Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/32/20230199

Enhanced diffusion model based on similarity for handwritten digit generation

Wenjing Kang[1,†] and Wenbo Li [2,3,†]

[1]School of electronic engineering, Xi'an University of Posts and Telecommunications, Xi'an, China
[2]College of computer science and cyber security, Chengdu University of Technology, Chengdu, China

[3]li.wenbo1@student.zy.cdut.edu.cn
[†]All the authors contributed equally and their names were listed in alphabetical order.

Abstract. In recent years with the rise of deep learning, there has been a major revolution in image generation technology. Deep learning models, especially the diffusion model. have brought about breakthrough progress in image generation. Various deep generation models have recently demonstrated a wide variety of high-quality sample data patterns. Although image generation technology has achieved remarkable achievement. There are still challenges and issues, such as quality control in generated images. In order to improve the robustness and performance of diffusion model in image generation, an enhanced diffusion model based on similarity is proposed in this paper. Based on the original diffusion model, the similarity loss function is added to narrow the semantic distance between the original image and the generated image, so that the generated image is more robust. Extensive experiments were carried out on the MINIST dataset, and the experimental results showed that compared with the other generation models, the enhanced diffusion model based on similarity obtained the best scores of IS=31.61 and FID=175.21, which verified the validity of the similarity loss.

Keywords: diffusion model, handwritten digit, similarity loss, generation.

1. Introduction

Image generation refers to the process of using artificial intelligence technology to generate images in single mode or cross-mode according to given data. According to different task objectives and input modes, image generation mainly includes image composition, image-to-image generation based on existing images, and text-to-image generation based on text description [1]. It is widely used in graphic design, game production, animation production and other fields. In addition, image generation also has great application potential in medical image synthesis and analysis, compound synthesis and drug discovery. Therefore, image generation has attracted more and more researchers' attention.

In order to realize image generation, many efficient generation models have been developed in recent years, such as generative adversarial model and autoregressive generation model [2]. Generative adversarial network (GAN) is the mainstream image generation model of the last generation. GAN continuously improves its generative ability and discrimination ability through game training of generator and discriminator, so that the data of generative network is more and more close to the real

DOI: 10.54254/2755-2721/32/20230199

data, so as to achieve the purpose of generating realistic images. However, in the process of development, GAN also has some problems, such as poor stability, lack of diversity of generated images, and mode collapse. The inspiration of autoregressive model for image generation comes from the successful experience of NLP pre-training mode. The self-attention mechanism in Transformer structure can optimize the training mode of GAN, improve the stability of the model and the rationality of image generation. However, the image generation based on autoregressive model has problems in reasoning speed and training cost. Make its practical application limited.

Due to the limitations of the previous Model in terms of performance, the Diffusion Model was proposed to solve these problems, and its effect on training stability and result accuracy was significantly improved, so it quickly replaced the application of GAN. The diffusion model is the process of gradually applying noise to the image in the forward stage until the image is destroyed into complete Gaussian noise, and then learning to restore the original image from Gaussian noise in the reverse stage. The diffusion model can restore the real data more accurately, and the processing ability of image details is stronger. However, the classical generation model only considers the noise loss in the process of diffusion, and does not consider the similarity and semantic consistency between image contents.

In order to ensure the quality and stability of the generated images, an enhanced diffusion model based on similarity is proposed in this paper. On the basis of the classical diffusion model, the semantic similarity between the original image and the generated image is described by comparing them, and it is taken as a part of the loss function and noise loss to guide the model optimization [3].

2. Related works

In fields including text-to-image translation and image creation, diffusion models (DMs), a revolutionary generative model based on deep learning and computational vision, have been put to use. Diffusion models have a number of benefits over conventional autoregressive models, energy-based models, and generative adversarial network (GAN) models, including the ability to create pictures with substantial variety and large details [4].

In Chen Li's Comparison of Image Generation methods based on Diffusion Models, IDDPM model is put forward (Improved Denoising Diffusion Probabilistic Models), By defining the goal function and enhancing the calculation's logarithmic likelihood function, which is used to compute variance variance learning, and lowering the degree of difficulty of the sampling step, accelerated sampling. The performance boost is modest, the Markov process is still used, it needs more processing power, and the sample steps are longer [5].

Therefore, Chen Li proposed a de-noising diffusion implicit generation model (DDIM) for effective sampling. This model is an implicit generation model that relies on edge distribution, so it only needs to let the sampling process meet the edge distribution conditions, rather than relying on Markov random process, so it does not need many sampling steps to get high-quality image samples, and the speed is faster [5].

At the same time, the loss function of the diffusion model most often adopts the simplified optimization objective based on the predicted noise. But there are other options, and prediction targets can be constructed based on raw data x0. In addition to the prediction target of the model, the loss function can also adopt different weight coefficients, which has a certain impact on the training of the diffusion model.

In the Progressive Distillation for Fast Sampling of Diffusion Models proposed by Tim Salimans [6]. the loss function based on the original data and the fitting data is adopted. SNR+1 and truncated SNR are used as the weight coefficients. The truncated SNR weight coefficient is designed to prevent the weight coefficient from being 0 when the SNR is close to 0, which is not conducive to distillation.

In addition, another weight coefficient, min-SNr-γ, is proposed by Efficient Diffusion Training via Min-SNR Weighting Strategy [7]. The main purpose of this paper is to avoid paying too much attention to the small noise level (that is, the number of diffusion steps t is small) during model training. One of its advantages is to accelerate the training process.

DOI: 10.54254/2755-2721/32/20230199

3. Method

3.1. Classical diffusion model

The Diffusion Model is a type of Generative model, which also includes the Variational Autoencoder (VAE) and the Generative Adversarial Network (GAN). Unlike other generative networks like GAN, the diffusion model progressively introduces noise to an image during the forward stage until the image is completely scrambled into Gaussian noise. It then learns to reconstruct the original image from this Gaussian noise during the reverse stage [8].

In particular, the forward stage progressively adds noise to the initial image x_0. The image x_t produced at each step is solely influenced by the result x_{t-1} from the preceding step until the image x_t at step T transforms into pure Gaussian noise [9-10].

The reverse stage involves the continuous process of noise reduction. Initially, Gaussian noise x_T is provided and gradually de-noised until the original image x_0 is completely restored. This process is guided by a loss function, as shown in equation (1).

$$Loss = E_{x_0,\varepsilon}\left(\left\|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)\right\|^2\right) \tag{1}$$

x_0 is the original image,ε is the real nois,ε_θ is the predicted noise,t represents the time step, $\bar{\alpha}_t = \sum_{s=0}^{t}\alpha_t$, $\alpha_t = 1-\beta_t$, $\beta_t$represents the variance of the Gaussian noise at time *t*.

3.2. Diffusion model based on similarity loss

In the traditional diffusion model, the reduction of noise loss equates to eliminating noise and enhancing the image, concentrating more on the clarity and reproducibility of the produced image. If the created image is noisy, discerning its content becomes challenging, thereby limiting its generalization capability. To boost the generalization and robustness of the produced model, it is crucial that the desired content remains clearly identifiable even when the image generation results are subpar. Hence, this paper emphasizes the semantic representation of the produced and original images, aiming to achieve maximum similarity in the semantic space. This approach ensures that a discernible target image can still be obtained even if the image is blurry.

With this in mind, this paper introduces an improved diffusion model that leverages similarity to generate more stable and realistic images. Unlike the traditional diffusion model that only concerns with noise or data, this paper develops loss functions that consider both noise and data as predictive targets. The model's structure is depicted in Figure 1.

The model is bifurcated into two phases: the diffusion generation phase and the similarity comparison phase. In the diffusion phase, noise is introduced and eliminated through the forward and reverse processes, culminating in the generation of the target image. Following that, in the similarity comparison phase, the semantic similarity between the created image and the original image is computed. The model's training is directed by these two phases.

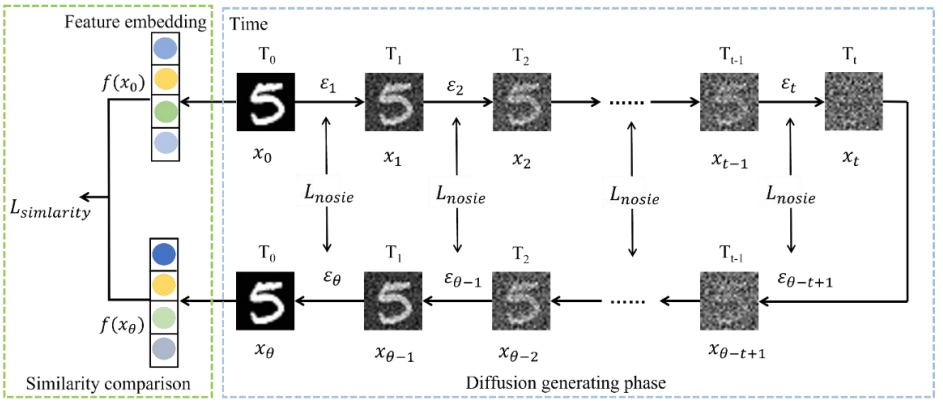


Figure 1. Enhanced diffusion model based on similarity.

DOI: 10.54254/2755-2721/32/20230199

Specifically, the loss function of the similarity-based enhanced diffusion model consists of two parts, namely, noise loss and similarity loss.

As shown in formula (2), noise loss follows the loss function of the classical diffusion model and takes noise as the prediction target.

$$L_{noise} = E_{x_0,\varepsilon}[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2] \tag{2}$$

The similarity loss, as depicted in formula (3), characterizes the content resemblance between the original and generated images. This paper ensures the semantic uniformity between the original and generated image by evaluating the quality of the produced image through the computation of the cosine similarity between the features of the two images.

$$L_{similarity} = \frac{|<f(x_0), f(x_\theta)>|}{|f(x_0)| \cdot |f(x_\theta)|} \tag{3}$$

<·>Represents inner product operation,f (·) represents feature extraction,|·|represents module.

The final optimization objective is shown in formula (4).

$$L = L_{noise} + L_{simliarity} \tag{4}$$

4. Experiment

In this section, in order to verify the effectiveness of the proposed model, we will demonstrate the improvement of the model's performance by comparing experimental data analysis and generating visual perception of images.

4.1. Experimental settings

The experiment was conducted in Ubuntu20.04, the programming language was Python3.9, the deep learning framework of the experiment was Pytorch2.0, CUDA12.1, the CPU processor was i9-13900KF, and the graphics card was 4090.

All experiments will be trained and tested on the MINIST dataset. We set the number of training rounds for all experiments to 50, the time step T to 500, the Batchsize to 256, the optimizer to adopt the Adam algorithm, and the learning rate to 0.0005. We set the forward process variance consistent with DDPM, increasing linearly from 0.0001 to 0.02. In the direction process, the Unet network structure is used as the common denoising structure. Unet takes the image as the entry point, finds the low-dimensional representation of the image by reducing the sampling, and then restores the image by increasing the sampling.

In order to verify the generalization ability of the model, all performance indicators will be calculated on the test set.

4.2. Analysis of experimental results

Figure 2 shows the results generated by the enhanced diffusion model based on similarity after training on the MNIST dataset. The figure shows the process of the image gradually becoming an image from random noise. It can be seen from the figure that the enhanced diffusion model based on similarity can generate high-quality clear pictures, only some numbers are unrecognizable. In the process of batch generation, it can be found that the number "7" generates the least amount, which may be caused by the unbalanced number of samples in the MINIST dataset. Some numbers have more samples than others, which can cause the model to be biased toward producing numbers that occur frequently.

DOI: 10.54254/2755-2721/32/20230199

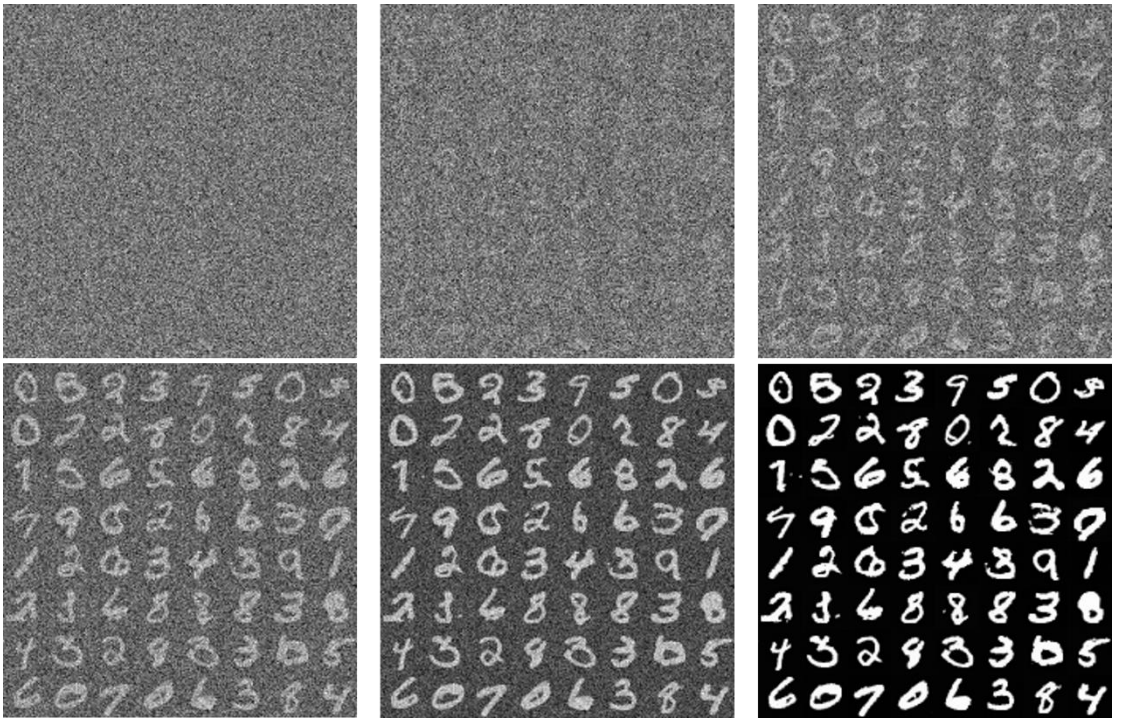


Figure 2. Image generation process of enhanced diffusion model based on similarity.

4.3. Contrast experiment

First, compared with the most basic diffusion model (the diffusion model with only noise loss function), the model performance before and after adding the similarity loss is compared. Secondly, the classical generation model is selected for comparison. The IS (Inception Score) and FID (Frechet Inception Distance score) were used as performance evaluation indexes. It IS worth noting that the IS and FID are calculated from the test set. The comparison results are shown in Table 1.

The following conclusions can be drawn from Table 1.

1) The enhanced diffusion model FID and IS based on similarity have achieved the best results in comparison.

2) IS measures the sharpness and variety of the generated image, while FID measures the distance of the generated image from the original image. The experimental results show that the FID value of the model proposed in this paper IS 31.61, and the IS value is 175.21. After adding the similarity loss to the enhanced diffusion model based on similarity, both FID and IS performance indicators have improved to some extent. Therefore, the addition of similarity loss is effective to improve the performance of the model, and can generate more realistic and clear images to a certain extent.

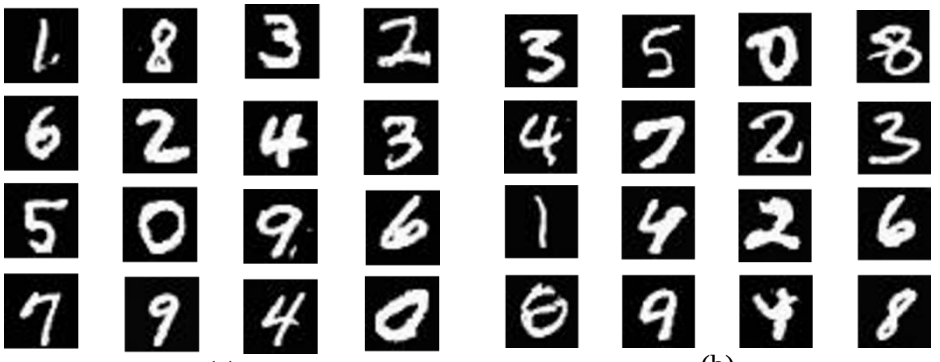
3) In comparison with other types of generative models, the enhanced diffusion model based on similarity also shows better performance and ranks first in the overall ranking.

Table 1. Comparative experimental results.

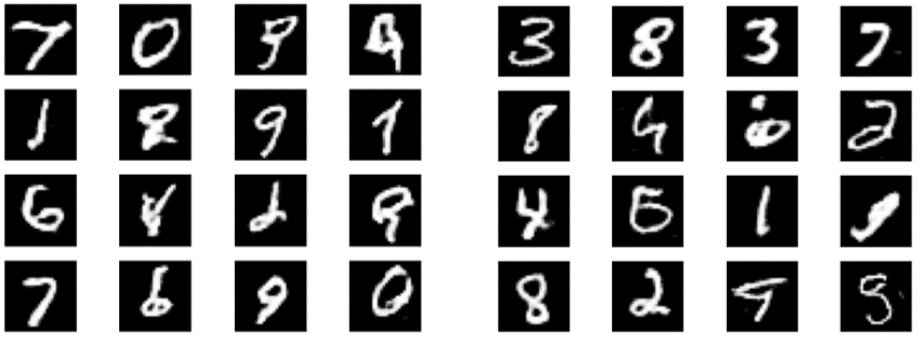
Model	Assessment Criteria	
	FID	IS
DDMP(L_{noise})	32.17	166.60
CGAN	35.20	148.68
DCGAN	36.19	132.98
VAE	39.98	121.67
Ours(L_{noise}+L_{sim})	**31.61**	**175.21**

DOI: 10.54254/2755-2721/32/20230199

In order to more intuitively feel the effect of image generation by the model, Figure 3 shows the generation effect of each comparison model. In the figure, each model shows 4*4 generated images, which are randomly sampled from the generated images. (a) is a similarity-based enhanced diffusion model, (b) is a diffusion model, (c) is a variational autoencoder (VAE), and (d) is a Deep Convolutional Generative Adversarial Networks (DCGAN). As can be seen from the figure, both the similarity-based enhanced diffusion model and the diffusion model can generate high-quality handwritten digital pictures, but the similarity-based enhanced diffusion model has better stability in image generation, and the image details are more clearly distinguishable.



(a) (b)



(c) (d)

Figure 3. Images generated by different comparison models.

5. Conclusion

In order to realize image generation, an enhanced diffusion model based on similarity is proposed in this paper. After reversing the diffusion process of the natural image, a new natural image can be gradually generated from a completely random noise image. Based on this, this paper improves the model by comparing the original image and the generated image to describe the semantic similarity between the two as part of the loss function, improves and optimizes the model, and uses the cosine similarity to measure the mass diffusion model of the generated image. Diffusion model has a strong development prospect. Diffusion model can be applied in various fields, such as image denoising, image restoration, super resolution imaging, image generation and so on. The simultaneous diffusion model is important for producing images with strong diversity and important details, and is a topic worthy of continue. In order to generate images, this paper proposes an enhanced Diffusion model based on similarity. After reversing the diffusion process of natural images, new natural images can be gradually generated from completely random noise images. Based on this, this paper improves the model by comparing the original image and the generated image, describes the semantic similarity between the two as part of the Loss function, improves and optimizes the model, and uses Cosine similarity to measure the quality of the generated image Diffusion model. The denoising diffusion Statistical model has achieved remarkable success in various image generation tasks, and can be applied to image denoising, image

DOI: 10.54254/2755-2721/32/20230199

restoration, super-resolution imaging, image generation and other fields. At the same time, Diffusion model is very important for generating images with strong diversity and important details, which is a subject worthy of further study.

Reference

[1] Aderhold J, Davydov V Yu, Fedler F, Klausing H, Mistele D, Rotter T, Semchinova O, Stemmer J and Graul J 2001 *J. Cryst. Growth* **222** 701

[2] Mingwen Shao, Wentao Zhang, Multi-scale generative adversarial inpainting network based on cross-layer attention transfer mechanism, *Knowledge-Based Systems*, 2020

[3] Jonathan Ho, Ajay Jain, Pieter Abbeel Denoising Diffusion Probabilistic Models arXiv:2006.11239v2 [cs. LG]

[4] Yanxi Wei, Yuru Kang, Fenggang Yao. Image Feature Understanding and Semantic Representation Based on Deep Learning, 2022 *International Conference on Artificial Intelligence of Things and Crowdsensing*, 2022

[5] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Efficient Diffusion Training via Min-SNR Weighting Strategy, *arXiv preprint arXiv:2303.09556*, 2023.

[6] Wijmans J G, Baker R W. The solution-diffusion model: a review. *Journal of membrane science*, 1995, 107(1-2): 1-21.

[7] Cao H, Tan C, Gao Z, et al. A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646*, 2022.

[8] C. Li, Y. Qi, Q. Zeng and L. Lu, Comparison of Image Generation methods based on Diffusion Models, 2023 *4th International Conference on Computer Vision, Image and Deep Learning* 2023, 1-4.

[9] Tianrui Huang, Yang Gao, Zhenglin Li, Yue Hu, Fuzhen Xuan. A Hybrid Deep Learning Framework Based on Diffffusion Model and Deep Residual Neural Network for Defect Detection in Composite Plates, *Applied Sciences*, 2023

[10] Pai Zhang, Hanqing Chen, Qinrui Li. Research on Vehicle Recognition Algorithm based on Convolution Neural Network, *Journal of Physics: Conference Series*, 2021

[11] Weilun Wang, Jianmin Bao, Wengang Zhou, Semantic Image Synthesis via Diffusion Models 2022