УДК 004.85

УГРОЗЫ И ПЕРСПЕКТИВЫ РАЗВИТИЯ СИНТЕЗИРОВАННЫХ МЕДИАФАЙЛОВ

Челпанов А.Д. (Университет ИТМО)
Научный руководитель – к.т.н. Югансон А.Д.
(Университет ИТМО)

Синтезированные медиафайлы становятся серьезной угрозой в руках злоумышленников. Для снижения риска финансовых и репутационных потерь разрабатываются системы детектирования синтезированных медиафайлов. В докладе рассматривается современное состояние сферы детектирования синтезированных медиафайлов и приводится список основных направлений для дальнейших исследований.

Введение. Технология DeepFake, позволяющая генерировать видеофайлы путем подмены лиц, является одним из самых известных приложений искусственного интеллекта. В руках злоумышленников рассматриваемая технология может нанести вред физическим и юридическим лицам. Для снижения рисков финансовых и репутационных потерь исследователи разрабатывают системы детектирования синтезированных медиафайлов.
В данной работе рассматриваются современные возможности технологии DeepFake, методы и системы детектирования синтезированных видеофайлов; выделяются основные направления для исследований в сфере детектирования синтезированных медиафайлов.

Основная часть. Подходы к синтезированию видеофайлов разделяются на синхронизацию участка рта человека, замене лица действующего человека с сохранением мимики, генерации нового видеофайла на основе участков головы и плеч человека. Современные методы создания DeepFake видео опираются на технологии глубокого обучения – автокодировщики и генеративно-состязательные сети. Инструменты, использующие генеративно-состязательные сети производят более реалистичные видеофайлы, однако более требовательны к настройке и обучению.
Злоумышленники могут применять синтезированные видеофайлы для обмана систем верификации биометрии, создания фальшивых новостей, создания компрометирующих видеофайлов и генерации контента для ботов в социальных сетях.
Первые подходы к детектированию синтезированных видеофайлов были нацелены на нахождение визуальных следов проведенных манипуляций: размытия, большие изменения между кадрами и мерцания. В это же время были исследованы способы основанные на поиске несоответствий в физическом поведении человека: частота морганий, позиции головы относительно тела, мимика. В данный момент развивается направление сравнения голосовых эмоций и мимики. Однако эффективность подходов с нахождением несоответствий будет быстро снижаться, учитывая темпы развития методов синтезирования видеофайлов.
Наиболее перспективным подходом к детекции синтезированных видеофайлов является использование интеллектуальных систем. Детектирование синтезированных видеофайлов может быть описано в виде проблемы бинарной классификации. Задачей классификатора является определение является файл синтезированным или естественным. Точность таких системы детектирования зависит от обучающего набора данных. В процессе обучения необходимо использовать все доступные наборы данных для достижения наибольшей точности.
Современные системы детектирования показывают высокую точность лишь на определенных наборах тестовых данных и уязвимы к состязательным атакам. Подавляющее большинство систем не доведены до состояния продукта и далеки от интеграции с информационными системами. В качестве основных направлений для совершенствования методов детекции предлагается обеспечить защиту моделей детектирования от состязательных атак, создать объемные наборы данных для обучения моделей детектирования с множеством методов генерации видеофайлов, интегрировать системы детектирования в информационные системы

социальных сетей и системы верификации биометрических данных, создать метод детектирования сгенерированных видеофайлов с подробным описанием результатов.

Выводы. Выделены основные угрозы использования синтезированных медиафайлов. Обозначены векторы для дальнейшей работы в сфере детектирования синтезированных видеофайлов.

Челпанов А.Д. (автор)

Югансон А.Н. (научный руководитель)