Research and Application Analysis of Correlative Optimization Algorithms for GAN

Tianmeng Wang*

Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, 98105, USA

*Corresponding author: tw78@uw.edu

Abstract. Generative Adversarial Networks (GANs) have been one of the most successful deep learning architectures in recent years, providing a powerful way to model high-dimensional data such as images, audio, and text data. GANs use two neural networks, generator and discriminator, to generate samples that resemble real data. The generator tries to create realistic looking samples while the discriminator tries to differentiate the generated samples from real ones. Through this adversarial training process, the generator learns to produce high-quality samples indistinguishable from the real ones.Different optimization algorithms have been utilized in GAN research, including different types of loss functions and regularization techniques, to improve the performance of GANs. Some of the most significant recent developments in GANs include M-DCGAN, which stands for multi-scale deep convolutional generative adversarial network, designed for image dataset augmentation; StackGAN, which is a text-to-image generation technique designed to produce high-resolution images with fine details and BigGAN, a scaled-up version of GAN that has shown improved performance in generating high-fidelity images.Moreover, the potential applications of GANs are vast and cross-disciplinary. They have been applied in various fields such as image and video synthesis, data augmentation, image translation, and style transfer. GANs also show promise in extending their use to healthcare, finance, and creative art fields. Despite their significant advancements and promising applications, GANs face several challenges such as mode collapse, vanishing gradients, and instability, which need to be addressed to achieve better performance and broader applicability.In conclusion, this review gives insights into the current state-of-the-art in GAN research, discussing its core ideas, structure, optimization techniques, applications, and challenges faced. This knowledge aims to help researchers and practitioners alike to understand the current GAN models' strengths and weaknesses and guide future GAN developments. As GANs continue to evolve, they have the potential to transform the way we understand and generate complex datasets across various fields.

Keywords: GAN, DCGAN, StackGAN, Image dataset augmentation, Text-to-image generation.

1. Introduction

As the development of artificial neural networks continues to progress, deep learning, an important field of artificial intelligence, is also constantly evolving [1]. Since its introduction in 2014, Generative Adversarial Networks (GANs) have become a notable focus in the machine learning domain [2]. Utilizing GANs has shown to be a successful approach in generating realistic synthetic data such as images, videos, and audio [3].

GANs are comprised of training neural networks that function like an adversarial game, drawing from the two-person zero-sum game theory concept [4]. The model is trained by having two neural networks competing with each other and reaches the Nash equilibrium through continuous optimization iterations [5]. GANs feature a value function where one agent aims to maximize it while the other agent seeks to minimize it, based on a minimax game rather than an optimization problem. GANs can generate novel data, like images or videos, that are similar to a given dataset. Consisting of two neural networks, a generator G and a discriminator D, GANs are trained together where the generator creates new data samples by mapping random noise to a data distribution, and the discriminator distinguishes between real and fake data [6]. A good GAN must have efficient training methods; otherwise, the output may not meet the desired standard due to the model's excessive

freedom. The training process continues until the generator can create indistinguishable data, as concluded by the discriminator. GANs have been applied in a diverse range of fields, including image synthesis, style transfer, data augmentation, and more..

2. Basic Principles of GAN

The core idea of GAN comes from the Nash equilibrium of game theory, assuming that the two agents participating in the game are a generator and a discriminator. The purpose of the generator is to learn the real data distribution as much as possible, while the purpose of the discriminator is to try to correctly distinguish whether the input data comes from real data or from the generator. In order to win the game, the two gamers need to continuously optimize and improve their generating ability and discriminating ability respectively. This learning optimization process is to find a Nash equilibrium between the generator and the discriminator [7]. D has the input of either true data distribution or false data distribution that can either come from other datasets or from the generator and has the output of either true (denoted by 1) or false (denoted by 0) to indicate the classification of the input. In order to generate new samples that are similar to the training data but are not exact copies of any particular sample, G has the input of some random noise vector, or latent vector. The use of random noise ensures that the generator can produce a wide range of diverse samples, rather than just memorizing and reproducing the training data [8]. In practice, the noise input is often a random gaussian distribution as it provides a convenient way to sample random noise that is easy to work with mathematically. Using a Gaussian distribution as the input, the generator may sample the noise in a form that is simple to understand and manage. Another benefit of employing a Gaussian distribution is that it gives the generator access to a comparatively smooth and continuous input space. In particular, for picture and video data, this might be crucial as it guarantees that the created data is also smooth and continuous. The output of the generator is a generated sample that is similar to the real data. The specific form of the output is dependent on the type of data being generated. Formally, the GAN objective, in its original form is

$$\min_D \max_G E_{x\sim q_{data}(x)}\ [log\ D(x)] + E_{z\sim p(z)}\left[log\left(1 - D\big(G(z)\big)\right)\right] \quad (1)$$

Where z is a latent variable drawn from distribution $p(z)$ such as $N(0, I)$.

Then, the purpose of training is to make D(z) be false and D(x) be true and make G produce data that are most similar to the real data distribution. Throughout each training iteration, the generator creates a collection of samples for the discriminator to categorize as either real or fake [9]. The discriminator modifies its settings in response to the ensuing mistake to increase its capacity to discriminate between authentic and fraudulent input. The mistake propagates backward through the generator as well, changing its settings to produce samples that are more difficult to differentiate from the true data [10]. After the generator can generate samples that are challenging for the discriminator to differentiate from real data, the training phase is complete. At this time, the generator has gained the ability to create samples that accurately reflect the statistical characteristics of the actual data, given that the discriminator can correctly differentiate most of the fake data distribution from the real data distribution.

3. Optimize and Improve the Algorithm

3.1. DCGAN in the Image Dataset Augmentation

Deep Convolutional Generative Adversarial Networks (DCGAN) is a type of GAN architecture that uses convolutional neural networks (CNNs) in the generator and discriminator. Since first introduced in 2015, DCGAN has become a popular and effective approach for synthesizing high-quality images. The DCGAN architecture was created to solve various issues, including instability during training and problems producing high-quality pictures, with prior GAN models. DCGAN is

better able to produce realistic and detailed pictures since it uses CNNs to extract spatial information from the input images. To create the final picture in the generator network, DCGAN commonly upsamples the input noise vector using transposed convolutional layers, often referred to as deconvolutional layers. The generator can produce high-resolution pictures from low-dimensional input vectors using these layers, which are effectively the opposite of convolutional layers. Convolutional layers are also used by DCGAN's discriminator network to extract characteristics from the input picture and categorize it as genuine or fraudulent. A single scalar number generally expressing the likelihood that the input image is real is the discriminator's output. The usage of normalization layers, such as batch normalization, which aids in stabilizing the training process and enhancing the quality of the produced pictures, is one of the fundamental components of the DCGAN design. DCGAN also frequently uses ReLU activation in the generator network and LeakyReLU activation in the discriminator network, which have been demonstrated to be effective in GAN designs. Maintaining GAN's capacity to produce outstanding data, DCGAN enhances the ability by including the benefits of CNN feature extraction, improving its capacity for image processing and analysis. Through training on actual large-scale datasets as CelebA, LSUN, and Google Image Net, DCGAN has produced good results. There are many near-duplicate pictures within the visual data, which seriously wastes the networks' scarce storage, computing, and transmission resources and degrades the experience of identification. Hence, the use of a DCGAN network structure is needed in the sample generation to eliminate the duplicates generated by normal GANs.

Data augmentation is a technique that involves utilizing various transformations on existing data to generate new training data, which can reduce overfitting and enhance the generalization capability of a model. Prior to the introduction of DCGAN, data augmentation in the deep learning industry was achieved by using either traditional augmentation methods or ordinary GANs, which were found to produce limited diversity in new data samples. Traditional augmentation methods include translation, flipping, rotation, and adding noise to original images to create new samples. However, such methods do not necessarily increase the diversity of the training data. In addition, when multiple augmentation methods are used simultaneously, it is possible for some methods to damage the semantic information of the targets, which could negatively impact the model and even lead to overfitting when data sets are small.

The main advantage of using a GAN for data augmentation is that the generated samples are based on the statistical properties of the original data rather than random noise, and are therefore more representative of the true data distribution, which can improve the performance of the model. DCGAN, as a specific type of GAN designed for image generation tasks, uses convolutional neural networks (CNNs) to extract spatial information from input images instead of fully connected layers used by conventional GANs in the generator and discriminator networks. Compared to ordinary GANs, DCGAN provides several benefits for enhancing picture datasets. One of the key advantages of DCGAN is its ability to produce high-quality and realistic images with more detailed textures and structures, owing to its capability of extracting spatial information from input images [11]. Another benefit of DCGAN is its ability to produce images of various dimensions and sizes since its generator can upsample low-dimensional noise vectors and produce high-resolution images using transposed convolutional layers. DCGAN also employs normalization layers such as batch normalization to stabilize the training process and enhance the quality of output images, which is critical for data augmentation since it assures that the generated samples are accurate representations of the actual data distribution.

3.2. StackGAN in Text-to-image Generation

Synthesizing high-resolution images from text descriptions may be used in several practical ways. Computer vision, where high-resolution images can be used to create realistic virtual environments for video games, movies, and simulations, is one of the instances where the images are most practically used. Designers and developers may also simply produce high-quality visual information by synthesizing pictures from text descriptions without the need for costly and time-consuming

manual design and rendering. Text-based synthesized high-resolution photographs may also be utilized in the world of online shopping to present goods to buyers in a more enticing manner. Retailers can rapidly and effectively produce a large number of product pictures by creating images from text descriptions, which may assist to enhance the consumer experience and boost sales. In the realms of art and design, the text text-based synthesized images may be applied to provide fresh and inventive visual material, and better resolution of the synthesized images often means a better understanding of the material. Artists and designers can swiftly experiment with new ideas and produce distinctive visual materials that would be either challenging or impossible to draw or produce manually.

StackGAN is a type of GAN architecture that is specifically designed for synthesizing high-resolution images from text descriptions. It was first introduced in a research paper published by Han Zhang et al. in 2017. In the first stage, StackGAN uses a conditional GAN to generate a 64x64 resolution image based on a given text description. This low-resolution image is then fed into the second stage, which uses a conditional GAN with an attention mechanism to generate a high-resolution image of size 256x256. Stage-I GAN produces low-resolution pictures that frequently lack vivid object details and may have form distortions. The first stage could also leave out certain elements from the text which is essential for creating photo-realistic graphics. In order to produce high-resolution pictures, Stage-II GAN is based upon the Stage-I GAN's result, depending on low-resolution photos and text embedding. In the second stage, the StackGAN completes the details of the object by reading the written description again, fixing flaws in the low-resolution picture from Stage-I, and creating a high-quality photo-realistic image. Based on the written description, the attention mechanism in the second stage enables StackGAN to concentrate on producing details in particular areas of the picture. For instance, if a bird with black feathers and a yellow beak is described in the text, the attention mechanism can concentrate on creating the bird's beak and feathers with the proper hues and textures [7]. Without the attention mechanism, the generator can produce the image paying less attention to the exact characteristics specified in the text description, producing a less specific and more generalized image. Without the attention mechanism, if the text description, for instance, states a person wearing a red shirt and blue slacks, the generator could produce a picture of a person wearing the proper colors of clothing, but with less attention to the finer features and textures of the clothing. The final image might not be as realistic or true to the provided text description.

3.3. BigGAN, as a Big GAN

GAN architectures have been known to be struggling to generate high-quality images with consistent visual quality and diversity, particularly for large datasets like ImageNet. As a result, researchers have been working on improving GAN architectures to generate high-quality images with more consistency and diversity. This has led to the development of several advanced GAN architectures, such as BigGAN, which address these issues by using techniques like class-conditional generation.

BigGAN is a type of GAN architecture, introduced by Andrew Brock et al. in 2018, that can scale up the size and complexity of GANs in order to generate high-quality, diverse images across a wide range of classes. Using a large-scale distributed infrastructure to train GANs with substantially more parameters than prior methods is one of BigGAN's primary contributions. This enables BigGAN to produce pictures with a resolution that is substantially greater than the previous state-of-the-art GAN designs, 128x128 or 256x256 pixels. Another important contribution of BigGAN is the implementation of a new class-conditional generator architecture, which enables the production of pictures with excellent visual quality and variety for a huge number of classes. BigGAN employs a truncated noise distribution throughout the generation process, which lessens mode collapse and increases the diversity of the pictures that are produced.

Andrew Brock et al. found out that simply raising the batch size for the baseline model would enormously benefit the model: increasing the batch size by a factor of 8 alone leads to a 46% improvement in the state-of-the-art IS. The authors then increase the breadth, or the number of

channels, of each layer by 50%, which result in almost doubling the number of parameters in both models. This results in an additional 21% improvement in IS, which is attributed to the model's enhanced capability in relation to the dataset's complexity.

As the result, BigGAN significantly outperformed the state of the art and set a new bar for performance among ImageNet GAN models. Moreover, the authors have examined the training behavior of large-scale GANs, described their stability in terms of the single values of their weights, and spoken about how performance and stability interact. BigGAN demonstrated impressive results in generating high-quality images across a wide range of classes, which has opened up new possibilities for computer graphics, e-commerce, art, and more.

4. Application Field

4.1. Medical Field

In deep learning industry, the standard augmentation technique is used for data augmentation, but it cannot generate very diverse new data samples. In medical context, standard data augmentation method could generate unrealistic pictures because they do not account for the biological variance of medical imaging data and instead create new examples of data by varying lighting, field of view, and spatial rigid transformations. GAN used by the author is DCGAN (deep convolutional GAN), with real X-ray images of the chest, and they use the trained generator to produce realistic images. Images were obtained from the prostrate, lung, colorectal, and ovarian dataset of the National Institute of Health (NIH PLCO). In specific, the authors used chest X-rays that were both normal and that had cardiovascular illness, and any patient who displayed visual indicators of cardiomegaly, congestive heart failure, or a cardiac anomaly in general was given one label as the study's identifier. Two GANs were trained concurrently, one using only the set of chest images of healthy patient to produce normal images and the other one using images identified as having cardiovascular disease to generate abnormal image set. After 500 epochs of iterations, the authors make the two GANS achieve the full convergence. Then, after the training was completed, the authors used the two generators to generate images by feeding it random normal distribution. As the result, by applying the test for cardiovascular abnormality detection on the three different training scenarios, namely no augmentation, traditional augmentation, and GAN augmentation, the authors concluded that the test accuracy of GAN augmentation outperformed the rest two methods. The author stated that the test accuracy for GAN data augmentation is 84.19%. Even though the improvement is not substantial compared to the test accuracy of 81.93% for no augmentation and 83.12% for traditional augmentation scenarios, it is still a promising result: it showed that using GAN to generate medical images is possible, and the produced images can not only enlarge the medical imaging dataset available for analysis, which is already a major advantage and the motivation of this research, but also have the potential to improve the accuracy of medical image analysis, which is an even more significant benefit.

4.2. Text Generation

Text data, having particular characteristics like high degree of complexity, variability, and ambiguity, is difficult to generate using conventional machine learning techniques like RNN. For example, unlike learning from the text data, which consistently uses ground-truth words, RNN generates words in sequence from previously generated words. As a result, errors grow in number proportional to the length of the sequence, and the quality of the phrase quickly degrades after the first few words. Text data can also have the property of long-term dependencies, which means that the context set by a few words or in a few sentences earlier in the text may influence the interpretation of a word or phrase way behind the text.

GANs have demonstrated promising results in producing text data that closely resembles human-written text primarily because after training on big sets of text datasets, GANs are able to recognize the underlying pattern and relationship within the texts. Due to the nature of adversarial training, the discriminator analyzes both fake and real sentences as opposed to just the words in real sentences,

which should, in theory, lessen the exposure-bias problem. A well-trained discriminator, for instance, can summarize the pattern in coherence, grammar, and overall similarity of the real data and differentiate the fake, generated texts from them.

Zhang et al. (2017) proposed TextGAN to address the problems of generating realistic texts using GAN. Their method makes use of a convolutional network as a discriminator and a long short-term memory network as a generator, offering an approach that compares the high-dimensional latent feature distributions of real and synthetic phrases using a kernelized discrepancy metric, departing from the traditional GAN objective [1].

Even TextGAN has presented a potential method for producing realistic texts using GAN, there are certain shortcomings and restrictions that could be solved in future studies. For instance, TextGAN could have trouble producing lengthy, complicated phrases with several clauses or dependencies. This is because the latent feature space is represented by a fixed-length vector in the generator, which might not be able to capture the nuance and complexity of longer phrases. To get around this restriction, future study might look into using more sophisticated methods of encoding the latent space, including hierarchical or variational approaches. Furthermore, due to the lack of variance in the training data, TextGAN also has the potential to produce output that is repetitive or unoriginal. The generator model can overfit to the training data with GANs, which results in outputs that are very similar to the original samples. Future studies could take a look into methods like reinforcement learning or using outside knowledge sources for enhancing the diversity and uniqueness of the generated content.

4.3. Music Production

Traditional music creation methods have been known to be time-consuming, resource-intensive, and have a relatively long production cycle. Conventional techniques are not always able to satisfy the demands of music creation. This is a result of the scarcity of talents and rising costs of the creative process. Deep learning, on the other hand, has the potential to streamline and improve the creative process in music production. Deep learning models are able to discover patterns in vast volumes of music data, learn from them, and create new content on their own. This can open up new channels for musical expression and significantly cut down on the time and effort needed to create music.

GANs have been employed for a variety of music-generation tasks, such as composing original works in a particular aesthetic, producing accompaniments, and coming up with new melodic arrangements. One benefit of using GANs in music production is their capacity to capture the high-dimensional temporal structure of music, enabling the development of music that progresses naturally and more closely resembles music made by humans. GANs have seen a number of recent successes in the music industry, including music transcription, genre classification, and style transfer. GANs has also been employed to produce new music, with the capacity to produce works that resemble those of real musicians. Even though GANs have the potential to be used in the creation of music, there are still several issues that need to be resolved. To overcome these difficulties and raise the caliber of the created music, researchers are still working to design and improve GAN models.

5. Conclusion

GAN has been extensively utilized in various fields and has proven to be a powerful tool for generating high-quality and realistic synthetic data, as well as improving data augmentation techniques. The previous studies reviewed in this paper have all made significant strides in the design and enhancement of GAN algorithms, as well as showcasing practical applications across different industries. These articles have advanced the state-of-the-art in GAN research by introducing new techniques and potential future directions for the field.

References

[1] Zhang J, Li S, Lin N, et al. Spatial identification and trade-off analysis of land use functions improve spatial zoning management in rapid urbanized areas, China[J]. Land Use Policy, 2022, 116: 106058.

[2] Wen C, Yang J, Gan L, et al. Big data driven Internet of Things for credit evaluation and early warning in finance[J]. Future Generation Computer Systems, 2021, 124: 295-307.

[3] Tian Z, Gan W, Zou X, et al. Performance prediction of a cryogenic organic Rankine cycle based on back propagation neural network optimized by genetic algorithm[J]. Energy, 2022, 254: 124027.

[4] Gan Y, Meng B, Chen Y, et al. An intelligent measurement method of the resonant frequency of ultrasonic scalpel transducers based on PSO-BP neural network[J]. Measurement, 2022, 190: 110680.

[5] Gan X, Pavesi G, Pei J, et al. Parametric investigation and energy efficiency optimization of the curved inlet pipe with induced vane of an inline pump[J]. Energy, 2022, 240: 122824.

[6] Chai X, Tian Y, Gan Z, et al. A robust compressed sensing image encryption algorithm based on GAN and CNN[J]. Journal of Modern Optics, 2022, 69(2): 103-120.

[7] Gan X, Pei J, Wang W, et al. Application of a modified MOPSO algorithm and multi-layer artificial neural network in centrifugal pump optimization[J]. Engineering Optimization, 2022: 1-19.

[8] Chai X, Fu J, Gan Z, et al. An image encryption scheme based on multi-objective optimization and block compressed sensing[J]. Nonlinear Dynamics, 2022, 108(3): 2671-2704.

[9] Gan V J L, Lo I M C, Ma J, et al. Simulation optimisation towards energy efficient green buildings: Current status and future trends[J]. Journal of Cleaner Production, 2020, 254: 120012.

[10] Gan C, Cao W H, Liu K Z, et al. A novel dynamic model for the online prediction of rate of penetration and its industrial application to a drilling process[J]. Journal of Process Control, 2022, 109: 83-92.

[11] Liu Y, Wu J, Qu L, et al. Self-supervised correlation learning for cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2022.