ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНЫЕ СЕТИ И ПРОБЛЕМЫ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Кадан М. А.
Кафедра системного программирования и компьютерной безопасности,
Гродненский государственный университет имени Янки Купалы
Гродно, Республика Беларусь
E-mail: kadan.maria@gmail.com

Рассмотрена концепция генеративно-состязательных нейронных сетей, применение которых позволяет эффективно решать многие актуальные прикладные задачи. Внимание акцентировано на возможности генерации цифровых объектов по образцу, преобразовании искаженных и зашумленных цифровых объектов, формированию поддельной (фейковой) мультимедийной продукции, применению генеративно-состязательных нейронных сетей в задачах информационной безопасности. Отдельно затронуты вопросы уязвимости обучения искусственных нейронных сетей, позволяющие злоумышленнику изменять результат предсказания сети.

Введение

Генеративно-состязательные сети (англ. Generative Adversarial Networks, GAN) рассматриваются как одно из наиболее перспективных современных направлений в области глубокого обучения.

Генеративно-состязательные сети, концепцию которых предложил Ян Гудфеллоу из компании Google в 2014 году [1], являются алгоритмом машинного обучения без учителя (алгоритмом неконтролируемого обучения), который построен на комбинации двух сетей. Одна из которых (сеть G, Generator, Генератор) генерирует образцы (генеративная модель) а другая (сеть D, Discriminator, Дискриминатор) старается отличить правильные («подлинные») образцы от неправильных (дискриминативная модель) (см. рисунок 1).

Так как сети G и D имеют противоположные цели – создать образцы и отбраковать образцы, между ними возникает антагонистическая (состязательная) игра. В ходе этого «состязания» сети G и D конкурируют и сотрудничают друг с другом и, в результате такого обучения, в конечном итоге учатся выполнять свои задачи.

I. Реализация генеративно-состязательных сетей

Сети GAN реализованы, в частности, в библиотеке TensorFlow [2] – открытой программной библиотеке для машинного обучения, разработанной компанией Google для решения задач построения и тренировки нейронной сети с целью автоматического нахождения и классификации образов. TensorFlow обеспечивает API для Python, а также C++, Haskell, Java и Go.

TensorFlow является продолжением закрытого проекта DistBelief. Изначально TensorFlow была разработана командой Google Brain[en] для внутреннего использования в Google, а потом (9 ноября 2015 года) была переведена в свободный доступ с открытой лицензией Apache 2.0.

Библиотека TensorFlow применяется как для исследований, так и для разработки продуктов, достигая при этом качества человеческого восприятия. TensorFlow может работать на многих параллельных процессорах, как CPU, так и GPU, опираясь на архитектуру CUDA для поддержки вычислений общего назначения на графических процессорах.

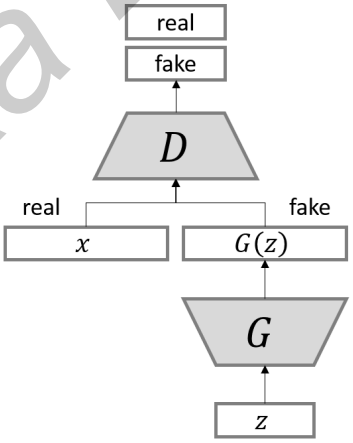

Рис. 1 – Схема работы GAN

II. Задачи, решаемые с использованием GAN

Использование GAN обычно объясняют на примере фальшивомонетчика (Генератора) и детектива (Дискриминатора). Первоначально, фальшивомонетчик предъявляет детективу фальшивые деньги. Детектив доказывает, что это подделка и объясняет фальшивомонетчику, почему деньги фальшивые. Фальшивомонетчик пытается сделать новую подделку на основе полученных объяснений. Детектив снова утверждает, что деньги поддельные и дает новый набор объяснений. И так до тех пор, пока детектив не примет фальшивые деньги за настоящие.

Генерация фейкового контента. Эксперты в области информационной безопасности

утверждают, что распространение фальшивых новостей — это лишь первая волна фейкового контента, нашествие которого стоит ожидать в ближайшие годы [3]. Генеративные нейросети уже позволяют создавать видеоролики, на которых люди делают то, чего не делали, и говорят то, чего никогда не говорили. По прогнозам, через три года YouTube заполнят фейковые ролики. Другие эксперты считают, что на налаживание процессов медиафальсификации уйдет больше времени, но рано или поздно это точно произойдет. С помощью нейросети изучают статистические характеристики аудиозаписи, а затем воспроизводят их в другом контексте. При этом улавливаются изменения в речи на очень коротких интервалах. В идеале, достаточно ввести текст, который нейросеть должна воспроизвести, и она сгенерирует правдоподобное выступление.

Корректировка искажений и удаление шума в цифровых объектах. Использование GAN позволяет, в частности, генерировать фотографии, которые человеческй мозг воспринимает как натуральные изображения. Например, известна попытка синтезировать фотографии кошек, которые вводят в заблуждение эксперта, считающего их естественными фото [4]. Кроме того GAN используются для улучшения качества нечётких или частично испорченных фотографий.

Восстановление суперразрешения. Несмотря на прорыв в точности и скорости однократного улучшения изображения с использованием глубоких сверточных нейронных сетей, центральная проблема остается в значительной степени нерешенной: как восстанвить мельчайшие детали текстур, чтобы сохранить качество изображения при больших коэффициентах масштабирования [5].

Наряду с существенными угрозами информационной безопасности, которые могут быть реализованы с использаванием GAN, можно выделелить ряд задач, в которых роль GAN несомненно позитивна. К их числу можно отнести задачи, связанные с

- улучшением/восстановлением размытых фото- и видеоизображений с камер видеонаблюдения;
- восстановлением трудноразличимого текста на фотографиях, рассматриваемых в качестве носителей криминалистически значимой информации;
- генерацией образцов клавиатурного почерка и голоса в задачах, использующих биометрические методы аутентификации.

III. Уязвимости обучения искусственных нейронных сетей

В 2017 году была опубликована работа, демонстрирующая, как имплантировать бэкдоры во время глубокого обучения нейронных сетей [6]. Это приводит к обучению модели, которая сохраняет точность, но неправильно классифицирует ввод при вставке бэкдор-триггера. В качестве примера приводится классификатор дорожных знаков для автономной системы вождения. Запуск бэкдора может происходить, к примеру, при распознавании наклейки на знаке остановки, который в результате определяется как знак ограничения скорости.

Активно поднимается вопрос о неприкосновенности частной жизни при выполнении глубокого обучения [7]. Модели обычно обучаются централизованно, при этом все данные обрабатываются одним и тем же алгоритмом обучения. Если данные представляют собой коллекцию личных данных пользователей, включая привычки, личные фотографии, географические позиции, интересы и т.д., централизованный сервер будет иметь доступ к конфиденциальной информации, которая потенциально может быть использована с нарушением конфиденциальности. Для решения этой проблемы предлагаются модели совместного обучения для совместной работы, где стороны на местах обучают свои глубокие сети и разделяют только подмножество параметров в попытке сохранить приватность собственных обучающих данных. Параметры также могут быть запутаны с помощью метода дифференциальной конфиденциальности [7], чтобы сделать извлечения информации еще более сложным.

1. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. «Generative Adversarial Networks», arXiv:1406.2661.
2. TensorFlow. An open-source software library for Machine Intelligence [Electronic resource]. – Mode of access: https://www.tensorflow.org/. – Date of access: 14.09.2017.
3. Fake news: you ain't seen nothing yet [Electronic resource] / The Economost. – Mode of access: https://www.economist.com/news/science-and-technology/21724370-generating-convincing-audio-and-video-fake-events-fake-news-you-aint-seen. – Date of access: 14.09.2017.
4. Salimans, Tim; Goodfellow, Ian; Zaremba, Wojciech; Cheung, Vicki; Radford, Alec Chen, Xi (2016), «Improved Techniques for Training GANs», arXiv:1606.03498.
5. Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi. «Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network», arXiv: 1609.04802.
6. Tianyu Gu, Brendan Dolan-Gavitt, Siddharth Garg (2017), «BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain», arXiv: 1708.06733.
7. Briland Hitaj, Giuseppe Ateniese, Fernando Perez-Cruz (2017). «Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning», arXiv: 1702.07464.