Левин, А. О. Использование генеративно-состязательных сетей для генерации изображений по тексту / А. О. Левин, Ю. С. Белов // Научное обозрение. Технические науки. – 2023. – № 2. – С. 11-15. – DOI 10.17513/srts.1427. – EDN FJZRPS.

УДК 004.89

ИСПОЛЬЗОВАНИЕ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНЫХ СЕТЕЙ ДЛЯ ГЕНЕРАЦИИ ИЗОБРАЖЕНИЙ ПО ТЕКСТУ

Левин А.О., Белов Ю.С.

Московский государственный технический университет имени Н.Э. Баумана, филиал в г. Калуга, Калуга, e-mail: iu4-kf@mail.ru

Нейронные сети используют практически во всех сферах человеческой деятельности, предлагая различные решения для многочисленного спектра задач. Распространенной задачей является генерация изображений. Дополнительным параметром в данной задаче является использование технологии генерации изображения по тексту (Text-To-Image). С основой на результаты многочисленных работ был произведен анализ нескольких моделей генеративно-состязательных сетей, использующих технологию Text-To-Image, с рассмотрением их архитектуры, процесса обучения и влияния на непосредственный процесс генерации изображений с его последующими результатами. Целью работы является исследование архитектуры, концепций обучения и принципов работы генеративно-состязательных сетей для генерации изображения по тексту (Text-To-Image). Основываясь на полученных в процессе данной работы данных, считаем, что генеративно-состязательные сети являются хорошим подходом к решению задач генерации изображений по тексту, подразумевающих использование технологии Text-To-Image, несмотря на небольшие затруднения как в процессе непосредственного обучения, так и в процессе последующего использования – генерации изображений, ввиду ограниченного разнообразия выборок, что зачастую приводит к большим временным затратам. Однако данные недостатки компенсируются точностью и качеством результирующих изображений, что подтверждает обоснованность их применения для такого рода задач.

Ключевые слова: генеративно-состязательные сети, генерация изображений, генерация изображений по тексту

APPLICATION OF GENERATIVE-ADVERSARIAL NETWORKS TO TEXT TO IMAGE GENERATION

Levin A.O., Belov Yu.S.

Bauman Moscow state technical University, Kaluga branch, Kaluga, e-mail: iu4-kf@mail.ru

Neural networks are used in almost all spheres of human activity, offering various solutions for a wide range of tasks. A common task is to generate images. An additional parameter in this task is the use of Text-To-Image technology. Based on the results of numerous works, an analysis was made of several models of generative adversarial networks using Text-To-Image technology, considering their architecture, learning process and influence on the direct process of image generation with its subsequent results. The aim of the work is to study the architecture, learning concepts and operating principles of generative adversarial networks for Text-To-Image generation. Based on the data obtained in the course of this work, generative adversarial networks are a good approach to solving the problems of generating images from text, using the Text-To-Image technology, despite minor difficulties both in the process of direct learning and in the process of subsequent use. – image generation, due to the limited variety of samples, which often leads to large time costs. However, these shortcomings are compensated by the accuracy and quality of the resulting images, which confirms their justification for applications for this kind of tasks.

Keywords: generative adversarial networks, image generation, text-to-image

В настоящее время нейронные сети стремительно развиваются, расширяя свое влияние практически на все сферы человеческой деятельности, предлагая различные решения для многочисленного спектра задач. Одной из таких задач является генерация изображений – сфера искусственного интеллекта, в рамках которой компьютеры обучаются интерпретировать визуальный мир. Дополнительным параметром в данной задаче является использование технологии генерации изображения по тексту (Text-To-Image). Ввиду этого применение генеративно-состязательных сетей (GAN) позволит успешно и стабильно решать данную задачу.

Цель исследования – исследовать архитектуру, концепции обучения и принципы работы генеративно-состязательных сетей для генерации изображения по тексту (Text-To-Image).

Общая информация о генеративно-состязательных сетях. Генеративно-состязательные сети (GAN) – это алгоритмические архитектуры, которые используют две нейронные сети, противопоставляя одну другой для создания новых синтетических экземпляров данных, которые можно принять за реальные данные [1]. Они широко используются в таких сферах, как: генерации изображений, генерации видео и генерации голоса.

Потенциал генеративно-состязательных сетей (GANs) огромен, они могут научиться имитировать любое распределение данных благодаря самому процессу состязания одной сети с другой. То есть генеративно-состязательные сети можно научить создавать изображения, максимально приближенные к реальности.

Общие принципы обучения моделей генерации изображений (T2I). Любые модели,

используемые для генерации изображений по тексту, обучаются на больших наборах данных, представляющих собой наборы пар (текст, изображение), то есть изображения с текстовыми подписями (описанием) [2].

Самым распространенным набором данных, используемым для обучения моделей генерации изображений, является упомянутый ранее набор COCO (Common Objects in Context), состоящий из 123 000 изображений различных объектов, с пятью текстовыми описаниями к каждому изображению [3].

Также в области генерации изображений были разработаны метрики количественной оценки (например, R-точность (R-precision), визуально-семантическое сходство и семантическая точность результирующего объекта), которые были введены специально для оценки качества моделей генерации текста в изображение, а именно: T2I – Text To Image [4] (рис. 1).

Принцип работы и обучения генеративно-состязательных сетей. Рассмотрим детально генеративно-состязательные сети. Они состоят из двух нейронных сетей: генераторной сети G(z) с шумом z ~ pz, выбранным из предыдущего распределения шума, и дискриминаторной сети D(x), где x ~ p_{data} реальные изображения, а x ~ pg – сгенерированные изображения соответственно [5].

Обучение представляет собой взаимодействие двух нейронных сетей, в которой одна нейронная сеть, называемая генератором, генерирует новые экземпляры данных, а другая, дискриминатор, оценивает их на подлинность. Другими словами, дискриминатор решает, принадлежит ли каждый экземпляр данных, которые он просматривает, фактическому набору обучающих данных или нет (рис. 2).

Обучение можно определить как минимаксную игру двух игроков (правило принятия решений, используемое в теории игр) с функцией V (D, G), где дискриминатор D(x) обучается максимизировать логарифмическую вероятность, присваивая ей правильный класс, в то время как генератор G(z) обучается минимизировать вероятность того, что дискриминатор log(1 − D(G(z)) классифицирует его как фальшивый [6]:

$$\min\max V(D,G) = \mathrm{E}_{x\sim p_{data}}[\log D(x)] + \mathrm{E}_{z\sim p_z}[\log(1-D(G(z)))]$$

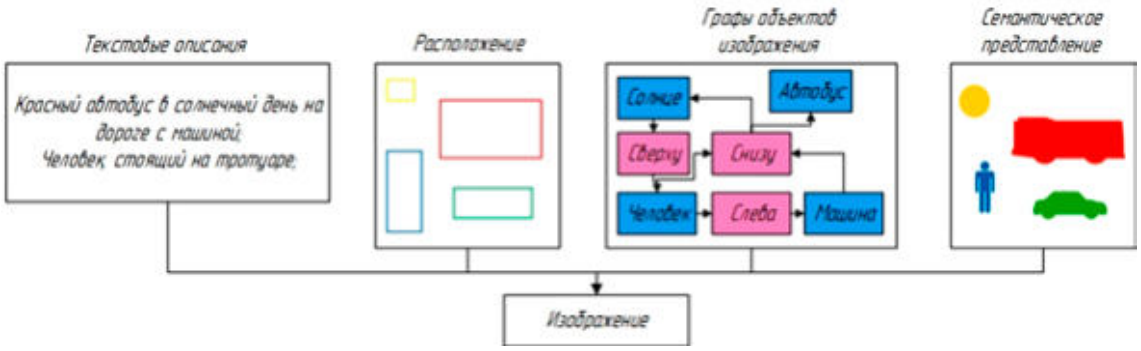

Рис. 1. Виды аннотаций, используемых для генерации изображений по тексту

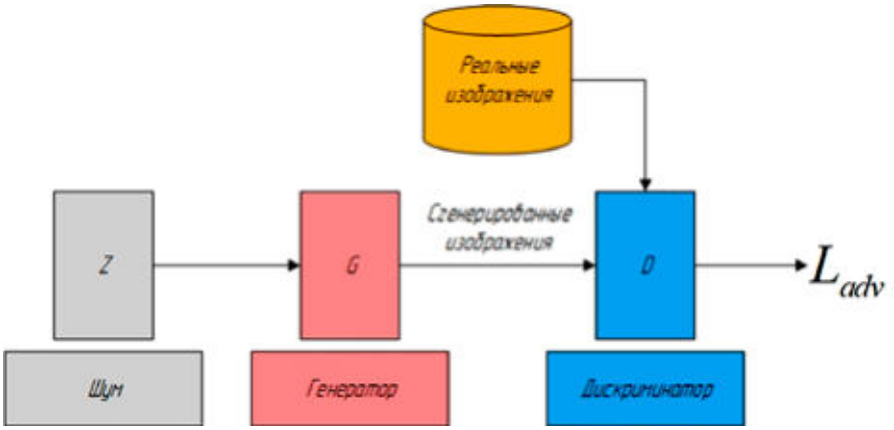

Рис. 2. Принцип обучения генеративно-состязательной сети

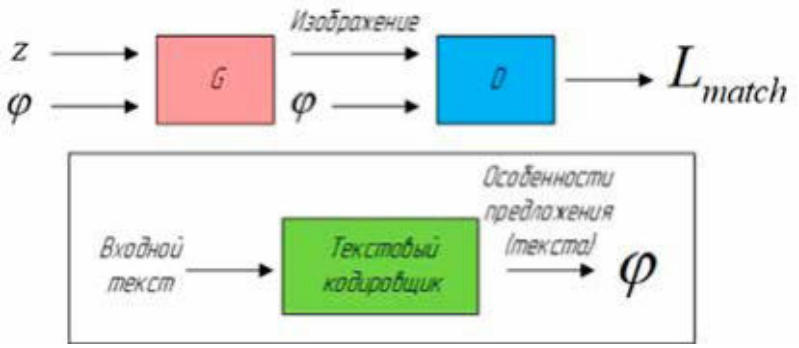

Рис. 3. Принцип работы технологии генерации изображения по тексту в генеративно-состязательных сетях

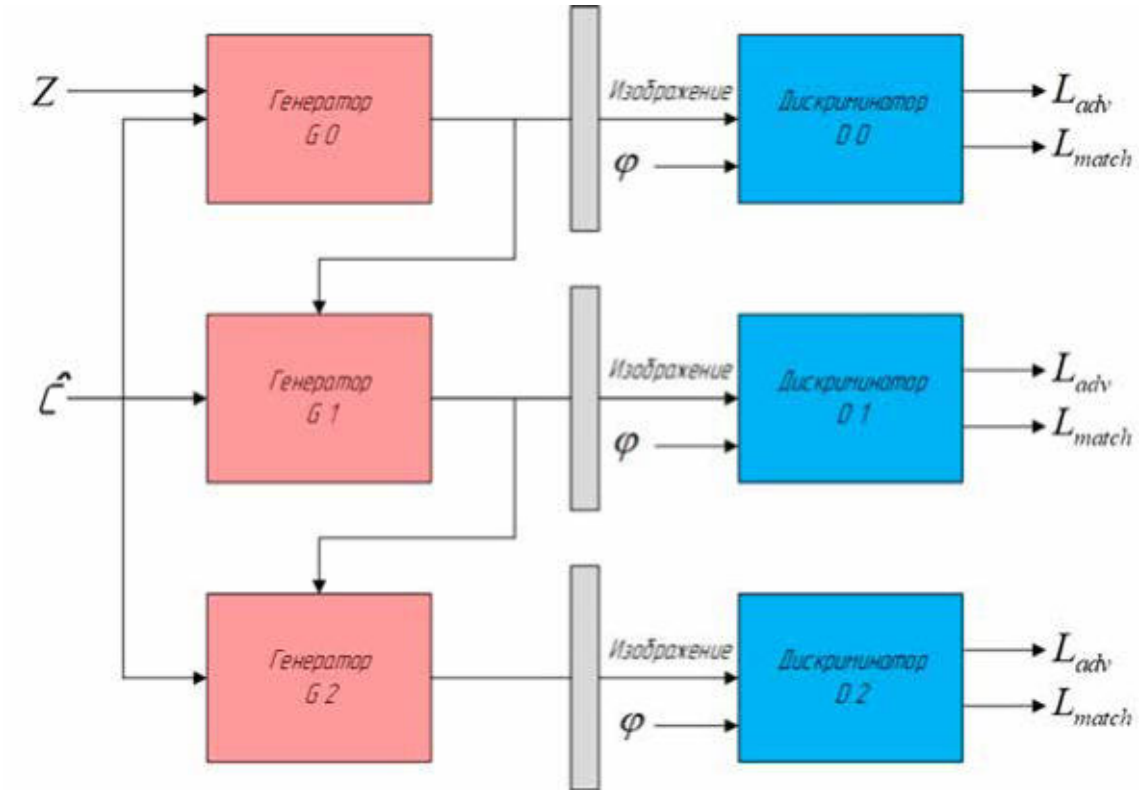

Рис. 4. Архитектура генеративно-состязательной сети StackGAN++

Использование технологии генерации изображения по тексту (Text To Image). Первый подход T2I обусловливает процесс генерации анализа всего предложения, полученного из предварительно обученного текстового кодировщика.

Дискриминатор обучен различать реальные и сгенерированные пары «изображение – текст». Следовательно, модель T2I является естественным расширением генеративно-состязательной сети в том смысле, что условие на метке класса Y просто заменяется текстовым вложением φ (рис. 3).

В глубокой сверточной генеративно-состязательной сети (GAN-INT-CLS) в качестве входных данных для дискриминатора используются три разные пары:

- реальное изображение с совпадающим текстом;
- сгенерированное изображение с соответствующим текстом;
- реальное изображение с несовпадающим текстом.

Такой подход заставляет и генератор, и дискриминатор не только фокусироваться на реалистичных изображениях, но и сравнивать их с входным текстом [7].

Однако ранние версии GAN-INT-CLS могли генерировать только изображения с низким разрешением 64 × 64 пикселя.

Чтобы модели, основанные на генеративно-состязательных сетях, могли генерировать изображения с более высоким разрешением, необходимо использовать обновленную архитектуру, включающую в себя несколько объединенных генераторов – StackGAN.

В StackGAN на первом этапе генерируется грубое изображение размером 64×64 пикселя с учетом вектора случайного шума и вектора обработки текста. Это исходное изображение и текст поступают во второй генератор, который выводит изображение уже размером 256 × 256 пикселей [8]. На обоих этапах дискриминатор обучается различать совпадающие и не совпадающие пары «изображение – текст».

StackGAN++ еще больше улучшил архитектуру с помощью сквозной структуры, в которой три генератора и дискриминатора совместно обучаются для одновременной аппроксимации многомасштабных, условных и безусловных распределений изображений [9] (рис. 4).

Применение генеративно-состязательных сетей для генерации изображения по тексту. Рассмотрим пример, в котором необходимо сгенерировать написанные от руки цифры, подобные тем, которые можно найти в открытом наборе данных. Цель дискриминатора при отображении экземпляра из истинного набора данных состоит в том, чтобы распознать те, которые являются подлинными (рис. 5).

Тем временем генератор создает новые синтетические изображения, которые он передает дискриминатору. Это делается с расчетом на то, что они тоже будут считаться подлинными, даже если они – сгенерированные нейронной сетью поддельные. Цель генератора – генерировать максимально приближенные к реальности рукописные цифры. Цель дискриминатора – идентифицировать изображения, поступающие от генератора, как фальшивые или же попросту неудачные, то есть некачественные [10] (рис. 6).

Сам же алгоритм данного процесса заключается в следующем:

1. Генератор принимает случайные числа и возвращает изображение.

2. Сгенерированное изображение подается в дискриминатор вместе с потоком изображений, взятых из фактического, достоверного набора данных.

3. Дискриминатор принимает как настоящие, так и поддельные изображения и возвращает число от 0 до 1, где 1 представляет собой прогноз подлинности, а 0 представляет подделку.

Вследствие этого данная модель обучается распознавать правдоподобные изображения, благодаря чему она сможет уже сама синтезировать необходимые изображения такого типа, при этом максимально приближенные к настоящим.

Такие результаты обеспечиваются за счет долгого и последовательного обучения нейронной сети, что отчасти является недостатком стандартных генеративно-состязательных сетей, поскольку процедура состязательного обучения внутри таких моделей проблематично масштабируется для моделирования сложных мультимодальных распределений.

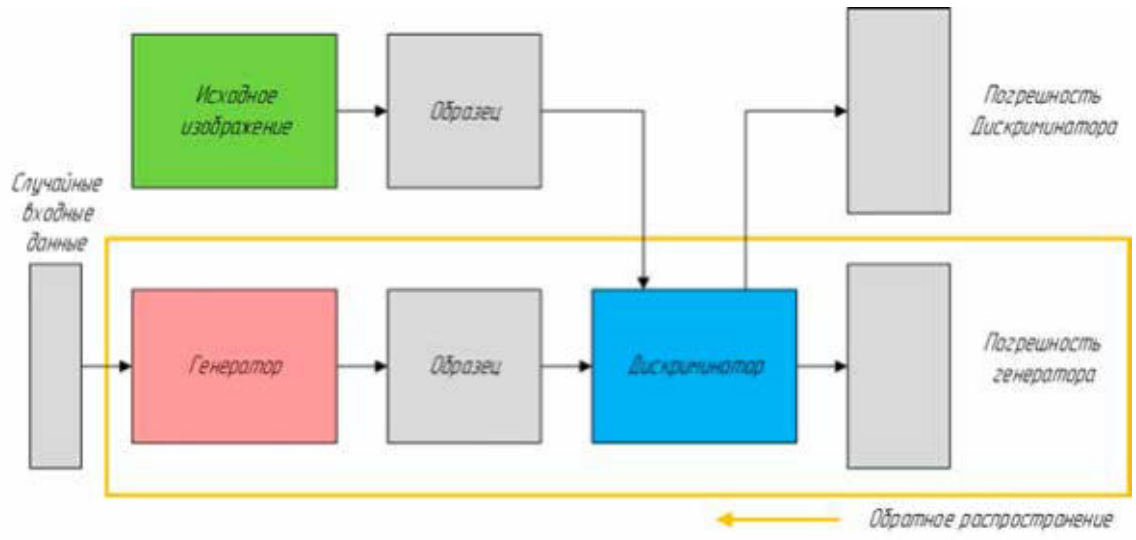

Рис. 5. Принцип работы генеративно-состязательной сети для генерации изображений

Рис. 6. Пример сгенерированных изображений цифр при помощи генеративно-состязательной сети

Заключение

Таким образом, генеративно-состязательные сети являются хорошим подходом к решению задач генерации изображений по тексту, подразумевающих использование технологии Text To Image, обеспечивая при этом достойные результаты в виде сгенерированных изображений, зачастую достаточно точно отображающих объекты, упомянутые текстовым описанием на этапе ввода данных.

Несмотря на внушительные результаты, большинство моделей, основанных на генеративно-состязательных нейронных сетях, зачастую имеют небольшие затруднения как в процессе непосредственного обучения, так и в процессе последующего использования – генерации изображений, ввиду ограниченного разнообразия выборок, что зачастую приводит к большим временным затратам, а в рамках масштабных нейронных сетей, которые обучаются на сотнях миллионов входных изображений, это приводит к замедлению процесса создания конечной модели. Однако данные недостатки компенсируются точностью и качеством результирующих изображений, что подтверждает обоснованность их применения для такого рода задач.

Список литературы

1. Zhang C., Peng Y. Stacking VAE and GAN for Context-aware Text-to-Image Generation. 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM). 2018. P. 1-5. DOI: 10.1109/BigMM.2018.8499439.

2. Dong H., Zhang J., Mcllwraith D. I2T2I: Learning text to image synthesis with textual data augmentation. 2017 IEEE International Conference on Image Processing (ICIP). 2017. P. 2015-2019. DOI: 10.1109/ICIP.2017.8296635.

3. Yanagi R., Togo R., Ogawa T. Scene Retrieval Using Text-to-image GAN-based Visual Similarities and Image-to-text Model-based Textual Similarities. 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE). 2019. P. 13-14. DOI: 10.1109/GCCE46687.2019.9015366.

4. Sun J., Zhang B. MCA-GAN: Text-to-Image Generation Adversarial Network Based on Multi-Channel Attention. 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). 2019. P. 1845-1849. DOI: 10.1109/IAEAC47372.2019.8997584.

5. Frolov S., Hinz T., Raue F. Adversarial Text-to-Image Synthesis: A Review. Neural Networks. 2021. V. 144. P. 187-209. DOI: 10.1016/j.neunet.2021.07.019.

6. Jeon E., Kim K., Kim D. FA-GAN: Feature-Aware GAN for Text to Image Synthesis. 2021 IEEE International Conference on Image Processing (ICIP). 2021. P. 2443-2447. DOI: 10.1109/ICIP42928.2021.9506172.

7. Liao W., Hu K., Rosenhahn B. Text to Image Generation with Semantic-Spatial Aware GAN. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 18166-18175. DOI: 10.1109/CVPR52688.2022.01765.

8. Zhang H. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019. V. 41. No. 8. P. 1947-1962. DOI: 10.1109/TPAMI.2018.2856256.

9. Zhang H. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV). 2017. P. 5908-5916. DOI: 10.1109/ICCV.2017.629.

10. Халтурин Е., Макарец А. Генеративно-состязательные сети: комбинирование нейронных сетей для стимулирования обучения и облегчения вычислительной нагрузки // Математика и математическое моделирование: сборник материалов XIII Всероссийской молодежной научно-инновационной школы. 2019. С. 297-299.