Comparison of improved variational autoencoder models for human face generation

Comparison of improved variational autoencoder models for human face generation

Lijie He

School of mathematics and statistics, Shandong University, Weihai, Shandong, 264200, China

201900820133@mail.sdu.edu.cn

Abstract. The variational autoencoder (VAE) model has evolved a number of VAE improved models in the past 10 years, including CVAE, WAE, NVAE, etc. These models have greatly improved the calculation speed of VAE and the resolution of generated images. The main goal of this paper is to compare the principles of these different models and the effect of generating images. The principle analysis method is mainly used to study the improvement direction of the VAE model in different papers. The main ideas for improving the VAE model include optimizing the loss function, optimizing the objective function, introducing other parameters, or improving code efficiency. Some models even add many algorithms in computer vision in for improving image effect. In the experiments in this paper, results show the image results processed by different models. In the current optimal model NVAE, this algorithm has solved most of the VAE image blur problems, and has achieved perfection in image details. The improvement of the VAE model in the future may require innovative ideas. Under the current principle and model structure, the room for improvement of the VAE model is relatively limited. The final experiment shows that to improve the quality of VAE generated images, it is required to optimize the objective function, optimize the algorithm and add image visual processing, and then the generated images will be significantly improved.

Keywords: Variational autoencoder, deep learning, image generation, principle analysis.

1. Introduction

Variational autoencoder was first introduced by Kingma in 2014 [1], who applied Gaussian distribution to hidden layer in Auto-encoder model. It is not only a very powerful deep generative model but a powerful neural network model. Variational autoencoders are often used in data set processing for data dimension reduction or feature extraction. Its structure mainly consists of two parts, namely Encoder and Decoder. Although VAE has many advantages because his sample comes from N (0, 1), VAE still has many shortcomings such as his image results are very blurred. Therefore, many algorithms are based on the improvement of VAE to improve image quality, such as NVAE, WAE, etc. This document is based on the theory analysis and comparison of the results of these enhanced VAE algorithms. Finally, this paper concludes that in order to improve the clarity of images generated by the VAE model, improvements need to be made from both the principle and the algorithm. The best example is NVAE.

Previously Sønderby [2] proposed the Ladder Variational Autoencoder. It is a novel model that could rectify the generative distribution by an approximate likelihood of data dependency. Doersch [3] described some empirical behavior. Dilokthanakul [4] studied a variant of the VAE aims at learning an unsupervised grouping. It takes Gaussian mixture as a pre-distribution. Mescheder [5] presented Adversarial Variational Bayes (AVB), a technique for Variational Autoencoders training with arbitrary expressive inference patterns. Ballé [6] describes an end-to-end training model for dimensional auto-encoder based image compression. Is it possible to move the progression to the graphics area? Simonovsky [7] proposed to get around the obstacles related to the linearization of these discrete structures by having a decoder produces a fully connected probability graphic of a predefined maximum size directly at the same time. Aim of Jin [8] is the direct production of molecular graphics, a task previously addressed in generating linear SMILES chains rather than graphics. Chen [9] broken down the lower end of the evidence to show the evidence that, between latent variables, there could be a term measuring the total correlation. Ball [10] describe an end-to-end training model for dimensional auto-encoder based image compression. Kingma [11] introduce variational auto-encoders and a few important extensions.

2. Method

2.1. Autoencoder

AE is short for autoencoder. Through self-supervised training, it can obtain a potential feature encoding from the original features, realize automatic feature engineering, and achieve the purpose of dimensionality reduction and generalization. Its network structure is very simple, consisting of two parts: encoding and decoding:

It can be seen that because the target of the network is the input itself, no additional label work is required. While AE is made up of two main components, which are the encoder and the decoder. It aims at obtaining the vector of this hidden layer as the potential feature of the input, which is a common embedding method. The decoding result, based on the training target, will be the same as the input if the loss is small enough. From this point of view, the decoded value has no practical significance, except to supplement and smooth some initial zero values by increasing the error or have some use. Because the entire process from input to output is based on the mapping of existing training data, although the hidden layer has less dimensions than the input layer, the probability distribution of the hidden layer still only depends on the distribution of the training data. That of the hidden state space is not continuous, so if the state of the hidden layer is randomly generated, it will probably no longer have the characteristics of the input features after decoding, so it is difficult to generate data through the decoder.

2.2. Variational autoencoder

The difference between VAE and the auto encoder is that the auto encoder reconstruction target is the fixed result vector itself, whereas the VAE target is the input distribution. In the implementation process, the only difference between the two is the Bottelneck vector in the auto-coder, which is split into two in the VAE—the mean vector and the SD vector. Then use the average variance setting to sample and send the sampled result to the decoder. Its conventional architecture is demonstrated in Figure 1.

VAE is also mainly used for image processing. In fact, the process of training a autoencoder structure is a process of finding ways to compress data. The most well-known application is based on the regeneration of face images. Other applications include image segmentation, removing noise points in images, completing images, removing image watermarks, etc.

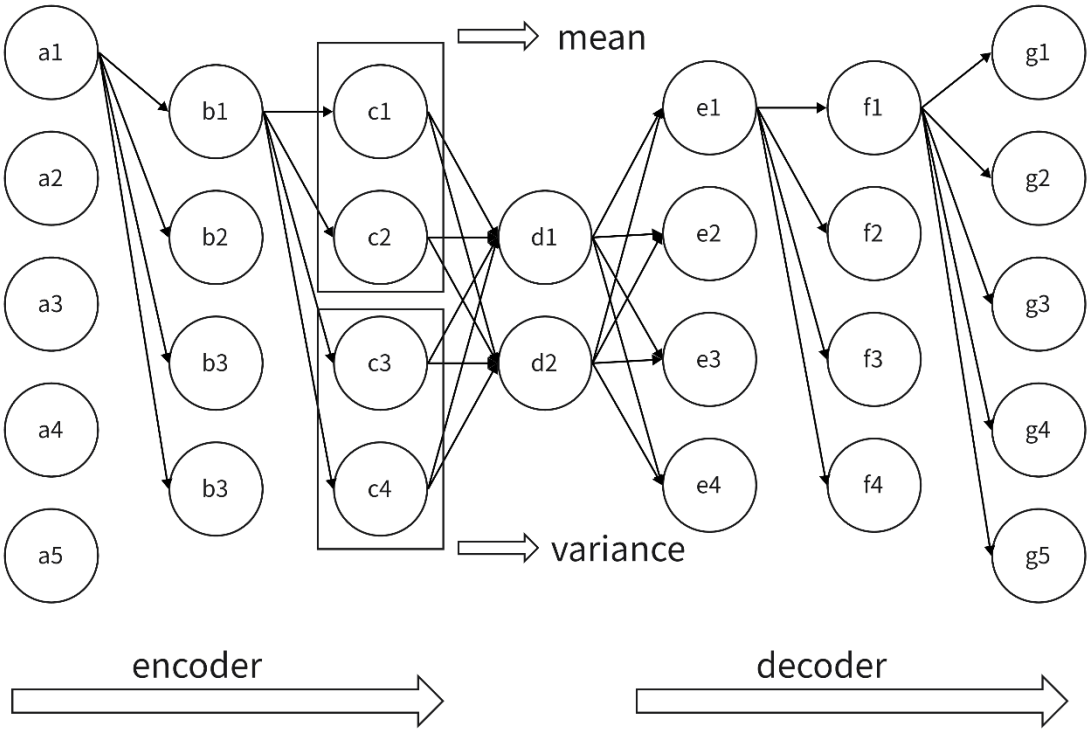

Figure 1. Architecture of the VAE.

2.3. Deficiencies of conventional VAE

When from Figure 2, VAE-generated pictures, it could be find that the face photos regenerated by VAE often have only approximate outlines. The boundary line between the face and the background is not clearly defined in the picture. Colors tend to transition slowly from light to dark without sharp borderlines. Therefore, looking at the pictures generated by VAE as a whole, except for the eyes that can be seen more clearly, the details of the rest of the face are very blurred as shown in Figure 2. In addition, the background of the picture is usually a solid color background. In contrast, images generated by GAN models tend to be clear and have sharp edges.

Figure 2. Blurry image generated by VAE.

The reason is because of L1 (or L2) reconstruction loss used in VAEs. As is discussed in [12], simply minimizing the Euclidean distance between predicted and terrestrial clues of truth cannot guarantee a clear result. Since the Euclidean distance does not focus on details but merely concentrated on minimising the average errors, which causes blurring. Moreover, GANs discriminates if the output image is real or not and blurry one will be punished as they are obviously fake.

The main reason for blurring images generated by VAE is discussed above. Therefore, the main idea for improving the image generated by VAE is to reduce the error in the encoding process and the error in the image generation process. In some improved models, the researchers chose to improve the probability function to make the target q(z|x=x) more reasonable; in some models, they changed the definition formula of the distance, and when the calculation loss function is minimized, make the result more optimized.

2.4. Improvements of VAE

The main reason for blurring images generated by VAE is discussed above. Therefore, the main idea for improving the image generated by VAE is to reduce the error in the encoding process and the error in the image generation process. In some improved models, the researchers chose to improve the probability function to make the target q(z|x=x) more reasonable; in some models, they changed the definition formula of the distance, and when the calculation loss function is minimized, make the result more optimized.

2.4.1. WAE. WAE has the same optimization goal as VAE, but simplifies the optimal transport (OT) in the process of computing minimization. where OT is defined as:

$$W_c(P_X, P_G) \coloneqq \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_G)} E_{(X,Y)\sim \Gamma}\,[c(X,Y)] \tag{1}$$

Among them, c(x, y) is the loss function. When p $\geqslant$ 1, if c(x, y) = dp(x, y), it is called the Wasserstein distance. And the final OT function is

$$D_{\mathrm{WAE}}(P_X, P_G) \coloneqq \inf_{Q(Z|X)\in\mathcal{Q}} E_{P_X} E_{Q(Z|X)}\left[c\big(X, G(Z)\big)\right] + \lambda \cdot \mathcal{D}_Z(Q_Z, P_Z) \tag{2}$$

Then the researchers proposed two kinds of regularizers: DZ(QZ, PZ). The former is GAN-based DZ. Let DZ(QZ, PZ) = DJS(QZ, PZ), (DJK is the JK divergence), and then use adversarial training to approximate it. The latter is MMD-based DZ. Let DZ(PZ, QZ) = MMDk(PZ, QZ). Among them, MMDk(PZ, QZ) is calculated by the following formula:

$$MMD_k(P_Z, Q_Z) = |\int_Z k(z,\cdot) d\,P_Z(z) - \int_Z k(z,\cdot) d\,Q_Z(z)|_{\mathcal{H}_k} \tag{3}$$

2.4.2. Beta-VAE. Beta VAE is an improvement of VAE with a particular focus to uncover disentangled latent factors. A hyperparameter called beta is introduced for balancing recon accuracy and latent channel capacity together with independence constraints. This work appropriately tune beta>1 such that beta-VAE is qualitatively superior to VAE (beta=1), and unsupervised (InfoGAN) and semi-supervised (DC-IGN) methods for use on various datasets (celebrities, face and chair) on disentangled factor learning. The idea is to keeping the distance between the actual distribution and estimated below a threshold while maximize the likelihood of generating real data. This could be written as a single equation using the Kuhn-Tucker condition:

$$\mathcal{F}(\theta, \phi, \beta; x, z) = E_{q\phi(z|x)}[\log p_\theta(x \mid z)] - \beta\left[D_{KL}\big(\log q_\theta(z \mid x)|p(z)\big) - \epsilon\right] \tag{4}$$

Where β of the KKT multiplier is a regularization coefficient z. It can limits the latent channel capacity and imposes an independence pressure p(z) on the learned posterior as a result of the isotropic nature of the Gaussian anterior. Finally get the Beta-VAE formula:

$$\mathcal{F}(\theta, \phi, \beta; x, z) \geq \mathcal{L}(\theta, \phi, \beta; x, z) = E_{q\phi(z|x)}[\log p_\theta(x \mid z)] - \beta D_{KL}\big(\log q_\theta(z \mid x)|p(z)\big) \tag{5}$$

2.4.3. NVAE. The most impressive model is NVAE [13]. The result in the paper is amazing for its high resolution and smooth faces, while normal VAE model generates blur images.

The reason why NVIDIA made such improvement is complicated. First of all, they improved the methods. They enhanced the probability p(x|z), p(z|x) and p(z) get a new KL dispersion. Secondly, they used many new techniques in CV field, such as Batch Normalization (BN), shared top-down model and improved residual cell.

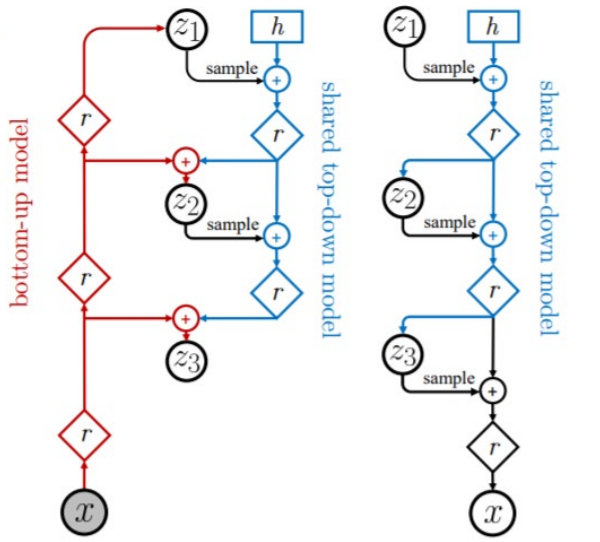

Figure 3. NVAE model. Figure from [13].

NVAE cleverly designed a multi-scale encoder and decoder. First, the encoder is encoded layer by layer to obtain the topmost encoding vector z1. Afterwards, it slowly goes down from the top layer to obtain the bottom layer features z2,…, zL step by step.

As for the decoder, it is naturally a process of using z1, z2, ..., zL from top to bottom, and this part happens to have something in common with the process of generating z1, z2, ..., zL by the encoder. All NVAE directly let the corresponding Part of the parameters are shared, which saves the number of weights, meanwhile improves the generalisation performance through the mutual constraints between the two.

This type of multi-scale design is reflected in the latest generation models, such as StyleGAN, VQ-VAE-2, BigGAN, etc., which shows that the effectiveness of multi-scale design has been fully verified. In addition, to improve results, it also borrowed the residual mechanism, as displayed in Figure 4.

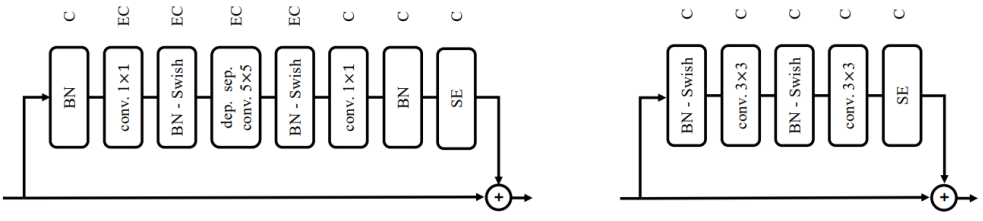

Figure 4. Architecture of the residual block in NVAE. Figure from [13].

3. Result

In this part, multiple pytorch-based VAE and VAE improved models are used. The goal of the experiment is to compare the image generation effects of multiple VAE improved models. 4 models

are compared: VAE, beta-VAE, WAE and NVAE. See how well they and the original VAE solve the image blur problem.

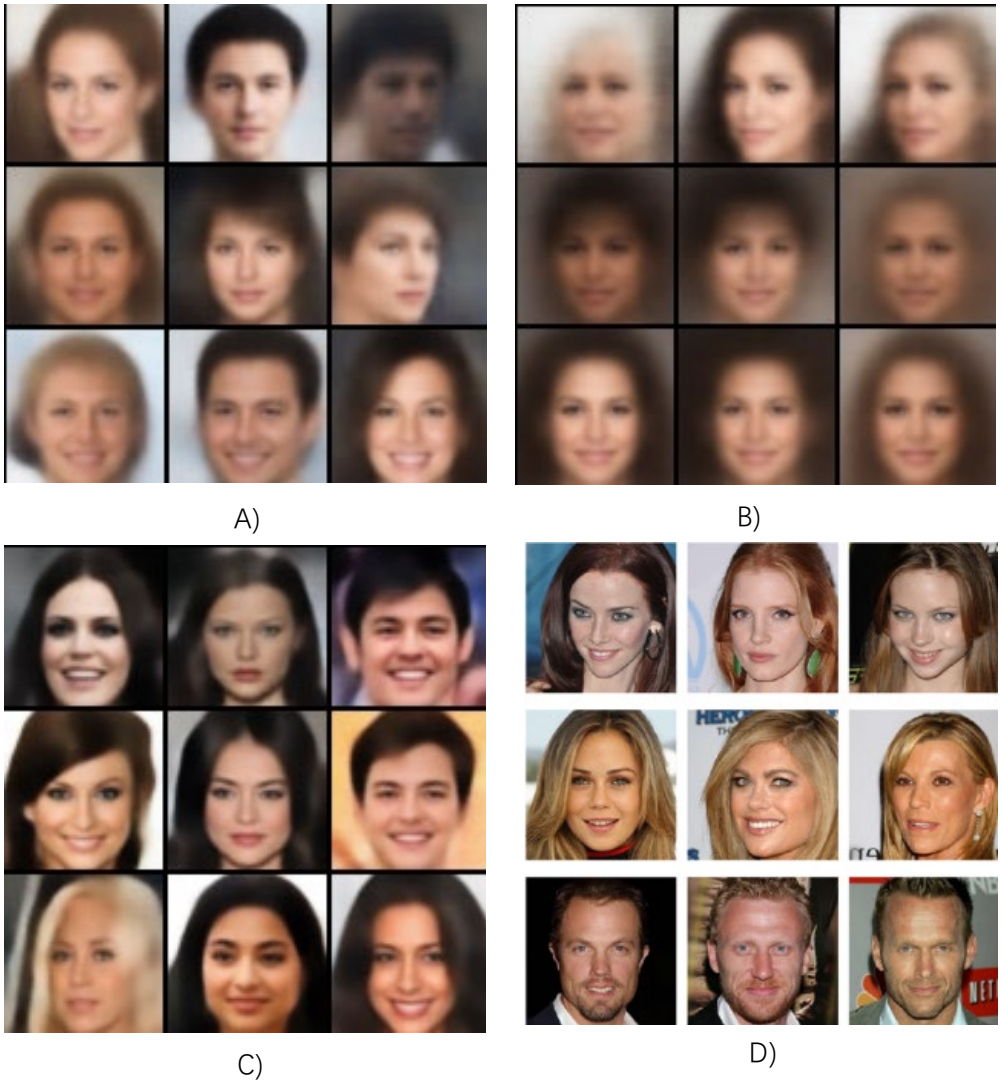

Figure 5. Visualization result generated by the VAE family models.

As the visualization results shown in Figure 5. A) is the face image generated by the original VAE model. B) is the image reconstruction result of beta-VAE. In betaVAE，the changing beta value is introduced into the loss function equation, yet the result of image generation is not much different from that of VAE. It can be seen that the background of even parts of the picture is more blurred. C) is the picture generated by the WAE model. On the basis of VAE, the Wasserstein distance is introduced to optimize the objective function, and two regularizer are introduced at the same time. In terms of image generation, it is much better than the original VAE effect. At the junction of image colours, there is a more obvious dividing line, and the image quality is also greatly improved. D) is the result obtained by the NVAE model, which introduces a lot of improvements, including principle improvements and algorithm improvements. So, the processing result of the image is perfect. Images are high resolution, have detailed colour variations and are no longer blurry.

4. Discussion

The experimental comparison mainly analyzes the different improvement ideas for the VAE model and summarized in Table 1.

Firstly, the optimization of the function. The experimenters optimized two objective functions and changed some probability formulas to make the calculation results more reasonable. For example, the structure and principle of the betaVAE model and the original VAE model are basically the same. But betaVAE changes the beta value, making it a variable between 0 and 1. In the VAE model, the beta value defaults to 1. The change of this variable optimizes the equation and improves the result.

Secondly, adding computer vision algorithm. Computer vision algorithms are mainly embodied in the NVAE model. The experimenters added many latest algorithms such as BN (Batch Normalization), shared top-down model and improved residual cell. These algorithms effectively improve the definition of generated images. But these algorithms make the NVAE model much more complicated than the previous VAE

Table 1. Summary of the VAE models.

	Optimize the function.	computer vision improvement
Beta-VAE	Introduce variable beta	None
WAE	Introduce Wasserstein distance	None
NVAE	Optimize objective functions	BN (Batch Normalization), shared top-down model and improved residual cell

5. Conclusion

The VAE model is an earlier generative model. It improved the AE model and introduced probability distribution. Although it has advantages in computing efficiency and other aspects, VAE still cannot get rid of the shortcomings of generating image blur. Therefore, many subsequent improved models are trying to improve the rationality of the calculation equations and incorporate the latest algorithms to improve the image generation quality of the VAE model. This article mainly discusses the different stages, the experimenter's improvement ideas for the VAE model, and the image generation results. It can be seen that early improvements such as Conditional-VAE and beta-VAE did not break away from the model structure of VAE, but only updated the equation of the loss function and the optimization goal. Make the result of the function more reasonable. This does improve image generation results in some cases, but the improvement is very limited. In order to make the results more perfect, subsequent researchers improved the VAE model from multiple perspectives, such as incorporating confrontation mechanisms, multi-scale design and residual modules. The most notable example is NVAE, which not only has the improvement of the above principles, but also incorporates the latest image processing algorithms to pre-process images. This makes the generated images have very high clarity, and these images have details such as colours and lines, but they also become very complex.

References

[1] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
[2] Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. (2016). Ladder variational autoencoders. Advances in neural information processing systems, 29.
[3] Doersch, C. (2016). Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908.
[4] Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., & Shanahan, M. (2016). Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648.
[5] Mescheder, L., Nowozin, S., & Geiger, A. (2017). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In International conference on machine learning, 2391-2400.

[6] Ballé, J., Minnen, D., Singh, S., Hwang, S. J., & Johnston, N. (2018). Variational image compression with a scale hyperprior. arXiv preprint arXiv:1802.01436.
[7] Simonovsky, M., & Komodakis, N. (2018, October). Graphvae: Towards generation of small graphs using variational autoencoders. In International conference on artificial neural networks, 412-422.
[8] Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. In International conference on machine learning, 2323-2332.
[9] Chen, R. T., Li, X., Grosse, R. B., & Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. Advances in neural information processing systems, 31.
[10] Ballé, J., Minnen, D., Singh, S., Hwang, S. J., & Johnston, N. (2018). Variational image compression with a scale hyperprior. arXiv preprint arXiv:1802.01436.
[11] Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. Foundations and Trends® in Machine Learning, 12(4), 307-392.
[12] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).
[13] Vahdat, A., & Kautz, J. (2020). NVAE: A deep hierarchical variational autoencoder. Advances in Neural Information Processing Systems, 33, 19667-19679.