Journal of Artificial Intelligence Practice (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2023.060110
ISSN 2371-8412 Vol. 6 Num. 1

AI Application to Generate an Expected Picture Using Keywords with Stable Diffusion

Zhan Shi[1]

[1]*Shanghai Guanghua Qidi College, Shanghai, China*

Keywords: Stable Diffusion; Diffusion model; Image processing; Ethical issues

Abstract: Nowadays, artificial intelligence (AI) has a big impact in the field of painting. In contrast to the hand-painting and challenging personal creativity, AI applications practices to add noise, remove noise, restore image and conserve the present process after converting to data. The various AI image generator models, Diffusion model is the latest application consist of two main models, word-image mapping and diffusion algorithm. This article is introduced the mechanism of how images are generated and makes arguments about ethical issues of AI application of image generation. As a result, sophisticated legal restrictions should be implemented to prevent possible negative effects of AI models.

1. Introduction

The potential of AI image generation is eye-opening, which can start with simple text descriptions and change the image of high quality. With the open-source code of Stable Diffusion, it illustrates a remarkable model for public. At the same time, this model remains rapid operating speed and lower memory requirement. Having used this amazing technology of Diffusion, you may be wondering how it works so well.

Introducing the foundations of Diffusion Model at first will help to have better understanding of how Stable Diffusion generate images. Diffusion model makes use of the U-Net in the field of image segmentation, the training loss is stable, and the model performs very well. Compared to GAN's requirement of training against a discriminator, or VAE, which needs a variational posterior, estimating the diffusion model's loss is really simple. Such simple and efficient training also allows the diffusion model to perform very well in many tasks and have broken long-term dominance of GAN in image synthesis. [1]

Following that, a deeper look at how the Diffusion Model works, using probability and examples can be presented. Briefly, Key point of all these approaches is to progressively perturb data with intensifying random noise (called the "diffusion" process), then successively remove noise to generate new data samples.[2] During this process, two Markov chains are used: a forward chain perturbs the original image to noises, a reverse chain converts the noises back to a new image.

Firstly, the AI models for painting are discussed firstly. Secondly, the result of these models is presented. Lastly, conclusion of these results is presented, and future works are mentioned.

2. Related Work

Before exploring how Stable Diffusion work, the distinct features of GAN and VAE may display the superiority of Diffusion model, which is very similar to how Stable Diffusion works. Figure 1 shows simple processes of different image generators.

GAN only needs a 'generator', which samples the Gaussian noise and then uses the 'generator' to map this Gaussian noise to the data distribution. We really only care about the generation, and in most cases, we don't care about what the posterior distribution actually is. But there are other drawbacks, such as the fact that it requires additional training of the discriminator, which makes training very difficult.

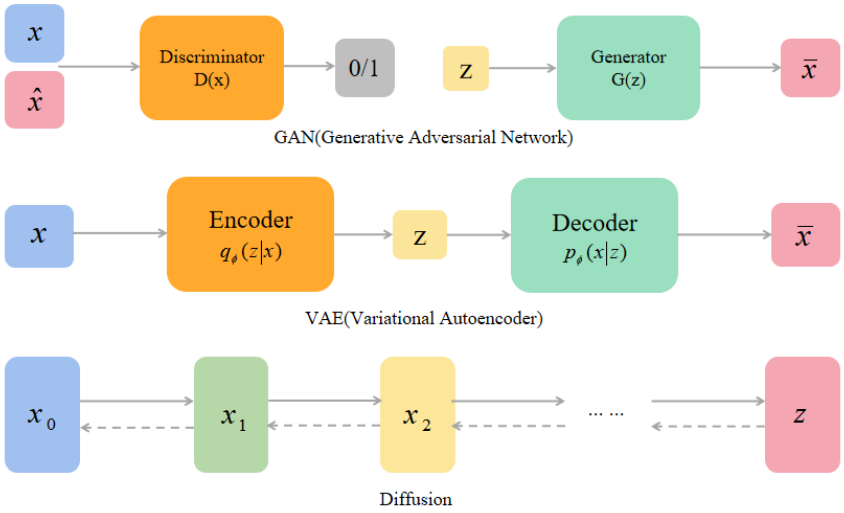

Figure 1: Structure of different generators

In VAE, it first defines a latent variable z that satisfied $p(z) = N(0, I)$, then define a condition distribution $P_\theta(x|z)$ thus defining the joint distribution of z and x. After successfully trained the model, generating data x can therefore follow 'ancestral sampling' to obtain $\bar{x}$. So here we can think of $P_\theta(x|z)$ as a "generator".

However, the training VAE is based on maximum likelihood, while the calculation of $\log p_\theta(x)$ is intractable, often we need to use the variational inference technique, introducing $q_\sigma(z|x)$ to approximate the true posterior. This leads to the VAE's most central issue in recent years which is the expressiveness of the variational posterior distribution and trade-offs for calculating costs. [6]

So, is there a model that only needs to train a single 'generator', has a simple training objective function, does not require training other networks, and has no restrictions on the generators? The answer is the Diffusion model.

We will describe how Diffusion model works, as Stable diffusion emerged to solve some of the problems it presented. For example, Diffusion model is too memory intensive, very computationally demanding and so on.

3. Methodology

3.1 Diffusion model

The overview steps of Diffusion model are to disturb original image (by adding noise that

follows Gaussian distribution) until the complete noise figure is obtained, then converts noise back to image. This makes use of two Markov chains: one for forward, another for reverse. The former is typically hand-designed with the goal to transform any data distribution into a simple prior distribution (e.g., standard Gaussian), while the latter Markov chain reverses the former by learning transition kernels parameterized by deep neural networks. [1]

New data points are subsequently generated by first sampling a random vector from the prior distribution, followed by ancestral sampling through the reverse Markov chain. [3]

Many essays have proved this process, but here displays a plainer method.

3.1.1 Forward chain [1]

Firstly, the forward chain is mainly dependent on the following formula

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t$$

or

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{\beta_t}\varepsilon_t$$

In this formula, x_t, x_{t-1} represent the present and previous image, α_t , β_t is the weight for previous image and noise respectively, ε_t represent the noise added where $\varepsilon_t \sim N(0, I)$ and the noise added at each moment is independent.

With the increment of ε_t, α_t is getting smaller. As $\alpha_t = 1-\beta_t$, β_t will increase from 0.0001 to 0.02[4]. This process allows Diffusion model to disturb the image continuously

$x_0, x_1, x_2 ..., x_t ..., x_T$ where the total step length required is T.

However, it would be very inefficient to keep trying to solve iteratively from x_0 onwards. In fact, since the noise added at each step is independent and follows a normal distribution, the following derivation can be made:

$$q_t(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t)I)$$

$$= \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t$$

By substitute x_{t-1} with x_{t-2}, we can get a new formula:

$$= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}}\varepsilon_{t-1}) + \sqrt{1-\alpha_t}\varepsilon_t$$

$$= \sqrt{\alpha_t\alpha_{t-1}} x_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\varepsilon_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t$$

Using the property of ε_t, we can combine the last two items.

$$= \sqrt{\alpha_t\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\varepsilon$$

$$= \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon$$

where α_t is known

Thus, this formula allows us to directly receive the noise map from original image.

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{\overline{\alpha}_t} x_0, (1-\overline{\alpha}_t)I)$$

Where $\alpha_t \approx 0$ and x_t is almost a Gaussian distribution, so we have:

$$q(x_t) := \int q(x_T|x_0)q(x_0)\,dx_0 \approx N(x_T;0,I)\,[1]$$

3.1.2 Reverse chain [5]

Briefly, the forward Markov chain allow model component to slowly adding noise to the original image until all the data in it has lost. However, what we actually want is visible and newly generated image. So, in this reverse process, it starts by first generating an unstructured noise vector from the prior distribution (which is typically trivial to obtain), then gradually remove noise therein by running a learnable Markov chain in the reverse time direction. Specifically, the reverse Markov chain is parameterized by a prior distribution $P(x_T) = N(x_T;0,I)$ and a learnable transition kernel $p_\theta(x_{t-1}|x_t)$.[4]

Therefore, we can define the Reverse step as the inverse step with respect to learned Gaussian transitions parameterized by θ [2]:

$$p(x_{t-1}|x_t) = N(x_{t-1};\ \mu_\theta(x_t,t), \sum_\theta(x_t,t))$$

Where the mean $\mu_\theta(x_t,t)$ and variance $\sum{}'_\theta(x_t,t)$ are parameterized by deep neural networks.

3.2 Stable Diffusion

Since every Markov chain need to do prediction in memory, which means the memory is required to store several enormous deep networks. This is a very high demand for memory. And the highly computational requirements of the diffusion model and the very large amount of memory and power required for training made it impossible for most researchers to implement the model in reality earlier.

One solution gave out earlier is to split the large image into a number of smaller resolution images for training, and then use an additional neural network to produce a larger resolution image (super-resolution diffusion).

However, Stable Diffusion, which is also called the Latent Diffusion model, released in 2021, gives a different approach. Stable Diffusion does not operate directly on the manipulated image, but in the latent space. By encoding the original data into a smaller space, it allows U-Net to add and remove noise on a low-dimensional representation. [7] The whole process is shown in Figure 2.

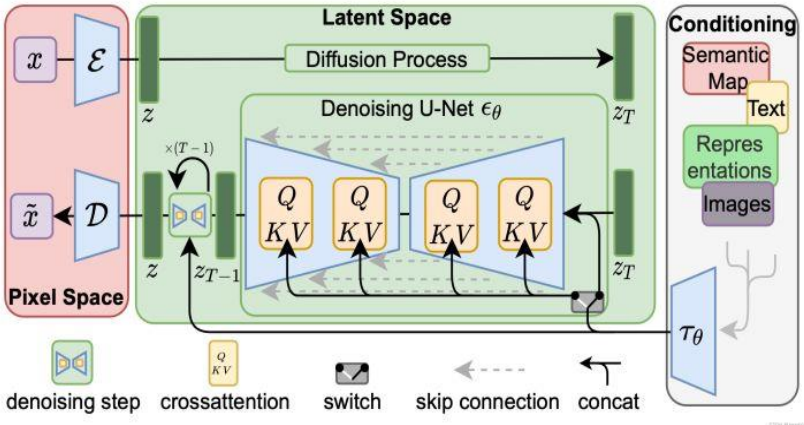

Figure 2: Stable Diffusion structure, from High-Resolution Image Synthesis with Latent Diffusion Models

3.2.1 Latent Space

Briefly in Stable Diffusion, the auto-encoder uses the Image Encoder to compress the image into latent space and then the Image Decoder to reconstruct the compressed information as Figure 3 shown.

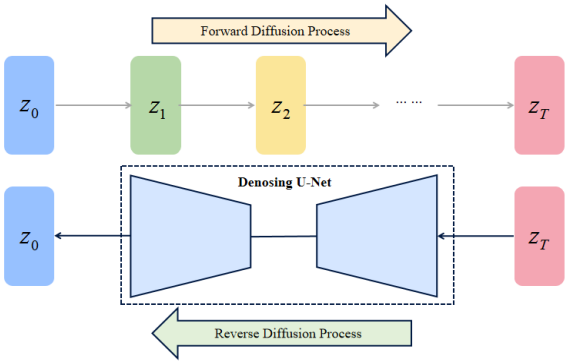

Figure 3: Simplified structure in latent space

In this process, the image is transformed from high to low dimensional, reducing redundant information. This also means that it allows us to get rid of any irrelevant information and focus only on the most important features.

Here comes the benefits of using Latent Space. It is able to transform more complex forms of raw data into simpler representations of data that are more beneficial to go about processing. Moreover, processing data with less information than in higher dimensions can also speed up the generation of images.

Forward diffusion is the use of Image Encoder to generate image data to train the noise predictor. Once the training is complete, reverse diffusion can be performed, using Image Decoder to generate the images.

However, if we want to integrate text into the image generation, we have to adjust the noise predictor to the input text.

3.2.2 Transformer Language Model

The Stable Diffusion model uses CLIPText (a GPT-based model). [7]

As CLIP model provides the token embedded and the image it corresponds to, this allows the system increases focus on text.

The text information is not processed directly by ResNet, a component of U-Net, but is incorporated into latents through the attention layer. In this way, the next ResNet can make use of the incorporated text information during processing.

In short, U-Net will receive three inputs, the text information, noisy image(latent) and noise amount and then produce a predicted noise sample which processed by many ResNets and layers.

Above process can be concluded as the reverse process of Stable Diffusion can denoise, at the same time, react to keywords user entered.

4. Evolution

Since the logic and algorithms of Stable Diffusion itself are now considered well established, I have focused on investigating how to train a better model with fewer graphs. In this case, the model refers to the style and character traits of the generated graphs (if the training graphs are character

specific). But this is not easy to accomplish, so I'm actively exploring it.

5. Conclusion and Future Works

This article describes a wonderful image generator: Stable Diffusion. Its operation principle is quite similar to Diffusion model but solves the problem arise with Diffusion model by using latent space.

Diffusion model is divided into two parts: forward and reverse diffusion. Forward diffusion can be calculated using a closed form formula. However reverse diffusion can be done with a trained neural network.

Stable Diffusion (latent diffusion model) is a diffusion process that takes place in latent space and is therefore much faster than Diffusion model. Attributed to CLIP model, Stable Diffusion can produce images by enter keywords.

References

[1] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Minghsuan Yang. 2022. Diffusion Models: A Comprehensive Survey of Methods and Applications. 1, 1 (October 2022), 39 pages.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, Vol. 33. 6840–6851.

[3] Daphne Koller and Nir Friedman. 2009. Probabilistic graphical models: principles and techniques. MIT press.

[4] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using non-equilibrium thermodynamics. In International Conference on Machine Learning. 2256–2265.

[5] Cao Hanqun, Cheng Tan, Zhangyang Gao, Guangyong Chen, Pheng-Ann Heng and Stan Z. Li. "A Survey on Generative Diffusion Model." ArXiv abs/2209.02646 (2022)

[6] Kingma Diederik P., Tim Salimans and Max Welling. "Improved Variational Inference with Inverse Autoregressive Flow." ArXiv abs/1606.04934 (2016)

[7] Rombach, Robin, A. Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 10674-10685.