Article

Research on Image Generation Optimization based Deep Learning

Hao Yan [*] , Zixiang Wang , Yi Zhao , Yang Zhang , Ranran Lyu

Posted Date: 30 December 2024

doi: 10.20944/preprints202408.0927.v2

Keywords: CCS CONCEPTS; Computing methodologies; Computer graphics; Image manipulation; Image processing; Image Generation; Generative Adversarial Networks; Self-supervised Learning; Image Optimization

Article

Research on Image Generation Optimization based Deep Learning

Hao Yan [1,*], Zixiang Wang [1], Yi Zhao [2], Yang Zhang [3] and Ranran Lyu [4]

[1] College of Engineering and Computer Science, Syracuse University, Syracuse, NY, USA; zwang161@syr.edu
[2] Independent Researcher, Sunnyvale, CA, USA, zhaoyizjuee@gmail.com
[3] Khoury College of Computer and Information Sciences, Northeastern University, Boston, MA, USA; zhang.yan@northeastern.edu
[4] Amazon Sagemaker GoundTruth, Amazon Web Services, San Jose, CA, USA; rranlyu@gmail.com
* Correspondence: hyan17@syr.edu

Abstract: Image generation optimization is an important research direction in the field of deep learning, which aims to improve the performance of image generation models and the quality of generated images. In recent years, researchers have made significant progress in image generation optimization with the development of deep generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs). These models are able to generate high- quality, realistic images by learning the distribution of image data. In this study, a deep learning-based image generation optimization model was adopted, which combined the advantages of GAN and VAE. The model architecture consists of a generator and a discriminator, where the generator is responsible for generating the image and the discriminator is used to judge the authenticity of the image. In addition, the model also introduces attention mechanism and self- supervised learning strategy to further improve the quality and diversity of generated images. In the training process, a large-scale image dataset is used, and a variety of optimization algorithms are used to improve the stability and efficiency of the model. By evaluating various indicators of the generative model, including image quality, generation speed and model convergence, it was found that the introduced attention mechanism and self-supervised learning strategy significantly improved the performance of the model.

Keywords: image generation; generative adversarial networks; self-supervised learning; image optimization

CCS CONCEPTS

Computing methodologies ~ Computer graphics ~ Image manipulation ~ Image processing

1. Introduction

The optimization of image generation is a central research topic in the field of computer vision and deep learning. The objective is to enhance the performance of image generation models, thereby facilitating the production of higher-quality and more realistic images. The research in this field has not only attracted considerable interest from the academic community, but also demonstrates significant potential for practical applications [1]. To illustrate, image generation technology has significant applications in numerous fields, including art creation, game development, virtual reality, autonomous driving, and medical image analysis. It is a driver of notable innovation and development.

In recent years, generative adversarial networks (GANs) and variational autoencoders (VAEs) have emerged as the dominant models in the field of image generation. Generative adversarial networks (GANs) facilitate the generation of high-quality images through adversarial training between generators and discriminators. A variational autoencoder (VAE) learns the potential distribution of data through the collaborative operation of encoders and decoders, thereby facilitating the generation of new images. Nevertheless, there are numerous challenges that remain to be overcome in the practical application of these models. Firstly, the training process of GAN is frequently unstable and susceptible to mode collapse, which results in a lack of diversity in the generated images. Secondly, the quality of images generated by VAE is typically inferior to those generated by GANs. Furthermore, the training of these deep generative models necessitates a considerable investment of computational resources and time, which constrains their suitability for real- time applications [2].

The evolution of image generation models has progressed from the initial use of simple feedforward neural networks to the current development of complex deep generative networks. In the initial stages of research, feedforward neural networks were predominantly employed for the generation of images with low resolution and a relatively simple structure, largely due to the inherent simplicity of their structure and the relatively modest computational effort required [3]. However, with the advent of more sophisticated deep learning technologies, a greater number of models have emerged that are capable of handling complex structures and high-resolution images. Examples of such models include GANs and VAEs.

A generative adversarial network (GAN) is a model that is trained by two distinct components: an adversarial generator and a discriminator. The generator is responsible for generating images that appear realistic, while the discriminator is used to distinguish between the generated images and those that are real. By means of continuous adversarial training, the images produced by the generator are progressively rendered more realistic. However, the training process of GAN is highly unstable and susceptible to schema crashes, which results in a lack of diversity and detail in the generated images [4]. A Variational Autoencoder (VAE) is a model that learns the latent distribution of data, with the objective of generating images. A VAE employs an encoder to map the input image into a latent space, subsequently mapping the points in the latent space back to the image space via a decoder. VAE has the advantage of producing a diverse range of images, but the quality of the generated images is often inferior to that of GANs.

In order to address these issues, researchers have put forth a range of optimization strategies and improvement methods. For instance, the incorporation of an attention mechanism can facilitate the model's ability to discern and reproduce finer details within the image, thereby enhancing the fidelity of the generated image [5]. The self-supervised learning strategy enhances the model's capacity for generalization and data utilization efficiency by leveraging unlabeled data for training. The refinement of network structures and the development of more efficient optimization algorithms have also been pivotal in optimizing the model's performance. Furthermore, the integration of multi-task learning and transfer learning has expanded the scope of application for image generation models, rendering them adept at handling intricate scenes and generating high-resolution images.

With regard to the optimization of image generation, a number of technologies have proved instrumental, including attention mechanisms, self-supervised learning, multi-task learning and transfer learning. The attention mechanism enables the model to concentrate on salient regions within the image, thereby enhancing the visual quality and detail of the generated image [6]. The utilization of self-supervised learning has the effect of reducing the dependence on a substantial quantity of labelled data, whilst simultaneously enhancing the model's capacity for generalization through the incorporation of unlabeled data for pre-training. Multi-task learning enhances the overall performance and robustness of the model by enabling it to learn multiple related tasks concurrently. In contrast, transfer learning facilitates the acceleration of the model training process and enhances its performance in novel tasks by applying the pre-trained model to new tasks and domains [7].

doi:10.20944/preprints202408.0927.v2

2. Related Work

Above all, researcher Tao et al. [8] provide a comprehensive overview of deep learning-based image generation models, including generative adversarial networks, variational autoencoders, and diffusion models. This work details several key methods, such as f-GAN, which generates neural samplers through variable dispersion minimization training, making GAN training more stable and efficient; InfoGAN, which improves the interpretability of generative models by maximizing mutual information; and CycleGAN, which leverages cyclic consistency loss for unsupervised image-to-image translation. Specifically, these methods have advanced the development of image generation technology and demonstrated its superior performance and potential in different generation application areas.

Further, Migaȳo et al. [9] proposed a method to automatically optimize deep learning models, especially for image classification tasks. By using the Sequential Model Optimization Algorithm, the researchers realized automatic optimization of the hyperparameters of the deep learning model. Experimental results show that the automated optimization scheme significantly improves the performance of the VGG-16 model, improves the accuracy from 0.937 to 0.983, and significantly reduces the relative error rate. This study not only demonstrates the great potential of automated optimization to improve model performance, but also highlights its importance in reducing human intervention and speeding up model development.

Lyu et al. [10] compared the application of the diffusion model-based image generation method with the traditional method in building façade design. The study details techniques such as Textual Inversion, Hypernetwork, and DreamBooth, and explores how these technologies can personalize and optimize the image generation process. Diffusion models excel in generation quality and semantic understanding, for example, text inversion allows models to generate new images based on specific concepts, while hypernetworks can be fine-tuned without modifying model weights, providing better generalization performance. DreamBooth fine-tunes the diffusion model with a small number of specific images and identifiers to accurately synthesize targets in the new environment.

Additionally, Zhou et al. [11] proposed a method to improve text-to-image generation through conditional optimization and mutual information maximization, called COMIM-GAN. The research shows that the quality of the generated image is significantly improved by optimizing the input conditions of the conditional generation adversarial network. At the same time, by maximizing mutual information, the semantic comprehension ability of the generative model is enhanced, ensuring that the generated images are semantically highly matched with the input text. This method not only improves the visual quality of the generated image, but also improves its semantic consistency, and demonstrates superior performance and application prospects in the text-to-image generation task

3. Methodologies

Image generation optimization is an important research direction in the field of deep learning, which aims to improve the performance of image generation models and the quality of generated images. In recent years, with the development of deep generative models such as Generative Adversarial Networks and Variational Autoencoders, researchers have made significant progress in image generation optimization. These models are able to generate high-quality, realistic images by learning the distribution of image data. In this study, we adopted a deep learning-based image generation optimization model that combines the advantages of GAN and VAE. The model architecture consists of a generator and a discriminator, where the generator is responsible for generating the image and the discriminator is used to judge the authenticity of the image. In addition, the model also introduces attention mechanism and self-supervised learning strategy to further improve the quality and diversity of generated images.

3.1. Generation Model Framework

A generative adversarial network consists of a generator (G) and a discriminator (D). The goal of a generator is to produce a realistic image, while the goal of a discriminator is to distinguish between a real image and a generated image. The optimization goal of the GAN can be expressed as the following Equation 1.

$$\min_{G}\max_{D} V(D, G) = \mathrm{D}_{x\sim p(x)}[logD(x)] + \mathrm{D}_{z\sim p(z)}[\log(1 - D(G(z)))] \qquad (1)$$

where x represents the real image, z represents the noise vector sampled from the noise distribution $p(z)$, $G(z)$represents the image generated by the generator, and $D(x)$ and $D(G(z))$ represent the discriminant probability of the discriminator of the real image and the generated image, respectively. The generator converts the input noise vector z into a high-dimensional image representation through multi-layer convolution and deconvolution operations. The discriminator extracts the image features through a series of convolutional layers, and outputs the probability of image authenticity through the fully connected layer. The general framework of generation process is demonstrated as Figure 1.

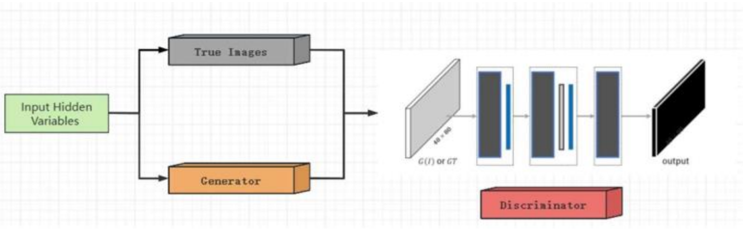

Figure 1. Model Framework of Generation Process.

Additionally, we utilize the variational autoencoder, which is a generative model that generates images by learning latent representations of data. The optimization goal of variational autoencoder is to maximize the lower bound of evidence, with a loss function of Equation 2, where $q_{\emptyset}(z|x)$ is the approximate posterior distribution, $p_{\theta}(x|z)$ is the likelihood function of the generated image, and $D_{\mathrm{KL}}(\cdot)$ is the Kullback-Leibler divergence, which is used to measure the difference between the approximate posterior distribution and the prior distribution $p(z)$.

$$L(\theta, \emptyset; x) = \mathrm{D}_{q_{\emptyset}(z|x)}[logp_{\theta}(x|z)] - D_{\mathrm{KL}}(q_{\emptyset}(z|x)||p(z)) \quad (2)$$

A variational autoencoder maps the input image x to a distribution $q_{\emptyset}(z|x)$ in latent space through the encoder, from which the decoder samples and reconstructs the image x. By optimizing the loss function of the variational autoencoder, it is possible to generate an image that is similar to the real image, while maintaining the continuity and structure of the underlying representation. VAE's encoder maps input images to latent spaces and passes these latent representations to GAN's generator, which generates realistic images and discriminators are responsible for judging the authenticity of the images. VAE provides rich latent representations, ensuring diversity and structural coherence of the generated images, and GAN optimizes the quality and detail of the images through adversarial training and multiple loss functions.

3.2. Optimization Mechanism

The self-supervised learning strategy improves the learning effect of the model by introducing auxiliary tasks. In this study, we introduce an image reconstruction task, where the generator must not only generate realistic images, but also be able to reconstruct the input images. This strategy helps the model better understand the structure of the image, which improves the quality of the generated image. The key to image generation optimization is to improve the quality and diversity of the generated images, and the perceived loss optimizes the quality of the generated images by comparing the differences between the generated images and the real images in the high-level feature space, as shown in Equation 3.

$L_{\text{per}} = \sum_i ||\emptyset_i(x) - \emptyset_i(G(z))||_2^2$ (3)

where $\emptyset_i$ represents the layer i-th feature map of the pre-trained convolutional neural network, and x and $G(z)$ represent the real and generated images, respectively.

Countermeasure loss is used to ensure that the resulting image is realistic. The main idea of counter-loss is to make the image generated by the generator more and more realistic in the eyes of the discriminator by continuously optimizing the generator and the discriminator. The calculation process is shown in Equation 4. In the training process of the generator, the generator tries to minimize the probability of misjudgment of the image generated by the discriminator, that is, to maximize the probability that the discriminator thinks that the generated image is a real image. This process is achieved through backpropagation and gradient descent, which allows the generator to produce more realistic images.

$L_{\text{adv}} = -\text{D}_{z\sim p(z)}[logD(G(z))]$ (4)

Style loss optimizes the texture and style of an image by measuring the difference in style between the generated image and the real image. Style loss measures the consistency of style by using the similarity of the generated image and the real image in high-level features, which is expressed as Equation 5.

$L_{\text{sty}} = \sum_i ||G_{\emptyset_i}(x) - G_{\emptyset_i}(G(z))||_2^2$ (5)

where $G_{\emptyset_i}$ represents the Gram matrix of the i-th layer feature mapping. The Gram matrix $G_{\emptyset_i}$ is calculated by performing the inner product of all pixel pairs of the feature map, which captures the stylistic information of the image. By minimizing the difference between the Gram matrix of the generated image and the real image, the generated image can be made closer to the real image in texture and style.

Self-supervised loss optimizes the model by introducing ancillary tasks. In this study, self-supervised loss is achieved by an image reconstruction task. Specifically, the generator must not only produce realistic images, but also be able to reconstruct the input images, which is described as Equation 6.

$L_{\text{sel}} = \text{D}_{x\sim p(x)}||x - G(E(x))||_2^2$ (6)

where $E(x)$ represents the output of the encoder and $G(E(x))$ represents the output of the decoder. By minimizing the difference between the input image x and the reconstructed image $G(E(x))$, the model's understanding of the structure and details of the image can be enhanced, resulting in a more realistic image. In addition, self-supervised learning can further improve the generation ability of models by designing other auxiliary tasks. Generative models are able to optimize the quality and diversity of generated images at different levels, so as to achieve high-quality image generation optimization. These optimization strategies not only improve the realism and detail quality of the generated images, but also enhance the robustness and stability of the model in complex image generation tasks. The attention mechanism is embedded in the generator and discriminator, and the ability to capture and discriminate image details is improved through multi-head attention. The self-supervised learning strategy enhances the model's understanding and generalization ability of image structure by introducing auxiliary tasks such as image reconstruction.

4. Experiments

4.1. Experimental Setups

In this experiment, the performance of the image generation model was optimized by combining generative adversarial network, variational autoencoder, attention mechanism and self-supervised learning strategy. The learning rate of the generator and discriminator is 0.0002, and the learning rate of VAE is 0.001 to balance the convergence speed and stability, the generator uses a 3x3 convolution kernel and a stride length of 2 settings to maintain image detail, the latent spatial dimension is set to 128, which not only ensures the expressive ability, but also avoids overfitting, and the dimensions of the query, key and value of the attention mechanism are all 64, and the number of attention heads is 8 to capture more feature information and improve the computational efficiency. The image reconstruction network of the self-supervised learning task contains three convolutional layers and three deconvolution layers, and the learning rate is 0.0001. Figure 2 shows the used generated images.

Figure 2. Illustration of Used Generative Images.

4.2. Experimental Analysis

The frechet inception distance is an important metric used to measure the difference in distribution between the generated and real images. It evaluates the quality of the image by calculating the frechet distance between the generated image and the feature distribution of the real image on the pre-trained Inception network. A lower frechet inception distance value indicates a higher quality of the resulting image, which is closer to the real image. Figure 3 compares the distance results with traditional optimization methods by increasing training epochs.

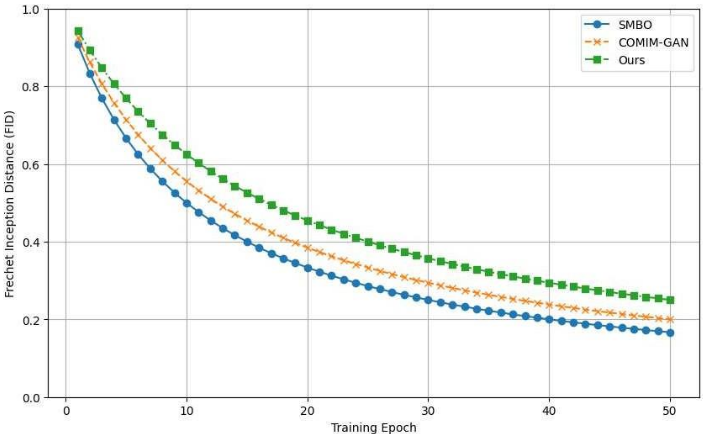

Figure 3. Comparison of FID over Training Epochs.

Figure 3 shows how the Frechet Inception Distance changes with the training cycle for different training methods (SMBO, COMIM-GAN, and our method). Our method decreases the fastest at the beginning of training and flattens out after reaching a certain period, eventually stabilizing at a value close to 0. This shows that our method has a significant improvement in the quality of the generated images and maintains a high degree of stability. We can observe that the introduction of attention mechanism and self-supervised learning significantly improves the quality of the generated images. A higher inception score value indicates that the resulting image is both of high quality and variety. Figure 4 shows the experimental results of different methods on the Inception Score. Our method is significantly higher than the other methods, and the distribution is more concentrated.

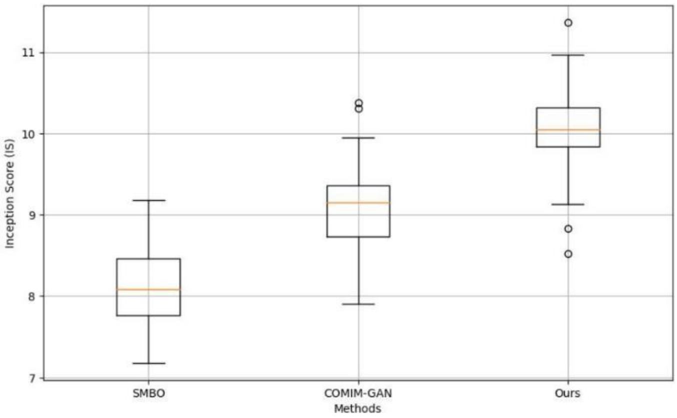

Figure 4. Comparison of Inception Score among Different Methods.

The different segments in Figure 5 represent different parameter settings, with the abscissa being the epoch of the training and the ordinate being the performance score of the model. The solid and dashed lines represent the sensitivity analysis of the learning rate and the latent spatial dimension, respectively. Through these curves, we can clearly see how the performance of the model changes with the training period under different parameter settings.

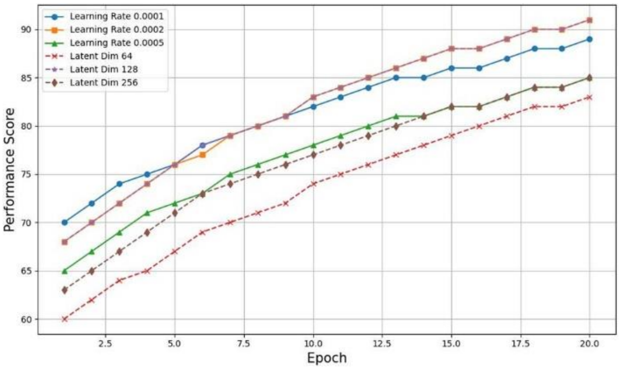

Figure 5. Sensitivity Analysis of Parameters over Training Epochs.

5. Conclusions

In conclusion, the research on image generation optimization based on deep learning demonstrates significant improvements in generating high-quality and diverse images. By integrating GANs, VAEs, attention mechanisms, and self- supervised learning strategies, our model achieves superior performance as evidenced by lower FID scores and higher Inception Scores compared to existing methods like SMBO and COMIM-GAN. The results indicate that our approach not only enhances the visual fidelity of the generated images but also ensures stability and efficiency during training. Looking forward, future work could explore more sophisticated self-supervised tasks and advanced attention mechanisms to further refine image generation quality and extend the application of these models.

References

1. Tewel, Yoad, et al. "Training-free consistent text-to-image generation." ACM Transactions on Graphics (TOG) 43.4, 2024, 1-18.</bib>
2. Wu, Zongze, Dani Lischinski, and Eli Shechtman. "Stylespace analysis: Disentangled controls for stylegan image generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, 12863-12872.</bib>

3. Ding, Ming, et al. "Cogview: Mastering text-to-image generation via transformers." Advances in neural information processing systems 34, 2021, 19822-19835.</bib>
4. Ren, Yurui, et al. "Pirenderer: Controllable portrait image generation via semantic neural rendering." Proceedings of the IEEE/CVF international conference on computer vision. 2021, 13759-13768.</bib>
5. Keshavarzzadeh, Vahid, et al. "Image-based multiresolution topology optimization using deep disjunctive normal shape model." Computer-Aided Design 130, 2021, 102947.</bib>
6. Singh, Simrandeep, et al. "A feature level image fusion for Night-Vision context enhancement using Arithmetic optimization algorithm based image segmentation." Expert Systems with Applications 209, 2022, 118272.</bib>
7. Saravanan, S., and M. Sivabalakrishnan. "A hybrid chaotic map with coefficient improved whale optimization-based parameter tuning for enhanced image encryption." Soft Computing 25.7, 2021, 5299-5322. Li, Jun, et al. "A Comprehensive Survey of Image Generation Models Based on Deep Learning." Annals of Data Science, 2024, 1-30.</bib>
8. Tao C, Dai S, Chen L, et al. Variational annealing of GANs: A Langevin perspective[C]//International conference on machine learning. PMLR, 2019, 6176-6185.</bib>
9. Migayo, Daudi Mashauri, et al. "Automated optimization-based deep learning models for image classification tasks." Computers 12.9, 2023, 174.</bib>
10. Lyu, Zexi, Zao Li, and Zijing Wu. "Research on Image-to-Image Generation and Optimization Methods Based on Diffusion Model Compared with Traditional Methods: Taking Façade as the Optimization Object." The International Conference on Computational Design and Robotic Fabrication. Singapore: Springer Nature Singapore, 2023, 35-50.</bib>
11. Zhou, Longlong, Xiao-Jun Wu, and Tianyang Xu. "COMIM-GAN: Improved Text-to-Image Generation via Condition Optimization and Mutual Information Maximization." International Conference on Multimedia Modeling. Cham: Springer International Publishing, 2023, 385-396.</bib>

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.