Research Advanced in Image Generation Based on Diffusion Probability Model

Yuhan Huang *

School of Computer and Information Engineering, Hubei University, Wuhan, Hubei province, 430064, China

* Corresponding Author Email: 202231116020148@stu.hubu.edu.cn

Abstract. Image generation has been a popular research task in the computer vision community, which aims to learn a distribution from a specific dataset and generate realistic images obeying this distribution. Thanks to the rapid development of deep learning technology, image generation models based on convolutional neural networks, especially generative adversarial networks (GANs) and variational autoencoders (VAEs), have become mainstream frameworks for image generation. However, in recent years, with the gradual deepening of the research on the denoising diffusion probability model (DDPM), the image generation technology based on DDPM has made new breakthroughs in accuracy and speed. Around the Diffusion Model, this paper introduces its latest research progress in image generation and derivative tasks. Specifically, this paper reviews the key techniques and basic theories of the diffusion model in detail. Then, the main research work, improvement mechanism and characteristics of the DDPM-based image generation method are summarized and summarized. This paper focuses on the basic structure and related applications of diffusion models, and evaluates some basic functions. Finally, the current problems and future development directions of image generation technology based on diffusion model are analyzed and summarized.

Keywords: Image generation; Diffusion probability model; Deep learning; Application.

1. Introduction

The level of image quality has important practical significance for whether the image content information can be transmitted to the outside world correctly and comprehensively, and plays an significant role in contemporary society. With the rapid development of multimedia technology, people are getting used to expressing information in the form of images or videos, which increases people's demand for high-quality images. Therefore, researching the creation and processing of high-quality photographs has enormous practical value.

The task of image generation is a fundamental and key research task in the computer vision community. Deep learning-based picture production models can now generate images with a range of scales and resolutions thanks to the recent rapid advancements in deep learning technology. According to the representation of the probability distribution, existing deep learning-based image generation methods can be further divided into likelihood-based generative models and implicit generative models. The likelihood-based generative models [1-2] learn the probability distribution of data, and representative methods include auto-regressive models, flow models, energy-based models, and variational autoencoders. In implicit generative models [3-4], there is no unequivocal representation of the probability distribution. For example, by using antagonistic learning, the generator may directly convert noise to samples thanks to Generative Adversarial Networks (GAN) [5]. Relying on its powerful image generation ability, GAN has become the mainstream choice in the field of image processing, and it has derived a variety of improved forms, which further improves the performance of GAN in the field of image processing [6].

A recently developed image generation model called the diffusion model [7] features a more adaptable model architecture and more precise logarithmic likelihood computation than GAN. The diffusion model includes forward diffusion and reverse diffusion. Forward diffusion introduces random noise into the sample, while reverse diffusion is used to create samples from the noise. The

diffusion model is becoming increasingly popular among researchers because of its outstanding performance in the area of high-quality image production.

By studying the basic principle of Diffusion Model and its related models, the quality of image generation is improved, which is helpful to apply artificial intelligence image generation to a wider range of fields. This paper focuses on the basic structure and related applications of Diffusion Model, and evaluates some basic functions. Finally, the current problems and future development direction of image generation based on Diffusion Model are analyzed and summarized.

2. Key Technology

2.1. Diffusion Model

An developing image generation model called the diffusion model has a more adaptable model architecture and a more precise estimate of the log-likelihood than the GAN does [8-12]. A Markov chain that has been trained via variational inference is the Diffusion Probabilistic Model, sometimes referred to as the Diffusion Model. Diffusion models contaminate the training data by incrementally introducing Gaussian noise, which gradually obliterates the original data's details until only noise remains. The neural network is then taught to reverse the entire destruction process, transforming pure noise into a high-quality image by gradually denoising the image until the noise is gone. As shown in Figure 1, the diffusion model includes forward diffusion and reverse diffusion.

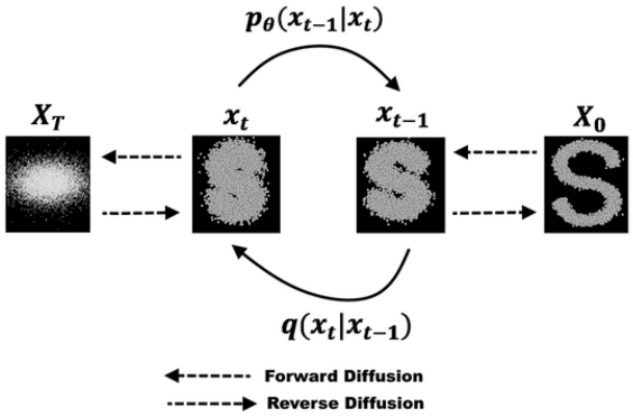

Fig. 1. The workflow of denoising diffusion probabilistic model

In the forward diffusion process, the diffusion model receives the input data from the image that contains complex information, completes the transformation from the ordered low-entropy state to the disordered high-entropy state, and then transforms the input data into noisy data. Unlike other models, the estimated posterior distribution $q(x_{1:T}|x_0)$ in the diffusion model is fixed as a Markov chain, and thus the joint probability distribution is:

$$q(x_{0:T}) = q(x_0)q(x_{1:T}|x_0) = q(x_0)\sum_{t=1}^{T} q(x_t|x_{t-1}) \quad (1)$$

The transition probability distribution, that is, the Markov diffusion kernel $q(x_t|x_{t-1})$, is defined as a normal distribution $q(x_t|x_{t-1};\ \beta_t) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$. When β_t is small enough, the mean and variance of the normal distribution are x_{t-1} and 0, respectively. This means x_t in the diffusion model is obtained by making small changes to x_{t-1}.

The reverse diffusion process is calculated repeatedly by the diffusion model to reduce entropy, that is,it can convert noise data into images by gradually denoising in the reverse diffusion process.Another Markov chain, the reverse process has the following joint probability distribution:

$$p_\theta(x_{0:T}) = p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \quad (2)$$

Where $p_\theta(x_T) \sim N(0, I)$. Since the forward probability transition distribution $q(x_t|x_{t-1})$ is pre-defined, the input data can be restored according to x_T if the inverse probability transition

distribution $q(x_{t-1}|x_t)$ is known. However, the backward transition probability distribution is unknown and cannot be solved. Therefore, the conditional probability distribution $p_\theta(x_{t-1}|x_t) = N(x_t; 1-\beta_t x_{t-1}, \beta_t I)$ is used in the diffusion model to estimate $q(x_{t-1}|x_t)$, where $p_\theta(x_{t-1}|x_t)$ also obey the normal distribution. To this end, the $p_\theta(x_{t-1}|x_t)$ can be calculated by:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x, t)) \quad (3)$$

The forward diffusion process's individual steps are all known, while the process of reverse diffusion is not known, the neural network must be trained to obtain the appropriate parameters $\mu_\theta(x_t, t)$ and $\sum_\theta(x, t)$. Then $p_\theta(x_{t-1}|x_t)$ is known, and the original image x0 can be reconstructed from it.

2.2. Contrastive Language-Image Pre-Training

Contrastive Language-Image Pre-Training (CLIP) is a used four hundred million images and text comparison and trained model. It compares an image to a vector of text generated by an encoder. If they match, they'll be close, if they don't match, they'll be far away from each other. With the pre-trained CLIP model, text and images can be well matched. In subsequent image generation model, the CLIP is widely used, which provides great help to the image generation based on text.

2.3. Extent Model

Based on the Diffusion Model, software such as Stable Diffusion and DALL·E were developed. Aiming at the problems of the original model, Stable Diffusion uses a new method to improve the efficiency of model training. Because before all the work is done in pixel space, efficiency is not high, so it introduced a potential space for diffusion process, this method greatly lowers the cost of training. The cross-attention layer is also added in the meantime, which makes new progress in text-to-image synthesis, unconditional image generation and super-resolution tasks.

OpenAI introduced its own CILP model in DALL·E. This model can stably capture the feature information and image style, and improve the image diversity with the minimum loss of photo authenticity. A decoder conditioned on an image representation can also generate an image that preserves both its semantics and its style, while changing non-essential details missing from the image representation. There is a two-stage model, the prior and the decoder. The diffusion model is also used in the decoder. Decoder's autoregressive and diffusion models were tested, and it was found that the use of diffusion model was more computationally efficient. Higher quality samples are produced, which greatly improves the diversity of generated images.

SDXL is a latent diffusion model for text-to-image synthesis. SDXL utilizes a three times larger UNet backbone compared to previous versions of Stable Diffusion. The increase in model parameters is mainly due to the fact that SDXL uses a second text encoder and therefore has more attention blocks and cross-attention layers. It adds a fine-tuning model to the basic model and retrains the VAE, which significantly improves the effect of the generated images and the fit to the text. Compared with previous versions of the Stable coursing together, SDXL reflects the performance of the remarkable progress.

2.4. Low-Rank Adaption

In order to get a more accurate and appropriate image, Low-Rank Adaptive (LoRA) model can be used for fine-tuning. Originally used for language models, now it has been applied in image generation. The amount of trainable parameters for downstream tasks is drastically decreased by LoRA, which incorporates trainable rank decomposition matrices into the Transformer architecture. LoRA model reduces the parameters in training, training time and computing power, and also gets better performance.

3. Application

3.1. Main applications

DDPM has been widely used in image generation and its various derivative tasks, including:

(1) Image generation. The most basic application is image generation. The model is passed an image and a paragraph of descriptive text. Images and texts are encoded by their corresponding encoders and fed into the latent space. A diffusion process takes place in this latent space and produces a lower resolution image. Finally, the low-resolution image is enlarged in pixel space to obtain the target image. Different models have different details. Some will add hypernetworks and cross-attention to the diffusion process, or change the VAE decoder to improve the text fit of the generated images.

(2) Modify image according to the text. By passing the model a base image and a text suggesting changes to the image, we can change some elements of the image. After using CLIP to parse text, lock need to modify the area, and local redrawing of the region to modify the effect of the picture.

(3) Keywords backtracking. Given an image, the model will recognize each element in the image and its degree of match using the CLIP which has been trained. This allows to parse each element in the image for keyword backtracking.

(4) HD repair and picture enlargement. HD repair and picture enlargement is still image-based image generation, which adds noise to the original image and sends it to the latent space to regenerate. In the process, you can change the model and tune parameters. Thus, the generated images can be both enlarged and details added on the original picture.

3.2. Application analysis

The image generation based on the diffusion model is supported by CLIP, which can generate different images with the same text. This is a good representation of the creativity of the generated images. According to the different description, different style of pictures can be produced. Both realistic style and animation style can be easily done. So far, images based on Diffusion Model have been very good, and the generated works have been making waves in NFT. In the future, training models with more accurate datasets can have more practical applications in fields such as art, industry and education.

According to the function of the text to modify picture can without changing the original case, other elements according to the text only modify some part of the picture. This function can be more widely used in the future development in picture modification. We can not only add or remove features, but also move or partially redraw features from the original image. Keyword backtrack can find both key visual features and small visual features, which can be widely used in image recognition. The use of HD Repair and Picture enlargement can reduce the cost of image generation. Because of the random nature of image generation, the quality of the generated image can be reduced in advance, and then the desired image can be retained as needed. Finally, the desired image is repaired and enlarged, while increasing the quality of detail.

4. Discussion

Although current image generation models based on DDPM models have already made great progress in improving image resolution and image diversity, there may be much room for future development of deep learning-based high-quality image models in terms of.

(1) Interpretability of neural networks. Currently, most neural networks are regarded as a "black box", i.e., only the inputs and outputs of the neural network are explicitly defined, while the internal neural network cannot establish a clear connection between the outputs and inputs using specific and definite expressions, i.e., the neural network lacks interpretability. This may lead to uncertainty about the output of the network, especially in some special fields with high security requirements, the interpretability of the neural network is more important.

(2) Image quality evaluation criteria. At present, there is not yet a strict image quality evaluation standard in the field of image processing, and most of them only rely on human vision to evaluate, and a reasonable and effective quality evaluation standard for the design and evaluation of the image generation model has a pivotal role, therefore, the future for the image quality evaluation standard is also a worthwhile research direction.

(3) Multimodal fusion. Images, text, audio, and other types of information can all be included in an image's description in the area of image production, and each type of information can include a variety of modalities. The study of how to effectively fuse the information of various modalities is also of great practical significance for the generation of high-quality images.

5. Conclusion

From Diffusion Model was applied to image generation, many experts and scholars constantly improve base Model. At the same time, new models are introduced from other generative models to improve the accuracy and variety of generated images. As new capabilities are added, the technology has the ability to expand beyond the realm of art to other areas. The cost and training set of the basic model are huge, and it is difficult for the average individual to train the model by himself. However, due to the use of other fine-tuning techniques, individuals can train some models to achieve the desired performance using a computer with average computing power and a training set prepared by themselves. The combination of open source large models and personalized models can meet the basic needs of general individuals in the image field. In this context, models are constantly refined and applied to more specialized fields.

References

[1] Yang K, Ding X, Yuan X. Bayesian empirical likelihood inference and order shrinkage for autoregressive models [J]. Statistical Papers, 2022: 1-25.

[2] Hou Y, Zhai J, Chen J. Coupled adversarial variational autoencoder [J]. Signal Processing: Image Communication, 2021, 98: 116396.

[3] Xia W, Zhang Y, Yang Y, et al. Gan inversion: A survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3121-3138.

[4] Wanle C, Choo Y H, Goh O S. Review of Generative Adversarial Networks in Image Generation [J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2022, 26(1): 3-7.

[5] Xu C, Lang W, Xin R, et al. Generative detect for occlusion object based on occlusion generation and feature completing [J]. Journal of Visual Communication and Image Representation, 2021, 78: 103189.

[6] Aggarwal A, Mittal M, Battineni G. Generative adversarial network: An overview of theory and applications [J]. International Journal of Information Management Data Insights, 2021, 1(1): 100004.

[7] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis [J]. Advances in neural information processing systems, 2021, 34: 8780-8794.

[8] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models [C]//International Conference on Machine Learning. PMLR, 2021: 8162-8171.

[9] Du C, Zhu J, Zhang B. Learning deep generative models with doubly stochastic gradient MCMC [J]. IEEE transactions on neural networks and learning systems, 2017, 29(7): 3084-3096.

[10] Vega-Márquez B, Rubio-Escudero C, Nepomuceno-Chamorro I. Generation of synthetic data with conditional generative adversarial networks [J]. Logic Journal of the IGPL, 2022, 30(2): 252-262.

[11] Khan Z N, Ahmad J. Attention induced multi-head convolutional neural network for human activity recognition [J]. Applied soft computing, 2021, 110: 107671.

[12] Rajee M V, Mythili C. Gender classification on digital dental X-ray images using deep convolutional neural network [J]. Biomedical Signal Processing and Control, 2021, 69: 102939.