В. А. Кулыба

ОБЗОР DEEPFAKE ТЕХНОЛОГИЙ И РЕШЕНИЙ В ОБЛАСТИ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Технология Deepfake использует глубокие нейронные сети для убедительной замены на видео одного лица другим. У этой технологии есть очевидный потенциал для злонамеренного использования, и она становится всё более распространённой. Есть все основания полагать, что Deepfake-техологогия в следующие года будет становиться только лучше, быстрее и дешевле. Ключевые слова: Deepfake, глубокое обучение, преобразование изображений и видео, Компьютерное зрение — теория и технология создания машин, которые могут производить обнаружение, отслеживание и классификацию объектов.

Введение

В последние годы в области глубокого обучения и компьютерного зрения произошло множество революций и появилось много новых технологий. У некоторых технологии, таких как Deepfake, есть очевидный потенциал для злонамеренного использования, и по поводу социальных и политических последствий этого тренда уже многое известно. В следующих пунктах представлены более подробные сведения о техническом устройстве Deepfake решений.

I. Вычислительные мощности и данные

Видео называемые Deepfake [«глубокими подделками»] получили свое название потому, как создаются с использованием глубоких нейросетей. За последнее десятилетие специалисты по информатике обнаружили, что нейросети становятся всё более мощными при добавлении дополнительных слоёв из нейронов. Но чтобы раскрыть весь потенциал глубоких нейросетей, нужно очень много данных и огромные вычислительные мощности. Под вычислительными мощностями рассматриваются GPU, с большим объемом графичекой памяти и наибольшим количеством CUDA-ядер. Для создание Deepfake решения также требуется множество видеоролоков с исходным объектом на кадрах видеозаписи и такое же множестов видеоролиокв с объектом используемым для замены.

II. Принцип работы

В основе ведущих программных пакетов для создания Deepfake – находится автоэнкодер. Это нейросеть, обученная принимать на вход изображение и выдавать идентичное изображение. Само по себе это умение может быть не таким уж полезным, но, как увидим далее, это ключевой строительный блок в процессе создания Deepfake.

Автоэнкодер берёт это компактное представление, известное, как «латентное пространство», и пытается развернуть его, получив изначальное изображение. С одной стороны сети находится энкодер, принимающий изображение и сжимающий его до небольшого числа переменных. С другой стороны нейросети находится декодер. Он берёт это компактное представление, известное, как «латентное пространство», и пытается развернуть его, получив изначальное изображение.

Искусственно ограничение количества данных, передаваемых от энкодера к декодеру, заставляет две этих сети разработать компактное представление человеческого лица. Энкодер – это что-то вроде алгоритма сжатия с потерями, который пытается сохранить как можно больше информации о лице при ограничениях на объём хранилища. Латентное пространство должно каким-то образом извлечь важные детали, например, в какую сторону смотрит субъект, открыты у него глаза или закрыты, улыбается он или хмурится.

Важно, что автоэнкодеру нужно сохранить только те особенности лица, которые меняются во времени. Ему не нужно хранить неизменные вещи тип цвета глаз или формы носа. Если на каждой фотографии человека у него голубые глаза, тогда декодер его сети обучится автоматически рисовать его лицо с голубыми глазами. Нет нужды запихивать в тесное латентное пространство информацию, не меняющуюся при переходе от одного изображения к другому.

Каждому алгоритму для обучения нейросети нужен какой-то способ оценить качество работы сети, чтобы его можно было улучшить. Во многих случаях это делается через обучение с учителем, когда человек обеспечивает правильный ответ для каждого элемента из набора обучающих данных. Автоэнкодеры работают по-другому. Поскольку они просто пытаются воспроизвести собственные входные данные, обучающее ПО может судить об их качестве работы автоматически, что принято называть обучением без учителя в терминоллогии машинного обучения.

Как и любая нейросеть, автоэнкодеры в Deepfake сетях обучаются при помощи обратного распространения. Обучающий алгоритм получает определённое изображение в нейросеть и смотрит, какие пиксели на выходе не соответствуют входу. Затем он подсчитывает, какие из нейронов последнего слоя внесли наибольший вклад в

ошибки и немного подправляет параметры каждого нейрона так, чтобы он выдавал результаты получше.

Затем эти ошибки распространяются обратно, к предыдущему слою, где параметры каждого нейрона подправляются вновь. Ошибки распространяются таким способом всё дальше назад, пока каждый из параметров нейросети – как у энкодера, так и у декодера – не окажутся исправленными.

Затем обучающий алгоритм скармливает ещё одно изображение сети, и весь процесс повторяется снова. Могут понадобиться сотни тысяч таких повторов для того, чтобы получился автоэнкодер, хорошо воспроизводящий собственный вход.

ПО для Deepfake работает, параллельно обучая два автоэнкодера, один для оригинального лица, а второй – для нового. Во время процесса обучения каждому автоэнкодеру выдают изображения только одного человека, и он обучается выдавать изображения, очень похожие на оригинал. Схема работы автоэнкодеров (Рис. 1).

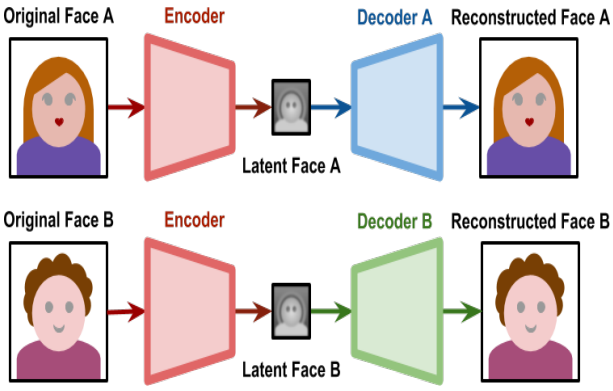

Рис. 1 – Схема работы автоэнкодеров

Есть, однако, некоторые трудности: обе сети используют один и тот же энкодер. Декодеры – нейроны в правой части сети – остаются раздельными, и каждый из них обучается выдавать разное лицо. Но нейроны в левой части сети имеют общие параметры, меняющиеся каждый раз, когда обучается любой из автоэнкодеров. Когда сеть обчается на оригинальном лице, это меняет половину сети, принадлежащую энкодеру и в сети с лицом для замены (fake-лицом).

В итоге у двух автоэнкодеров есть один общий энкодер, способный «считывать» либо настоящее лицо, либо поддельное. Цель энкодера в том, чтобы использовать одинаковое представление таких вещей, как угол наклона головы или расположение бровей. А это, в свою очередь, означает, что когда вы сжали лицо при помощи энкодера, его можно распаковать при помощи любого декодера.

После обучения таким способом пары автоэнкодеров, остаётся последний шаг для создания Deepfake: изменение декодеров местами (Рис. 2).

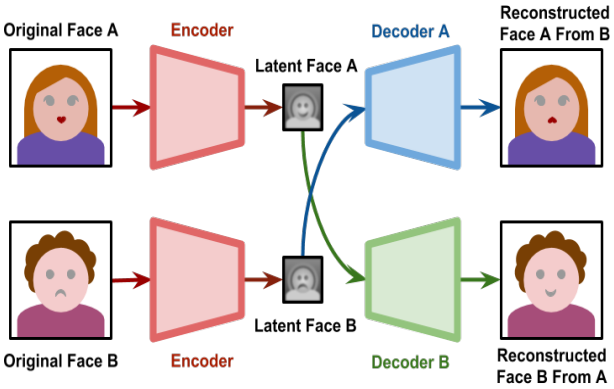

Рис. 2 – Схема изменения декодеров местами

В итоге получается реконструированная fake-фотография – но с тем же положением головы и выражением лица, как и у оригинальной фотографии.

III. Заключение

Резонанс Deepfake технологий очевидно вызывает беспокойство. До недавнего времени люди могли с достаточной легкостью принимать видеозапись с человеком за чистую монету. Появление ПО для создания Deepfake и других цифровых инструментов привело к тому, что теперь люди относимся к видеозаписям со скептицизмом. В настоящее время ролик, в котором человек утверждает что-то скандальное – или раздевается – стоит рассмотреть возможность того, что некто подделал это видео с целью дискредитации того человека.

IV. Список литературы

1. Основы современного искусственного интеллекта: как он работает, и уничтожит ли наше общество уже в этом году? `https://habr.com/ru/post/451214/` Часть 1.
2. Эти новые уловки пока ещё способны перехитрить видеоролики от Deepfake `https://habr.com/ru/post/429192/`
3. Я создал свой собственный Deepfake за две недели — `https://habr.com/ru/post/482684/`.

Кулыба В. А., магистрант факультета информационных технологий и управления БГУИР, kulyba.vadim@gmail.com

Научный руководитель: Гуринович Алевтина Борисовна, кандидат физ.-мат. наук, доцент, gurinovich@bsuir.by