Behind the AI Art Creation: A Study of Generative Models for Text-to-Image Generation

Lege Zhao[1, a], Han Zhang[2, *]

[a]zhaolege@dufe.edu.cn; [*] Corresponding Author. hanzhang@dufe.ed.cn

School Humanities and Communication, Dongbei University of Finance and Economics, Dalian, China[1], School of Data Science and Artificial Intelligence, Dongbei University of Finance and Economics, Dalian, China[2]

Abstract. The advancement of deep learning has greatly facilitated computer vision and natural language processing. Among its applications is text-to-image generation, which involves creating images from textual descriptions. Recent text-to-image techniques offer a compelling yet straightforward ability to convert text into images, making them a prominent research topic in both AI and art creation. Image generation from text holds a myriad of practical and creative applications in computer design and the creation of digital art. This paper conducts a comprehensive study to review three types of generative models for text-to-image generation, aiming to provide a foundational understanding of the principles underlying these models.

Keywords: AI Art Creation; Test-to-Image Generation; Generative Models

1 Introduction

Generative models have recently garnered attention due to their use in producing counterfeit images. The rise of AI-generated deceptive images, commonly referred to as "deep fakes," presents a host of challenges, encompassing the creation of realistic synthetic images, images featuring complex objects, and the establishment of dependable evaluation metrics that align with human discernment. Nonetheless, novel technologies offer promising prospects for image generation. The development of a system capable of generating images that accurately convey a given textual description, mirroring human perception, marks a pivotal advancement in computational intelligence. Before the development of general models, the process of image generation from text relied on image querying. This intricate task entails meticulously selecting an optimal collection of images from an image database to illustrate text descriptions. However, recent strides in artificial intelligence and computer vision have notably streamlined the generation of images based on textual prompts. The purpose of text-image synthesis is to generate images from text descriptions. Owing to its diverse applications across domains such as arts, marketing, business, and education, text-to-image research has garnered substantial attention. Several frames and enhancements have been suggested to create more realistic visuals.

The depiction of an image through numeric pixel values is termed as image data distribution. Consider an image consisting of pixels. It can be presented as a vector of size , corresponding to a distribution of dimensional data. The multidimensional distribution of image data

ADDT 2023, September 15-17, Xian, People's Republic of China

DOI 10.4108/eai.15-9-2023.2340842

introduces complexity to the task of image generation. In the realm of full-text image synthesis, the generated images must adhere to both visual realism and semantic accuracy. Semantic accuracy refers to the coherence between image content and textual description. The model operates by providing it with a description of the intended output, which it then translates into an image. These models have the ability to autonomously learn from input data, reproducing it dynamically and precisely.

There are many different types of generative models, such as Auto-encoders[1], Generative Adversarial Networks (GANs)[2], and Diffusion Models[3] designed for text-to-image generation. A comparative analysis of these models has been conducted based on criteria such as visual realism, diversity, and semantic alignment, aiming to discern the optimal model for generating text-associated images. Notably, DALL-E[4], a prominent text-to-image generator, was made accessible to the public by OpenAI shortly after introducing the broadcast model. Diffusion modeling has demonstrated its capability to yield high-quality images. DALL-E's initial release transpired in January 2021, followed by DALL-E 2 in 2022. OpenAI embraced this approach as the foundation for DALL-E 2 due to its utilization of straightforward image-denoising networks, leading to convex regression loss reduction, rather than relying on minimax optimization. DALL-E and other versatile AI imaging tools represent the forefront of innovation that captures the attention of venture capitalists. Text-to-image generation algorithms hold the potential to automatically create and colorize characters based on textual descriptions, thereby yielding images suitable for books and educational materials, enhancing visual learning's accessibility. Such generative models can be harnessed by short story writers to complement their narratives with illustrative imagery. Artists, too, stand to gain inspiration from these models, which can aid in the creation of original artwork. Developers can leverage these tools to craft and manipulate images for their websites or applications, thereby mitigating concerns related to image costs and copyright issues. Consequently, the potential applications of future text-to-image generation are boundless. In this paper, we conduct a study of the Variational AutoEncoder (VAE)-based text-to-image generative models in Section 2 and the GAN-based text-to-image generative models in Section 3. The conclusion is made in Section 5.

2 VQ-VAE-based text-to-image generation model

A significant challenge with VAEs lies in their reliance on a fixed prior followed by the utilization of continuous intermediate representations, resulting in limitations in image generation diversity and controllability. Addressing this concern, the Vector Quantized Variational Autoencoder (VQ-VAE) opts for discrete intermediate representations coupled with an auto-regressive model to refine the prior estimation (e.g. PixelCNN or Transformer). Within the framework of VQ-VAE, the intermediate representation remains stable and diverse, exerting influence over the decoder output and facilitating the generation of more intricate and varied images. Consequently, numerous models for text-driven image generation are grounded in the principles of VQ-VAE[5]. The architecture of VA-VAE is presented in Figure 1. The architectural blueprint of VQ-VAE is depicted in Figure 1. A pivotal component of VQ-VAE is the Codebook query operation, adept at augmenting the controllability and richness of image generation by virtue of coherent codebooks, thereby circumventing convoluted intermediate representations. The procedure of the VA-VAE algorithm is as follows: (1) Set K

vectors as queryable codebook; (2) The input image is passed through the encoder CNN network to obtain intermediate representations , followed by querying the Codebook using the nearest neighbor algorithm to identify the vector most akin to the N intermediate representations; (3) By placing the similar vectors queried in Codebook at the corresponding positions of $z_e(x)$, we obtain $z_q(x)$; (4) The decoder reconstructs the images from $z_q(x)$.

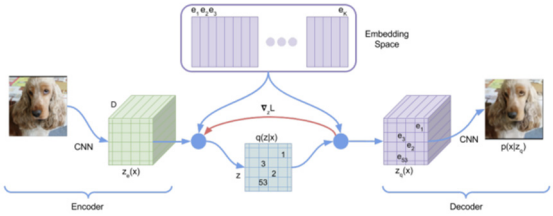

Fig. 1. Architecture of VQ-VAE.

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

Fig. 2. Two examples of images generated by DALL-E. (a) and (b) are two the input text description of the generated images.

The core part of VQ-VAE is the Codebook query operation, which can effectively improve the controllability and richness of image generation by using highly consistent codebooks instead of confusing intermediate representations. The loss function of VQ-VAE is defined as:

$$L_{total} = \log p(x|z_q(x)) + \|sg[z_e(x)] - e_k\|_2^2 + \beta\|z_e(x) - sg[e_k]\|_2^2 . \quad (1)$$

Where, $sg[\cdot]$ is the gradient pause operation, the module in which $sg[\cdot]$ is located will not perform gradient updates.

The first term of the loss function in equation (1) is mainly for the encoder and decoder. In this process, since the intermediate codebook query operation is discrete, here the gradient of $z_q(x)$ is directly copied to $z_e(x)$, thus forming the gradient backpropagation. The second term, called VQ loss, aims to train the codebook module e_k, which will force the codebook module e_k towards closer to $z_e(x)$since it is fixed here. The third term is called commitment loss, where only the gradient of $z_e(x)$ is changing, which in turn aims to get $z_e(x)$ closer to the codebook module e_k, thus making the outputs of the encoder module more stable.

Developed by OpenAI, DALL-E's initial version incorporates VQ-VAE and ranks among the prominent text-to-image models. As of now, the initial version of DALL-E is not accessible, but there is a DALL-E-mini version available for exploration. A distinctive attribute of the first-generation DALL-E lies in its remarkable grasp of semantics, enabling the generation of various unconventional but semantically consistent images[6]. The generation module within the DALL-E model employs the principles of VQ-VAE for its operations. As illustrated in Figure 2, two examples of generated images exemplify this process. Notably, the distinction lies in its prior learning approach, which involves mapping from text to intermediate discrete representations. The training protocol encompasses the subsequent steps: (1) Training a discretized VAE (referred to as dVAE in the paper, although it functions as a VQ-VAE), with the utilization of 8192 codebooks; (2) Training an auto-regressive model, specifically a Transformer in this instance, to anticipate intermediate representations by inputting the text. In the generation process, the text is directly input, subsequently projected through the Transformer to predict the intermediate representation, which in turn serves as input to the Decoder module of dVAE for generating the final image. For instance, it posits the utilization of the Contrastive Language-Image Pre-training (CLIP) model to identify the model with the highest text similarity for image selection during the final stages. Moreover, the paper underscores methodologies such as distributed training and mixed precision training. For comprehensive insights, readers are referred to the original paper.

3 VQGAN

Vector Quantized Generative Adversarial Network (VQGAN)[7] represents one of the variants of the GAN paradigm (Figure 3.), which is inspired by VQ-VAE and uses a codebook to learn discrete representations. Specifically, a predefined vector functions as a lookup table for discrete features. As part of its operation, when an image is inputted into the CNN Encoder, an intermediate representation $\hat{z}$ with n_z images is extracted. $\hat{z}$ is subsequently matched against the Codebook to identify the most analogous vector, thereby procuring z_q, which consists of n_z representations. Then, the CNN decoder will reconstruct the image based on the obtained representation z_q. The above process can be formulated as in equation (2):

$$z_q = \boldsymbol{q}(\hat{z}) \coloneqq \left(\arg\min_{z_k \in Z} \left\| \hat{z}_{ij} - z_k \right\| \right). \tag{2}$$

The training process of VQGAN closely mirrors that of VQ-VAE. Diverging from VQGAN, the procedures inherent in VAGAN encompass solely the generator facet within the GAN structure; the presence of a discriminator is imperative to assess the quality of each generated image patch.

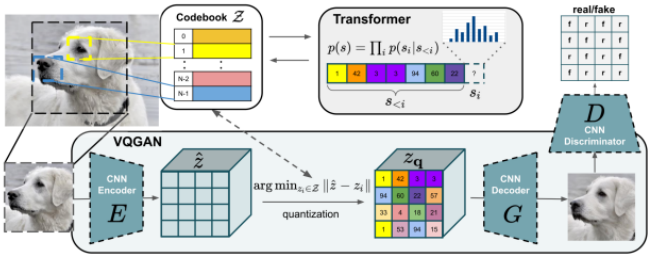

Fig. 3. Architecture of VQ-GAN.

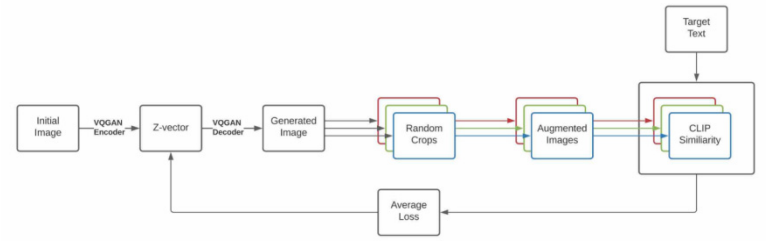

Fig. 4. Training process of VQ-GAN-CLIP.

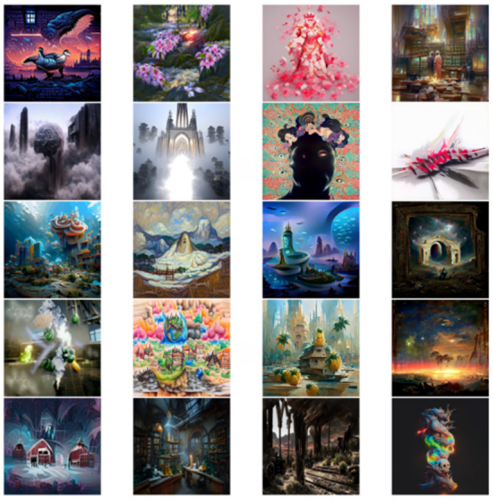

Fig. 5. Examples of Generated Pictures by the VQ-GAN-CLIP.

VQGAN-CLIP, a widely acclaimed model for generating images from textual prompts, is harnessed by certain open platforms specializing in text-based image generation[8]. It capitalizes on textual descriptions to guide the VQGAN model in producing images closely aligned with the textual prompts. The specific training process is elucidated in Figure 4, showcasing that complex descriptions pave the way for generating high-quality images through VQGAN-CLIP. Notably, Figure 5 provides illustrative examples of such high-quality image generation.

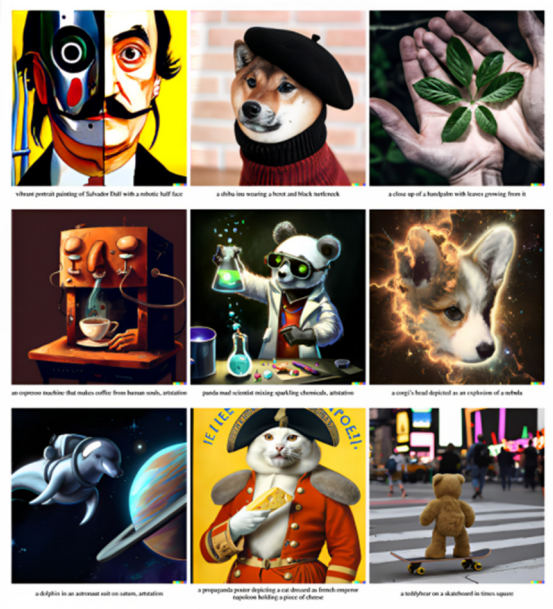

Fig. 6. Pictures Generated Pictures by the DALL-E2.

4 Diffusion-based models

DALL-E2 is OpenAI's latest AI-generated image model. Its standout feature lies in its exceptional comprehension and creativity. Its parameters are about 3.5B. Employing a manual evaluation approach, the author enlisted volunteers to assess 1000 images. 71.7% of respondents perceived enhanced alignment between the images and accompanying textual descriptions, while 88.8% found the images to be more aesthetically appealing than the previous version[9] (Images generated by DALL-E2 are shown in Figure 6).

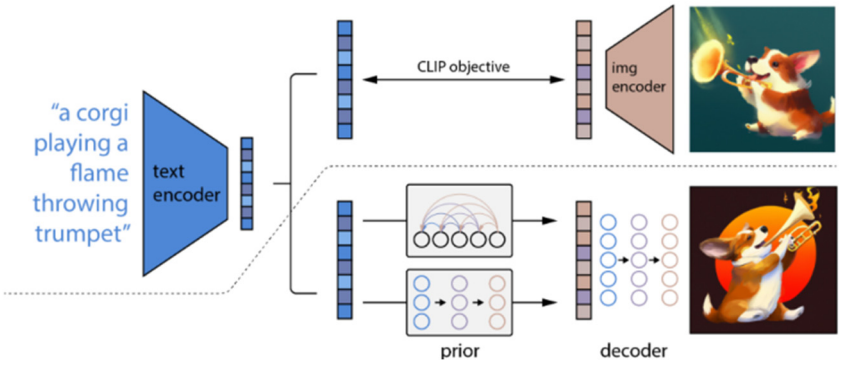

Fig. 7. Modules of the DALL-E2.

DALL-E2 is structured around three core modules, as elucidated in Figure 7: (1) The CLIP model, instrumental in aligning image-text representations; (2) The prior model, designed to receive text input and subsequently transform it into the corresponding CLIP image representation; (3) Diffusion models, responsible for generating complete images through the utilization of the image representation.

5 Comparisons

The COCO (Microsoft Common Objects in Context) dataset, which was funded and annotated by Microsoft in 2014. Along with the ImageNet competition, MS COCO is regarded as one of the most watched and authoritative competitions in computer vision and it is a large and rich dataset for object detection, segmentation, and captioning dataset. This dataset include 91 classes of objects, 328,000 images, and 2,500,000 labels. The COCO joint point detection data set constructed by Microsoft includes a training set, a validation set and a test set. It contains 200,000 images and 250,000 human targets with 17 joint point information labeled.

We compare the above four models: VQVAE, VQGAN, VQGAN-CLIP and DALL-E2 on the above dataset COCO with their FID (Frechet Inception Distance) score, which measures the similarity of two sets of images from the statistical aspect of the computer vision features of the original image, and is a measure of the distance between the real image and the generated image feature vectors. The FID results are shown in the Table 1. Among which, VQGAN-CLIP and DALLE-E2 have similar FID scores, which means they can generate images that are closer to the real images than the other two methods.

Table 1. FID Score of VQVAE, VQGAN, VQGAN-CLIP and DALL-E2

Methods	FID Score
VQVAE	35.49
VQGAN	20.79
VQGAN-CLIP	18
DALL-E2	15

6 Conclusion

The existing text-to-image generation models are mainly based on three basic algorithms: VQ-VAE, VQ-GAN, and diffusion-based models. As the diffusion-based models can generate rich, diverse, and high-quality graphics, it has become the core method in field of text-to-images generation. The training of diffusion-based model are slower compared to the other two types of models due to the extra number of iterations required for each generation, which is the most significant problem that limits the widespread uses of diffusion-based models. However, with the emergence of some new technologies, the generation time of the diffusion model has been gradually shortened. In the future, the diffusion model will bring a new revolution to the field of AI art generation.

Acknowledgments. This work is supported by the Applied Basic Research Program of Liaoning Province (2023JH2/101600040) and the Basic Scientific Research Project of Colleges and Universities of the Educational Department of Liaoning Province.

References

[1] Kingma, D. and Welling, M. Auto-encoding variational bayes. ICLR. pp. 1312-6114 (2013)

[2] Gregor K, Danihelka I., Graves A., Rezende D., et al. Draw: A recurrent neural network for image generation. In International conference on machine learning. pp. 1462-1471 (2015)

[3] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In International conference on machine learning. pp. 1278-1286 (2014)

[4] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems. Vol. 34, pp. 8780–8794 (2021)

[5] Van Den Oord A, Vinyals O., et al. Neural discrete representation learning. Advances in neural information processing systems, Vol. 30 (2017).

[6] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. International Conference on Machine Learning. pp. 8821-8831 (2021).

[7] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. The IEEE/CVF conference on computer vision and pattern recognition. pp. 12873-12883 (2021)

[8] Crowson K, Biderman S, Kornis D, et al. Vqgan-clip: Open domain image generation and editing with natural language guidance. European Conference on Computer Vision. pp. 88-105 (2022)

[9] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems. Vol. 35, pp. 36479-36494 (2022)