Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

A Comparative Analysis Between GAN and Diffusion Models in Image Generation

Yingying Peng

School of Economics and Management, Beijing Jiaotong University, Weihai, Shandong, 264400, China

22711058@bjtu.edu.cn

Abstract. In the field of artificial intelligence, image-generation techniques have been a hotspot for research. Two generative models that have garnered a lot of attention are diffusion models and generative adversarial networks (GANs). This review paper aims to compare and analyze GAN and Diffusion Models in the field of picture generation, as well as to give a thorough discussion of their features, applications, benefits, and drawbacks. Firstly, the work related to the working principle of GAN and diffusion models are introduced, and then their applications and results in image generation are reviewed. By comparing the existing research results, the author find that GAN performs well in generating realistic images but suffers from problems such as pattern collapse and unstable training, while the diffusion model has better stability and controllability. Combining the advantages of the two methods, this paper explores the possible fusion methods and looks forward to the future development direction in the field of image generation. These research results provide important references and insights to further enhance the level and application scope of image generation technology.

Keywords: Image generation; generative neural network; diffusion model.

1. Introduction

Image generation is an important branch in the field of computer science and artificial intelligence, which is dedicated to learning and generating new, realistic images from a given dataset. With the continuous progress of technology, image generation not only occupies an important position in academic research but also plays a key role in many practical applications, such as art creation, entertainment industry, medical image analysis, virtual reality and so on.

Present-day picture-creation techniques fall into two primary categories: Diffusion Models and Generative Adversarial Networks (GANs). Since its inception, GANs have attracted widespread attention for their powerful generative capabilities and adversarial training strategies. By constructing a game process of generators and discriminators, GANs can generate highly realistic images, but they also face problems such as unstable training and pattern collapse [1,2]. Diffusion models, especially Stable Diffusion, show advantages in image quality and stability by introducing diffusion process and conditional guidance in potential space.

The purpose of this paper is to comprehensively explore the applications and characteristics of GANs and Diffusion models in the field of image generation, and to provide in-depth references for researchers in related fields by comparatively analyzing the principles, strengths, and weaknesses, as well as the performance of both in practical applications. In addition, this paper will also explore the future development trend and possible fusion direction of these two methods, in anticipation of promoting the further development of image generation technology in the future.

2. Generative Adversarial Network

A generative model known as a Generative Adversarial Network operates on the tenet that two neural network models should be trained independently of one another in order for them to learn from one another. It is made up of two models: one discriminative (D) and the other generative (G). To ensure

that the discriminator cannot tell the difference between created and actual data, the generative model must capture the data distribution, implicitly project a random noise vector into the data space, and produce fictitious data samples. To ascertain whether the sample data originate from the genuine data or the false data, on the other hand, the discriminative model must calculate the likelihood that the samples originate from the training data rather than the G [3]. The two networks compete with one another, learning and improving continuously until the generator's data approaches the real data more closely and the discriminator's ability to distinguish between them gets more precise. Adversarial generative networks are frequently employed in the production of images because of their adversarial learning methodology, which typically produces high-quality data samples.

2.1. Representative Works

2.1.1. CycleGAN

CycleGAN is a model for unsupervised image conversion that has two generators and two discriminators competing and cooperating with each other and ensures the consistency and realism of the image conversion through adversarial networks and cyclic consistency loss [4]. This means that CycleGAN model is different from ordinary GAN in that it can convert one type of image into another type of image with high quality without the need of paired training data.

It has several advantages. (1) No need for paired training data: CycleGAN can perform image transformations without paired image data, which makes it suitable for tasks where large amounts of paired data are difficult to obtain [4]. (2) Cyclic consistency loss: Through the cyclic consistency loss function, CycleGAN can ensure that the learned mapping functions G and F will not produce contradictory results, and the generated images can maintain consistency during the conversion process, thus improving the quality of the generated images and avoiding the appearance of illogical images [4]. (3) Robustness: CycleGAN shows good robustness in dealing with image transformations in different domains and can accurately and efficiently handle complex image transformation tasks in the face of noise, distortion, occlusion, etc., without failing due to changes in the input data [4]. (4) Adversarial Loss: CycleGAN uses Adversarial Loss to make the generated image match the data distribution in the target domain, thus improving the realism of the generated image [4].

In terms of practical applications, CycleGAN solves the problem of lack of pairwise training data in image transformation, making the task of image transformation more flexible in practice and able to be extended to a wider range of applications. The method employs a specific generative network structure to ensure that the generated images are of high quality and realistic. Since CycleGAN does not rely on predefined task-specific similarity functions and does not require inputs and outputs to be in the same embedding space, it has the potential to be a generalized solution for a wide range of visual and graphical tasks [4]. The emergence of CycleGAN enriches the application scenarios of image processing techniques and brings new ideas and methods to the field of image research.

2.1.2. StyleGAN

StyleGAN is an image generation model based on generative adversarial networks, which introduces the concept of potential space to control the style and content of a generated image by manipulating potential vectors. By mapping the input noise vectors to a potential space and operating in that space, StyleGAN enables fine control of the style and content of the generated image [5]. This mechanism makes the generated images more diverse, realistic, and of higher quality.

In addition to this, StyleGAN has the following features: (1) Style blending: StyleGAN employs hybrid regularization means, trained using two random latent codes [5]. This allows the user to control the style of the generated image, resulting in more personalized image generation. (2) High-Resolution Generation: StyleGAN is capable of generating high-resolution images and can handle complex image generation tasks such as face generation, landscape generation, etc., generating images with high clarity and richness of details. This avoids the blurring and distortion problems common in traditional GAN algorithms. (3) Highly controllable: StyleGAN can control the style and

attributes of the generated images by manipulating specific directions in the latent space. This enables users to achieve more personalized image generation. (4) Stochastic variation: StyleGAN introduces a noisy input that provides the generator with a way to generate random (diversity) details. This noisy input is a single channel of data consisting of uncorrelated Gaussian noise that is fed to each layer of the generative network. This allows the generated image to have some random variations in detail while maintaining overall structural consistency, increasing the diversity and realism of the image [5].

StyleGAN improves the performance of the generative adversarial network generator and solves the problem of blurring and unrealism that occurs in traditional GAN models when generating high-resolution images, as well as some unnatural artifacts and distortions that occur in the generated images.

2.1.3. BigGAN

BigGAN is a hybrid model that combines Generative Adversarial Networks (GAN) and Variational Autocoders (VAE) [6]. Unlike the original GAN structure, it has an additional encoder network. Also, BigGAN sends noise to multiple layers of the generator, unlike traditional GANs that embed noise vectors as input directly into the generative network. The quality and diversity of the generated images are improved by scaling up the model and influencing the features that do not allow resolution and hierarchy levels.

BigGAN differs from traditional GAN and other GAN variants in the following aspects: (1) Large-scale training: BigGAN improves performance by increasing the model parameters and batch size. Large-scale training is the ability of the model to be trained at higher resolutions without using explicit multi-scale methods [7]. (2) Truncation Technique: BigGAN introduces a stage technique to control the diversity and fidelity of the generated samples. The model generates images with better controllability and flexibility by staging certain portions of the noise vector [7].

BigGAN not only improves the performance of GAN in image generation but also makes the images generated by GAN close to real-world images in terms of fidelity and diversity. Moreover, BigGAN identifies the instability problem in large-scale GAN training and improves conditional performance through simple architectural changes and regularization schemes [7].

2.2. Summary

Successful applications of GANs in the field of image generation are mainly characterized by their high-quality image generation, diverse inputs, and unpaired data. Among them, StyleGAN stands out with high-quality and detail-rich face image generation, CycleGAN demonstrates its strengths in unpaired image transformation tasks, and BigGAN generates high-quality, high-resolution images through large-scale training. These models drive the versatility and controllability of GAN in the field of image generation.

However, since GAN has both a generator and discriminator, it is difficult to train both networks at the same time during the training phase. It is difficult for the user to observe the loss during training, and the loss fluctuates back and forth, requiring feedback through multiple passes during the training of multilayer deep learning models. However, in practice, the loss function often does not converge to the saddle point, and the result of training is not easy to converge, so the stability of the model is poor, especially when dealing with complex and diverse data, the saddle point is more difficult to deal with [2].

In addition, GAN is an unsupervised model, although it can be used in unsupervised and semi-supervised domains, the focus of the model training may run away, so it may learn some things that the user doesn't want it to learn or learn some phenomena that the user can't control when it learns. In this way, the model learning process is unstable.

Meanwhile, the training process of GAN is usually complicated and requires careful adjustment of the network structure, loss function training parameters, etc. Moreover, even if the training is successful, it is difficult to guarantee that the generated images fully meet the expectations.

Besides, training high-quality GAN models usually requires a lot of computational resources and time, which is a challenge for many researchers and practitioners.

3. Diffusion Model

Noise contamination is gradually introduced into an image until a completely random noise is generated and the ability to recover data from Gaussian noise is acquired. The process is divided into two phases: a forward phase and an inverse phase [8].

In the forward phase, noise is gradually added to the data until the data becomes completely Gaussian noise. This process can be viewed as continuously adding noise to the input data until a purely noisy picture is finally obtained. The whole process of adding noise operation can be seen as the process of constructing labels.

In the inverse stage, learning is reduced from Gaussian noise to the original data. This process can be viewed as starting from a purely noisy image and gradually denoising it until the original data is restored.

The Diffusion Model learns the inverse diffusion process (inverse diffusion process) to generate the desired data samples from the noise. It is inspired by nonequilibrium thermodynamics and defines a Markov chain of diffusion steps (the current state is only related to the state of the previous moment) [8].

3.1. Representative Works

3.1.1. Stable Diffusion Model

The fundamental idea of the Diffusion Model is still adhered to by Stable Diffusion, which is to extract the original data from the noise by introducing noise to the data progressively and recovering the original data from the noise through an inverse process. However, Stable Diffusion introduces some improvements and optimizations in the diffusion process to improve the quality and stability of the generated image. Stable Diffusion performs the diffusion process in the potential space instead of directly in the pixel space [9]. This means that it first uses an encoder to encode the input image into a low-dimensional latent representation, and then performs forward and backward diffusion in the latent space. Stable Diffusion can guide the generation process by introducing conditional information (e.g., text descriptions, category labels, etc.). This conditional information can be used as a guiding signal in the backward diffusion process to help the model generate images that meet specific conditions. Also, to increase the generation speed, Stable Diffusion employs an efficient sampling strategy that uses a predictive model to estimate the noise that needs to be removed at each step, instead of generating the image step-by-step directly from the noise. This approach can significantly reduce the time required to generate an image [10].

The features of stable diffusion mainly include: (1) Versatility: The Stable Diffusion model is a powerful AI painting software drawing model that produces realistic depictions and gradients. This model has the dual nature of smooth and sharp boundaries for a variety of forms and processes, making the depiction very realistic. (2) High-quality image generation: The Stable Diffusion model has been trained on a large number of high-quality images and can generate images with incredible detail and texture. At the same time, the model can handle a variety of styles and subjects, from surrealism to realism, from landscapes to portraits, all can get high-quality generation results. (3) Flexibility: The Stable Diffusion model can be used to generate various types of images, such as faces, objects, etc., providing developers with a wide range of application scenarios. In addition, the model can also be based on the user providing text prompts to generate images that meet specific conditions, realizing the text-to-image conversion. (4) Highly adaptable: Stable Diffusion model is highly adaptable and can adapt to various user-input text descriptions and image contents to generate images that meet the requirements. At the same time, the model can also be customized according to the user's needs, adjusting the image quality, resolution artistic style, and other parameters. (5) Ease of

use: The Stable Diffusion model has a simple and easy-to-understand interface that users can quickly learn to use. In addition, the model also provides a wealth of options and parameter settings, allowing users to fine-tune the model according to their own needs to obtain the best generation results. (6) Large-scale datasets: Stable Diffusion uses large-scale datasets to ensure that it can quickly and accurately output the images users want in a variety of situations.

Stable Diffusion significantly improves the performance and stability of the Diffusion Model in image generation tasks by introducing improvements and optimizations such as potential spatial diffusion, conditional guidance, and efficient sampling [10]. With the deepening of research and the continuous expansion of application scenarios, Stable Diffusion is expected to play a greater role in the future.

3.1.2. Conditional Diffusion Model

The basic principle of the Conditional diffusion model is to add spatially localized input conditions to a large pre-trained text-to-image diffusion model via the ControlNet structure [11]. ControlNet injects additional conditions into the neural network block by locking the parameters of the original model and simultaneously creating a trainable copy [11]. This structure uses a zero-convolution layer to connect the trainable copy to the original model, ensuring that no harmful noise is added to the deep features of the large-scale pre-trained model at the start of training. This approach protects the backbone of the pre-trained model while building a deep, robust, and powerful encoder for learning specific conditions.

Features of Conditional diffusion models include: (1) Spatial control: Allows the user to finely specify the spatial composition of image generation by adding conditional controls, such as edges, depth, segmentation, human pose, etc., improving the accuracy and versatility of image generation. (2) Model Protection: Protects the quality and functionality of large pre-trained models through the ControlNet structure, while providing a strong backbone for learning specific conditions, avoiding overfitting and forgetting problems. (3) Training Efficiency: The ControlNet structure improves the training efficiency of the model and enables effective control of the Stable Diffusion model by effectively controlling the image generation process, supporting the application of single or multiple conditions with or without cue text.

Unlike the traditional Diffusion Model, the Conditional Diffusion Model introduces conditional information into the diffusion and counter-diffusion process. This conditional information can be in the form of textual descriptions, sketches, labels, or other forms of input that provide additional guidance to the generation process. In the diffusion process, condition information is used to adjust how and how much noise is added to ensure that the generated intermediate state matches the condition requirements. In the inverse diffusion process, the condition information is then used to guide the denoising process to ensure that the final generated image conforms to the condition description.

Conditional diffusion models address the limitations of text-to-image generation models in terms of spatial control and conditional learning. By adding conditional control, users can specify the spatial composition of image generation more finely, such as edges, depth, segmentation, and human pose. Meanwhile, the image generation process is effectively controlled by the ControlNet structure, which improves the training efficiency and image generation quality. This approach has a wide range of application extensibility, provides more possibilities for control and customization of image generation models, and solves the problem of text-to-image generation models in dealing with image generation under specific conditions.

3.2. Summary

Diffusion models, especially Stable Diffusion models, exhibit several compelling advantages, however, they are also accompanied by some potential limitations. The following is an in-depth analysis of the advantages and disadvantages of Diffusion models:

The Diffusion model is capable of generating images with high resolution, high quality, and variety in a way that is often indistinguishable from real images. The model can handle all kinds of text and image inputs, including simple descriptions, complex narratives, abstract concepts, and concrete requirements, showing great flexibility. By introducing stability measures, the Diffusion model effectively avoids common image generation problems such as blurring, artifacts, repetition, and unnaturalness. In addition to this, by reducing the number of noises and steps, the model reduces the training time and computational cost, making it more efficient and feasible when dealing with large-scale datasets.

The model has fine-tuned parameters, such as noise vector size, step size, and number of steps, which makes the model easier to adjust and optimize for better training results. Not only that, but the model also applies to all kinds of image types and processing tasks, including image denoising, edge-preserving smoothing, image segmentation, and so on.

However, despite the optimization in training speed, the Diffusion model still requires a large number of computational resources to support its training and inference process. Moreover, due to the multiple steps involved and the complex network structure, the implementation and maintenance of Diffusion models are relatively complex. As deep learning models, Diffusion models are often regarded as black boxes whose decision-making process is difficult to explain, which may limit applications in certain domains that require a high degree of interpretability.

4. Discussion

In the field of image generation, GANs, and Diffusion Models are two generative models that have attracted much attention. By comparing the advantages and disadvantages of these two models, this paper aims to explore in depth their application prospects in image generation tasks.

First of all, GANs are known for their adversarial training framework, which realizes image generation through the game of generator and discriminator. GANs are capable of generating realistic images with high generative effectiveness and diversity. In terms of generation speed, GANs still have a clear advantage. However, GANs suffer from the problems of unstable training, pattern crashes, and pattern collapse, which lead to inconsistent quality and stability of the generated images. In addition, GANs require complex techniques and adjustments to solve the pattern collapse problem, which increases the complexity and training difficulty of the model.

In contrast, Diffusion Models avoid adversarial training in GANs by generating images through stepwise denoising, which improves the stability and controllability of training. Diffusion Models are capable of generating more realistic and detail-rich images, especially when dealing with complex backgrounds and textures, and perform well when dealing with large-scale datasets. However, Diffusion Models still require a large number of computational resources to support the training and inference process, and the high complexity of the model requires specialized technical knowledge and experience.

Combining the experimental results and model characteristics, it is believed that future research directions can focus on the following aspects: (1) How to further improve the generation speed of Diffusion models to achieve faster response time while maintaining high quality. (2) Exploring the combination of GANs and Diffusion models to fully utilize the advantages of both, e.g., utilizing the fast generation capability of GANs to accelerate the backpropagation process of Diffusion models. (3) Expand the application of the Diffusion model in other image processing tasks, such as image super-resolution and image denoising.

In summary, GANs and Diffusion models have their advantages in the field of image generation. Future research can further explore the way of combining these two models and work on improving the quality, stability, and interpretability of generated images. By continuously optimizing the model structure and training strategy, as well as exploring new application scenarios, it is expected to achieve higher-quality image generation and wider application prospects.

5. Conclusion

In conclusion, the GAN and diffusion models in the area of picture generation have been thoroughly examined in this review work. In terms of producing realistic pictures, both models have advanced significantly; nonetheless, the Diffusion Model is superior in terms of stability and controllability, while GAN excels in terms of realism. By comparing their strengths and weaknesses, this work discusses future research directions that can combine the advantages of both models to enhance image generation performance. Looking ahead, research can focus on exploring innovative fusion techniques to further improve image quality, efficiency, and stability. Meanwhile, with the continuous advancement of technology, the application prospects in the field of image generation will be broader. This review provides an important research foundation for the application of image-generation techniques in various fields and points out the direction for future development.

References

[1] Wang, Lei, Wei Chen, Wenjia Yang, Fangming Bi, and Fei Richard Yu. A state-of-the-art review on image synthesis with generative adversarial networks. IEEE Access, 2020, 8: 63514-63537.

[2] Gui, Jie, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. IEEE transactions on knowledge and data engineering. 2021, 35(4): 3313-3332.

[3] Creswell, Antonia, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. Generative adversarial networks: An overview. IEEE signal processing magazine, 2018, 35(1): 53-65.

[4] Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[5] Karras, Tero, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.

[6] Kipf, Thomas N., and Max Welling. Variational graph auto-encoders. ArXiv Preprint. 2016:1611.07308.

[7] Brock, Andrew, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. ArXiv Preprint, 2018: 1809.11096.

[8] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems. 2020, 33: 6840-6851.

[9] Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.

[10] Stable Diffusion 1 vs 2 - What you need to know. Tutorials, AI Research. Last Accessed: Mar. 02, 2024. URL: https://www.assemblyai.com/blog/stable-diffusion-1-vs-2-what-you-need-to-know/

[11] Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 3836-3847.