Research on the Application of Variational Autoencoder in Image Generation

Jianing Liu

Ulster College, Shaanxi University of Science and Technology, Xi'an, 710021, China

Abstract. The rapid development of artificial intelligence and deep learning has significantly influenced the domain of image creation, finding extensive applications in applications in fields like medical imaging, computer vision, and entertainment. Despite these advancements, challenges remain, especially in enhancing the quality and variety of produced images. This paper concentrates on applying Variational Autoencoders (VAEs) to image generation, a topic of increasing importance due to the model's theoretical interpretability and stability. Through a detailed analysis of VAE principles, architecture, and applications, this research underscores the model's capabilities in producing high-quality, varied images and its effectiveness in tasks such as image denoising and enhancement. The study also analysis the limitations of VAEs, like the inclination to generate blurry images, and discusses potential improvements, including hybrid models and enhanced loss functions. The results of this research enhance the comprehension of VAE's capabilities and provide a foundation for future research aimed at advancing image generation technologies.

1 Introduction

Rapid advances in the field of artificial intelligence as well as deep learning have greatly broadened the scope of applications for image generation technology in various domains, including medical imaging, computer vision, and entertainment. Image generation is a technology that automatically generates new images through algorithms, which can play a key role in tasks such as data enhancement, image completion, and synthesis. Variational autoencoder (VAE), as a generative model, has good theoretical interpretability and stability and is an important direction in image generation research. A thorough investigation into using VAE for image generation can drive advancements in the technology, enhance the quality and variety of produced images, and offer technical support for related applications.

The basic theory of VAE was first proposed by Kingma and Welling. They introduced the variational inference method to effectively solve the problem of difficulty in calculating the posterior distribution in the generative model, laying a solid foundation for subsequent research [1]. On this basis, Sohn introduced the Conditional Variational Autoencoder (CVAE), expanding the application of VAE to the generation task of structured output representation by introducing conditional variables [2]. This method has shown wide

Corresponding author: 202115030413@sust.edu.cn

application potential in complex generation tasks such as image generation and text generation. Further research has expanded the boundaries of VAE. For example, Razavi proposed VQ-VAE-2 in 2019, a vector quantized variational autoencoder that achieves a diversified generation of large-scale images by introducing vector quantization technology, especially showing significant advantages in high-fidelity image generation [3]. Another major advance was that Dai and Wipf proposed a framework for diagnosing and improving VAEs by analyzing the latent space, addressing issues such as mode collapse, and enhancing the performance of the model, as evidenced by a significant reduction in reconstruction error on the MNIST dataset [4]. Child's hierarchical VAE achieves higher log-likelihood than pixel CNN on natural image benchmarks such as CIFAR-10, ImageNet-32, and FFHQ. In addition, the VAE uses 39M parameters, while PixelCNN++ uses 53M parameters, showing higher efficiency [5].

Although VAE has made significant progress in image generation, there are still certain challenges in the detail processing and diversity of generated images. To further improve the performance of VAE in complex generation tasks, researchers continue to explore new improvement methods and application scenarios.

This paper analyzes the basic principles, architecture, and application of VAE in image generation, explores the advantages of VAE in improving image generation effects, and summarizes and evaluates the current research progress. This paper aims to advance the field by exploring how VAE can be applied to image generation technology, offering reference and guidance for future research.

2 Principles of VAE

2.1 Basic concepts

The VAE is a type of generative model that creates new data with similar characteristics by optimizing the marginal likelihood within the latent space. The fundamental concept behind VAE involves employing variational inference to estimate the intricate posterior distribution. Unlike traditional autoencoders where the encoder reduces the input data to a latent representation and the decoder then reconstructs the original data from this latent representation, VAE enhances this process with probabilistic modeling. VAE introduces a probabilistic model that assumes the latent variables follow a specific distribution, typically a multi-dimensional Gaussian distribution. It then quantifies the discrepancy between the latent representation's distribution and the true distribution using Kullback-Leibler divergence.

2.2 Architecture of VAE

The structure of VAE is composed of three main components: the encoder, the latent space, and the decoder, as shown in Figure 1.

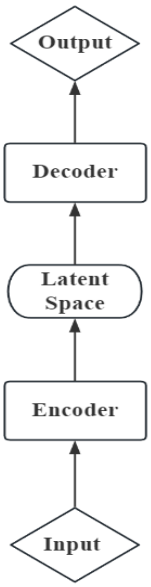

Fig. 1. Architecture of VAE (Picture credit: Original).

Encoder: The encoder is a neural network designed to transform the input data x into parameters of a distribution in the latent space, usually mean μ and variance σ^2. The encoder's output can be expressed as $q(z|x)$, where z is the latent variable.

Latent Space: The latent space is a very important part of the VAE model, capturing the underlying patterns of the data. During training, the VAE generates new data by sampling latent variables z. To promote variability in the generated data and maintain model robustness, the latent space is typically modeled as a normal distribution $z \sim N(\mu, \sigma^2)$.

Decoder: The decoder, which is also a neural network, creates data from samples z within the latent space. The resulting distribution can be denoted as $p(x|z)$. The decoder's objective is to accurately reconstruct the original input data, so its structure is usually mirror-symmetric to the encoder.

2.3 Loss function

The VAE's loss function is consistent in two main parts: reconstruction loss and KL divergence.

Reconstruction Loss: This loss quantifies the difference between the original input data x and the reconstructed data $\hat{x}$ generated by the model. It is commonly calculated using Mean Squared Error (MSE) or binary cross-entropy loss. This component of the loss function indicates how similar the VAE-generated samples are to the original input, aiming to minimize the discrepancy between the reconstructed and input data.

KL Divergence: This term is a key aspect of variational inference, measuring the difference between the approximate posterior distribution $q(z|x)$ and the prior distribution $p(z)$. The KL divergence encourages the encoder's output distribution to align closely with the standard normal distribution $N(0, I)$. where $N(0, I)$ denotes a multivariate normal distribution with a mean of 0 and an identity covariance matrix.

The overall loss function for the VAE can be formulated as:

$$L(\theta, \phi) = E_{q\phi(Z|X)}[logp\theta(x|z)] - KL(q\phi(z|x)||p(z)) \quad (1)$$

Among them, the first term, $E_{q\phi(Z|X)}[logp\theta(x|z)]$, is the reconstruction loss. And the second term, $KL(q\phi(z|x)||p(z))$, is the KL divergence. These two components jointly form the loss function of a VAE. The reconstruction loss guarantees the model's ability to generate data accurately, while the KL divergence term imposes regularization on the latent space and enhances the model's generalization capability. By optimizing the overall loss function $L(\theta, \phi)$, the VAE can generate high-quality new data that captures the underlying data distribution, ensuring that the generated samples are both diverse and stable.

2.4 Evaluation system

The following indicators are usually used to evaluate the quality of VAE models:

- Reconstruction Error: This metric assesses how well the model can recreate the input data. It is usually evaluated by computing the mean squared error or cross-entropy loss between the original data and the data generated by the model.
- KL divergence: This metric evaluates the disparity between the distribution in the latent space and the prior distribution. A KL divergence value approaching 0 signifies a greater similarity between these distributions, indicating that the VAE is making efficient use of the latent space.
- Evidence Lower Bound (ELBO): This is the optimization target of the objective function during VAE training. It is comprised of both the reconstruction error and the KL divergence, together indicating the model's overall performance.
- Generation quality and diversity: The quality and variety of the generated images can be assessed through visual inspection of the samples or by using quantitative metrics like the Fréchet Inception Distance (FID). A lower FID score suggests superior generation quality, representing a closer match between the distributions of the generated and real images, and thus indicating higher fidelity and diversity in the generated outputs.

By utilizing these evaluation metrics, one can comprehensively assess how well the VAE model performs and determine its suitability and benefits for specific tasks.

3 Applications of VAE in image generation

3.1 Image generation

VAE is widely used in image generation due to their ability to model complex data distributions in a continuous latent space. The primary advantage of VAEs in image generation lies in their ability to create a wide variety of high-quality images by drawing samples from the learned latent space. This generative process allows VAE to create entirely new images that resemble the original data, making them useful in various creative and scientific applications.

For example, Ramesh proposed the DALL·E model, which combines VAEs with the Transformer architecture for zero-shot text-to-image generation [6]. This model demonstrated a remarkable capacity to generate realistic and diverse images based on textual descriptions, achieving significant FID scores on benchmark datasets. Additionally, Esser combined variational autoencoders (VAEs) with transformers for high-resolution image synthesis, showcasing the potential of VAEs in generating detailed and high-quality images [7]. VQGAN demonstrated superior generation quality with an FID score of 4.9 on the CelebA-HQ and FFHQ datasets, compared to VQVAE-2's score of 10. On ImageNet data reconstruction, VQGAN achieved an FID score of 10.7, further underscoring its advantage in generating complex visual content, making it particularly valuable for virtual environment creation and data augmentation. The child made additional contributions to the development

of VAEs by introducing a hierarchical VAE framework, which showed enhanced performance over conventional autoregressive models in a range of image generation tasks [5]. In benchmarks such as CIFAR-10, ImageNet-32, and ImageNet-64, the very deep VAE achieved a negative log-likelihood (NLL) of 2.87, 3.80, and 3.52, respectively, outperforming all GatedPixelCNN and other non-autoregressive models. Moreover, this architecture excelled in generating high-resolution images, achieving an NLL of 0.61 on FFHQ-256 and 2.42 on FFHQ-1024. These results highlight the advantages of VAEs in efficiently representing and generating complex images, showcasing their superiority over autoregressive counterparts.

Through these advancements, VAEs have established themselves as a vital tool in creative content generation, demonstrating significant potential across various applications in fields such as entertainment, advertising, and virtual reality. The continued exploration of VAE architectures and their integration with other models promises to further enhance image generation capabilities and application diversity.

3.2 Image denoising

In image denoising, VAEs are employed to generate clean images from noisy ones. The advantage of using VAEs for this purpose stems from their capacity to learn a probabilistic representation of the data, enabling them to effectively separate noise from the true image content. This makes VAEs particularly effective in reducing noise while preserving important image details.

Recent research by Zheng illustrates that VAEs can improve image denoising by mapping complex noisy distributions to a latent space where the assumption of Additive White Gaussian Noise (AWGN) holds [8]. This transformation facilitates the effective application of traditional Gaussian denoisers, such as BM3D, thereby merging generative models with classical denoising techniques. The method proposed by Zheng incorporates a neural network to accurately estimate the noise distribution, resulting in improved denoising results. Notably, the combination of the neural network with BM3D achieves a PSNR of 38.26 and an SSIM of 0.9606 on the CC dataset, significantly outperforming many conventional denoising methods. Moreover, the analysis of noise distributions reveals that the KL divergence in the latent space (0.3821) is lower than that in the image space (0.4701), indicating that the latent space is more suitable for noise modeling, further corroborating the effectiveness of the proposed approach. Moreover, Krull et al. (2020) have explored the use of VAEs specifically designed for noise reduction in microscopy images [9]. Their approach involves explicitly modeling the noise distribution, allowing for a more tailored and effective denoising process. The results indicate that VAEs can generate denoised images that preserve fine details while significantly reducing noise levels, showcasing their robustness in medical imaging applications.

In conclusion, the application of VAEs in image denoising not only illustrates their versatility as generative models but also emphasizes their role in enhancing the quality of images across various fields, including biomedical imaging and general photography. The ongoing research in this area continues to propel advancements in denoising techniques, with VAEs serving as a pivotal tool in achieving higher fidelity in image restoration.

3.3 Image enhancement

VAE is applied in image enhancement to improve the visual quality of images by adjusting aspects like resolution, contrast, and color balance. The main advantage of using VAEs in this context is their ability to generate enhanced images that retain the original structure while improving overall image quality.

Recent studies highlight the advantages of Variational Autoencoders (VAEs) in image enhancement. For example, Pucci proposed the UW-CVGAN model, which integrates VAE into underwater image enhancement by leveraging capsule networks for feature quantization and clustering [10]. This innovative approach not only improves image quality and preserves fine details, but also significantly reduces storage space requirements, compressing the input image to one-third of its original size (from 3×256×256 to 256×16×16), achieving a compression factor of 3 times. Comparative studies show that UW-CVGAN outperforms existing methods in multiple metrics; for example, on the EUVP dataset, it achieves a UCIQE score of 6.47, a UIQM of 2.91, a PSNR of 29.11, and an Inception score of 4.39±0.5. These results highlight the model's superior performance in image reconstruction over alternatives such as VQ-GAN, and emphasize the effectiveness of VAE in adapting to environmental changes and enhancing underwater images. Zeng mainly explored the task of compressed dark image enhancement and proposed a multi-latent space mapping network based on VAE to solve the problem that existing methods will amplify compression artifacts when processing compressed dark images [11]. By adopting multi-layer VAE to learn latent features of different resolutions, the method can effectively maintain detail information and improve image quality. In addition, experimental results show that the proposed method performs well on multiple datasets (such as LOL, LSRW, and SICE). When QF is 90%, the proposed method obtains a PSNR of 19.16, an SSIM of 0.8373, and a PSNR-B of 19.06 on the LOL dataset, which are better than other comparison methods, fully demonstrating the effectiveness and advantages of the method. Additionally, Xu introduced a VAE-based approach for contrast enhancement in low-light imaging, demonstrating significant improvements in visual clarity and performance metrics, such as PSNR and SSIM, compared to conventional techniques [12]. This advancement is particularly crucial for surveillance applications, as it effectively enhances low-light images, making them more discernible and actionable.

These case studies highlight the efficacy of VAE in image enhancement, demonstrating their versatility and robustness in different enhancement tasks. This makes VAEs an important tool for advancing image quality improvement technology, enabling better visual experience in a variety of fields such as medical imaging, photography and graphic design, surveillance and security, autonomous driving, virtual reality and augmented reality, and remote sensing image processing. In summary, VAE provides significant benefits for image enhancement by effectively modeling complex distributions and generating high-fidelity images, and their broad application prospects point the way for future research and development in image processing technology.

4 VAE discussion and analysis

4.1 Advantages of VAE

- Theoretical Interpretability and Stability: VAEs are well-grounded in theory through variational inference, making them more interpretable and stable compared to other generative models. The KL divergence regularization contributes to consistent and reliable image generation.
- Versatility in Applications: VAEs prove to be efficient in a range of image generation tasks, including denoising and enhancement. They have shown success in reducing noise while preserving details and improving image resolution and quality.
- Capturing Complex Data Distributions: VAEs excel in modeling complex data distributions within the latent space, facilitating the creation of varied and high-caliber

images. This capability has been demonstrated in models that integrate VAEs with other architectures to generate diverse images from text or other inputs.

4.2 Limitations of VAE

- Blurriness of Generated Images: VAEs often produce blurry images due to the averaging effects of the mean squared error used in reconstruction loss, which can be problematic in tasks requiring fine detail.
- Limited Diversity: While VAEs are designed to generate diverse images, they sometimes struggle with variability, leading to less diverse outputs compared to other models like GANs.
- Balancing Reconstruction and Regularization: VAEs face a trade-off between reconstruction accuracy and latent space regularization. Overemphasis on one aspect can lead to underfitting or poor generalization.

4.3 Improvements and future prospects

- Hybrid Models: Combining VAEs with other models, such as GANs, can improve image sharpness and realism, addressing issues like blurriness.
- Enhanced Loss Functions: Developing more sophisticated loss functions, like perceptual loss, can help generate sharper images while maintaining the structural benefits of VAEs.
- Application-Specific Adaptations: Tailoring VAEs for specific tasks, such as medical imaging, can enhance their performance in specialized applications.
- Exploration of CVAEs: Conditional VAEs (CVAEs) offer the potential for controlled image generation, making them valuable for tasks that require structured outputs.

In summary, while VAE has strengths in stability and versatility, ongoing research is essential to address limitations like image blurriness and limited diversity. Advances in hybrid models, loss functions, and application-specific adaptations are likely to enhance VAE performance in the future.

5 Conclusion

This paper comprehensively explores the application of VAEs to image generation, covering the fundamental principles, architectures, and practical applications of VAEs. By delving into the strengths of VAEs, such as their theoretical interpretability, stability, and versatility, this study highlights their effectiveness in various image generation tasks, including denoising, enhancement, and generation of complex visual content. Throughout the study, this paper highlights the excellent performance of VAEs in modeling complex data distributions, thereby facilitating the creation of varied and superior-quality images. The analysis also discusses the successful integration of VAEs with other architectures such as Transformers, which further expands their potential for applications in tasks such as text-to-image generation. However, this paper also critically examines the limitations of VAEs, particularly the tendency to produce blurry images and the challenge of balancing reconstruction accuracy with latent space regularization. Despite these challenges, this study underscores the notable advancements in the area, including the creation of hybrid models that integrate VAEs with GANs to enhance image clarity and realism. In summary, this paper thoroughly investigated the role of VAEs in advancing image generation technology, providing valuable insights into their strengths, limitations, and prospects. The findings from this research enhance the overall

comprehension of generative models and pave the way for continued exploration and innovation in this domain.

References

1. D. P. Kingma, M. Welling, Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).
2. K. Sohn, H. Lee, X. Yan, Learning structured output representation using deep conditional generative models. Adv. Neural Inf. Process. Syst. 28 (2015).
3. A. Razavi, A. Van den Oord, O. Vinyals, Generating diverse high-fidelity images with vq-vae-2. Adv. Neural Inf. Process. Syst. 32 (2019).
4. B. Dai, D. Wipf, Diagnosing and enhancing VAE models. arXiv preprint arXiv:1903.05789 (2019).
5. R. Child, Very deep vaes generalize autoregressive models and can outperform them on images. arXiv preprint arXiv:2011.10650 (2020).
6. A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, I. Sutskever, Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning (pp. 8821-8831), PMLR, (2021, July).
7. P. Esser, R. Rombach, B. Ommer, Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12873-12883), (2021).
8. D. Zheng, S. H. Tan, X. Zhang, Z. Shi, K. Ma, C. Bao, An unsupervised deep learning approach for real-world image denoising. In Proceedings of the International Conference on Learning Representations (2021).
9. M. Prakash, A. Krull, F. Jug, Fully unsupervised diversity denoising with convolutional variational autoencoders. arXiv preprint arXiv:2006.06072 (2020).
10. R. Pucci, C. Micheloni, N. Martinel, UW-CVGAN: UnderWater image enhancement with capsules vectors quantization. arXiv preprint arXiv:2302.01144 (2023).
11. Y. Zeng, Z. Wang, Y. Liu, T. Zeng, X. Liu, X. Luo, B. Zeng, Multiple latent space mapping for compressed dark image enhancement. arXiv preprint arXiv:2403.07622 (2024).
12. X. Wu, Z. Lai, J. Zhou, X. Hou, W. Pedrycz, L. Shen, Light-aware contrastive learning for low-light image enhancement. ACM Trans. Multimedia Comput. Commun. Appl. (2024).