MINDSMITH

Дипфейки и способы их распознавания

Подзарядка

26 февраля 2024 года

MINDSMITH

Что такое дипфейк?

Дипфейк — поддельный медиафайл, созданный с использованием нейросетей для максимальной правдоподобности.

В связи с доступностью технологии, все чаще возникают ситуации, где она используется злоумышленниками.

MINDSMITH

Люди не могут надежно вычислять дипфейки

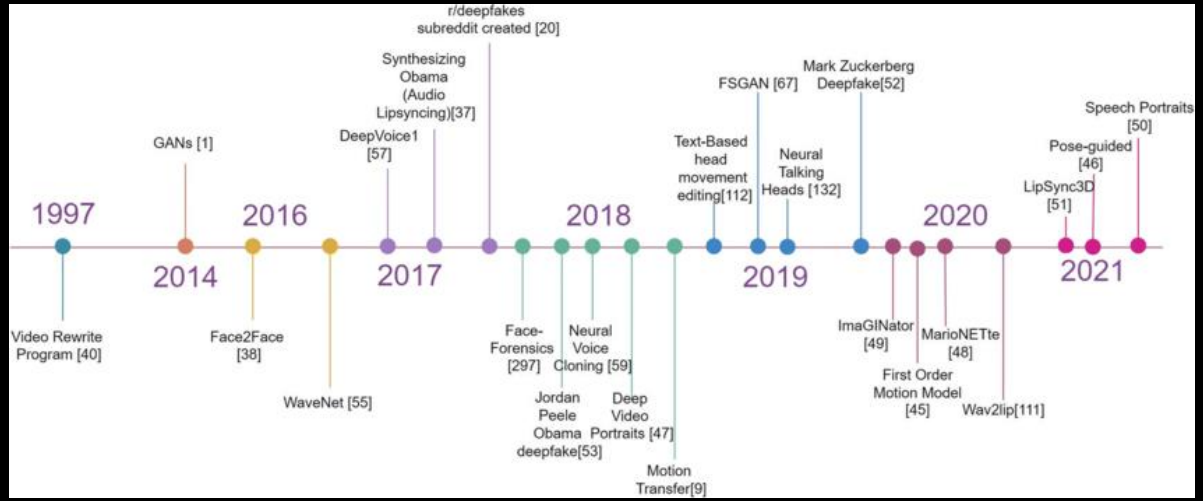
— Только 73% людей смогли корректно определить дипфейки в эксперименте, поставленном учеными из Университетского колледжа Лондона.

— Представьте ситуацию, в которой 27% ваших клиентов могут быть обмануты злоумышленниками, которые не тратят на атаку буквально ничего.

Источник: Kimberly T. Mai. «Warning: Humans cannot reliably detect speech deepfakes», PLOS ONE, 2023.

MINDSMITH

От липсинка к манипуляции положения лица



Источник: Masood, M. «Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward», Applied Intelligence, 2022.

MINDSMITH

SORA — новая ступень в генерации видео

MINDSMITH

Какие бывают дипфейки?

Под дипфейком больше не подразумевается только фотография — это может быть также текст, видео и аудиодорожки.

Некоторые системы способны производить дипфейки в реальном времени, натуралистично изменяя лицо и голос согласно параметрам, заданным пользователем.

Из-за развития больших языковых моделей, становится все сложнее разделить текст, написанный человеком от текста, написанного машиной.

MINDSMITH

Как производят дипфейки?

Сейчас с созданием дипфейка справится даже ребенок. Буквально.

Нужно всего лишь найти в интернете одну из многочисленных бесплатных утилит и передать ей файл с лицом жертвы и видео или фото, в которые это лицо будет монтироваться.

Для продвинутых пользователей, в открытом доступе на GitHub и других сайтах можно найти исходный код ПО и уже обученные GAN. Именно так в руки мошенников попадают инструменты для создания дипфейков.

MINDSMITH

10 секунд работы нейросети

MINDSMITH

Generative-Adversarial Network

GAN принципиально отличаются от других нейросетей тем, что по сути состоят из двух нейросетей: «генератора» и «дискриминатора», которые первично обучаются на одном датасете.

«Генератор» создает контент, в то время как «дискриминатор» оценивает, является ли контент реальным и дает свой фидбек.

«Генератор» все время пытается обмануть своего оппонента, создавая все более правдоподобные образы.

MINDSMITH

Как распознать дипфейк?

Определить низкокачественный дипфейк можно и невооруженным глазом: непостоянство черт лица, визуальные артефакты, неестественные движения мышц, глаз, общая скованность и др.

Или просто попросить подозреваемого в видеоконференции встать или произнести неестественно длинное или сложное слово.

В случае с более продвинутыми подделками человеку уже становятся необходимы особые технические средства для глубокого анализа, например, для изучения тепловой карты лица.

MINDSMITH

Технические методы защиты

В сети существует множество утилит, позволяющих определить, является ли контент дипфейком. Простейший пример — бесплатные утилиты DeepFake-Detect и Deepware.

Помочь защитить пользователей от дипфейков могут также и создатели контента: при помощи цифровых вотермарок, распознаваемых только специальным ПО.

MINDSMITH

Гибкий инструмент

Дипфейки используются во многих отраслях, включая кино, образование, игры и другие виды развлечений.

Сама модель может быть продуктом для дальнейшего оказания услуг — по сбору данных пользователей, для черного рынка или для нужд медиаконгломератов.

MINDSMITH

Три главных угрозы

— Атаки на клиентов: мошенник представляется родственниками или друзьями

— Атаки на институты: мошенник представляется начальством или клиентами банковскому служащему, подделывает документы

— Изменение инфополя: мошенники запускают инфоповод, который может привести к принятию неверных решений.

Источник: Dan Milmo. «Company worker in Hong Kong pays out £20m in deepfake video call scam», The Guardian, 2024

MINDSMITH

Как злоумышленники применяют дипфейки

В январе 2024 г. гонконгская полиция запустила расследование о хищении мошенниками 200 млн гонконгских долларов (2,343 млрд рублей). От сотрудницы неназванной компании поступила жалоба о том, что мошенники заставили ее совершить 15 транзакций по пяти местным банковским счетам в ходе видеоконференции.

Оказалось, что мошенники создали дипфейки нескольких старших должностных лиц компании.

Источник: Dan Milmo. «Company worker in Hong Kong pays out £20m in deepfake video call scam», The Guardian, 2024

MINDSMITH

Вы больше не можете верить новостям

Если ваше любимое СМИ не использует систем для проверки дипфейков — частота ошибок редакции будет расти вместе с развитием генеративного ИИ.

Вы не можете знать точно, насколько тщательно СМИ проверили видео- и аудиофайлы, представленные в контенте.

Желательно сомневаться, критически оценивать их и проверять самостоятельно.

MINDSMITH

Дипфейки уже меняют инфополе и культуру

В конце января 2024 года в интернете появились порнографические дипфейки с Тейлор Свифт, что привело к публичным высказываниям гендиректора Майкрософт, политиков США и НКО.

Видео, в которых авторы обыгрывают манеры знаменитостей или используют их голос для подачи контента, набирают миллионы просмотров.

MINDSMITH

На дипфейках зарабатывают не только мошенники

В январе 2023 г. в Китае вступили в силу релевантные законы. Китайское правительство обязало все сервисы, производящие дипфейки, делать очевидной их фальшивость.

А в ноябре 2023 г. Microsoft заявили о формировании команды, которая будет предоставлять поддержку по работе с дипфейками. Компания создаст так называемый «Центр предвыборной коммуникации».

Источник: Giulia Interesse. «China to Regulate Deep Synthesis (Deepfake) Technology Starting 2023», China Briefing, 2022; Emilia David. «Microsoft offers politicians protection against deepfakes», The Verge, 2023

MINDSMITH

Банк как продавец правды

— Банки владеют деньгами и компетенциями на разработку продуктов в ИТ.

— У банков есть практики обеспечения надежности при проверке, обмене и хранении информации.

— В условиях деформации информационного поля, клиентам будут нужны надежные источники с проверенной информацией.

Источник: Dan Milmo. «Company worker in Hong Kong pays out £20m in deepfake video call scam», The Guardian, 2024

MINDSMITH

Руслан, что с этим делать?

— Разработка систем для распознавания дипфейков и последующая продажа государствам и бизнесу.

— Страховка и предоставление медиаконгломератам лиц актеров, моделей.

— Создание, внедрение и продажа новых систем, подходов к антифроду, KYC.

— Продажа новых подходов к маркетингу и производству медиа.

Источник: Dan Milmo. «Company worker in Hong Kong pays out £20m in deepfake video call scam», The Guardian, 2024