Research on image generation technology based on deep learning

Jinchen Li
School of Information Science and Technology, Yunnan Normal University, Yunnan, China

Abstract. In the realm of image creation, deep learning stands out as an effective and valuable machine learning technique. Deep learning can automatically learn the intrinsic features of images， reaching the goal of generating high-quality images by utilizing multi-layer neural network models. In recent years, deep learning-based image generation technology has made significant progress. This paper mainly introduces two main methods: generating adversarial network (GAN) and variational autoencoder (VAE). GAN has been widely used in image generation, image repair and other aspects. VAE has a good performance in image generation, image classification and so on. However, current image generation technologies still face problems such as diversity and insufficient authenticity. Based on the above problems, this paper analyzes the methods of improving and optimizing the mainstream image generation algorithm from the perspectives of improving and optimizing the loss function, improving the space modeling, revising the structure of both the generator and discriminator, while speeding up the training process. Furthermore, the performance of these methods in image generation tasks is compared, and the strengths and weaknesses of each approach are evaluated. Image generation has emerged as a prominent research area in contemporary academia, with a high possibility of exploration and practice.

1 Introduction

With the rise of big data, remarkable advancements have been achieved in the realm of artificial intelligence. Deep learning technology, serving as a key driving force, has played a significant role in accelerating the development of image generation technology [1]. Deep learning technology has demonstrated impressive outcomes across various domains, including computer vision, speech recognition, and natural language processing [2]. In the area of image creation, deep learning technology not only improves the quality of generated images, but also expands the range of potential applications for image generation [3]. The technology of image generation based on deep learning holds significant academic importance and presents promising prospects for practical applications. Therefore, it warrants thorough research and exploration in the scholarly community [4].

The generative adversarial network is a revolutionary deep learning framework proposed by Goodfellow and other researchers [5] in 2014. The fundamental concept is to produce

Corresponding author: sogeni@ldy.edu.rs

authentic visual representations [6] via the competitive interaction between two neural networks - a generator and a discriminator. The generator is responsible for taking in a random noise vector and using a multi-layer neural network to convert it into data points that fit within the image space. Conversely, the discriminator's role is to assess whether the input image is authentic or if it has been generated by the aforementioned generator.

The variable autoencoder is a deep learning technology based on probabilistic generation model proposed by Diederik Kingma et al. [7]. Its core idea is to express the data distribution as a set of variables, and to encode and decode these variables into data. The encoder's role is to transform the input image into a distribution in a latent space, often represented as a multi-dimensional Gaussian distribution. On the other hand, the decoder is in charge of converting the points in the latent space back to their original image representation. The workflow of variational autoencoder (VAE) includes encoding, sampling, decoding, and loss functions. Unlike generating adversarial network (GAN), VAE is based on a probabilistic generation model, which has a solid theoretical basis and can generate images that match the true distribution of the data.

Due to the swift advancement of deep learning technology, significant strides have been made in the realm of image generation. In the second chapter, the main emphasis of this paper will be on the GAN and VAE. The data set and related performance comparison will be explained in the third chapter. In the fourth chapter, the present concerns are under scrutiny to explore the potential progress of image creation, and the fifth chapter provides an overview of the entire document.

2 Image generation method based on deep learning

2.1 Background overview of deep learning

Receiving extensive recognition as an effective machine learning technique, deep learning has garnered significant interest due to the enhancement of computer capabilities and the surge in data quantity. By building multi-layer neural network models, deep learning has the capability to autonomously acquire the high-level characteristics of the input data, so as to realize the modeling and processing of complex tasks. In the field of image generation, deep learning techniques especially show a powerful potential. Convolutional neural network (CNN) is an important branch of deep learning, which shows excellent performance in image processing tasks. CNN can effectively extract image features through convolution layer, pooling layer and fully connected layer, and achieve breakthrough results in a variety of image recognition tasks.

The progress in image generation using deep learning can be credited to a number of crucial elements, including the accessibility of large datasets, the enhancement of computational capabilities and algorithmic innovation. The models used in deep learning in image generation mainly include: GAN, VAE, AutoRegressive Model (AR Model), etc.

2.2 Image generation based on GAN

The GAN consists of a generator and a discriminator. The main role of the generator is to produce lifelike images, while the discriminator's task is to differentiate between generated and real images. The core idea behind GAN is to facilitate the generation of increasingly realistic images by pitting the generator against the discriminator in adversarial training, as illustrated in Fig 1.

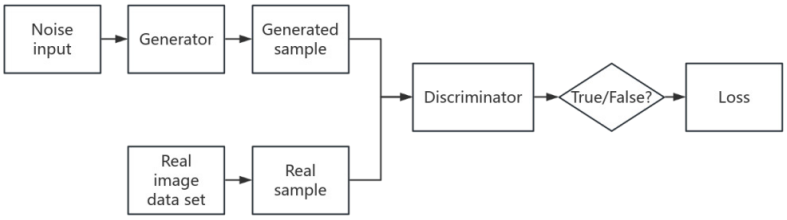

Fig. 1. Network structure of GAN. (Photo/Picture credit: Original)

The functioning of GAN can be summarized in the following steps:

(1) Initialization: randomly initialize the parameters of the generator and the discriminator.

(2) Generator training: the generator generates a set of images, and the discriminator scores these images, and the generator adjusts its parameters according to the score to generate a more realistic image.

(3) Discriminant training: the discriminator differentiates between the image produced by the generator and an authentic image, and adjusts its parameters according to the differentiation results to improve the differentiation ability.

(4) Iterative training: continuing the training process for both the generator and discriminator as described above, until the generator is capable of producing images that are sufficiently realistic.

Building on the original GAN, deep convolutional GAN (DCGAN), proposed in 2016, has become one of the most important early innovations in the GAN field since its inception. This is the first time that the researchers have successfully integrated the ConvNet directly into the full GAN model. To address the issue of excessive freedom in the original GAN, it is a logical approach to introduce additional limitations, on this basis then have the conditional GAN (CGAN), CGAN in the generation model (D) and discriminant model (G) modeling introduce conditional variable y, with the use of additional information y to increase the model condition, can guide the data generation process [8]. Nowadays, there are a considerable number of variants of GAN. The characteristic of GAN is its powerful generation ability, which can generate high quality and high resolution images. There are some problems in the original GAN, such as training instability, mode collapse, etc., and different GAN architectures can solve these problems well. Relevant to the quality, speed, and stability of image generation, various GAN architectures and methods each have their own strengths and weaknesses. Here are some of the typical GAN architectures and their comparisons.

The original GAN architecture is simple and contains a generator and a discriminant. The image is made by the generator, while the discriminator evaluates its authenticity. Despite the theoretical attraction of primitive GAN, it often has problems with training instability and pattern collapse in practical applications. The Deep Convolutional Generative Adversarial Network (DCGAN) is an enhanced version of the original generative adversarial network. It employs deep convolutional neural networks to enhance the quality of generated images and the performance of the discriminator. DCGAN significantly improves the sharpness of the generated images by using the convolution and deconvolution layers to extract and generate the features of the images. Horizontal adversarial GAN (HGAN) introduces horizontal adversarial learning, which enhances the diversity and authenticity of the generated images by introducing additional horizontal adversarial loss between the generator and the discriminant. This approach helps the generator to explore a wider distribution of the data. These methods have advantages in generating image quality, stability, training speed, etc. By comparing these methods, GAN variants can best fit a specific task.

2.3 Image generation based on the VAE

VAE is another popular technique for generating images using deep learning. The VAE is composed of an encoder and a decoder. The image is represented by a low-dimensional latent space due to the transformation effect of the encoder, and the decoder's role is to translate the representation of the latent space back to the original image space. In contrast to GAN, VAE introduces the concept of a probabilistic generative model, making the generative process more flexible and interpretable. The fundamental concept of VAE involves conceptualizing image generation as a probabilistic process, wherein images are produced through the sampling of latent variables from the latent space. The network structure is illustrated in Fig. 2.

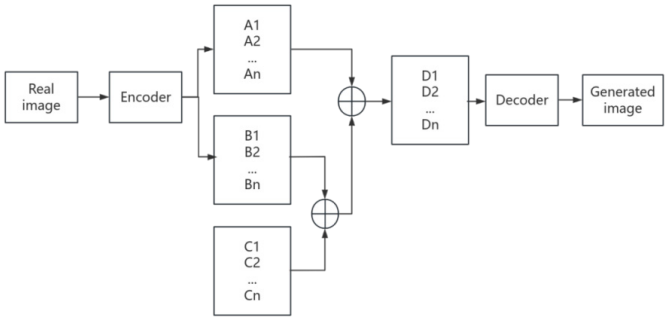

Fig. 2. The VAE network structure. (Photo/Picture credit: Original)

The working principle of the VAE includes the following several steps:

(1) Ending: the encoder receives the input image and encodes it into a latent variable, usually expressed as a pair of mean and variance.

(2) Sampling: Based on the average and spread of the hidden variable, a re-parameterization method (such as Gaussian sampling) is used to sample a latent variable from the latent space.

(3) Decoding: the decoder receives the sampled latent variables and decodes them into an image in the original image space.

(4) Loss function: the two parts of reconstruction loss and KL divergence together form the loss function of VAE. The difference between the decoder output and the original image is measured by the reconstruction loss, while the difference between the distribution of the latent variables and the prior distribution (usually assumed to be the standard normal distribution) at the encoder output is measured by the KL divergence.

Based on the standard VAE, β -VAE improves the diversity of generated images by the introduction of a hyperparameter β to regulate the coherence and distinctiveness of the latent space while maintaining the data structure. VAE-GAN improves the quality and diversity of generated images by introducing adversarial training on the basis of VAE. The advantage of VAE is that its generated images have a clear probability distribution, which makes the generation process more stable and interpretable. However, when standard VAE-generated images have relatively low quality and slow generation speed, and processing complex images may not capture all the details, they can also be solved by using different VAE architectures. Here are some typical VAE methods and their comparisons.

The standard VAE is the first proposed VAE architecture, which assumes that the data in the latent space follows a normal distribution. During training, the standard VAE optimizes the parameters by maximizing a lower boundary of the marginal likelihood (ELBO). Despite the theoretical appeal, a standard VAE may have limitations in dealing with complex distributions. Condition VAE (CVAE) enables the generated image to meet specific

conditions, such as category labels or text descriptions. This method performs well in the image-to-image transformation task, such as generating the corresponding images given the category label. The expressive capacity of the model is enhanced by Deep VAE through the augmentation of network depth. This architecture often contains multiple hidden layers, allowing the model to learn more complex latent space representations. However, the depth may also lead to training difficulties and gradient disappearance problems. Different VAE methods have different advantages in image generation tasks, so researchers should select the appropriate VAE architecture according to the specific application requirements and scenarios. With the continuous technological progress, VAE is anticipated to achieve further advancements in the realm of image synthesis.

3 Dataset and performance comparison

3.1 Dataset

The choice of dataset is crucial for the success of the image generation task. Datasets commonly used in deep learning include ImageNet, CIFAR-10, and the details of the relevant datasets are shown in Table 1 [9]. These datasets provide a rich sample of images, and serve as the foundation for training and testing image generation algorithms. The rapid development of deep learning algorithms is also largely due to the emergence of these large-scale datasets. With the continuous progress of technology, the utilization of datasets in the realm of image generation is poised to further proliferate, offering robust backing for the advancement of image generation technology.

Table 1. Common datasets for deep learning.

Name	Total number of images	Number of categories	Image size / pixel	Start time / year
Caltech101(Lin et al，2014)	9145	101	300×200	2004
	characteristic			
	There are relatively few training images; only one target per image; no noise in most images; less suitable for actual evaluation.			
PASCAL VOC 2007(He et al，2015)	9963	20	375×500	2005
	characteristic			
	Only 20 categories are common in daily life; extensive training images; images approaching the real world; large intra-class changes; targets in the corresponding scene; one image containing multiple images; many different samples; create standardized precedent.			
TinyImages(Torralba et al，2008)	About 79000000	53	46432×32	2006
	characteristic			
	The largest number of images and categories; low-resolution images; not manually verified.			
Caltech256(Griffin et al，2017)	30607	256	300 × 200	2007
	characteristic			
	Similar to Caltech101, there are more target classes than Caltech101.			
ImageNet(Nielsen，2018)	About 14000000	21841	500×4002009	2009
	characteristic			
	More number of images and target categories; more challenging than PASCAL VOC; subset ImageNet1000 popular.			
MS COCO(Havard et al，2017)	About 328000	91	640×480	2014
	characteristic			
	Closer to real scenarios; each image contains more target instances and richer target annotation information.			

3.2 Comparison of different methods for generating adversarial networks

GAN and VAE have advantages in image generation, the results are shown in Fig 3 and 4. Through the comparison of the images generated by GAN and VAE, it is clear that the GAN approach is able to demonstrate the rich details of the images, the image quality is superior, despite the subpar quality of images produced by VAE, it can better understand the probability of the structure and distribution of the image. In practical applications, more appropriate image generation methods can be selected according to the requirements of specific tasks. As technology continues to advance, both methods will continue to evolve, bringing more innovations and breakthroughs to the field of image generation.

(a) (b)

Fig. 3. Comparison of Image Generation Effects Between GAN and VAE on the MNIST Dataset. (a) Images Generated by GAN, (b) Images Generated by VAE [10].

(a) (b)

Fig. 4. Comparison of Image Generation Effects Between GAN and VAE on the CelebA Dataset. (a) Images Generated by GAN, (b) Images Generated by VAE [10].

4 Common problems and solutions

GAN and VAE are two mainstream methods, each of which has different performance in the accuracy, quality and diversity of image generation. Today, although significant progress has been made in generation areas, there are still some problems and challenges.

4.1 Model training stability

The training process of GAN and variational VAE often requires a large amount of computing resources and time, during the training process, potential instability may manifest, leading to fluctuations in the quality of the generated image. The solution includes gradient penalty, learning rate adjustment and pre-training. Introduce gradient penalty Wasserstein GAN (WGAN) and LSGAN in GAN, which can reduce the gradient disappearance and gradient explosion problems in the training process and make the model converge effectively. Leveraging a pre-trained model as the foundation for both components reduces early-stage instability, mitigates mode collapse, and enhances overall model stability.

4.2 Failure to meet standards in variety and quality

While GAN and VAE can generate high-quality images, in some cases the generated images may lack realism or have limitations in diversity. Solutions have conditional generation and latent space exploration, and by introducing conditional information, such as category labels or text descriptions, it can guide the generator to generate images with specific properties, improving the realism and diversity of images. Higher quality images can be generated by exploring more complex latent spatial structures in VAE, such as using flow models, or autoregressive models.

4.3 Against the attack and defense of the sample

Image generation models are vulnerable to adversarial samples, i.e., misleading the model by adding small perturbations to the input data. The solution includes adversarial training and model regularization, which introduces adversarial samples during the training process to improve the resilience of the model to adversarial attacks. The model's defense against antagonistic samples can be improved, by including regularization, such as Dropout or Batch Normalization. Although there are some challenges in the field of image generation, these problems can be solved gradually, through continuous technological innovation and improvement, driving the further development of image generation technology. Future research could also investigate the potential of integrating generative models with other deep learning methodologies to enhance the fidelity, variability, and interpretability of synthesized images.

5 Conclusion

This paper provides a comprehensive review of the two main recognition methods of deep learning: GAN and VAE. First, the study introduce the two methods and show the network structure, and illustrate the advantages and disadvantages of the method process and GAN and VAE. Then some representative datasets are collected for list presentation, discussing the advantages and disadvantages of existing methods from the results, and proposing a series of possible solutions to these problems. Finally, this study provides a summary of deep learning-based image generation techniques and prospects for future work. Future research

should aim to improve the stability, interpretability, diversity and efficiency of models, while ensuring data privacy and security. Anticipated by this research is the continuous advancement of deep learning technology, which is expected to lead to more substantial advancements in the field of image generation and introduce more creative applications to computer vision and other related fields.

References

1. C. Wang, J. Song, L. Wang, P‐2.9: A review of image generation methods based on deep learning. SID. **54**, 507-512 (2023).
2. D. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inferencein deep generative models, in Proceedings of the International conference on machine learning, (2014).
3. Y. Gao, Y. Song, X. Yin, Deep learning-based digital subtraction angiography image generation. Int. J. Comput. Assist. Radiol. Surg. **14**, 1775-1784 (2019).
4. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutiona lgenerative adversarial networks. arXiv preprint arXiv:1511.06434 (2015).
5. I. Goodfellow, P.-A. J, M. Mirza, Generative Adversarial Networks. NeurIPS. **3**, 2672-2680 (2014).
6. T. Salimans, I. Goodfellow,W.Zaremba, Improved techniques fortraining gans. NeurIPS. **29,** (2016).
7. D. Kingma, M. Welling, Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).
8. M MIRZA, S. OSINDERO, Conditional generative adversarialnets. arXiv preprint arXiv:1411.1784 (2014).
9. Y. Zhao， Y. Rao， S. Dong, Survey on deep learning object detection． J. Image Graph. **25**, 0629-0654 (2020).
10. F. Chen， Research on image generation technology based on GAN and VAE fusion network. Hebei GEO University. **08**， (2019).