Researches Advanced in Image Generation based on Deep Learning

Jianing Duan
Faculty of Engineering, Chinese University of Hong Kong, Hong Kong, 999077, China
1155141464@link.cuhk.edu.hk

Abstract. Image generation has always been a study hotspot in machine learning, which aims to build models to learn specific semantic distributions from massive image data to generate realistic simulated images. Thanks to the deep learning technology's quick development, generative models are constantly being developed and huge success has been achieved in image generation tasks. According to difference between generative models, the existing image generation methods based on deep learning can mainly be separated into three models: image generation based on Variational Autoencoder (VAE), image generation based on Generative Adversarial Network (GAN) and image generation combined the VAE and GAN. Focusing on the three frameworks, in this paper, the development process and related principles of each type of generation model are described respectively. After that, the different generation results of different generation models for the agreed training set are compared intuitively, the advantages and problems of various models are proposed, and reasonable improvement measures are proposed for some problems. Finally, the development prospects of various models are prospected.

Keywords: Image Generation; VAE and GAN; Deep Learning.

1. Introduction

In 1956, artificial intelligence was formally born as a new discipline. It is a new technology science that uses computer program to simulate and realize human intelligence. Since 1980, machine learning has really become an independent direction. As the research core in artificial intelligence, machine learning is weighted in using data and algorithms to simulate human learning behaviour, and continuously improves its accuracy and performance from learning. As an important branch of machine learning, the basic of deep learning is deep neural networks, which are used to learn the internal rules and representation levels of sample data. Deep learning makes it possible for machines to simulate human activities, and then analyse and learn like humans. Over these years, deep learning has been used extensively in lots of areas. According to the principle of deep learning, image generation models are proposed.

The key idea of image generation model based on deep learning is to find a representation potential space of low dimensional, where any point can be mapped into a realistic image. Train through a great quantity of input image sets, and then output the generated image. At present, there are three kinds of image generation models, which are based on deep learning:

(1) The first is the variational auto-encoder (VAE) image generation model. The main feature of VAE is that it converts the input image into statistical distribution parameters, mean and variance, which is called the encoding process. After the encoding process, the input image is converted into a hidden vector z, which is received by the decoder. Then randomly sample an element from the normal distribution and decode the element to the original input. This process is called decoding process.

(2) The second type is the generative adversarial network (GAN) image generation model. The main feature of GAN is that it consists of discriminator and generator. The discriminator network uses an image (real or synthetic) as input to predict whether it comes from the training set or the generating network. The generator network takes a random vector (a point in the potential space) as the input and decodes it into a composite image.

(3) The third type is the image generation model combining VAE and GAN. This kind of model takes the VAE as the generator of the GAN. The first half of the network is a complete VAE, and the second part is a GAN, both of which share a decoder.

Focusing on the above three image generation models, in this paper, firstly we introduce the principles and basic process of each generation model in detail. We also analyse and compare the generation results of the three generation models. Moreover, we summarize the existing problems in image generation and discussion the possible solutions to looks forward to their future development.

2. Deep Learning based Image Generation

2.1 Image Generation based on VAE

Deep learning models often need large data sets for training, and each data point usually has continuous latent variables, which makes it difficult for previous deep learning models to effectively reason and learn about. Diederik Kingma and Max Welling put forward a variational auto-encoder in 2014, and used SGVB estimator to effectively approximate a posteriori inference of continuous latent variables of data points, avoiding the complex iterative reasoning of the previous model and improving the learning efficiency [1].

Due to the continuous expansion of modern data sets and the difficulty in obtaining label information, semi-supervised learning has become an important research direction. When a small part of the observed values has labels, but most of the data sets are in the unlabelled state, it is difficult to get the total data set's labels. Diederik Kingma et al. proposed the conditional variational auto-encoder on the basis of VAE [2]. This model converts the above problems into the missing data induction task of the classification problem. Using the data density estimator, it allows generalization from small labelled datasets to large unlabelled datasets.

In the variational auto-encoder, noise is usually injected into the random hidden layer. In noise reduced automatic encoders, noise is often injected into the input layer, and the standard VAE model is difficult to train when the input is destroyed [3]. Daniel et al. showed through research that noise injection in both the input and hidden layers is beneficial to training [4]. An improved lower bound of variation is proposed as the objective function of this setting. Even if the input is destroyed, the model can still reflect a feasible bound.

In addition, βVAE model introduces a super parameter β based on the previous VAE [5]. This parameter can be adjusted appropriately. The experimental data shows that in many cases, the model ($\beta > 1$) is qualitatively superior to the standard VAE ($\beta = 1$).

2.2 Image Generation based on GAN

As the deep learning model continues to develop, a generative adversarial network has been proposed by Goodfellow et al. in 2014 [6]. The generator and discriminator make up the majority of the GAN model. While the discriminator is being trained to accurately differentiate between the true image and the image which is generated by the generator, the generator is being trained to make the generated image be judged as true by the discriminator as much as possible.

Based on the principle of GAN, the conditional GAN allows manual construction by adding appropriate condition data to the generator and discriminator [7]. Compared with the general GAN, the conditional generative adversarial network can guide the data generation process through additional information. As an extension of the generative adversarial network, infoGAN can learn in a completely unsupervised state [8]. InfoGAN generates samples of the same category by maximizing mutual information between potential variables and observations.

Besides, the discriminator is modified in the generative adversarial network model. EBGAN regards the discriminator in the standard GAN model as an energy function [9]. In EBGAN, generators are trained to generate comparison samples with minimum energy, and discriminators are trained to distribute high energy to these generated samples. The unsupervised learning of the generative adversarial network can be realized; however, the loss function of the conventional GAN discriminator may lead to the problem of gradient disappearance in the learning process. LSGAN uses the least squares loss function to avoid this problem, which can make the generated image quality higher and the learning process more stable [10].

Aiming at the new improved training method of conventional GAN, Augustus et al. constructed a GAN variant ACGAN using marking conditions [11]. Different from CGAN, ACGAN specially learned the image structure under a category, added category information to the real data of discriminator, and further told the generator the network sample structure information, so that it can generate better results for samples of the same category. For unsupervised learning, WGAN can train a stable GAN without sophisticated network design and training process [12]. WGAN updates the discriminator network stably by cutting the discriminator network parameters. However, sometimes we blindly cut the weight parameters to ensure the training stability, which may lead to the generation of low quality and low-definition images. In order to solve the problem that WGAN sometimes generates low-quality images, WGAN-GP abandoned the method of cutting discriminator network weight parameters, but adopted the method of cutting discriminator network gradient (cutting according to input data) [13]. DRAGAN is also a gradient clipping in essence, and its discriminator and generator value functions are similar to WGAN-GP [14]. DRAGAN's gradient penalty scheme can avoid the sharp gradient of discriminator function displayed around some real data points by local balance.

In general, conventional GAN is not applicable to discrete data. For this, R Devon et al. proposed a boundary search generative countermeasure network [15]. BGAN has the advantage of generating discrete samples. BGAN generator aims to continuously generate samples on the decision boundary.

2.3 Image Generation Combining VAE and GAN

If VAE is used for generation, a common problem is that very vague results will be obtained; If standard GAN is used, the training will be unstable. GAN based on auto-encoder hopes to strengthen auto encoder through GAN to make the generated results clearer. The generation model standard combining the two is VAE-GAN. The principle of VAE-GAN is to add a discriminator to the original VAE, which forces the output image to be more realistic [16]. In the training process of conventional GAN, the generator cannot contact the real image, so it may output meaningless generated images. By adding the VAE structure before GAN, the generator can directly contact the real image while training to cheat the discriminator, thus making the learning process more stable.

CVAE-GAN is generated by adding appropriate conditions on the basis of VAE and GAN combined to generate the model VAE-GAN [17]. In the process of encoder and generator training, relevant condition information can be added to guide the learning process, so that the generated results are more relevant.

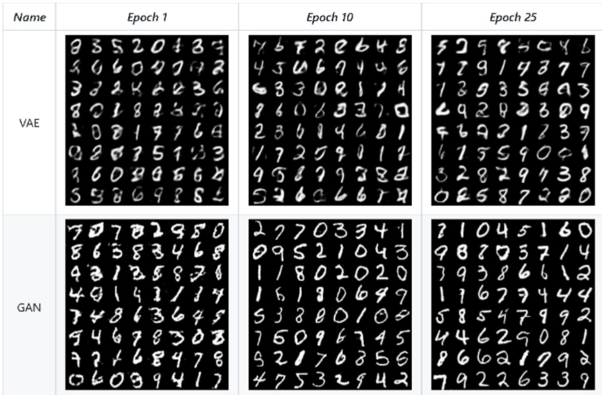
Another generation model that combines VAE and GAN is BiGAN [18]. In BiGAN, encoder and decoder are not together. For the encoder, its input is a real image and a vector is obtained. For the decoder, which is also the generator in the model, the input is a randomly sampled vector, and the output generates an image. For the discriminator, its input is an image and its paired vector, and then it can determine whether the information is from the encoder or decoder. Although encoder and decoder are not directly connected, they can form an ideal auto-encoder through discriminator.

Because of its encoding and decoding structure, VAE has good reasoning ability. GAN has good generalization ability because of its discriminator generator structure. InfoGAN skilfully combines the advantages of the two and adds a new algorithm to the field of deep learning [19].

3. Performance Comparison of Difference Methods

As the most important two stages in machine learning, training and testing need a large number of data sets to support. MNIST dataset is a subset of NIST dataset. Each MNIST image is a gray handwritten digital image with 28×28 pixels of 0~9, with white characters on a black background. The image pixel value is 0~255. The larger the pixel value, the whiter the point.

For different depth learning generation models, by inputting the same MNIST training set, the generated results are also different. As the Figure 1(a) shown, the following are the random sampling results of the VAE, GAN generation models1. For VAE and CVAE models, after increasing epoch within a certain range, it can be found that some of the generated results are improving, while some of the generated results are declining in quality. For GAN and CGAN models, when we increase the epoch twice within a certain range, the meaningless images in the generated results are reduced, and better generation results are obtained. We also compare the generation results of VAE and GAN under the condition generation status, which can be seen in Figure 1(b). The noise vector of each row is the same, and the label condition of each column is the same[1].



(a) Generation results of VAE and GAN

(b) Results under the condition generation status

Figure 1. Comparison of different methods' generation results on MNIST dataset

Finally, we compare the image results generated by the combined generation model of VAE and GAN. The former (see Figure 2) is VAE-WGANGP, and the latter (see Figure 3) is CVAE-GAN[2].

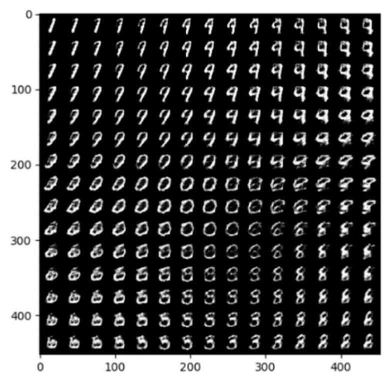


Figure 2. Generation results of VAE-WGANGP on MNIST dataset

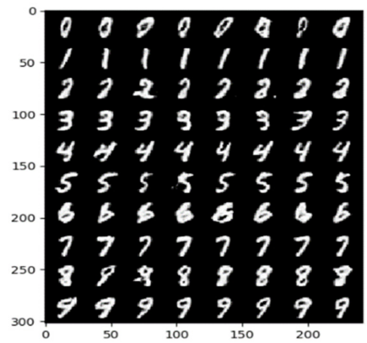


Figure 3. Generation results of CVAE-GAN on MNIST dataset

By comparing the image generation results, it can be observed that the images generated by VAE are smooth and fuzzy, while the ones generated by GAN are clear and artifact. The combination of VAE and GAN produces the best result without labels, and the generated image is very clear and overly uniform.

4. Discussion

For the VAE image generation model, the images generated by the VAE model are often fuzzy and not clear enough because there is no use of the countermeasure network. In this regard, we can improve the output quality by modifying the VAE appropriately without using the variational Bayesian method.

For the GAN image generation model, first, from the working principle of the model, the image generated by the GAN image generation model should be identified as true by the discriminator as far as possible. As long as the discriminator identifies the image as true, no matter whether the content is meaningful or not, the image will be learned and retained. Therefore, images that are judged to be true but meaningless may appear in the generated results. In this regard, negative samples can be added during training to improve the discriminator's judgment accuracy for training and irrelevant samples, thus reducing the generation of meaningless images. Second, from the structure of the model, the noise vector used by GAN for training each time is random, and cannot produce some specific images. In this regard, during the training process, specific noise vectors can be added to the training set through manual processing, so as to generate specific images.

For the VAE-GAN image generation model, since the VAE extracts the corresponding features of the training set image and reconstructs them in the decoder, the generated image can be operated accordingly. It can not only see the extracted feature vector in the generation process, but also directly compare the training image with the generated image, and ensure that the generated image is meaningful. Besides, the addition of GAN also allows the image generated by the decoder to be judged by the discriminator, avoiding the problem that the image generated only by the VAE model is not clear enough. It can also be observed from the experimental results that the generation model of VAE combined with GAN is the best among the three. However, the implementation of this model is more complex than the former two, and it is still in the improvement stage.

5. Conclusion

Focusing on the topic of image generation, this paper divides the current popular image generation models into three categories: VAE, GAN, and VAE-GAN, and expounds the evolution process and relevant generation principles of image generation models from these three perspectives. Then, by comparing the different generation results of the same training set in different models, the generation results of the three generation models are compared to obtain the advantages and disadvantages of the three methods. Finally, the corresponding improvement methods are proposed for these disadvantages. It is believed that with deep learning technology's continuous development, more generation models and algorithms will be proposed, and various generation models will become more perfect, thus playing their roles in many fields.

References

[1] Kingma D P, Welling M. (2013) "Auto-encoding variational bayes". arXiv preprint arXiv:1312.6114, 2013.

[2] Kingma D P, Mohamed S, Jimenez Rezende D, et al. Semi-supervised learning with deep generative models[C]//Advances in neural information processing systems, 2014, 27.

[3] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th international conference on Machine learning. 2008: 1096-1103.

[4] Im Im D, Ahn S, Memisevic R, et al. Denoising criterion for variational auto-encoding framework [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017, 31(1).

[5] Higgins I, Matthey L, Pal A, et al. (2016) "beta-vae: Learning basic visual concepts with a constrained variational framework".

[6] Goodfellow I, Pouget-Abadie J, Mirza M, et al. (2020) "Generative adversarial networks". Communications of the ACM, 63(11): 139-144.

[7] Mirza M, Osindero S. (2014) "Conditional generative adversarial nets". arXiv preprint arXiv:1411.1784, 2014.

[8] Chen X, Duan Y, Houthooft R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets[J]. Advances in neural information processing systems, 2016, 29.

[9] Zhao J, Mathieu M, LeCun Y. (2016) "Energy-based generative adversarial network". arXiv preprint arXiv:1609.03126, 2016.

[10] Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2794-2802.

[11] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans[C]//International conference on machine learning. PMLR, 2017: 2642-2651.

[12] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]//International conference on machine learning. PMLR, 2017: 214-223.

[13] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[C]//Advances in neural information processing systems, 2017, 30.

[14] Kodali N, Abernethy J, Hays J, et al. (2017) "On convergence and stability of gans". arXiv preprint arXiv:1705.07215, 2017.

[15] Hjelm R D, Jacob A P, Che T, et al. (2017) "Boundary-seeking generative adversarial networks". arXiv preprint arXiv:1702.08431, 2017.

[16] Larsen A B L, Sønderby S K, Larochelle H, et al. Autoencoding beyond pixels using a learned similarity metric [C]//International conference on machine learning. PMLR, 2016: 1558-1566.

[17] Bao J, Chen D, Wen F, et al. CVAE-GAN: fine-grained image generation through asymmetric training [C]//Proceedings of the IEEE international conference on computer vision. 2017: 2745-2754.

[18] Donahue J, Krähenbühl P, Darrell T. (2016) "Adversarial feature learning". arXiv preprint arXiv: 1605. 09782, 2016.

[19] Ye F, Bors A G. InfoVAEGAN: learning joint interpretable representations by information maximization and maximum likelihood [C]//2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021: 749-753.