Research on Anime-Style Image Generation Based on Stable Diffusion

Mingbo Yang*

Hubei University of Education, Wuhan, Hubei Province, 430000, China

Abstract. Animation style generation technology based on artificial intelligence is gaining increasing attention in the current society, and it has shown a wide range of application prospects in creative design, game development, and advertising. Based on the Animefull-follow pre-training dataset, the effect of using Stable Diffusion technology combined with LoRA model to generate a single anime character image was discussed. The experimental results show that the generated image is highly consistent with the original image in terms of style and detail, and successfully captures the unique characteristics of animation art. Although the Fréchet Inception Distance (FID) value of the generated image is 77. 29 when calculating the FID value, indicating that there are visual differences to a certain extent, these differences usually do not significantly affect the user's perception in the animation field, and the generated images still show excellent visual effects. Combined with the advantages of the LoRA model, the resources and time required for training are significantly reduced, enabling high-quality image generation even in resource-constrained environments.

1 Introduction

Stable Diffusion technology is a generation method based on the diffusion model, which has made significant progress in the field of image generation in recent years. The technology produces high-quality images through a step-by-step denoising process and is widely used in fields such as art creation, advertising design, and game development. The study showed that Stable Diffusion obtained lower Fréchet Inception Distance (FID)values on multiple datasets. For example, on the CelebAHQ dataset, the FID value of LDM-4 of the Stable Diffusion model is 5. 11, which is significantly better than the generative model based on the Generative Adversarial Network (GAN) [1]. At the resolution of 256×256 on the FFHQ dataset, the FID of Stable Diffusion is 4. 98 [1]. In addition, the FID of the unguided LDM-KL-8 model was 23. 31 in the text-to-image generation task of the MS-COCOdataset, and the FID of the LDM-KL-8-G decreased to 12. 63 after optimization without classifier guidance [1,2].

Compared with generative models such as DALL-E 2, Stable Diffusion performs better in the stability, detail expression, and diversity of image generation. For example, in the 4x upsampling task of ImageNet, the FID values for Stable Diffusion LDM-4 are 2. 8 (validation

* Corresponding author: Zhangailin@asu.edu.pl

set) and 4. 8 (training set). The FID value of Stable Diffusion is better than that of the traditional pixel spatial diffusion model SR3 (FID of 5. 2) [3]. The stepwise denoising method can better capture the details of the image, making it highly adaptable and widely applicable in scenarios with limited computing resources.

To further improve the generation efficiency and reduce the training time and resource consumption, the Low-rank Adaptation LoRA model is also introduced in this study, and the low-rank decomposition method is used to achieve efficient model adjustment, to optimize the generation performance without significantly increasing the computational cost [4]. This combination achieves good results in Anime image generation tasks and significantly reduces the training cost and resource requirements. For example, the model performance of LoRA on the GLUE benchmark is only 0. 02% lower than that of the original model, but the number of parameters is reduced by about 99% [4]. In art creation and animation design, the application of Stable Diffusion can greatly shorten the creation cycle, help designers quickly generate diverse images, and improve production efficiency. In addition, combined with the optimization scheme of LoRA, Stable Diffusion has good adaptability in scenarios with high resource requirements, such as game development and advertising design, and can meet the needs of high-quality customized image generation.

In recent years, some studies have comprehensively reviewed the diffusion model, pointing out that this model not only shows unique advantages in image generation but also has broad potential in the direction of speed optimization and multimodal generation [5]. This study further expands the application value of Stable Diffusion in small studios and individual creators, so that it can achieve high-quality image generation even under the condition of limited resources, and provides new research ideas and practical references for the future field of Anime image generation and creative design.

2 Method

2.1 Stable diffusion principle

Stable Diffusion is a diffusion-based generative model that generates high-quality images through stepwise denoising. The basic principles include the following:

First, Stable Diffusion relies on the Diffusion Process, which is trained by simulating a forward process that progressively adds noise from a clean image until a purely random noise is generated, a process called the Forward Diffusion Process [6].

At the same time, the model learns the inverse diffusion process and gradually restores the high-quality image from the noise. This process is implemented by a denoising autoencoder, and the model is gradually recovered using the U-Net network structure during the denoising process, to ensure that the features at different resolutions can be effectively extracted [1].

Second, a significant difference between Stable Diffusion and traditional diffusion models is that it operates in Latent Space. With a pre-trained VAE (Variational Autoencoder), the image is compressed into latent space for diffusion and denoising [7]. Computing in latent space significantly improves build efficiency and reduces the consumption of computing resources [1].

In the generation phase, Stable Diffusion is generated through Reverse Diffusion Generation, which starts with random noise and gradually restores the image. In this process, the model goes from blurry to clear to produce an image, removing noise step by step with each iteration and finally restoring the full details of the image [1].

In addition, Stable Diffusion can use Text-to-Image Generation to generate plug-ins that are relevant to the input text description. This method uses WD 1. 4 Tagger to provide clear semantic guidance for image generation.

2.2 Principles of the lora model

Low-Rank Adaptation (LoRA) is a lightweight fine-tuning technology designed to reduce the training and storage costs of large-scale pre-trained models [4]. The core idea is to make only low-rank updates to the weight matrix for a particular layer when fine-tuning while keeping most of the structure of the model unchanged. Specifically, LoRA decomposes the weight matrix $W0 \in Rd \times k$ of the model into two low-rank matrices: $A \in Rr \times k$ and $B \in Rd \times r$, where r is the low-rank dimension parameter and represents the rank number after dimensionality reduction. This decomposition form can be expressed as:

$$W \approx A \cdot B \tag{1}$$

In this formula, the original weight matrix W is approximated as the product of two low-rank matrices. This low-rank decomposition significantly reduces the number of parameters that need to be updated, which greatly reduces the consumption of computational resources [8]. At the same time, this approach also reduces storage requirements, making model fine-tuning more efficient [4]. LoRA is particularly well-suited for scenarios that require fine-tuning for a specific task, such as generating a specific character in an Anime image. In this kind of task, LoRA can produce high-quality images with fast and efficient fine-tuning without sacrificing model generation performance.

2.3 Advantages and disadvantages of the stable diffusion model

Stable Diffusion is a diffusion model-based generation method that generates high-quality images through a stepwise denoising process, which performs well in the stability and detail preservation of image generation. Compared with the traditional GAN model, Stable Diffusion obtained lower Fréchet Inception Distance (FID) values on multiple datasets, showing higher quality and realism of the generated images. Stable Diffusion operates in latent space and relies on a pre-trained variational autoencoder (VAE) for image compression, which significantly improves the generation efficiency and reduces the consumption of computing resources, making it have good application potential in resource-constrained environments [1]. Studies have shown that Stable Diffusion performs well at generating photorealistic images, with its FID score performing best across multiple scenarios, especially when it comes to generating real faces, where the resulting facial feature detail is less error-prone [9]. At the same time, improved diffusion models (such as Shifted Diffusion) significantly improve the training efficiency of the model by optimizing the diffusion process. For example, using only 1. 7% of the labeled data, a FID score comparable to DALL-E 2 was obtained, further demonstrating the potential of the diffusion model in resource-constrained environments [10].

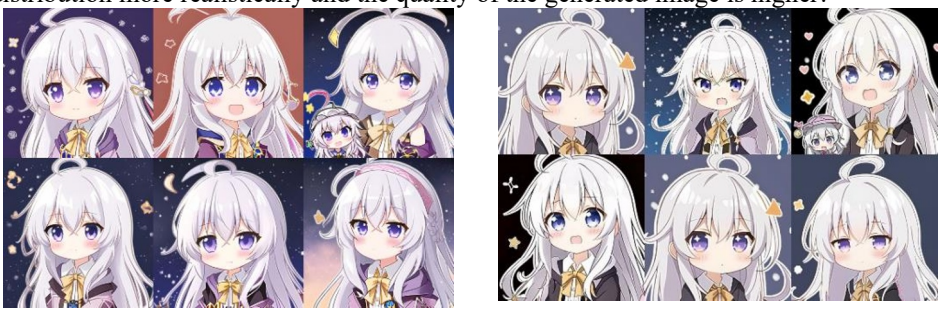
However, there are some limitations to Stable Diffusion. It is sensitive to the quality and quantity of training data, especially when the training sample is insufficient or the data quality is not high, the FID value of the generated image may be high, which affects the quality of the generated image [6]. In addition, Stable Diffusion can be challenging when generating complex scenes and highly detailed images, where some details may be missing, resulting in a degraded visual impact of the image.

2.4 Modeling steps

In this study, the following parameter settings were selected in Lora training, Animefull-latest was selected in the pre-trained model name, iA3-Prodigy-sd15 was selected in Presets, LyCORIS/iA3 was selected for Lora type, and the maximum training step was selected as 10000 steps, and the training time was 40 minutes.

3 Results

As shown in Fig. 1, the Fréchet Inception Distance (FID) value sum was used to evaluate the effect of the generated image. In this paper, 100 generated images and 28 original images were used for calculation, and the FID value was 77. 29. FID is a metric commonly used to evaluate the quality of generative models, especially in the field of image generation, where FID can objectively measure the similarity of generated and real images in feature space. Compared with traditional pixel-level metrics, FID can more accurately reflect the visual quality and content similarity of the generated image by comparing it in a more advanced feature space. The lower the FID, the closer the feature distribution between the generated image and the real image, indicating that the generative model can simulate the data distribution more realistically and the quality of the generated image is higher.



(a) (b)

Fig. 1 The results are compared with the images, (a) is the sample image generated, and (b) is the sample of the original image (Picture credit: https://www. douyin. com and searching for elaina emojis).

The model in this study learns the specific visual features and styles in the dataset very well and especially excels in visualizations. Examples include lines, color blocks, and background styles in anime. As a result, even with high FID values, the resulting image will still be able to show the same style as the real image. Especially in areas such as anime, small visual differences (such as slight changes in color distribution) do not significantly affect the user's visual perception. This is consistent with what the study has pointed out that small visual changes are often not captured by the human eye, but can enhance the artistic effect in certain situations.

The calculation of the FID value is based on the statistical characteristics between the generated image and the original image, such as the mean value and the covariance matrix, which can effectively reflect the quality of the generated image. However, the small number of 28 original images used in the experiment may affect the stability and accuracy of the FID values. Studies have shown that smaller sample sizes can lead to misjudgments about the true image distribution [11, 12]. Therefore, to obtain a more reliable FID value, it is recommended to increase the number of original images in subsequent experiments to improve the representativeness and reliability of the calculation.

The advantage of this model lies in its good ability to learn the features of a specific dataset, which makes the generated images have a high similarity in style and content.

However, the disadvantage of the model is that the interpretation of FID values is limited by the quality and quantity of training samples. When the number of original images is insufficient, the FID value may not adequately reflect the actual difference between the generated image and the real image. In addition, models can face challenges in generating highly complex images, resulting in the loss of some detail that can affect the final visual effect.

Overall, although the results of this experiment show the potential of the model to generate specific styles of images, it still needs to be optimized in terms of dataset selection and training parameters to improve the practical application value of the model.

4 Conclusion

Based on the Animefull-latest dataset, this paper uses Stable Diffusion and LoRA models to generate and optimize Anime-style images. The final generated image had a (Frechet Inception Distance) FID value of 77. 29 throughout the experiment. The results show that Stable Diffusion can maintain the quality of details well when generating Anime-style images, and the training efficiency and resource consumption can be effectively improved through LoRA optimization.

Experimental results show that the LoRA model successfully optimizes the image generation process without significantly increasing the computational complexity. The introduction of this model enables us to complete high-quality image generation in a relatively short time, which verifies its feasibility and effectiveness in practical application. Compared with traditional generation methods, Stable Diffusion combined with LoRA shows stronger generation capabilities, which not only performs well in maintaining image diversity and detail integrity, but also greatly reduces training time and resource requirements. This provides us with a reference and efficient solution for our future research in the field of animation image generation.

On the one hand, by using a combination of Stable Diffusion and LoRA, this study proves that this method can generate high-quality images with limited resources. It provides a feasible solution for small and medium-sized creators and studios to achieve efficient image generation with limited computing resources. On the other hand, the method proposed in this paper has made remarkable progress in the field of animation image generation, especially for generating complex scenes or images with rich details. The method in this paper can also be used as a general framework for further application and promotion in other styles or complex image generation tasks. Finally, this study points out the direction for the further development of image generation. First of all, there is still room for further optimization in the combination of Stable Diffusion and LoRA, such as through finer model tuning (e.g., adjusting learning rate, text encoder learning rate, Unet learning rate, maximum training steps) and dataset expansion, which can further improve the quality and speed of image generation.

References

1. AR. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, arXiv preprint arXiv:2112.10752 (2022).
2. L. Tsung-Yi, M. Michael, J. B. Serge, D. B. Lubomir, B. G. Ross, H. James, P. Pietr, R. Deva, D. Piotr, and C. Lawrence Zit nick. Microsoft COCO: common objects in context. CoRR. 6, 7, 27 (2014).
3. C. Saharia, J. Ho, W. Chan, D. J. Fleet, M. Norouzi, Image Super-Resolution via Iterative Refinement, arXiv preprint arXiv:2104.07636 (2022).

4. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, arXiv preprint arXiv:2106.09685 (2021).
5. H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P. A. Heng, S. Z. Li, A survey on generative diffusion models, IEEE Transactions on Knowledge and Data Engineering (2024).
6. J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, Advances in Neural Information Processing Systems 33, 6840-6851 (2020).
7. D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, Proceedings of the International Conference on Learning Representations (ICLR) (2014).
8. Y. Zhou, et al., Efficient Fine-tuning of Pre-trained Models with Low-Rank Adaptation, In Proceedings of the AAAI Conference on Artificial Intelligence (2022).
9. A. Borji, Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2, arXiv preprint arXiv:2210.00586 (2022).
10. Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen, J. Xu, Shifted diffusion for text-to-image generation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10157-10166 (2023).
11. M. Chhabra, K. Manasa, A. Devraj, Impact of Data Sample Size on Machine Learning Model Accuracy, International Journal of Data Science and Analytics 7(4), 150–162 (2020).
12. A. Koshti, Challenges of Small Sample Sizes in Deep Learning Models, Proceedings of the Neural Information Processing Symposium (NIPS) 34, 3001–3012 (2022).