Evaluation of Natural Image Generation and Reconstruction Capabilities Based on the β-VAE Model

Honghao Zhang

Software Engineering, Nanchang Hangkong University, Nanchang, 330063, China

Abstract. Natural image generation models are crucial in computer vision. However, the Variational Autoencoder (VAE) has limitations in image quality and diversity, while β-VAE achieves a balance between the decoupling of latent space and generative quality by adjusting the coefficient β. This article evaluates the performance of β-VAE in natural image generation and reconstruction tasks, comparing it with Conditional VAE and Information VAE. Train the model on the CelebA dataset, using three metrics: Mean Squared Error (MSE), Structural Similarity Index (SSIM), and Fréchet Inception Distance (FID) for matrix evaluation, to analyze the generation quality and reconstruction capability of each model. The results indicate that β-VAE performs well in reconstruction tasks, but due to the strong constraints of the latent space, the realism of the images is lower in generation tasks. Conditional VAE and Info VAE perform more balanced in terms of generation quality and task balance. This article suggests optimizing β-VAE by introducing adversarial training and flexible latent space modeling to enhance its generative capabilities. Future research can further validate the potential of β-VAE in different application scenarios.

1 Introduction

Natural images, as important information carriers, are widely used in various fields. It covers multiple modern computer vision fields, including medical image analysis, autonomous driving, virtual reality, and augmented reality. In the medical field, natural image generation and reconstruction techniques are often used for the synthesis and enhancement of medical images to assist in diagnosis and treatment [1]. In the field of autonomous driving, these technologies are often used to synthesize simulated training data, thereby enhancing the robustness of autonomous driving systems [2].

With the development of deep learning technology, generative models have made significant progress in natural image processing. By learning the distribution of data, generative models can generate realistic images and perform tasks such as image reconstruction and editing. Since its introduction by Ian Goodfellow and his colleagues in 2014, Generative Adversarial Networks (GANs) have become an important tool in the field

Corresponding author: 21201732@stu.nchu.edu.cn

of image generation, achieving significant success in producing high-quality and realistic images [3]. However, the training process of GAN models is complex, and the adversarial relationship between the generator and the discriminator makes convergence difficult. This often leads to mode collapse, and the resource consumption during the training process is substantial [4]. According to the research findings of Lu Mi and his colleagues, the entropy value of images generated by the GAN model on the MNIST dataset is 0.0, indicating that the GAN model indeed suffers from the issue of mode collapse [5]. In contrast, Variational Autoencoders (VAE) generate and reconstruct input data by learning the probability distribution of the latent space. VAE also has unique advantages in terms of interpretability and controllability of the latent space; however, its ability to generate high-quality and diverse images is relatively lower [6].

To enhance the performance of VAE, Irina Higgins and her colleagues first proposed the β-VAE model in 2017. By introducing a regularization term β, the model achieves a balance between reconstruction loss and Kullback-Leibler (KL) divergence, thereby improving the decoupling of the latent space and the quality of generated images [7]. The β-VAE model has shown potential on several low-dimensional datasets. However, its performance on high-dimensional natural image datasets remains underexplored. This study systematically evaluates the generative and reconstruction capabilities of the β-VAE model using the CelebA dataset, aiming to address this research gap.

This study systematically evaluates the performance of the β-VAE model in natural image generation and reconstruction tasks and compares it with Conditional VAE and Information VAE. To achieve this goal, this paper uses the CelebA dataset to train these three models, which contain 200,000 celebrity facial images, all resized to 64x64 pixels. At the same time, unify the training configuration and design an evaluation matrix that includes three metrics: Mean Squared Error (MSE), Structural Similarity Index (SSIM), and Fréchet Inception Distance (FID). These steps allow for a comprehensive analysis of the differences in performance among various models in terms of generation quality, reconstruction ability, and the structural characteristics of the latent space. Ultimately reveals the unique advantages and shortcomings of β-VAE in terms of generating image diversity, reconstruction quality, and the structural characteristics of the latent space, providing a theoretical basis and practical reference for the selection and further optimization of generative models.

2 Methods

2.1 Model description

β-VAE is an extended version of VAE that introduces a scaling factor β to establish a balance between reconstruction loss and KL divergence loss, thereby controlling the decoupling of the latent space. The innovation of β-VAE lies in its ability to adjust the β value, allowing for a balance between enhancing the [7]. In natural image generation and editing, the β-VAE model enables fine-tuned editing of generated images by making small adjustments to the latent variables, due to the decoupling of its latent space [8]. For example, by adjusting a latent variable, one can alter a specific feature in an image (such as facial expression) without affecting other features. This capability makes β-VAE widely applicable in fields like automated image editing and style transfer. In an example using the CelebA dataset, the β-VAE model can control whether the generated images feature glasses or a smile by adjusting a specific latent variable. This ability to edit based on decoupling in latent space allows the model to perform multidimensional free editing of images without introducing noticeable artifacts or distortions.

2.2 Dataset description

The CelebFaces Attributes (CelebA) dataset is a widely used image dataset in the field of computer vision, released by the Multimedia Laboratory of the Chinese University of Hong Kong. This dataset contains over 200,000 facial images of celebrities, each annotated with 40 attribute labels such as "smiling," "wearing glasses," "having a beard," and so on. The original resolution of the image is 178x218 pixels, but it is usually adjusted to 64x64 pixels before processing to meet different research needs. The CelebA dataset is widely used in tasks such as facial attribute classification, image generation, and editing due to its rich label information and high-quality images [9]. Using the CelebA dataset to evaluate the natural image generation and editing capabilities of the β-VAE model has many advantages [10]. The rich label information allows researchers to utilize these tags for supervised or semi-supervised learning when generating and editing images, enhancing the model's generation and editing capabilities. At the same time, the images in the dataset have high quality, and the resolution of each image is uniform, providing a solid data foundation for model training. In addition, the CelebA dataset includes facial images of different genders, races, and age groups, which helps the model learn a broader range of features, thereby enhancing its ability to generate and edit various types of facial images. Since the CelebA dataset has been widely used in various studies of generative models, conducting experiments with this dataset can make the research results more comparable and generalizable.

2.3 Experimental design

To ensure the fairness and scientific rigor of the evaluation, the training configurations for β-VAE, Conditional VAE, and Information VAE were standardized during the experimental design. Specifically, all models were trained on the CelebA dataset, with training parameters such as learning rate, batch size, and number of training epochs kept consistent to ensure that the experimental results can objectively reflect the differences in model structures.

Before training the model, first normalize the images in the CelebA dataset to a resolution of 64x64 and perform normalization. After that, train each model separately, recording the training logs and evaluation results for each model, ensuring that the models are trained under the same conditions. Finally, a quantitative analysis of the generation and reconstruction capabilities of the three models will be conducted using the designed evaluation matrix. The metrics MSE, SSIM, and FID will be calculated and model comparisons will be made based on the weighted scores of these metrics.

3 Evaluation matrix

To systematically and scientifically evaluate the performance of the β-VAE model in natural image generation and reconstruction tasks, this paper designs a comprehensive evaluation matrix that includes three key metrics: MSE, SSIM, and FID. Each metric has different levels of importance in the specific tasks of different models. Therefore, this study allocated weights for each model separately and provided reasonable explanations.

MSE: Used to measure the pixel-level error between the reconstructed image and the original image. The lower the MSE, the higher the reconstruction quality [11]. MSE mainly reflects the model's performance in reconstruction tasks, particularly suitable for evaluating the model's ability to preserve details of the original image.

SSIM: It is used to measure the structural similarity between the reconstructed image and the original image, taking into account brightness, contrast, and structural information

[11]. The higher the SSIM, the closer the reconstructed image is to the original image in terms of perceived quality. SSIM not only focuses on pixel differences but also on the overall visual quality of the image.

FID: Used to measure the distribution difference between generated images and real images in feature space. The lower the FID, the higher the quality of the generated images [12]. FID is particularly suitable for assessing the realism and diversity of generative models and is an important metric for measuring the quality of generated images.

3.1 β-VAE evaluation matrix

Weight distribution: MSE (30%), SSIM (30%), FID (40%).

Reason for weight allocation: β-VAE balances the structure of the latent space and the quality of reconstruction by adjusting the β parameter. To accurately reflect its generative capabilities and the advantages of potential space, this paper raises the weight of FID to 40%. MSE and SSIM account for 30% each, to ensure the importance of reconstruction quality and structural preservation.

3.2 Conditional VAE evaluation matrix

Weight distribution: MSE (25%), SSIM (30%), FID (45%).

Reason for weight allocation: Conditional VAE needs to generate images that are consistent with the input conditions, so the quality of generation (FID) is crucial for the evaluation of Conditional VAE. To reflect this, the weight of FID is set at 45%. The weight of MSE is relatively low, as conditional consistency often affects reconstruction quality, while SSIM is maintained at 30% to balance the importance of structural preservation.

3.3 Information VAE evaluation matrix

Weight distribution: MSE (30%), SSIM (35%), FID (35%).

Reason for weight allocation: The design purpose of the Information VAE is to strike a balance between generative capability and information transmission. To reflect this balance, the weight distribution of MSE, SSIM, and FID is relatively even, with SSIM and FID slightly higher than MSE, highlighting the unique advantages of the Information VAE in terms of generation quality and information retention.

4. Experimental results and analysis

Table 1. β-VAE evaluation matrix

Metrics	Numeric	Score (1-10)	Weight (%)	Weighted Score
MSE	0.15	4.00	30	1.20
SSIM	0.34	3.00	30	0.90
FID	192.55	2.00	40	0.80
Total score			100	2.90

Table 2. Conditional VAE evaluation matrix

Metrics	Numeric	Score (1-10)	Weight (%)	Weighted Score
MSE	0.35	3.00	25	0.75
SSIM	0.30	3.00	30	0.90
FID	186.52	3.00	45	1.35
Total			100	2.95

score				

Table 3. Information VAE evaluation matrix

Metrics	Numeric	Score (1-10)	Weight (%)	Weighted Score
MSE	0.37	3.00	30	0.90
SSIM	0.34	3.00	35	1.05
FID	190.90	3.00	35	1.05
Total score			100	3.00

4.1 Analysis of β-VAE metrics

In the evaluation of the β-VAE model, various metrics demonstrated its characteristics and limitations in reconstruction and generation tasks.

First of all, in terms of the MSE metric, the β-VAE scored 4.00, indicating that the model can effectively recover the pixel information of the original images in the reconstruction task, reflecting its relatively robust reconstruction capability. The SSIM score is 3.00, indicating that β-VAE shows moderate performance in structural preservation and perceptual quality, possibly due to the model losing some global structural information and details during the reconstruction process. What is more concerning is that the FID score is 2.00, reflecting that β-VAE performs poorly in terms of the realism of generated images and their consistency with the distribution of real images.

This result indicates that although β-VAE can retain pixel information well in reconstruction tasks, its excessive constraints on the latent space lead to poor image quality in generation tasks, making it unable to effectively mimic the distribution of the training data. In summary, β-VAE shows certain advantages in reconstruction quality, but it has significant shortcomings in the realism of generated images and structural preservation, indicating that it is more suitable for reconstruction tasks rather than high-quality image generation tasks.

4.2 Results analysis

Under the same training configuration, the evaluation results of the three models on the CelebA dataset are as follows:

The evaluation results of β-VAE are shown in Table 1, demonstrating relatively good performance in the reconstruction task, particularly excelling in the MSE metric. However, its performance in generation tasks is rather average, especially in terms of the realism of generated images, showing a significant gap. This situation may be related to the decline in the quality of generated images caused by the overly decoupled latent space of β-VAE [7]. To enhance the generative capability of β-VAE, future research may consider improving its generative ability by adjusting the β parameter or introducing new regularization methods.

The evaluation results of Conditional VAE are shown in Table 2, which slightly outperforms the β-VAE in terms of generation quality, with an FID value of 186.52, indicating better image realism. However, the Conditional VAE still has significant room for improvement in reconstruction quality, with an MSE value of 0.35, slightly higher than that of β-VAE. This may be related to the fact that Conditional VAE fail to effectively preserve the details and structural information of images when handling complex conditional information [13]. Future improvement directions could consider using more powerful conditional encoding methods or combining adversarial networks to enhance the expressive capability of conditional information.

The evaluation results of the Information VAE are shown in Table 3, indicating that its performance in generation and reconstruction tasks is relatively balanced. The MSE value of the Information VAE is 0.37, the SSIM value is 0.34, and the FID value is 190.90. Although the overall scores are slightly higher than those of other models, its regularization strategy has limited effectiveness in practical applications. This indicates that the Information VAE has certain limitations in structurally controlling the latent space, which may lead to fluctuations in the quality of generated and reconstructed images. Future research could attempt to introduce more complex latent space regularization strategies to further enhance the latent space representation capabilities of Information VAE.

5 Conclusion

After a systematic evaluation of the natural image generation and reconstruction capabilities of the β-VAE model, this paper conducts an in-depth discussion and analysis of the results. The performance of β-VAE in reconstruction tasks is relatively outstanding, as reflected in the MSE scores. The β-VAE model can accurately restore the pixel details of input images, primarily due to the strong regularization constraints it imposes in the latent space, which enhances reconstruction accuracy. However, this strict constraint on the latent space also limits the generative capability of the β-VAE model, resulting in a lower FID score in generative tasks and a significant distribution difference between the generated images and the real images. This phenomenon may be because β-VAE tends to prioritize reconstruction quality during the optimization process, resulting in a decrease in the diversity and realism of the generated images.

Based on the aforementioned research findings, this paper proposes some improvement suggestions. First, to address the shortcomings in generation quality, one could consider introducing adversarial training mechanisms (such as GAN) into the VAE architecture to enhance the diversity and realism of the generated images. In addition, more flexible modeling of the latent space, allowing for greater expressive capacity of the latent variable distribution, may help achieve a better balance between generation and reconstruction tasks, thereby enhancing the overall performance of the model.

Although this study provides valuable insights into the natural image generation and reconstruction capabilities of the β-VAE model, there are still some limitations. First, since the experiment is solely based on the CelebA dataset, the findings may have certain limitations and may not fully reflect the model's applicability to other types of datasets. Secondly, the evaluation metrics used in this paper mainly focus on pixel-level errors and image structural similarity, which may not fully capture the high-level features that influence human visual perception. Future research can comprehensively validate the performance of β-VAE by introducing more diverse datasets and more complex evaluation metrics, as well as exploring its performance in different application scenarios. Through these extensions and improvements, the application potential of the β-VAE model in natural image generation and reconstruction tasks will be more comprehensively demonstrated.

References

1. Y. Sun, J. Ortiz, Rapid review of generative AI in smart medical applications. Int. J. Comput. Sci. Inf. Technol. 3, 80–93 (2024).
2. J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, C. Xu, H. Xu, One million scenes for autonomous driving: ONCE dataset. Proc. Neural Inf. Process. Syst. Conf. (2021).

3. H. Alqahtani, M. Kavakli-Thorne, G. Kumar, Applications of generative adversarial networks (GANs): an updated review. Arch. Comput. Methods Eng. 28, 525–552 (2019).
4. D. Saxena, J. Cao, Generative adversarial networks (GANs) challenges, solutions, and future directions. ACM Comput. Surv. (CSUR) 54, 1–42 (2021).
5. L. Mi, M. Shen, J. Zhang, A probe towards understanding GAN and VAE models. arXiv preprint arXiv:1812.05676 (2018).
6. D. Shen, Z. Xu, X. Zhang, Z. Xu, Regularizing variational autoencoder with diversity and uncertainty awareness. arXiv preprint arXiv:2110.12381 (2021).
7. I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework. ICLR (Poster) 3 (2017).
8. C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, A. Lerchner, Understanding disentangling in β-VAE. arXiv preprint arXiv:1804.03599 (2018).
9. Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild. Proc. IEEE Int. Conf. Comput. Vis. (ICCV) (2015).
10. L. Sun, Z. Fan, X. Fu, Y. Huang, X. Ding, J. Paisley, A deep information sharing network for multi-contrast compressed sensing MRI reconstruction. IEEE Trans. Image Process. 28, 6141–6153 (2019).
11. A. Borji, Pros and cons of GAN evaluation measures. Comput. Vis. Image Underst. 179, 41–65 (2019).
12. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium. arXiv (Cornell Univ.) 30, 6626–6637 (2017).
13. K. Sohn, X. Yan, H. Lee, Learning structured output representation using deep conditional generative models. Adv. Neural Inf. Process. Syst. 28, 3483–3491 (2015).