30 апреля 2021

Что с нами будет? Письма ученых о самом важном

Это первое письмо нового сезона. В нем речь пойдет о дипфейках

Привет!

Мы возобновляем научную рассылку. Вместе с учеными и специалистами из разных областей продолжим разбираться, какие открытия, технологии и явления меняют настоящее и формируют будущее.

Недавно в TikTok появились видеоролики, на которых актер Том Круз показывает фокусы с монеткой, обнаруживает жвачку в леденце и играет в гольф. Сомнений в их подлинности не возникает, несмотря на то, что аккаунт называется deeptomcruise, что является отсылкой к deepfake — методике создания изображений, основанной на машинном обучении. Что это такое, как производится и как отличить реальные изображения и видео от поддельных?

В этом письме инженер-кибернетик Алексей Егоров из Института интеллектуальных кибернетических систем НИЯУ МИФИ рассказывает, как одни нейросети генерируют дипфейки, а другие учатся их отличать от реальных изображений.

Что такое дипфейк?

Слово «дипфейк» возникло из двух понятий: fake — подделка и deep learning — глубокое обучение нейросетей. Говоря простыми словами, чтобы нейросеть смогла сгенерировать дипфейк, она должна обработать большой массив фото,

изучить особенности лица и мимики человека и за счет этого научиться воспроизводить его.

С технической стороны всё сложнее. Глубокое обучение нейросетей — это область машинного обучения, основанная на искусственных нейронных сетях. С помощью него алгоритмы учатся принимать решения и предсказывать результат по набору данных, например определять, что изображено на фото или видео, и самостоятельно генерировать визуальный контент.

Одна из моделей глубокого обучения — генеративно-состязательные сети (Generative Adversarial Network, GAN), именно они используются, например, для создания фотографий на сайте thispersondoesnotexist.com. Это архитектура, состоящая из двух сетей: генератора и дискриминатора, настроенных на работу друг против друга. То есть, для того чтобы создать дипфейк, одна сеть должна быть обучена генерировать данные, а другая — отличать смоделированные данные от реальных.

Схема обучения такая: обе сети учатся на полученных результатах и с каждым разом всё лучше справляются со своими задачами. Через некоторое время генеративная сеть становится способной создавать такие фейковые изображения, которые другая нейронная сеть с тем же уровнем развития уже не может отличить от настоящих.

Главная сфера применения дипфейков — кино и реклама. С их помощью можно создать Халка, сделать моложе Роберта де Ниро или снять видео, на котором Дэвид Бекхэм говорит на девяти языках.

Как алгоритмы учатся отличать дипфейки от реальных изображений?

Технология производства изображений не может обойтись без ошибки: в нейронных сетях она всегда есть. Получается, нужно научить другую нейронную сеть эту ошибку вычислять. В итоге, по сути, возникает противостояние пули и брони: одни сети учатся создавать всё более искусные дипфейки, а другие — использовать технологию для их определения.

Обучение происходит по той же схеме, что и раньше. Опишу ее подробнее: у нас есть бинарная классификация «ложь/

истина» и изображения — настоящие и созданные искусственно, — подавая их на вход нашей сети, мы вынуждаем ее проводить классификацию.

Например, мы создали сеть, запустили ее, и она нам говорит: всё истина. Мы проверяем, так ли это, находим ошибки и затем с помощью математических манипуляций, которые называются методом обратного распространения ошибки, понимаем, как именно нужно модифицировать параметры, чтобы на выходе стало получаться нечто более похожее на то, что нам нужно. К слову, в нейронной сети может быть, например, 100 миллионов параметров. Все они меняются автоматически с помощью алгоритма обратного распространения ошибки, и сети вновь подается входная обучающая выборка. Она ее анализирует, мы проверяем, меняем параметры и так далее, пока не удается добиться определенной точности — 95 % правильных ответов считаются вполне нормальным результатом. Таким образом получается обученная сеть, способная распознавать дипфейки.

Кто лучше распознает искусственные изображения: алгоритмы или человек?

Примерно три года назад обнаружить дефекты на фото или видео, созданных при помощи машинного обучения, можно было невооруженным глазом. Однако с развитием вычислительных технологий разобрать, где правда, а где ложь, становится сложнее.

С помощью алгоритмов можно решать узкие задачи. Например, вы создали сеть, способную распознавать дипфейк-видео на изображениях из фильмов, но, как только вы подадите ей картинку с городских камер, она не сможет ничего определить, и ее придется переучивать.

При этом создать алгоритм, который бы учитывал все возможные типы изображений невозможно — вариантов бесконечное множество, а всё, что мы подаем в модель машинного обучения, всегда ограничено. Поэтому, когда нужно решить узкую задачу в точно зафиксированных условиях, алгоритм ее решит, но, как только условия

изменятся хотя бы на полпроцента, сеть придется переучивать.

Человек, в свою очередь, способен определять разные искусственно созданные изображения, исходя из здравого смысла, а также обращая внимание на следующие признаки:

- искажения на стыке текстур: например, размытый фон или, наоборот, чрезмерная резкость изображения;
- неправильные тени;
- несовпадение движения губ с речью;
- разное освещение на разных кадрах;
- человек на видео странно моргает или не моргает вообще.

Даже на, казалось бы, идеальных видео deeptomcruise есть огрехи. Например, если внимательнее рассмотреть видео, где актер надевает очки, заметен сбой в области глаз и рта.

Что еще почитать о дипфейках?

- Материал о том, как появились, менялись и развивались дипфейки, а также чем они могут быть полезны рекламной индустрии.
- Инструкцию о том, как создать дипфейк в программе DeepFaceLab.
- Материал о том, какие бывают дипфейки и как их распознать.

На этом всё.

Хороших выходных!
