

Analysis of Generative Adversarial Networks (GANs) and Their Variants Based on Encoders and Decoders

Jiteng Fan ^a

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China

Keywords: Generative Adversarial Networks (GANs), Conditional GAN, Deep Convolutional GGAN (DCGAN).

Abstract: Generative Adversarial Networks (GANs), are among the most noteworthy advances in machine learning. GANs successfully utilize game concepts to train neural networks. With the deepening of research, a large number of GAN variants have been proposed, which greatly improve the performance of GAN in various aspects. To further analyse GAN, this paper provides a detailed overview. The core objective of this paper is to study the basic ideas of GAN and to explore the principles and performance of some GAN variants in depth. Additionally, the paper evaluates the strengths and weaknesses of each model as well as possible future directions. Based on MNIST and Cifar-10 datasets, this paper analyses the GAN, Conditional GAN (CGAN), Deep Convolutional GGAN (DCGAN) and Big GAN (BigGAN) models using quantitative and qualitative methods. Among them, Inception score (IS), a widely used metric to assess the quality of GAN model generation, was used to compare model performance quantitatively. Based on the experimental results, this study critically compares each GAN variant. In addition, this study discusses the existing limitations of GAN and future research directions.


1 INTRODUCTION

With the Internet's ongoing expansion in recent years, huge data has emerged in a variety of industries. Artificial intelligence has experienced swift growth due to the expansion of data availability and the relentless advancements in hardware computational capacity. The core area of artificial intelligence is machine learning, which is divided into supervised learning and unsupervised learning according to supervision. The former method requires a large amount of labeled data during its learning process, which is, however, difficult to get: automatically collected data are usually messy, and manually labeling data is very time-consuming. Generative Adversarial Networks (GANs) are originally proposed to solve the problem, which is able to produce samples that almost match the distribution of actual data.

Since the proposal of GAN by Goodfellow in 2014 (Goodfellow, 2014), a lot of research has been conducted, yielding remarkable success. The field of generative modeling has witnessed significant growth and diversification. The Generative Multi-

Adversarial Network (GMAN) was proposed by Durugkar et al. in 2016 (Durugkar, 2016), which expanded on the original GAN concept by introducing multiple discriminators. This multi-discriminator approach allowed for a more robust training process, as each discriminator provides a different perspective on the generated data, leading to more stable convergence and less susceptibility to modes collapse.

Following this advancement, Arjovsky et al. introduced the Wasserstein GAN (WGAN) in 2017 (Arjovsky, 2017). This model further enhances training stability by utilizing the Wasserstein distance as a loss function, resolving several issues related to the conventional GAN loss. Another notable development is the introduction of conditional GAN (CGAN) (Mirza, 2014), which allows the generation of data conditioned on additional information, such as class labels. This has created new opportunities for regulated data production, with text-to-image synthesis and image-to-image translation among the possible uses. Furthermore, the exploration of latent space through models like Variational Autoencoders (VAEs) has been combined with GANs to create

^a <https://orcid.org/0009-0001-5076-3810>

hybrid models such as the VAE-GAN (Larsen, 2016). The purpose of these hybrids is to combine the advantages of both architectures: the strong generative capabilities of GANs and the organized latent space of VAEs.

The objective of this study is to offer a detailed overview of image generation based on GAN. This paper focuses on introducing the fundamentals of the generator and the discriminator, providing a comprehensive overview of the basic ideas and historical background of GANs. After providing this basic overview, the paper analyzes complex GAN design variations and explains the theoretical underpinnings of these variations and improvements. Based on the framework provided by standard GANs, the paper compares the empirical performance of these improved networks and analyzes their respective effectiveness in detail. The paper additionally provides a comparative assessment of generative artificial neural network paradigms, emphasizing unique advantages and possible drawbacks. The article's narrative gradually shifts into a survey of predictions, exploring potential paths and future developments for generative artificial neural networks and their offshoots. In summary, this article not only explains the operating principles of GANs and their variant frameworks, but also compares their performance criteria. The paper concludes with a field study of their future development paths.

This chapter gives an introduction. Chapter 2 analyzes the core concepts and principles of GAN, as well as the principles of several variants of the network. Then in chapter 3 the experimental performance of several networks is compared side by side, analyzed and discussed, and finally summarized in chapter 4.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

Due to the high generalizability of GAN models across multiple domains, the number of datasets to which they are applicable is vast. In the field of image generation alone, GAN and its variants can present excellent results on dozens of datasets. This paper will only introduce the more commonly used datasets. The simplest and most commonly used dataset is MNIST, a large handwritten digit recognition dataset created by the National Institute of Standards and Technology (Goodfellow, 2014). This dataset is

designed to provide researchers with a benchmark test set for evaluating various handwritten digit recognition algorithms. Ten thousand handwritten digits in the test set and sixty thousand in the training set make up the MNIST dataset. Each image is a 28x28 pixel grayscale image with a corresponding label, i.e., the corresponding real number (0-9).

CIFAR-10 is also a commonly used dataset. It was created by the CIFAR project initiated by the Canadian Institute for Advanced Research (CIFAR) (Durugkar, 2016). This dataset, designed for object recognition, comprises internet-sourced images divided into 10 distinct categories, each corresponding to specific objects like airplanes, cars, birds, cats, etc. A total of 60,000 32x32 pixel RGB three-channel images make up the dataset; 50,000 were utilized for training and 10,000 for testing. CIFAR-10 is more complex than the MNIST dataset because it contains RGB color information and more detail, which makes the model need to learn from more complex data.

2.2 Proposed Approach

This paper centers on GAN and its variants, discussing its background, technical principles, and development directions. As shown in Fig. 1 this section describes the relevant concepts and background of GAN in detail and discusses its performance in the field of image generation. In this chapter, the basic idea of GAN is introduced first. Then, the improvement ideas and network structures of CGAN, DCGAN, and BigGAN are discussed. In the following chapter, the experimental results of each model on various tasks and datasets are displayed, which also examines and evaluates each model's features, benefits, and drawbacks based on structural analysis. Among them, the last introduced BigGAN shows great advantages in terms of accuracy of image generation due to the large scale of its model and the use of multiple techniques. Finally, in the conclusion section, the authors analyze the improvement directions of GAN based on the experimental results and discuss its development prospects.

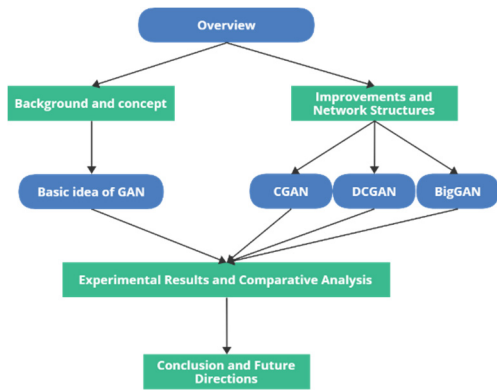


Figure 1: Flowchart of the paper (Picture credit: Original).

2.2.1 Generative Adversarial Network(GAN)

GAN, an iconic model in the realm of unsupervised machine learning, was introduced by Goodfellow and colleagues in 2014. The core idea of GAN is derived from game theory, and its basic idea is to generate data through two neural networks playing with each other, which are called the generator and the discriminator. In this process, the generator's goal is to generate an output that closely resembles the distribution of real data by taking in a random noise (also known as a latent space vector). In an image generation scenario, the generator usually tries to create images that look like real photographs. Real data or data produced by the generator are fed into the discriminator, and the result is a scalar that indicates the likelihood that the input data is real. That is, the discriminator tries to distinguish whether the input is a real distribution from the training set or a fake distribution created by the generator.

Mathematically, the above idea can be converted into a simple 'two-player minimax game', i.e., the equation $V(G,D)$. The paper define p_{data} and p_g to be the real distribution and generator's distribution over data x , respectively, and p_z to be the input random noise distribution. $G(z)$ represents the output of the generator after accepting the noise, while $D(x)$ represents the probability that x comes from the real data instead of p_g . So $V(G,D)$ can be written as (Goodfellow, 2014):

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (1)$$

In the training process, the discriminator aims to maximize $D(x)$ while minimizing $D(G(z))$, and the generator aims to maximize $D(G(z))$. Thus based on the value function, the generator and discriminator can be trained using gradient descent.

2.2.2 Conditional Generative Adversarial Network (CGAN)

Although GAN has good performance, the output is often uncontrollable due to its inputs being random noise. To enhance the controllability of the model, Mirza and Osindero proposed CGAN. By supplementing the inputs with additional conditional information, CGAN makes the generative process controllable, thus generating data with specific characteristics. For example, CGAN can generate images of handwritten digits, on-demand digits, which cannot be done with GAN. CGAN's network architecture is identical to that of GAN, with the difference that CGAN's inputs include not only data but also conditional variables. These conditions can be category labels, partial data information, or any other form of auxiliary information, most commonly category labels. This condition information is fed into both the generator and the discriminator to guide the data generation process so that the generated data is not only realistic but also satisfies specific condition constraints. This does not, however, imply that CGAN becomes a supervised training because the generator does not treat conditional constraints, like category labels, as "standard answers"—that is, there is no direct correlation between the data produced and the category labels. CGAN only finds out the commonality under a certain label category and outputs an image that satisfies the commonality.

Mathematically, it can be assumed that the generation and discrimination are done under the condition of knowing the real label, so the properties of $D(x)$ and $G(z)$ change from ordinary probability to conditional probability $D(x|y)$, $G(z|y)$. $v(D,G)$ thus changes to (Mirza, 2014):

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (2)$$

2.2.3 Deep Convolutional Generative Adversarial Network (DCGAN)

CNN creatively proposes a method for processing high-dimensional data, which has had a profound impact on the field of deep learning. DCGAN is the model combining CNN and GAN, which was proposed by Radford, A. et al. in 2016 (Radford, 2015). DCGAN inherits CNN's advantages in data feature extraction and drastically improves the level of detail of the pictures produced and the training stability. Compared with GAN, its network structure has the following characteristics. First, every pooling layer is removed, and in the generative network, the transposed convolution is utilized for upsampling,

while in the discriminative network, stride convolution is employed in place of pooling. Since the stride convolution and the transposed convolution both involve learnable parameters, the model's flexibility and learning ability are greatly improved, and the decision boundary is smoother. Second, the majority of DCGAN's network layers employ the Batch Normalization technique to address the training issue brought on by inadequate initialization. Furthermore, Batch Normalization lessens overfitting and eases the issue of Internal Covariate Shift. Thirdly, The discriminator uses LeakyReLU as its activation function, and the generator's last layer uses the tanh activation function, thus DCGAN avoids the dead ReLU problem and maintains the gradient flow.

2.2.4 Big Generative Adversarial Network (BigGAN)

As computational power continues to improve, GAN saw a breakthrough in 2018. DeepMind proposed the BigGAN model (Brock, 2018), which significantly improves the performance of image generation tasks by using larger model sizes, larger batch sizes, and a series of improved training techniques. The BigGAN model, as its name implies, improves the generator and discriminator's gradients and significantly boosts performance by increasing the batch size from 256 in SAGAN to 2048. Moreover, to match the increase of batch, BigGAN also increases the number of channels in every network tier, to expand the model capacity. In addition, BigGAN improves on the embedding of the prior distribution z , not only as the initial layer input to the generator but also transmitted to multiple layers separately, which improves the training speed by 18%. Not only that, BigGAN sets a threshold on the sampling process of the prior distribution z as a way to truncate the sampling range. As the threshold continues to drop, the sampling range becomes narrower and the model output becomes more accurate, although diversity is also negatively affected.

3 RESULT AND DISCUSSION

The authors have investigated the experimental results and performance of the above-mentioned GAN, CGAN, DCGAN, and BigGAN are shown below. This section begins with a discussion of the images produced on MNIST by GAN, CGAN, and DCGAN. Secondly, the accuracy of these three models on the same task is qualitatively analyzed. Then, the experimental results of BigGAN are

discussed, demonstrating far superior performance to the first three, along with equally massive computational resources and model parameters. In conclusion, the paper summarizes the comparison of the experimental performance of GAN, CGAN, DCGAN, and BigGAN, analyzes their feature based on the principles, and discusses the degree of adaptation of each model to different tasks. Based on the above discussion, the future research directions of GAN are analyzed, and several potential application areas of GAN are listed at the end of this chapter.

3.1 Qualitative Analysis on the MNIST Dataset

GAN and its variants are usually used for low-resolution dataset tasks and show good performance. The experimental outcomes of GAN, CGAN, and DCGAN are qualitatively analyzed in this section, based on MNIST (Cheng, 2020). The handwritten digit images produced by the GAN, CGAN, and DCGAN when the epoch is set to 10000 are shown separately in Figure. 2, Figure 3, and Figure 4 (Cheng, 2020). It is observed that DCGAN shows better performance than the other two individual models and produces the most recognizable pictures. There is a lot of noise in the output of GAN and CGAN compared to the clear pictures generated by DCGAN. For CGAN, since the input contains conditional variables, it can generate images based on label orientation. As shown in Figure. 3, the conditional variable for CGAN is the label y . The simplest model, GAN, produces the noisiest and most unreadable images. In summary, DCGAN exhibits the greatest performance on the MNIST dataset after 10000 epochs.



Figure 2: Digits generated by GAN (Cheng, 2020).

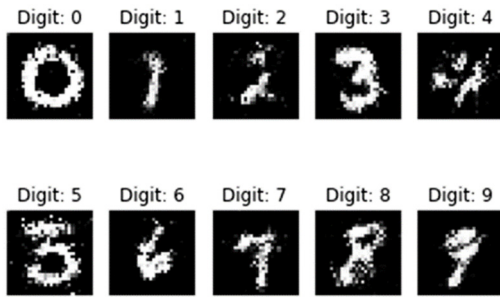


Figure 3: Digits generated by CGAN (Cheng, 2020).

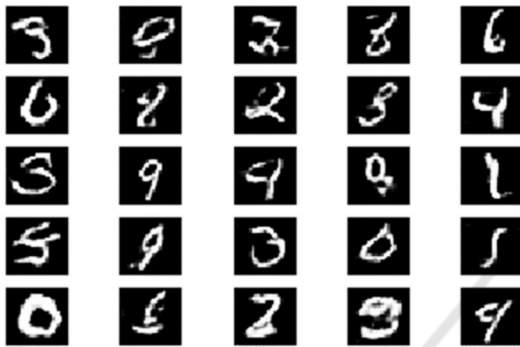


Figure 4: Digits generated by DCGAN (Cheng, 2020).

3.2 Quantitative Analysis on the MNIST Dataset

Based on qualitative analysis, quantitative analysis of the MNIST dataset is carried out in this paper. To ensure the accuracy of the comparison, the experiments set the model hyperparameters uniformly. The experiments use the Adam optimizer and the MNIST dataset to train the model for 10,000 epochs with 16 samples per batch at a learning rate of 0.0001, and set the model performance to be evaluated every 200 epochs (Cheng, 2020). The experimental results of GAN, CGAN, and DCGAN are displayed in Table 1, where accuracy is defined as the degree of similarity between the generator's output and the original images. This is in line with the results of the qualitative analysis, and DCGAN did produce clearer figures. Also, DCGAN took the shortest time, 7 minutes, indicating that it has higher efficiency. CGAN had the lowest accuracy, 55.02%, and the longest training practice, 15 minutes.

The experiment result illustrates that there is no statistically significant correlation between accuracy and loss in the discriminator and generator. Compared with other deep networks, GAN networks have special characteristics. in GAN, the discriminator and generator compete with each other, so their losses often conflict with each other, one

decreases while the other increases. This may be the main factor leading to the instability of the loss.

Table 1: Numerical performance metrics for GAN, CGAN, DCGAN.

Model	GAN	CGAN	DCGAN
Accuracy	65.62	55.02	68.12
Discriminator loss	0.65	0.67	0.57
Generator loss	0.98	0.85	1.07
Calculation time	~7 minutes	15 minutes	7 minutes
Loss function	Binary Cross Entropy		

3.3 Evaluating the Performance Results for BigGAN

BigGAN is an advanced GAN that aims to produce high-resolution, high-quality images. BigGAN is not discussed in the first two sections above due to its extremely complicated model structure and enormous parameter count. Also, the use of BigGAN on MNIST is prone to serious overfitting. So, qualitatively, this section shows the images generated by BigGAN on the ImageNet dataset (Brock, 2018); quantitatively, it shows the performance of BigGAN on the Cifar-10 dataset under the quantitative metric Inception score using Table 2 (Yinka, 2020). To compare with the above three models, additionally displayed is the DCGAN Inception score from Cifar-10.

According to Figure 5, it can be observed that the BigGAN model generates high-resolution images that are far clearer than the MNIST handwritten digit images produced by the three models, such as GAN, CGAN, and DCGAN. This is further supported by the quantitative study, where BigGAN receives an Inception score of 9.22, far higher than DCGAN's 6.58.



Figure 5: Class-conditional samples generated by BigGAN (Brock, 2018).

Table 2: Inception score of BigGAN and DCGAN (Yinka, 2020).

Dataset	Model	Inception score(IS)
CIFAR-10	BigGAN	9.22
CIFAR-10	DCGAN	6.58

In this section experimental results of GAN, CGAN, DCGAN, and BigGAN are analyzed

qualitatively and quantitatively. Despite being one of the most widely used neural network structures, GAN still faces several difficulties. For instance, the training process of GAN is very unstable, and maintaining the balance between the generator and the discriminator can be difficult, leading to the problem of non-convergence. In addition, the conflict between generators and discriminators also causes the problem of mode collapse, which significantly reduces the diversity of generated images. WGAN provides a solution to this issue, but it still performs poorly on high-resolution datasets.

The future of GAN is still promising despite the numerous challenges that still need to be overcome. BigGAN, which appeared in recent years, has made great breakthroughs in high-quality image generation compared with early GAN, CGAN, and DCGAN. In addition, GAN has been extensively applied in different fields. For example, NVIDIA uses GAN to convert graffiti into highly realistic landscapes or scenes (Park, 2019); AC Duarte et al. developed Wav2Pix to produce high-precision photographs of speakers' faces from voice sounds (Duarte, 2019). The potential of GAN is still far from being fully developed.

4 CONCLUSIONS

GAN and its variants are one of the most popular and promising generative models for applying game concepts to generative problems. This study provides a detailed introduction to the history and basic concepts of GAN. The article then reviews the basic principles of GAN and its three variants - GAN, DCGAN and BigGAN. Based on their principles, the article then discusses the properties, advantages, and disadvantages of model structure generation. After that, the article analyzes and compares the four models using qualitative and quantitative analysis methods based on MNIST and CIFAR-10 datasets. According to the experimental results, the performance of the four models, in descending order, is BigGAN, DCGAN, CGAN, and GAN, with BigGAN showing much higher performance than the remaining three models in both qualitative and quantitative experiments. BigGAN performs significantly better than the other three models in both qualitative and quantitative tests. It is worth noting that CGAN produces targeted results despite its poor performance. In the future, the limitations of GAN such as training stability and so on will be considered as the research objective for the next stage. The research will focus on providing feasible solutions to

the above problems. In addition, the latest GAN variants have many breakthroughs in the form of data, and model performance. The potential of GAN is far from being fully explored.

REFERENCES

- 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.
- Durugkar, I., Gemp, I., & Mahadevan, S. (2016). Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214-223). PMLR.
- Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016, June). Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning* (pp. 1558-1566). PMLR.
- Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.
- Cheng, K., Tahir, R., Eric, L. K., & Li, M. (2020). An analysis of generative adversarial networks and variants for image synthesis on MNIST dataset. *Multimedia Tools and Applications*, 79, 13725-13752.
- Yinka-Banjo, C., & Ugot, O. A. (2020). A review of generative adversarial networks and its application in cybersecurity. *Artificial Intelligence Review*, 53, 1721-1736.
- Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2337-2346).
- Duarte, A. C., Roldan, F., Tubau, M., Escur, J., Pascual, S., Salvador, A., ... & Giro-i-Nieto, X. (2019, May). WAV2PIX: Speech-conditioned Face Generation using Generative Adversarial Networks. In *ICASSP (Vol. 2019, pp. 8633-8637)*.