How to Imagine the World with Text? From Text-to-image Generation View

Jingyi Liu *

Department of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin, China

* Corresponding author email: 205219@stu.hebut.edu.cn

Abstract. Words are an effective and convenient way to describe the world, but sometimes what the texts convey may be misunderstood by readers. The expression of pictures is more vivid, easy to understand and has no borders, but creating a painting often takes a long time. Text-to-image makes the two expressions complement each other: It makes every ordinary person a “painter”, so that they can feel the world, express themselves, and create more whimsy through many rich pictures. For this vision, technologists are trying their best to improve image generation models, which enables computers to generate high quality images with texts better. And they are solving some technical defects, for instance, sometimes the content of generated images is strange. In the future, text-to-image can be adapted to applications in AI such as computer-aided design, image editing, and be employed in the field of art such as movies and artworks, and then it may even make a big difference on people's life, enriching the public's spiritual world and conveying information by vivid images.

Keywords: Text-to-image; Image Generation; GANs; cGAN; CLIP.

1. Introduction

Text-to-image is a kind of “magic” that can turn the user's wild imagination into visual images through rich text descriptions, which is the result of a combination of computer vision and machine learning. The conversion process basically depends on the correlation of text and image to find the best alignment between them [1].

There are two main tasks to implement text-to-image: Apply natural language processing to understand descriptions in inputs and the other one is to generate accurate and natural images to express the text [2]. Initially, Text-to-image was implemented by search and supervised learning. We employ keywords to identify informative and “picturable” text units and search for related images based on the units, and finally we optimize the picture layout to get the results [3]. The major limitation of the traditional methods is that it can only synthesize and change given images by searching for existing image parts instead of generating new content. Later since deep learning has progressed, the advent of deep generative models made it possible to create visually realistic images through trained neural network models. Over the next few years, there has been significant progress through a combination of methods, and we are always improving the quality and accuracy of the generated images. The optimized models can not only make the generated images accomplish the original effect, but also can generate more matched and coordinated images, for example, by using the machine's ability to deal with long and complex sentences. Simultaneously, images are also more clear. Now we mainly exploit models such as GANs and CLIP to complete the current text-to-image tasks. The paper will focus on basic architecture and principles of generative adversarial networks, conditional GAN and CLIP.

Structure of the rest of the paper: In the second part, we mainly explain something about the models, such as their background, functions and limitations. In the third part, the achievements and features of text-to-image will be presented. The conclusion will be pointed out in section four.

2. Method

2.1 GANs

Ian Goodfellow et al. were the first to proposed GANs in 2014 [2]. GANs are applied in the image generation process since the goal of adversarial training is to enable generators to create images that are comparable to the real images. GANs consist of two "adversarial" model, a generative model and a discriminative model [4]. Although the two are trained with conflicting aims, just like enemies, they can work together.

The generator attempts to trick the discriminator by creating sham samples that are relevant to the original data distribution, while the discriminator attempts to calculate the probability of a sample coming from the true data distribution rather than the generator, so that it can tell which samples were from the generator and which ones were from the training data. The generator and the discriminator both improve their abilities when each defeats the other throughout the iterative training process of GANs. The generator's task is to reduce the discrepancy between real and fake images, whereas the task of the discriminator is to enhance its ability of differentiating between true and fake images. As iteration goes on, each model becomes more proficient at its task and eventually a threshold is reached. The entire procedure between the two models resembles a game of min-max.

$$\min_{\theta_g}\max_{\theta_d} V\left(D_{\theta_d}, G_{\theta_g}\right) = E_{x\sim Pdata(x)}\left[\log\left(D_{\theta_d}(x)\right)\right] + E_{x\sim P_z(z)}\left[\log\left(1 - D_{\theta_d}\left(G_{\theta_g}(z)\right)\right)\right] \quad (1)$$

In Equation (1), x is a multidimensional sample such as an image, and Dθd() is a discriminator function which can categorize a sample into a binary space, so Dθd(x) designates a binary output for an image x, such as true/false. Z is a multidimensional latent space vector and the generator function called Gg () seeks to create a sample from z vector, so Gθg(z) intends to generate a fake image using a latent vector z. For GAN, the equation shows the relationship between images, the discriminative model and generative model. The output from Dθd(x) should be 1 if the discriminator is given a true image, while the ideal output from Dθd(Gθg(z)) should be 0 if that is given a result created from the generator Gθg(z).

GANs have been already applied in several applications such as image super-resolution, image in-painting, data augmentation, style transfer, image-to-image translation, and representation learning [5].

2.2 Conditional GAN

One of the disadvantages of GANs is that its outputs are uncontrollable and random, which means that we cannot anticipate and eliminate some images. To solve the problem, Mirza and Osindero put forward conditional GAN in 2014, whose central idea is to be able to control images generated by GANs [6]. Compared with GANs, cGAN's main technical innovation is to add a class label y into both the discriminator and the generator. Y may contain attribute information such as the category of the image and the facial expression of the face image. The following is the equation for cGAN:

$$\min_{\theta_g}\max_{\theta_d} V\left(D_{\theta_d}, G_{\theta_g}\right) = E_{x\sim Pdata(x)}\left[\log\left(D_{\theta_d}(x\backslash y)\right)\right] + E_{x\sim P_z(z)}\left[\log\left(1 - D_{\theta_d}\left(G_{\theta_g}(z\backslash y)\right)\right)\right] \quad (2)$$

Furthermore, cGAN's condition vector can not only be the class label, but also some other forms such as texts. And only images generated by the generator that meet conditions can pass the discriminator, which makes outputs more targeted.

2.3 GAN Frameworks

Based on several key points that mainly affect outputs in the image generation process, we classified GAN frameworks into four categories:

1. Semantic enhancement GANs are employed to assure that the created images are semantically related to the text, which is improved by encoding texts as dense features through a neural work and then feeding the features to a second network to generate images that match the descriptions well [2];
2. Resolution enhancement GANs aim to enhance the level of visual quality of generated images, which is realized by a multistage GAN framework.
3. Diversity enhancement GANs are designed to increase the variety of outputs, for instance, different styles of generated images such as pixel style and abstract art style.
4. Motion enhancement GANs are able to add a temporal dimension to the outputs so that they can form meaningful actions related to the text, which are mostly suitable to generate videos [2].

2.4 CLIP

Contrastive Language Image Pre-training (CLIP) is an efficient method to directly learn from the raw text about relevant images [7]. It aims to show matching rate of text and images, learning the link between textual and visual representations of the same abstract object. And it outperforms the best available ImageNet model with better computationally efficiency [8]. The following figure show its training process (as shown in Figure 1). The model architecture is divided into two parts, text encoder and image encoder. Sentences pass through the text encoder to get N vectors, denoted as TN, and in a similar way, images pass through the image encoder which also gets N vectors, denoted as IN. Then we build up a matrix, with their inner products, and the training objective is to maximize the inner products of vectors of correct image/caption pairs, which are the elements on the diagonal of the matrix, while minimizing the inner products of vectors of unrelated image/caption pairs.

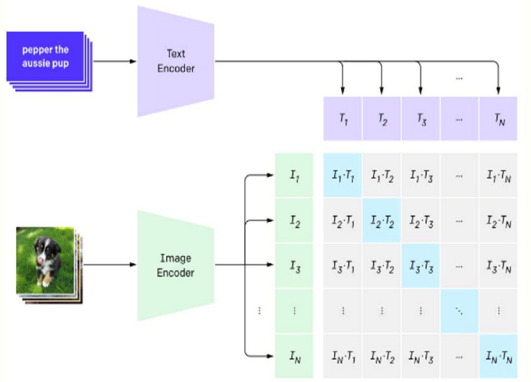

Fig 1. The schematic of CLIP

However, CLIP also has several shortcomings. For instance, it does not perform well on some systematic or abstract tasks such as counting. And sometimes when meeting complex tasks such as forecasting distance between the two objects, its results may be not satisfactory [9].

3. Current Achievements:

Text-to-image has made major breakthroughs in recent years. Let's take a look at some of the best models.

Google recently unveiled a new model called Parti, which is considered to be the most advanced text-to-image technology. Its process of image synthesis is similar to that of machine translation, and the most significant difference between them is that Parti's ideal results are image tokens rather than text tokens [8]. The following are its powerful functions:

Long vocabulary and rare words identifiability: Parti has a large vocabulary and can identify rare words. For example, if we input "pneumonoultramicroscopicsilicovolcanoconiosis" (a lung disease), which is recognized as the longest word in the world, it will generate multiple reasonable lung disease images (shown in Figure 2).

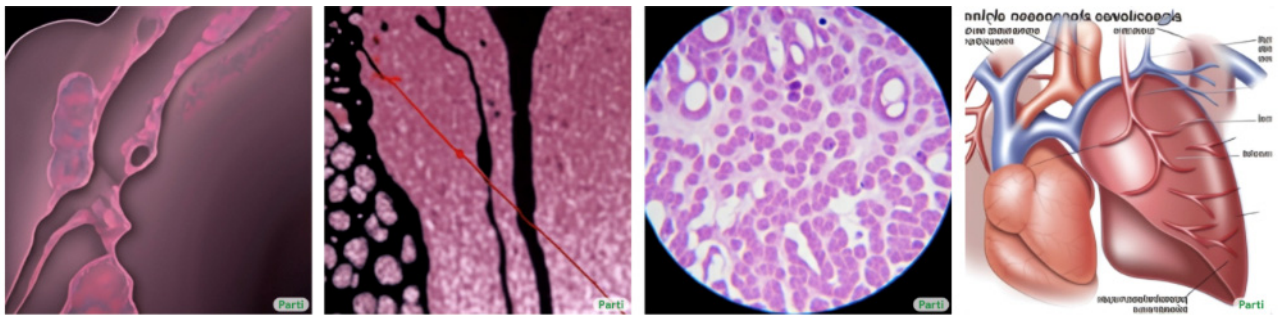

Fig 2. The illustration of "pneumonoultramicroscopicsilicovolcanoconiosis"

Strong comprehension: Parti understand long paragraphs accurately. "Oil-on-canvas painting of a blue night sky with roiling energy. A fuzzy and bright yellow crescent moon shining at the top. Below the exploding yellow stars and radiating swirls of blue, a distant village sits quietly on the right. Connecting earth and sky is a flame-like cypress tree with curling and swaying branches on the left. A church spire rises as a beacon over rolling blue hills." We fed it the 64-word paragraph and it generated the night sky perfectly (as shown in Figure 3).

Fig 3. Night sky generated by Parti

Fig 4. Generated raccoon in different kinds of painting

Variety of styles of generated images: Images Parti generates are in a variety of styles, such as Van Gogh style, Egyptian pharaoh style, pixel style, abstract art style and so on. For instance, we told Parti that "Raccoon in formal suit and top hat, leans on crutches and holds a garbage bag", and images it generated contains many kinds of paintings(shown in Figure 4).

High quality images generation: Generated images are very realistic and in high definition(shown in Figure 5). Especially its treatment of light and shadow makes generated images look like a real photograph. And the magnification of objects is also very clear.

Fig 5. Realistic images for "The back of a violin"

DALL-E is also a very representative model for text-to-image. In an experiment to assess DALL-E 2's abilities, we found lots of impressive advantages. For example, the visual quality of its generated images is quite wonderful and it is certainly outstanding in terms of image generation [10].

Rich "imagination": DALL-E can expand its "imagination" to design realistic details that are not mentioned in the text. For example, we input "A road sign with an image of a blue strawberry", and then DALL-E generated different kinds of images that all match the description of the text(shown in Figure 6).

Fig 6. Road signs in different scenes generated by DALL-E

That's because we did not give information about details such as the background, color, shape of the sign, so DALL-E created them randomly. Just like different people have different imaginations of the same description, it can improvise on details that are not mentioned.

Strong ability to deal with peculiar sentences: Researchers surprisingly found that it could complete well even on peculiar sentences or complex sentences that are difficult to understand. For example, we input the sentence, "A snail made of harp. A snail with the texture of a harp", and got the following images (as shown in figure 7).

Fig 7. DALL-E's combination of a snail and a harp

The description of the sentence may be even difficult for human to understand and imagine because it describes a quiet uncommon thing that we have hardly seen before. But DALL-E combined these two unrelated concepts perfectly. It created something new based on something close to reality.

So, DALL-E can understand your wild imagination and irrational texts, and give bold but logical results.

Ability to edit images through text descriptions: Outpainting is an advanced function. It means that DALL-E can not only generate images out of thin air, but also edit existing images according to the text. We can apply the function in many ways such as video expansion and panorama creation. [11].

Look at the famous painting (as shown in Figure 8), someone edited the painting in the way he or she wanted it to be.

Fig 8. "Mona Lisa" edited by DALL-E

4. Conclusion

Text-to-image can open our minds and expand our imagination, especially for those of us who work in art and design. In the meantime, its development has greatly boosted AI, which also play a key role in science and technology. That is mainly because the improvement of image generation technique not only can make text-to-image more perfect, but also it will be a major breakthrough in computer science and related fields.

However, there are still some issues about text-to-image, such as copyright. Deep generative models employ massive data sets to generate images, which inevitably involve some unauthorized works. So the copyright of generated images trained with unauthorized works is controversial. In addition, some people feel that the popularity of text-to-image will bring some social problems and the authenticity of pictures cannot be guaranteed which makes it difficult for the police to take evidence when people with ulterior motives edit pictures maliciously to fabricate facts. Anti-social images generated by outlaws or terrorists may also have a negative impact on human life. And we cannot avoid the inappropriate treatment of pictures by uncivilized people, especially famous paintings. So we should conduct something like legislating to avoid its improper influence before it is widely utilized.

So far, the technology of image generation has not been as successful as imagined. For example, researchers found that currently applied evaluation metrics are not suitable for assessing text-to-image synthesis models, and sometimes generated images look strange and scary. In the future, if I have the opportunity to continue to explore and learn this professional field, I am willing to work hard to understand the crux of the problem, and then execute investigate and simulations to upgrade the technology. In a word, text-to-image still has a long way to go.

References

[1] Lee H, Ullah U, Lee J S, et al. A Brief Survey of text driven image generation and manipulation [C]//2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia). IEEE, 2021: 1-4.

[2] Agnese J, Herrera J, Tao H, et al. A survey and taxonomy of adversarial neural networks for text-to-image synthesis[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, 10(4): e1345.

[3] Zhu X, Goldberg A B, Eldawy M, et al. A text-to-picture synthesis system for augmenting communication [C]// AAAI. 2007, 7: 1590-1595.

[4] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[5] Frolov S, Hinz T, Raue F, et al. Adversarial text-to-image synthesis: A review[J]. Neural Networks, 2021, 144: 187-209.

[6] Zhou Rui, Jiang C, Xu Qi. A Review of Text-to-Image Synthesis Based on Generative Adversarial Networks [J]. Neural Computing, 2021, 451: 316-336.

[7] Li Y, Liang F, Zhao L, et al. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm[J]. arXiv preprint arXiv:2110.05208, 2021.

[8] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[9] Yu J, Xu Y, Koh JY, et al. Scaling autoregressive models for content-rich text-to-image generation[J]. arXiv preprint arXiv:2206.10789, 2022.

[10] Marcus G, Davis E, Aaronson S. A very preliminary analysis of DALL-E 2[J]. arXiv preprint arXiv: 2204. 13807, 2022.

[11] Sabini M, Rusak G. Painting outside the box: Image outpainting with gans[J]. arXiv preprint arXiv: 1808. 08483, 2018.