Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/97/20241294

Application of Stable Diffusion and LoRA Models in AI Drawing

Shunkai Gong

School of Data Science, City University of Macau, 999078, China

D23090103911@cityu.edu.mo

Abstract. This paper explores the application of Stable Diffusion model and LoRA (Low-Rank Adaptation) model in AI-generated artwork. The authors introduce the foundational principles of Stable Diffusion model and LoRA, as well as their application in high-quality image generation. Using three popular datasets — ImageNet, COCO, and CelebA — we apply various image quality assessment metrics, including PSNR (Peak Signal-to-Noise Ratio), IS (Inception Score), and FID (Fréchet Inception Distance), and further validate their potential in artistic creation through subjective evaluations. By comparing the performance of these two models across different datasets, we examine their strengths, weaknesses, areas for improvement in image generation tasks, along with user experience considerations. The experimental results show that the Stable Diffusion model excels in terms of image quality and diversity, while the LoRA model offers significant benefits in computational efficiency and resource usage. Through comprehensive experimental evaluations, this paper provides a scientific basis for model selection in AI art creation and offers insights for the future development of hybrid models.

Keywords: AI Plotting, Stable Diffusion, Diffusion Model, Lora.

1. Introduction

With the development of deep learning technology, generative models have made significant progress in the field of image generation. Models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have demonstrated strong capabilities in image generation tasks. However, the emergence of new generative models such as Stable Diffusion and LoRA provides a new possibility for image generation[1][2]. In recent years, AI drawing technology has developed rapidly, with Stable Diffusion, as a text-to-image generation model based on latent space diffusion models, receiving widespread attention due to its ability to generate high-resolution, high-quality images. The LoRA (Low-Rank Adaptation) model, an important extension of Stable Diffusion, achieves fast and efficient fine-tuning of the model by introducing trainable low-rank matrices, while maintaining the high quality of generated images. Previous research has mainly focused on the technical implementation and optimization of the Stable Diffusion and LoRA models, whereas this paper explores the practical applications and effects of these two models in AI drawing, especially their performance in different scenes and styles, based on the evaluation of three commonly used datasets: ImageNet, COCO, and CelebA, based on evaluations using FID, IS, and PSNR metrics. This paper aims to compare the performance of these two models on different datasets and analyze their applicability in AI drawing tasks. Studying the application of Stable Diffusion and LoRA models in AI drawing not only promotes

DOI: 10.54254/2755-2721/97/20241294

technological progress and innovation but also brings significant benefits to society, professional fields, and various user groups. The quality and effectiveness of picture generation can be enhanced by these research findings, which will greatly advance both social growth and technical advancement.

2. Related work

GANs and VAEs: GANs use adversarial training of the generator and discriminator to produce high-quality images, whereas VAEs use probabilistic models to produce a wider variety of images [1]. Stable Diffusion Model: The Stable Diffusion Model progressively reduces noise in photos by using the diffusion process [2]. This model produces images with excellent stability and quality. LoRA Model: By using low-rank decomposition and parameter compression, it dramatically lowers computational complexity and storage needs without sacrificing image quality [3].

3. Method

3.1. Model description

3.1.1. Stable diffusion model overview. The Stable Diffusion model generates images by progressively removing noise, with the core idea of gradually transforming noise into high-quality images. Specifically, the model performs small denoising operations at each step, gradually approaching the target image. The training process of the Stable Diffusion model includes the following steps:

- Noise Injection: Adding random noise to the input image.
- Denoising Process: Gradually removing noise through multiple denoising steps, using a denoising network at each step.
- Loss Function: Optimizing the parameters of the denoising network using a specific loss function (e.g., mean squared error).

In simple terms, the Stable Diffusion model is like a patient sculptor, gradually removing noise (unwanted parts of the stone surface) to ultimately create a high-quality image (an exquisite statue) from a random noise image (a rough stone block), tailored to your needs.

3.1.2. LoRA model: efficient adaptation technique. LoRA, or Low-Rank Adaptation, is a model adaptation technique that reduces computational complexity and storage requirements by decomposing high-dimensional parameter matrices into the product of low-rank matrices. The training process involves three key steps:

- Parameter Decomposition: Decompose the high-dimensional parameter matrix into the product of two low-rank matrices.
- Model Training: Train the model on the decomposed low-rank parameter matrices.
- Parameter Update: Update the low-rank parameter matrices using optimization algorithms such as gradient descent.

In simple terms, consider a sound engineer adjusting audio effects for a movie. There is a generic preset (pre-trained model) available, but it does not perfectly fit the movie. The engineer needs to fine-tune the existing preset to better suit the movie, similar to how LoRA fine-tunes models through adaptation.

3.2. Dataset selection

This study selected three commonly used datasets for experimentation:

- ImageNet: Contains a vast collection of natural images, suitable for evaluating the image generation capabilities of models [4]. Created by Professor Feifei Li from Stanford University, it is one of the largest image recognition databases in the world, comprising over 14 million images across more

DOI: 10.54254/2755-2721/97/20241294

than 20,000 categories. This dataset is primarily used for tasks such as image classification and object detection, serving as a benchmark for assessing the performance of image classification algorithms.

- COCO: Consists of images with various objects and scenes, suitable for evaluating the diversity and complex scene generation capabilities of models [5]. With the full name Microsoft Common Objects in Context, it is a large-scale dataset for object detection, segmentation, and captioning. It includes over 330,000 images with more than 2 million annotated objects, covering 80 categories. The COCO dataset emphasizes contextual information, making it suitable for complex scene understanding.
- CelebA: Contains a large number of face images, suitable for evaluating the model's performance in face generation tasks [6]. It is a massive face attribute dataset comprising over 200,000 images of celebrities, with each image accompanied by 40 attribute annotations such as gender, age, and hairstyle. The CelebA dataset is widely used in tasks such as face attribute recognition, face identification, and face detection.

3.3. Evaluation indicators

To comprehensively evaluate the performance of the model, we use the following evaluation metrics:

- FID (Fréchet Inception Distance): Used to assess the quality and diversity of generated images [7]. It measures the distance between generated images and real images. Feature vectors are extracted using the Inception V3 model, and the Fréchet distance is calculated, considering both mean and covariance, to evaluate the difference between the two distributions. A lower FID value indicates better model performance.
- IS (Inception Score): Used to evaluate the quality of generated images. It is commonly used to evaluate GANs' ability to generate images [8]. The generated images are fed into the Inception model to obtain label vectors, and the entropy value is calculated to measure image quality and diversity. A higher IS value indicates better model performance.
- PSNR (Peak Signal-to-Noise Ratio): Used to evaluate the sharpness of generated images [9]. It assesses the performance of tasks such as image restoration and super-resolution. PSNR calculates the pixel-level difference between the generated image and the target image, evaluating the amount of image noise. A higher PSNR value indicates better model performance.

3.4. Experimental setup and fairness

Hardware Configuration is same:

- Used to ensure fairness and comparability.
- NVIDIA RTX4090 GPU cluster employed.

Adequate Resources:

- Sufficient memory and storage allocated.
- Ensured models completed training and testing within a reasonable time.

Training and Testing Process:

- Two models were selected for detailed training and testing.
- Experiments conducted under identical conditions to maintain fairness.
- Completion of training and testing phases monitored for timely execution.

4. Experimental results

4.1. Stable diffusion vs. LoRA on imagenet

Stable Diffusion:

- Generates high-quality images on ImageNet, with excellent FID and IS scores.
- Image examples showcase clear details and natural color transitions.

DOI: 10.54254/2755-2721/97/20241294

LoRA:

- Also performing well on ImageNet but is slightly inferior to Stable Diffusion in complex image generation tasks.
- Outstanding in computational efficiency, offering faster generation speed.

Comparative Analysis:

- Image Quality & Details: Stable Diffusion excels, producing images with more natural details and color transitions.
- Computational Efficiency & Resource Use: Faster generation and less resource use are two of oLoRA's many benefits.

4.2. Stable diffusion vs. LoRA on COCO dataset

Stable Diffusion:

- Continues to generate high-quality images on the COCO dataset, demonstrating strong diversity.
- Produced images showcase a wide range of objects and scenes, with rich and natural details.

LoRA:

- Also performs exceptionally well on the COCO dataset, matching Stable Diffusion in diversity and detail.
- Generated images exhibit various objects and scenes, with equally rich and natural details.

Comparative Analysis:

- Performance on COCO: Both models perform similarly, generating diverse and detailed images.
- Computational Complexity & Resource Use: LoRA continues to be advantageous, providing better computing efficiency and less resource usage. In particular, LoRA performs far better in terms of computation speed and resource consumption, but Stable Diffusion is superior in producing a wide variety of intricate scenarios.

4.3. Stable diffusion vs. LoRA on CelebA dataset

Stable Diffusion:

- Generating high-quality face images.
- Detailed facial features and natural expressions.

LoRA:

- Also performing well in face generation.
- Comparable quality to Stable Diffusion.
- Clear features and natural expressions.

Comparative Analysis:

- Both models excel in generating detailed facial features and natural expressions.

LoRA Advantages:

- Superior in computational efficiency.
- Faster generation and reduced resource use.

DOI: 10.54254/2755-2721/97/20241294

5. Discussion

5.1. Summary of results

The experimental results indicate that the stable diffusion model can generate detailed and realistic images. It performs well in generating high-quality images. This model provides artistic flexibility, allowing artists and designers to create diverse and complex images for new forms of artistic expression and experimentation. This makes it important for applications in digital art, realistic image synthesis, and high-resolution graphics. This model is suitable for fields where image fidelity is crucial. It also demonstrates the diversity of outputs, generating multiple different images from the same input. This model has become an ideal choice for creative industries that value creativity and diversity.

In terms of computational efficiency, the LoRA model works effectively. In particular, this model uses low rank decomposition to reduce hardware needs and increase resource efficiency. Because of this capability, it is very appropriate for contexts with limited resources, particularly for real-time applications and mobile devices. It uses low rank decomposition to increase resource efficiency and decrease hardware requirements. This model's effective use of resources also gives it scalability, which makes it appropriate for the widespread implementation of interactive design tools and instructional software.

In conclusion, the Stable Diffusion model's clear advantage in generating high-quality photographs makes it a suitable choice for applications requiring detailed photography. However, the LoRA model excels in resource efficiency and utilisation, which makes it suitable for contexts with limited resources. Each model has unique benefits that can be applied to various AI painting scenarios. Their combined use is expected to produce AI-driven art tools that are more efficient, better, and accessible, opening up new possibilities in the digital art industry.

5.2. Improvement

One possible approach to further improving the performance and user experience of these two models is to optimize response speed, hardware resource requirements, image quality, generation diversity, as well as user interface and feedback mechanisms. The main direction of future research is to combine the advantages of both, explore the possibility of hybrid models, and develop a generative model that can generate high-quality images and high computational efficiency. At present, there have been some explorations in this research, such as combining the denoising process of stable diffusion models with the low rank decomposition method of LoRA models, to make AI image generation more in line with the artist's intention and simplify the artistic creation process. The purpose of this combination is to achieve more efficient image generation[10].

6. Conclusion

Overall, artificial intelligence painting will play an increasingly important role in future art. It allows people to choose different models according to their specific artistic needs to meet different painting needs. This article evaluates the advantages and disadvantages of stable diffusion and LoRA models in terms of image quality, computational complexity, and resource utilization. This comparison was made based on their image generation abilities on different datasets. The research results indicate that the stable diffusion model has certain advantages in image quality and diversity, but the LoRA model outperforms it in terms of computational efficiency and resource utilization. To meet different user needs, the advantages of these two models can be utilized to select the most suitable model for different application scenarios. At the same time, this application can further promote the development and application of artificial intelligence painting technology.

Acknowledgment

First and foremost, I would want to express my sincere gratitude to my university's instructors and professors, who have given me invaluable advice at every step of the thesis writing process. I also want

DOI: 10.54254/2755-2721/97/20241294

to express my gratitude to my parents and all of my friends for their support and encouragement. I could not have finished my thesis without all of their insightful guidance and amazing kindness.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[2] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239.

[3] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

[4] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[5] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[6] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

[7] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.

[8] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. Advances in neural information processing systems, 29.

[9] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

[10] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. arXiv preprint arXiv:1406.2661.