Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241100

Investigation related to application of Generative Adversarial Networks in text-to-image synthesis

Yun Xu

School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China

1053458869@shu.edu.cn

Abstract. Recent research attention has been captivated by the advent of Generative Adversarial Networks (GANs) in the realm of generating visuals from textual descriptions. Within a GAN framework, the interplay between the discriminator and generator components facilitates the production of lifelike visuals. This method proves to be versatile and user-friendly, allowing for the generation of authentic, diverse, and semantically faithful conditional images. However, the field still has to solve two issues: the development of high-resolution images with multiple elements and the construction of proper evaluation criteria that correlate with human perception. This paper contextualizes a number of adversarial text-to-image generation models and their core principles. This article engages in a comprehensive examination of the current methodologies employed in the analysis of text-to-image generation models, emphasizing their limitations and proposing avenues for future advancements. The discussion within this article centers on the utilization of generative adversarial networks in text-to-image synthesis, offering researchers both a comparative analysis and a benchmark for their text-to-image generation studies.

Keywords: Generative Adversarial Networks, Text-to-Image Synthesis, Deep Learning.

1. Introduction

In recent years, the intersection of deep learning and generative models has made remarkable progress in various fields of artificial intelligence. One of the compelling frontiers is image synthesis from textual descriptions via Generative Adversarial Networks (GANs) [1].

The process of creating realistic visuals from written descriptions is known as text-to-image generation. The issue has received a lot of attention due to its applications in augmented reality [2], multimedia content creation [3], and even assisting artists and designers in visualizing concepts [4]. GANs have become the main technique for generating high-quality images. The discriminator separates genuine data from created samples, while the generator attempts to create synthetic images that are comparable to real data [1]. Adversarial training between these two modules enables the generator to produce increasingly realistic images.

The field of text-to-image synthesis has been established since the work of Reed et al. was published [5]. On constrained datasets and relatively low image resolution, it showed how conditional GANs may be extended to create lifelike visuals from text descriptions. Significant progress has been made in the application of GANs to generate images from text. The multi-stage process employed by StackGAN [6] has set a new standard for generating high-quality images that closely align with textual inputs. In

DOI: 10.54254/2755-2721/55/20241100

subsequent improvements, AttnGAN [7] included an attention mechanism to the text-to-image creation process, enabling the model to focus only on particular sections of the text's description during picture synthesis. A unique deep fusion block that makes the fusion process deeper and permits adequate fusion of text and image information is also proposed by Deep Fusion Generative Adversarial Network (DF-GAN) [8]. While these models demonstrate impressive results, challenges such as fine-grained control, mode collapse, and diversity preservation remain [8].

This article is divided into the following parts. This paper first reviews the current field of image generation from text, and discusses the development of GANs methods. It then introduces the concept of GANs and their application to generating images from text. Then an in-depth analysis of different versions of GANs is given, focusing on StackGAN, AttnGAN, DF-GAN, while emphasizing their strengths and limitations. The potential applications of GANs in generating images from text are then explored. Finally, the main conclusions are discussed, and potential research fields are offered.

2. Methods

2.1. Preliminaries of GAN

Two deep neural networks, a generator and a discriminator, each of which is trained separately, make up the unsupervised architecture known as GAN [9]. As shown in Figure 1, random noise vectors are input into the generator to generate images. The discriminator's job is to receive both fake and genuine data and make an effort to tell them apart. The goal of the discriminator is to identify as accurately as possible which data is real and which is generated. With the intention of making the discriminator incapable of telling the difference between created data and genuine data, the product of the generator is sent to it for review. In order to achieve the goal of fooling the discriminator, the distribution of the images generated is continuously close to the genuine image distribution using loss function optimization [10]. To increase its resolving capacity, the discriminator makes a distinction between actual and artificial images. The two modules constantly compete and optimize one another throughout the training process. Its objective purpose is indicated by formula (1):

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1-D(G(\mathbf{z})))] \quad (1)$$

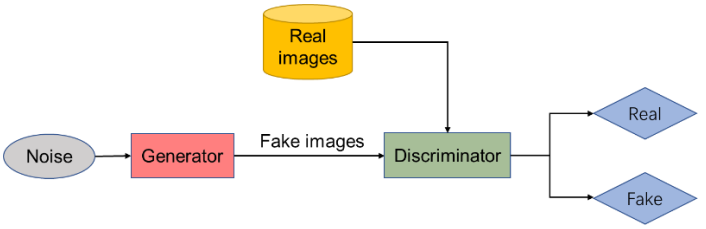

Figure 1. GAN framework (Photo/Picture credit: Original).

2.2. GAN-based text-to-image generation

Text and images serve as two vital channels through which individuals perceive the world, and investigating their interrelation constitutes a significant research endeavor. While it is feasible to extract rich and intricate semantic features from images for text-based analysis, synthesizing images directly from textual descriptions is an exceedingly intricate task [11]. The emergence of GAN provides an unsupervised model [12] to generate images. By extracting important attributes in the text (such as space, relationship between things, state of things, etc.), and then using the state of the game between the generator and the discriminator in GAN, it becomes possible to embed attributes into the image. As a result, numerous GAN variations have had outstanding applications and sustained study in the domain of text-to-image synthesis in recent years.

2.2.1. StackGAN. Cascaded GANs are used by StackGAN to enhance the detail and quality of the generated images [6]. As depicted in Figure 2, Stage-I and Stage-II make up the majority of StackGAN.

DOI: 10.54254/2755-2721/55/20241100

The Stage-I portion only produces low-resolution graphics and achieves text description information by text embedding of several conditional variables; the generated images do not pay much attention to image details, and only need to include rough outlines such as object outlines and colors information. Compared with the previous method that uses the Stage-I part to directly use noise to generate images, StackGAN directly inputs the images generated by Stage-I into the Stage-II part, and also inputs some neglected details in the text. Some errors and deficiencies in the Stage-I part have been re-corrected, thus the quality and resolution of the photographs generated are enhanced. At the same time, StackGAN proposes a conditional enhancement technique by inputting the original conditional variables and additional conditional variables generated in a Gaussian distribution into the generator. The stability of the training process is improved by adding KL divergence during the training process. The stability of the image generating process and the diversity of the created images are both enhanced by conditional enhancement technology.

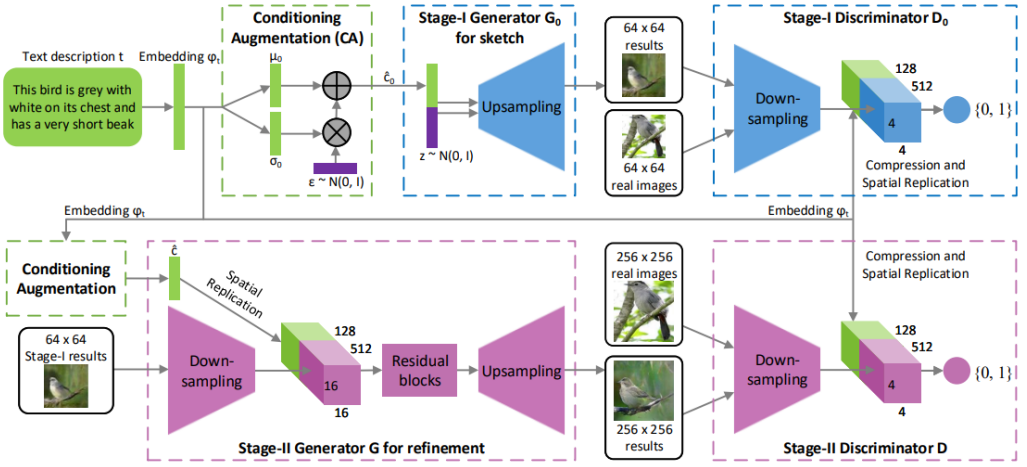

Figure 2. The structure of StackGAN [6].

2.2.2. AttnGAN. Text-to-image creation is improved by the attention-driven cascade network known as AttnGAN [7]. On StackGAN++ [6], AttnGAN's general structure shown in Figure 3 has been improved. and some special components are added. One of the important innovations of AttnGAN is the attention generation network. Establishing a global sentence vector for the entire text involves adding an attention layer to query the word vector at each stage after generating the first low-resolution image, resulting in the formation of a context vector that is then combined with the sub-region image. The vectors are combined to form multi-modal context vectors to improve the detailed features of the generated images. Another important innovation of AttnGAN is the proposed deep attention multi-modal similarity model (DAMSM). The resemblance between the image that is generated, and the associated text can be determined using this module at either a phrase level or the more precise word level, and the corresponding matching loss can be generated during the training process. The addition of the DAMSM module greatly improves the quality of synthesized images and text-image matching.

DOI: 10.54254/2755-2721/55/20241100

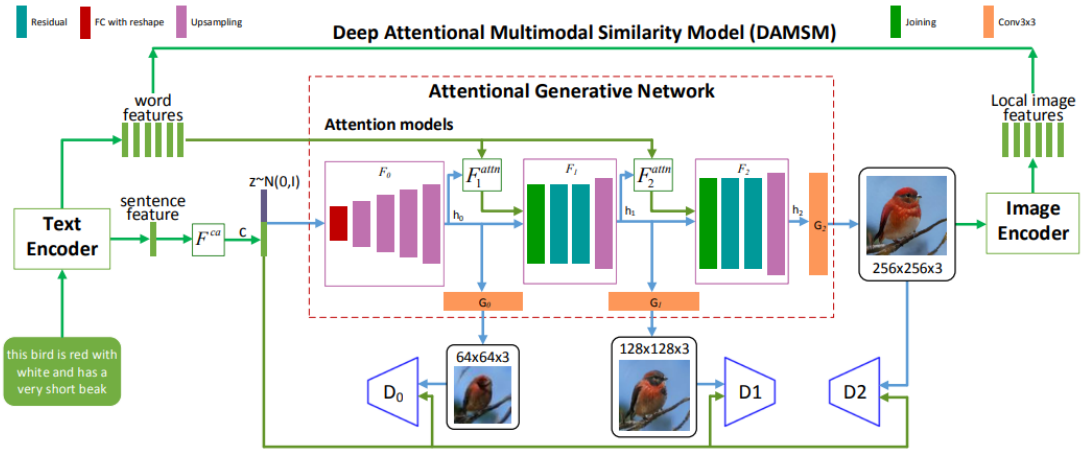

Figure 3. The structure of AttnGAN [7].

2.2.3. DF-GAN. Figure 4 shows the DF-GAN's elements, which include a generator, a discriminator, and a text encoder that has already been trained [8]. The text encoder's phrase vector and a noise vector are the generator's two inputs, respectively. This ensures the diversity of the output images. First, a fully linked layer receives the noise vector and reshapes it. The image features are then upsampled using a series of UP-Blocks. UP-Block is composed of a further sampling layer, a residual block, and a DF-Block for the purpose of combining text and picture features during the creation of images. The image characteristics are then transformed into an image via a convolutional layer.

Through a succession of DownBlocks, the discriminator turns images into image characteristics. The text vector will then be copied and joined with the features of the image. To assess the input's visual realism and semantic coherence, an adversarial loss will be projected. The discriminator makes it easier for the generator to create images with improved quality and semantic agreement with text images by separating created images from genuine samples [8].

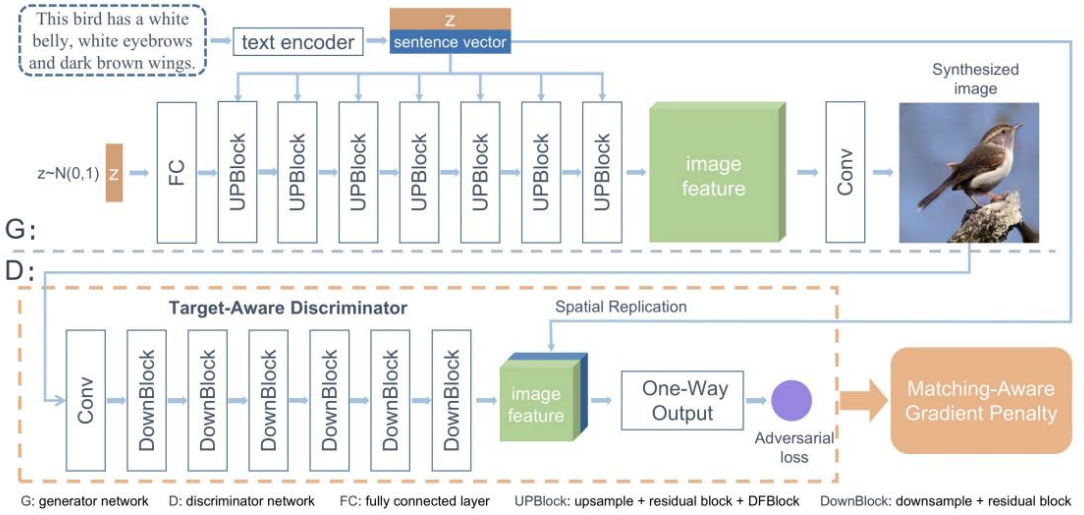

Figure 4. The structure of DF-GAN [8].

3. Applications and discussion

Text-to-image generation has many applications, such as image generation and augmentation: the process of generating an image can automatically create an image from a textual description. This is very useful in areas such as creative work, advertising design, special effects for movies, etc. For example, create advertising posters, movie scenes, or product prototypes from text descriptions. In

DOI: 10.54254/2755-2721/55/20241100

virtual and augmented reality applications, text-to-image generation can be used to create virtual environments, virtual items, and AR markers [2]. This provides a wealth of content for games, training simulations and virtual tours. Text-to-image generation can be used as an auxiliary tool for natural language processing models, converting text descriptions into images to help analyze, understand, and generate more natural text [13]. This has applications in chatbots, virtual assistants and automatic documentation generation. Within the realm of data science and visualization, the process of text-to-image generation finds application in crafting graphical data reports [14]. Users are empowered to furnish data interpretations, while generators subsequently produce corresponding charts and visualizations.

Text-to-image generation is a cross-modal research involving natural language processing and image generation [9]. Although the uniformity of images and semantics as well as the overall quality of images generated have significantly improved with the use of the present methods. However, there are still many difficulties to be studied outside of some issues that have reached a consensus (such as insufficient interpretability in the process of text image synthesis, and how to use better word embedding models to process text, etc.) [9]. In addition, most of the research on image generation from text remains at the theoretical stage, and there are not many models that can be applied in the field. The main reasons are as follows: (1) The generalization performance of the small model is too poor, and the generation of complex scene graphs is difficult to meet human expectations. (2) Only a super-large model with nearly 100 billion parameters can generate good-quality images, and the training of the model will consume huge resources. (3) The inference process of the large model is time-consuming and challenging to deploy to the terminal.

Although the application of generating images from text is still immature, existing multi-stage models based on AttnGAN [7] and single-stage models based on DF-GAN [8] have achieved the goal of generating simple images. However, it is challenging for these models to produce pictures of complicated settings with satisfactory quality, and these models can only input short text descriptions. Therefore, there is still significant work to be undertaken for practical applications when it comes to generating images from text.

4. Conclusion

One of the study areas that has advanced quickly in recent years is the cross-modal text picture generation work, which connects the computer vision and natural language processing tasks. In-depth assessments of various GAN-based text generating image techniques and network structures are provided in this article. After that, the structure, guiding principles, and major contributions are presented. Finally, possible shortcomings and potential future research initiatives in this area are reviewed objectively. While there have been notable advancements in the field of text-to-image synthesis, there remains ample room for enhancement. This pertains to the creation of higher-quality visuals that more accurately align with the input text's intended meaning, guiding user research, and affording developers greater flexibility in crafting user-friendly interfaces. The belief is that this research will aid scholars in gaining a deeper understanding of the current state of the art within this domain and in identifying the persisting challenges that demand resolution.

References

[1] Goodfellow I et al 2014 Generative adversarial nets Advances in neural information processing systems 27

[2] Liu D et al 2020 Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 8139-8148

[3] Kumar L and Singh D K 2023 A comprehensive survey on generative adversarial networks used for synthesizing multimedia content. Multimedia Tools and Applications 1-40

[4] Shahriar S 2022 GAN computers generate arts? a survey on visual arts, music, and literary text generation using generative adversarial network Displays 73 102237

DOI: 10.54254/2755-2721/55/20241100

[5] Reed S et al 2016 Generative adversarial text to image synthesis. In International conference on machine learning pp 1060-1069 PMLR
[6] Zhang H et al 2017 Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks In Proceedings of the IEEE international conference on computer vision pp 5907-5915
[7] Xu T et al 2018 Attngan: Fine-grained text to image generation with attentional generative adversarial networks In Proceedings of the IEEE conference on computer vision and pattern recognition pp 1316-1324
[8] Tao M et al 2022 Df-gan: A simple and effective baseline for text-to-image synthesis In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 16515-16525
[9] Wang W et al 2022 A Review of Generative Adversarial Networks and Text-to-Image Synthesis Computer Engineering and Applications (19) 14-36
[10] Wei Z and Ma L 2022 Bidirectional Network for Image Generation and Classification In 2022 International Conference on Computer Network, Electronic and Automation (ICCNEA) pp 229-233 IEEE
[11] Frolov S et al 2021 Adversarial text-to-image synthesis: A review Neural Networks 144 pp 187-209
[12] Pan Z et al 2019 Recent progress on generative adversarial networks (GANs): A survey IEEE access 7 pp 36322-36333
[13] Karuna E N et al 2022 Generative adversarial approach in natural language processing In 2022 XXV International Conference on Soft Computing and Measurements (SCM) pp 111-114 IEEE
[14] Gui J et al 2021 A review on generative adversarial networks: Algorithms, theory, and applications IEEE transactions on knowledge and data engineering 35(4) pp 3313-3332