Dean&Francis
ISSN 2959-6157

Research and Analysis of VAE, GAN, and Diffusion Generation Models

Xiang Li[1,*],

Yang Peng[2],

Hao Zheng[3]

[1]Shanghai Concord Bilingual School, Shanghai, China
[2] Computer Science and Technology, Hunan International Economics University, Hunan, China
[3]College of Artificial Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China
*Corresponding author: 202324090008@concordschool.com.cn

Abstract:

Generative models constitute a vital component of machine learning, having evolved to find extensive application in both image processing and natural language processing domains. This paper systematically reviews the three predominant generative models: Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and Diffusion Models (DM). It provides a comprehensive analysis of their fundamental principles, respective strengths and limitations, alongside their practical applications. From the aforementioned generative approaches, it is evident that: the VAE method innovatively incorporates probabilistic distributions to enable controlled sampling, though the generated quality tends to be somewhat blurred; the GAN method employs a generative-adversarial framework, yielding higher-quality images, yet numerous challenges persist, such as training instability; Diffusion models employ a forward noise addition and reverse denoising approach, effectively addressing both aforementioned issues. They demonstrate strong pattern coverage capabilities and excellent stability, though they suffer from slow generation speeds and high computational demands. This comparative analysis provides valuable reference points for understanding the developmental trajectory of generative models and guiding their future evolution.

Keywords: Generative Models; Variational Auto Encoders; Generative Adversarial Networks; Diffusion Models; Deep Learning

1. Introduction

The core objective of generative models is to discover patterns and distributional characteristics within existing datasets, thereby generating new data that closely approximates true samples. With the rapid advancement of deep learning, these models have become pivotal technologies in computer vision (CV), natural language processing (NLP), and speech recognition. The introduction of Variational Autoencoders (VAE) has popularized probabilistic modelling in both latent space representation learning and data

ISSN 2959-6157

generation [1]. Generative Adversarial Networks (GANs) [2], grounded in game theory, have achieved a significant leap in generative quality, giving rise to numerous high-quality image generation algorithms. Additionally, the emergence of Diffusion Models (DMs) in recent years has introduced finer details and enhanced stability to image generation outcomes, making them a current research focus [3].

However, each approach possesses distinct strengths: VAE emphasises structural integrity and controllability in the latent space, yet compromises generative quality; GAN delivers superior visual fidelity but suffers from training instability; diffusion models exhibit both high generative quality and broad pattern coverage, albeit at the cost of substantial computational overhead and low operational efficiency. Thus, comparative analysis of these methodologies assists researchers in selecting appropriate models according to task requirements, facilitating future optimisation and refinement of generative frameworks. For each specific method within different categories, their respective definitions, methodological descriptions, and characteristic features are presented. These are illustrated and contrasted through practical examples, culminating in a summary and outlook for representative methods across different categories.

2. Variational Autoencoder

Variational Autoencoders (VAE) are generative models that utilise latent variables to learn data distributions [1]. They can generate new sample data with identical statistical properties to the original data. As a probabilistic model, VAE combines deep learning with variational inference principles to address the limitation of traditional autoencoders in sample generation.

2.1 Fundamental Principles

Conventional autoencoders compress input x into a latent vector z, then reconstruct a similar output. However, their latent space is discrete and uncontrollable, rendering it unsuitable for direct sample generation.

The innovation of the VAE lies in representing the latent space using a probability distribution and requiring the latent vector to adhere to a specific distribution (typically Gaussian) during training. This enables arbitrary sampling of z from the probability distribution to generate data.

2.2 Advantages and Limitations

VAEs offer extensive advantages: the latent space is regularised into a known distribution, allowing direct sampling to obtain new data; they possess continuous controllability, enabling interpolation, feature editing, and conditional generation; and numerous variant models can be easily derived (e.g., Beta-VAE [4], CVAE [5], VQVAE [6], etc.). However, VAE exhibits certain drawbacks. For instance, images generated by VAE tend to be relatively blurred, lacking the sharpness of GAN outputs. The KL divergence constraint imposes limitations on the latent space, resulting in potential loss of image detail. Furthermore, the dual-network architecture (encoder + decoder) leads to substantial computational demands and slower processing speeds.

2.3 Applications of VAEs

Beyond data generation—sampling from the latent space and decoding data matching the training distribution—VAEs serve data compression by reducing high-dimensional data into low-dimensional latent variable representations that accurately capture key features. These latent variables are then decoded back into the observable space. Finally, they enable imputation of missing data values and purification of noisy datasets.

3. Generative Adversarial Networks

Generative Adversarial Networks (GANs)represent one of the classic deep learning architectures currently employed for high-quality generation [2]. They comprise two competing participants: a Generator and a Discriminator. Their adversarial interaction constitutes a game-theoretic process where mutual constraints drive joint improvement.

3.1 Training Process

The training process primarily comprises three steps [7]. First, the discriminator is trained: the generator is fixed, meaning real data is labelled as „1“ and generated data as „0“. The discriminator‘s parameters are then trained to distinguish between „1“ and „0“. This process employs the following formula:

$$\min_{G}\max_{D} V(D,G) = E_{x\, p_{data}(x)}\left[logD(x)\right] + E_{z\, p_{z}(z)}\left[log(1-D(G(z)))\right] \quad (1)$$

Next comes training the generator: the generated data is fed into the discriminator to be classified as „1“, while the generator‘s parameters are trained to minimise the discriminator‘s recognition capability. The formula employed in this process is:

$$\min_{G} E_{z\, p_{z}(z)}\left[log(1-D(G(z)))\right] \quad (2)$$

The final step involves iteratively repeating steps 1 and 2 until equilibrium is achieved between the two datasets.

3.2 Classification and Advantages

The first variant of DCGAN incorporates convolutional structures [8], significantly enhancing generation quality and training stability. The second variant is WGAN [9], which employs Wasserstein distance to effectively address vanishing gradient issues and unsupervised learning challenges; the third category is SARA-GAN [10], utilising self-supervised representation learning (SS) to resolve discriminator forgetting; the final category is StyleGAN [11], proposing a novel generator architecture based on style transfer principles to achieve higher-quality, more controllable generation.

3.3 Applications and Outcomes

In healthcare, it can distinguish chronic diseases, predict human movement patterns, and classify tumours; it has also effectively addressed historical challenges of insufficient medical data and difficult-to-treat conditions. In the arts, it can generate photorealistic paintings and special photographic works; GANs can remove blemishes from original images, transforming noisy and chaotic scenes into cleaner, more refined visuals. Within architecture, it facilitates the generation of construction blueprints, synthesises concrete crack imagery, and processes construction and design drawings, thereby advancing modular construction and enhancing safety standards in both construction and design. Applications extend to everyday life, encompassing facial recognition and autonomous driving technologies.

4. Diffusion Models

Diffusion Models (DM)are generative models constructed through a sequential process of noise addition and gradual restoration [3]. Their fundamental principle involves progressively introducing noise to original data via a forward process, causing it to approximate a Gaussian distribution. From this basis, the reverse denoising process is learned to obtain sample data. As the generation process relies on probability distribution transformation, diffusion models theoretically excel at fitting data distributions while avoiding issues like pattern collapse.

4.1 Fundamental Principles

The training process of diffusion models comprises two stages:

Forward Diffusion Process: Starting from the true sample x_0, Gaussian noise is progressively added over T time steps to obtain a sequence of intermediate states x_1, x_2, …, x_T. This process is defined by a fixed Markov chain:

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mid) \tag{3}$$

where βt denotes the noise intensity at time step t.

Reverse Process: Starting from pure noise xT~N(0,I), noise is progressively predicted and removed to reconstruct the data sample:

$$p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)) \tag{4}$$

This training approach is equivalent to a form of the Score Matching method [12], enabling direct approximation of the gradient of the data distribution.

4.2 Representative Methods

To date, numerous outstanding methodological and theoretical advances have emerged in diffusion models. For instance, DDPM was among the earliest approaches to gain significant attention [13]. It constructs the generative process through a noise-addition and denoising mechanism, theoretically ensuring generation quality while robustly establishing a probabilistic framework for image generation. Subsequently, the DDIM method was proposed [14], featuring customisable time steps. which similarly enhances sampling speed while preserving quality, thereby improving the practicality of diffusion models. Subsequently, scholars redefined the diffusion process using Score-based Models and their Stochastic Differential Equation (SDE) formulation [15,16], enabling generative modelling within a unified probabilistic framework. Building upon this, Latent Diffusion Models further advanced diffusion processes by constructing them within latent spaces [17], significantly reducing model training times and propelling the development of large-scale text-to-image generation systems. The rise of diffusion models, exemplified by the aforementioned approaches, addresses fundamental limitations of traditional diffusion models—such as computational inefficiency and scalability challenges—and represents the foundational technology currently enabling image, speech, and multimodal generation.

4.3 Advantages and Limitations of Diffusion Models

Diffusion models demonstrate remarkable efficacy, exhibiting lower FID values than GANs in high-resolution image tasks while producing finer, more realistic and natural samples. As a stepwise modelling process, diffusion models theoretically cover all data without pattern collapse issues. Training occurs without adversarial mechanisms, requiring only a single minimised objective function, yielding greater stability than GANs. However, these models possess certain drawbacks: Firstly, the classical DDPM requires hundreds of steps of backward inference during sampling to obtain the final result [13], rendering

ISSN 2959-6157

it relatively time-consuming. Secondly, training demands substantial computational resources and time, whilst also consuming significant GPU memory during operation. Thirdly, diffusion models prove challenging to control during the initial stages, lacking flexible conditional generation methods; subsequent approaches such as Classifier Guidance have progressively addressed this limitation [18].

4.4 Application Domains

Due to its excellent generation performance and high stability, diffusion models have shone in many fields. Among them, diffusion based image generation algorithms such as Stable Diffusion , DALL·E2 , and Imagen have been applied to image generation [17,19,20]; In addition, there are also algorithms such as WaveGrad and DiffWave that use diffusion modeling to generate audio signals for speech and audio generation [21, 22]; Diffusion models are also used to accomplish various tasks related to scientific computing and molecular design, such as molecular generation, protein structure prediction, and so on; In addition, it can also be seen that diffusion models are beginning to emerge in generating videos, such as Google's VideoPoet which can generate videos [23]. Under the good performance demonstrated by the diffusion model, it can be found that it can be applied in more scenarios. On the one hand, the effectiveness of the model itself provides a good foundation for us to expand our applications; On the other hand, with the support of powerful distributed parallel computing power, the fields of multimodal generation and scientific research will benefit greatly.

5. Future Outlook

Building upon the preceding discussion, future generative models are anticipated to evolve in multiple directions. Firstly, architectural innovations may seek to combine the latent space modelling strengths of Variational Autoencoders (VAEs), the high-quality output of Generative Adversarial Networks (GANs), and the stability of diffusion models, forming complementary hybrid generative frameworks. Secondly, sampling and inference efficiency will continue to advance, potentially through optimising diffusion steps, designing more efficient explicit generation pathways, or employing model distillation to accelerate sampling while reducing computational costs. Additionally, attention must be paid to controllability and interpretability. Future models should demonstrate enhanced semantic control to generate multi-dimensional, fine-grained samples, alongside heightened transparency and explainability. Beyond this, research into cross-modal and multimodal generation is essential, enabling models to fuse information from diverse media types—text, images, speech, video—for unified modelling and generation. These holds promise for delivering natural human-machine interaction and diverse content creation methods. The expanding practical application of generative models has also exposed certain challenges, such as integrating content safety checks and copyright protection into generation systems. Risks of online misuse and infringement are increasingly prominent, necessitating focused consideration on preventing the generation of exploitable content. Moving forward, as computational resources and theoretical methodologies continue to evolve, the application scope of generative models can expand towards cross-domain and multimodal directions, enabling multi-objective and even multi-task applications. Throughout this process, continuous iteration of their operational mechanisms and ongoing refinement to address existing shortcomings must be prioritised. Only thus can artificial intelligence development be provided with more robust technical support and directional guidance.

6. Conclusion

As representative generative models, Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models fall within this category. A comparative analysis of their respective strengths and weaknesses reveals: VAEs excel in latent space modelling and controllable generation; GANs produce higher-quality samples through adversarial training mechanisms; Diffusion Models demonstrate superior pattern coverage and generative stability. From an application perspective, VAEs are suitable for latent feature analysis and controlled generation, such as data compression and missing value imputation; GANs are more commonly applied in image synthesis, style transfer, and other domains demanding high fidelity; diffusion models are primarily used for high-resolution image generation and multimodal generation tasks. All three approaches still face trade-offs between generation quality and speed, training stability, and computational cost, indicating that the development of more efficient models remains a significant challenge.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

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

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020,

63(11): 139-144.
[3] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics// International conference on machine learning. pmlr, 2015: 2256-2265.
[4] Higgins I, Matthey L, Pal A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework// International conference on learning representations. 2017.
[5] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems, 2015, 28.
[6] Van Den Oord A, Vinyals O. Neural discrete representation learning. Advances in neural information processing systems, 2017, 30.
[7] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in neural information processing systems, 2014, 27.
[8] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.
[9] Adler J, Lunz S. Banach wasserstein gan. Advances in neural information processing systems, 2018, 31.
[10] Yuan Z, Jiang M, Wang Y, et al. SARA-GAN: Self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction. Frontiers in Neuroinformatics, 2020, 14: 611666.
[11] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.
[12] Hyvärinen A, Dayan P. Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 2005, 6(4).
[13] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020, 33: 6840-6851.
[14] Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.
[15] Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 2019, 32.
[16] Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.
[17] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.
[18] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 2021, 34: 8780-8794.
[19] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022, 1(2): 3.
[20] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 2022, 35: 36479-36494.
[21] Chen N, Zhang Y, Zen H, et al. Wavegrad: Estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713, 2020.
[22] Kong Z, Ping W, Huang J, et al. Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761, 2020.
[23] Kondratyuk D, Yu L, Gu X, et al. Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125, 2023.