявления дипфейков, и создана нейросеть определяющая дипфейки не только в реальном времени, но и в ранее созданных видео. Получившуюся нейросеть можно будет использовать для проверки просматриваемых видео, в судебных экспертизах для защиты чести и достоинства, и определении искусственно сгенерированного контента.

Список использованных источников:

1. Mittal G., Hegde C., Memon N., Gotcha: Real-Time Video Deepfake Detection via Challenge-Response https://export.arxiv.org/abs/2210.06186 , 2023
2. Mirsky Y. DF-Captcha: A Deepfake Captcha for Preventing Fake Calls https://arxiv.org/abs/2208.085247 , 2022
3. Shang J., Wu, J. Protecting Real-time Video Chat against Fake Facial Videos Generated by

Face Reenactment https://cis.temple.edu/~jiewu/research/publications/Publication_files/FakeFace__ICDCS_2020.pdf , 2020