Image Generation Based on Diffusion Models: Quality, Efficiency, and Controllability

Xiyao Xu

School of Management, Hefei University of Technology, Anhui, 230009, China

dabinahte200405133@outlook.com

Abstract. Image generation is a crucial technology in computer vision. Currently, deep learning-based image generation models are widely used in commercial, entertainment, and industrial applications. With the rapid transformation of the digital economy, higher demands are being placed on image generation. At a time when traditional image generation models are reaching their limits, diffusion models, with their unique advantages, have gained widespread attention in academia and have now become the mainstream paradigm in the field. This paper reviews the core technological achievements of diffusion models in recent years, outlines their evolution in terms of quality, efficiency, and controllability, and summarizes the core issues of each technology. However, due to the inherent constraints of the diffusion model generation paradigm, there is a natural zero-sum game among the three dimensions of quality, efficiency, and controllability—a rigid conflict between them. Finally, this paper analyzes the current state of diffusion models and provides an outlook on future research, with a focus on integrating diffusion models with large language models.

Keywords: Diffusion model, Sampling acceleration, Conditional generation, Image generation, Large language model.

1. Introduction

The development of image generation technology can be divided into two periods, with the application of deep learning as the dividing line. Before the widespread application of deep learning in the field of image processing, generating images that conform to the laws of the real world often required the human definition of rules such as physical geometry and material properties. This was essentially a splicing of existing materials with extremely weak generalization ability. With the development of statistical learning, researchers tried to model the pixel distribution of images using probabilistic models, but it was still difficult to achieve high-complexity image generation. In 2012, Krizhevsky et al. achieved excellent results on the ImageNet problem using Convolutional Neural Networks (CNN) [1], proving the potential of deep learning in the field of image processing. Subsequently, traditional image generation models such as Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) were proposed, widely studied, and developed.

Entering the AI era, data has shifted from passive collection to active creation, and its value, circulation, and boundaries have been redefined. As a high-density information carrier, the rapid generation of personalized, high-quality images to meet the needs of commercialization, industrial production, and other fields has become a bottleneck for traditional image generation models. For example, images generated by VAEs are usually blurry [2], and although GANs can generate high-definition images, they are prone to pattern mode collapse and training is extremely unstable [3].

Although the concept of the diffusion model was proposed a long time ago [4], it only gradually came into people's view after Denoising Diffusion Probabilistic Models (DDPM) laid the foundation for the diffusion model [5]. Subsequently, in the ImageNet high-resolution image generation task [6], Dhariwal et al first used the diffusion model to beat GAN and achieved state-of-the-art results, revealing the potential of the diffusion model in image generation tasks and attracting more and more researchers to participate in the research and development of the diffusion model. With the improvement of technology in recent years [7-11], the diffusion model has become the cornerstone of the field of image generation today and has been widely used and commercialized.

In the field of image generation, quality, efficiency, and controllability are often difficult to balance simultaneously. Specifically, how to generate realistic, clear, and detailed images while ensuring consistency between text and images, and how to reduce training costs or image generation time—these three conflicting goals are problems that researchers must consider. Therefore, this paper systematically reviews the main technological evolution of diffusion models and further discusses the irreconcilable contradictions among them. Furthermore, this paper provides an outlook on future research based on the current state of diffusion models and focuses on the integration of current diffusion models with large language models.

2. Progress in Generative Quality

Sohl-Dickstein et al. first proposed the concept of the diffusion model and constructed Diffusion Probabilistic Models (DPM) [4]. This model transforms complex data distributions into simple known distributions and naturally satisfies the Markov property; its reverse diffusion process constructs a Markov chain to fit the real Markov reverse diffusion process, gradually denoising and restoring the data from Gaussian noise. The training objective of the DPM reverse diffusion process is to minimize the Evidence Lower Bound (ELBO). The objective function is complex and requires thousands of iterations, resulting in extremely low sampling efficiency. At the same time, it is limited by the network architecture at that time, which makes the model weaker than GAN in terms of sampling speed and generation quality. Therefore, DPM did not attract widespread attention when it was first proposed.

Ho et al. used the Reparameterization Trick to prove that predicting the Gaussian noise ε added to the image at time t is equivalent to predicting the mean of the backward distribution, thus transforming a complex generation problem into a denoising problem [5]. In the forward diffusion process, DDPM uses a fixed forward noise schedule, which is approximated as a constant term independent of the input, and ignores the weighting coefficients that change with time step, resulting in a simplified mean squared error loss function, thus transforming model training into a simple supervised regression problem. In terms of architecture, DDPM uses a convolutional U-Net combined with a self-attention mechanism. Although it has shortcomings in global modeling ability and scalability, the U-Net was still the mainstream choice before the full introduction of diffusion models in the Transformer architecture.

Diffusion Transformers (DiT) divide the latent feature map into a Token sequence [7] and use a pure Transformer architecture to replace the mainstream architecture, the U-Net of the diffusion model, proving the effectiveness of the Transformer architecture in the diffusion model. In addition, DiT verifies that the diffusion model satisfies the scaling law, and the image generation quality can be improved by continuously increasing the model size(number of parameters) and computational cost (Gflops). In high-resolution generation tasks, since the self-attention mechanism of the Transformer core tends to learn low-frequency information and has insufficient ability to model high-frequency information, and is limited by the performance of the VAE used at that time, DiT finds it difficult to restore high-precision details and texture.

3. Progress in Generation Efficiency

Song et al. found that the training objective of DDPM depends only on the edge distribution at each time step t, and not on the joint distribution between adjacent time steps [12]. That is, as long as the edge distribution is determined, the forward diffusion process can be redefined. Accordingly, Denoising Diffusion Implicit Models (DDIM) do not require retraining the network. It only needs to define the forward process of DDPM as non-Markovian to obtain a deterministic reverse denoising process. At the same time, the sampling process is approximately equivalent to solving a smooth Ordinary Differential Equation (ODE), and the generation speed can be greatly improved by skipping steps in sampling. Although DDIM compresses the number of steps to 20-50, it is still slower than

the single-step generation speed of GAN. It is difficult to achieve ideal efficiency in real-time generation scenarios. If the number of steps is further compressed, the quality of the generated image will be significantly reduced.

Both DDPM and DDIM operate directly in pixel space, and the massive data dimension processed by the model consumes a great deal of memory and computing power. To address this, Latent Diffusion Models (LDM) trains an autoencoder to process the high-frequency details of the image [8], compresses the image into a low-dimensional latent space, and then trains a diffusion model in the latent space based on the semantic parts preserved in the image. Although memory usage and computational efficiency have been optimized, and the number of sampling steps can be further compressed while ensuring quality, LDM still cannot meet the needs of real-time applications and lightweight adaptation to mobile devices.

Lent Consistency Models (LCM) is based on the consistency model theory. It learns an end-to-end mapping function in the latent space to map the noise representation at any time of the forward process to a noise-free representation, thereby achieving high-quality generation in one step or with a few steps. During the training phase, LCM proposes Latent Consistency Distillation (LCD) [9]. By minimizing the difference (consistency loss) in the prediction results of the teacher model (pre-trained LDM) and the student model (LCM) for the same probability flow ordinary differential equation, the solution trajectory of the teacher model is distilled into the consistency function of LCM. However, the core of LCM is to optimize the model sampling process. In real-time interactive scenarios, the fixed overhead generated by the entire link during each interaction and the post-processing overhead to compensate for the quality problem of generation, with a few steps, will seriously dilute the speed improvement of the sampling process.

4. Progress in Generative Control

Dhariwal et al. proposed Classifier Guidance (CG) to control the image generation process. By adjusting the size of the guidance scale(S), the degree of intervention of the classifier in the generation can be controlled, and a trade-off between the diversity and fidelity of the generated images can be achieved [6]. Combining the improved architecture and the classifier guidance strategy, the diffusion model has, for the first time, surpassed the image generation quality of SOTA GAN on mainstream large-scale image benchmarks. However, the control achieved by a single parameter S has a coupling between controllability and diversity, and cannot achieve multi-dimensional fine control. At the same time, the guidance effect of CG depends on the classifier trained on the closed set labeled dataset, which is difficult to adapt to unlabeled open scenes, and the model has poor generalization ability and limited controllability.

Classifier-Free Diffusion Guidance (CFG) does not require training an additional classifier. Instead, during the training phase, it randomly replaces the conditional input with an unconditional empty embedding with a fixed probability, enabling a diffusion model to jointly learn the conditional and unconditional distributions [10]. During the sampling phase, the diffusion model performs two forward propagations in parallel at the same time step to obtain the predicted noise with and without conditions. The final noise is calculated by guiding the scale weighting [10], and then back-diffusion is performed. CFG only provides coarse-grained guidance at the global semantic level, making it difficult to achieve fine control over spatial structures such as graph composition.

Zhang et al. proposed a pluggable control framework, ControlNet [11], which realizes multi-dimensional conditional control of model image generation. ControlNet freezes all parameters of the pre-trained diffusion model backbone and constructs a structurally symmetrical trainable control branch by copying the encoder of the backbone and the weights of the intermediate blocks. After inputting the spatial conditions into the control branch, the dimension alignment is completed through a zero-initialized 1x1 convolutional layer [11]. The features are added to the decoder part of the frozen backbone in the form of residual links, so that the architecture can capture the spatial structure of the conditional input with high robustness and realize fine-grained spatial constraints for image

generation. During sampling, ControlNet applies global conditional constraints to the image and cannot realize independent local control. When the input text prompts have strong semantic conflicts with the spatial condition map, the generated image will produce serious structural artifacts and even lose control of the spatial conditions.

5. The Contradiction Between Quality, Efficiency, and Controllability

The essence of the diffusion model's reverse process is to gradually denoise and map the Gaussian distribution to the real data distribution, thereby obtaining samples that conform to the real data distribution. This determines that improving the image generation accuracy (quality) requires a higher number of sampling steps, which is fundamentally contradictory to improving efficiency. DDPM sacrifices inference efficiency to achieve high-quality generation with 1000 dense sampling steps. Although DDIM subsequently compresses the number of steps to 20-50, further compression of the number of steps will result in a serious decline in quality. LCM significantly improves sampling efficiency through single-step sampling, enabling real-time interaction of the model, but the generated images are often accompanied by significant quality loss.

The core of controllability in diffusion models lies in introducing conditional constraints into the sampling process. By progressively correcting the sampling trajectory to align it with the target conditional distribution, the fewer the sampling steps, the worse the correction effect on the sampling trajectory, and the weaker the controllability. Simultaneously, the classifier or training branch introduced to achieve controllability reduces model efficiency. LCM and CFG have extremely poor compatibility, exhibiting significant guidance failure issues. ControlNet achieves spatial control through additional control branches, increasing the computational overhead during sampling. Furthermore, in practice, multiple ControlNets are often stacked, further reducing sampling efficiency.

High-quality image generation requires the model to minimize the distance between the generated distribution and the real data distribution, while controllability requires aligning the generated distribution with the target conditional distribution. The difference between the optimization objectives of the two is the root cause of the contradiction. When the CFG is increased, the model's conditional constraints on image generation are strengthened, but the generated images will have obvious quality defects; conversely, when the CFG is decreased, the generated images have higher fidelity, but the controllability will be significantly weakened. DiT achieves multi-dimensional controllable generation through the attention mechanism of the Transformer, but if the control weights (or guiding weights) are set too high, it will lead to severe overfitting of the model, causing the generated images to lose detail, diversity, and fidelity significantly.

6. Future Research Directions for Diffusion Models

Currently, diffusion models have achieved high-fidelity generation that approximates or even surpasses real images, covering most application scenarios. While the optimization of generation quality is showing significant diminishing returns, there is still considerable room for improvement in high-dimensional generation tasks. In terms of efficiency, the sampling process has been compressed to millisecond levels, enabling widespread application of real-time interaction and batch generation. Furthermore, to adapt to the computational bottlenecks of mobile devices, the model has undergone lightweight adjustments by sacrificing some quality and controllability. Controllability is the core bottleneck for further commercialization of the model. The inherent randomness of diffusion models makes the generated images almost unreproducible, and it is difficult to adapt to the strong controllability requirements of professional fields such as medicine and engineering. The black box nature of input to output means that controllability can only be achieved by narrowing the generation space through external constraints, failing to achieve precise control from the root. In summary, research directions for diffusion models include: spatiotemporal consistency modeling for high-dimensional generation tasks; innovation of traditional diffusion paradigms; research on the

correspondence between the model's denoising process and the generated semantics; and native embedding of knowledge graphs, such as professional rules.

Furthermore, the fusion of diffusion models and large-scale models is a current research hotspot. Diffusion models excel at high-fidelity multimodal generation, but their generation is often limited by weak alignment capabilities with complex text, poor long-range consistency, and poor controllability of black-box generation. Large-scale models are adept at processing and generating discrete text sequences, but even with the development of multimodal large-scale models, it remains difficult to generate continuous visualizations. To break down the barriers of single-technology capabilities, fusing the two has become a core research direction in recent years. Currently, researchers are attempting to construct a unified model architecture to handle multimodal information, but this direction still has many issues that require further refinement in future research.

Cross-modal alignment is difficult to achieve. Large models deal with discrete token sequences, while diffusion models deal with continuous pixel data. The conflict between discrete and pixel data, i.e., how to achieve seamless interaction between discrete linguistic features and continuous visual features in the same space, remains a current challenge.

Output length rigidity is a concern. Large models based on autoregressive models can naturally terminate generation via EOS (Earth Array Transmission) and dynamically adjust the output length. However, diffusion models require pre-setting the output length before generation. This can lead to simple tasks being forced into long sequences, wasting computational resources; while complex long tasks are forcibly truncated due to insufficient pre-set length, resulting in information loss. While breakthroughs have been made in controlling the length of text generation, length rigidity in multimodal scenarios such as images and videos remains a direction for future research.

There is a lack of unified evaluation criteria. How to cover core dimensions such as semantic understanding ability, generation fidelity, cross-modal alignment, efficiency, and cost remains a key focus of future research.

7. Conclusion

This article reviews the main technological evolution of diffusion models from three aspects: quality, efficiency, and controllability. With the improvement of technology, diffusion models have become the core of the field of image generation, occupying a State-of-the-Art (SOTA) technological position. They can generate high-fidelity images with complex textures. The compression of sampling steps has enabled the widespread application of real-time interaction. Controllability has also achieved a leapfrog development from early coarse-grained guidance to multimodal and fine-grained approaches. However, due to the inherent constraints of the diffusion model's generation paradigm, namely the contradiction between continuous multi-step sampling to fit the real data, the compression of sampling steps, and the gradual injection of control signals to align with the target distribution, the diffusion model often struggles to achieve optimality in all three dimensions simultaneously.

Optimization of diffusion models is in a stage of rapidly diminishing marginal returns, but multidimensional generation and paradigm innovation remain research directions that require breakthroughs. Controllability is a core challenge for the further application and promotion of diffusion models; the irreproducibility of black-box generation and insufficient controllability in specialized domains remain problems that urgently need to be addressed. Furthermore, the integration of diffusion models with large-scale models achieves complementary advantages, providing new ideas for model optimization, which is also a current research direction.

References

[1] Krizhevsky Alex, et al. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2012, 60: 84-90.

[2] Zhai Zhengli, et al. A review of variational autoencoder models. Computer Engineering and Applications, 2019, 55(03): 1-9.

[3] Wang Kunfeng, et al. Research progress and prospects of generative adversarial networks (GAN). Acta Automatica Sinica, 2017, 43(03): 321-332. doi:10.16383/j.aas.2017.y000003.

[4] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, PMLR, 2015: 2256-2265.

[5] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[6] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems, 2021, 34: 8780-8794.

[7] Peebles W, Xie S. Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 4195-4205.

[8] Rombach R. High-Resolution Image Synthesis with Latent Diffusion Models [Internet]. arXiv [cs.CV], 2021.

[9] Luo S, Tan Y, Huang L, et al. Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378, 2023.

[10] Ho J, Salimans T. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

[11] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 3836-3847.

[12] Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.