The 36th IPPR Conference on Computer Vision,Graphs, and Image Processing (CVGIP2023),
August 20~22, 2023,National Quemoy University, Kinmen

Survey of Acceleration Techniques for Text-to-Image Generation using Diffusion Models

[1]*Augustine Tsai*（蔡岳廷）, [1]*Kuo-Hua Wu*（吳國華）, [2]*Chiou-Shann Fuh*（傅楸善）

[1]Institute for Information Industry, Taipei, Taiwan
[2]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan,

[*]E-mail: atsai@iii.org.tw , khwu@iii.org.tw, fuh@csie.ntu.edu.tw

ABSTRACT

Text-to-image generation is a challenging task in the field of artificial intelligence, where the goal is to synthesize visually plausible images from textual descriptions. Diffusion models have recently emerged as a promising approach for this task, offering iterative refinement processes that progressively enhance the generated images based on the provided text. However, the application of diffusion models to text-to-image generation can be computationally demanding, hindering their real-world scalability and practicality. This survey paper presents a comprehensive analysis of acceleration techniques for diffusion models applied to text-to-image generation, aiming to address these challenges and pave the way for more efficient and effective image synthesis.

Keywords: *Diffusion Models, Text-to-Image*

1. INTRODUCTION

Text-to-image generation is a challenging task in the field of artificial intelligence, where the goal is to synthesize visually plausible images from textual descriptions. Diffusion models have recently emerged as a promising approach for this task, offering iterative refinement processes that progressively enhance the generated images based on the provided text. However, the application of diffusion models to text-to-image generation can be computationally demanding, hindering their real-world scalability and practicality. This survey paper presents a comprehensive analysis of acceleration techniques for diffusion models applied to text-to-image generation, aiming to address these challenges and pave the way for more efficient and effective image synthesis.

The survey starts by providing an overview of the foundational concepts of diffusion models and their application to text-to-image synthesis. It explores the architecture and mechanics of diffusion models and highlights their capabilities in generating coherent images from textual prompts. Additionally, the survey examines the challenges associated with the computational complexity of diffusion models and the necessity for acceleration techniques to improve their scalability.

The core of the survey focuses on a systematic review of diverse acceleration methods tailored specifically for text-to-image generation using diffusion models. We will be discussing three approaches: training-based, training-free sampling, and parallel computing. They are explored for their potential to expedite the generation process and alleviate the computational burden. The survey also delves into the use of hardware accelerators, such as GPUs and TPUs, as well as specialized processing units designed to optimize the performance of diffusion models in text-to-image synthesis tasks.

To quantitatively evaluate the performance of the acceleration techniques, a set of comprehensive experiments is conducted using benchmark text-to-image datasets. Metrics like image quality, generation speed, and computational resource consumption are employed to assess the effectiveness of each acceleration method. By comparing the results obtained from various techniques, the survey identifies the most promising approaches to enhance the efficiency of diffusion-based text-to-image generation.

The survey also discusses the implications of acceleration on diffusion models, addressing potential trade-offs between generation speed and image quality. It highlights the importance of striking a balance between computational efficiency and preserving the fidelity of the generated images to ensure the practical viability of diffusion models for real-world applications.

In conclusion, this survey serves as a comprehensive guide to researchers and practitioners seeking to accelerate diffusion models. By exploring a wide range of acceleration techniques and their implications, we aim to promote the widespread adoption of diffusion-based methods in real-world applications, where real-time, high-quality image synthesis from textual descriptions is crucial.

2. GENRAL BACKGROUND

Diffusion Probabilistic Models (DPM) is the generation process starts from a fixed prior distribution (usually a simple distribution like Gaussian noise) in the latent space. This noise is transformed through a series of learnable transformations, often modeled using neural networks. The transformed latent variables are then mapped to the data space, generating the final output (e.g., images). The noise acts as the starting point, and the transformation progressively enhances the noise into the desired data distribution.

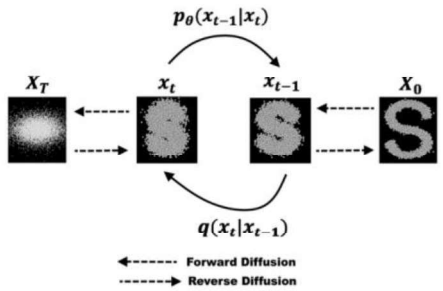


Figure 1. Diffusion Probabilistic Model (Source: [7])

De-noising Diffusion Probabilistic Models take (DDPM) a different approach to the generation process. Instead of starting from a fixed prior distribution, DDPM begins with a noise tensor. This noise tensor is conditioned on the data (image) and then de-noised through a series of transformations, similar to DPM. However, in DDPM, the noise tensor is gradually refined, and this de-noising process converges into the final image. The key difference is that DDPM performs a reverse diffusion process, starting from noise and refining it to generate the image, while DPM starts from noise and progressively transforms it to the image. Diffusion is composed of forward and backward process, equations (1) and (2)

Forward Process

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = N(\mathbf{x}_t;\mathbf{x}_{t-1},\beta_t\mathbf{I})$$
$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \quad (1)$$

Reverse Process:

$$p_\theta(\mathbf{x}_0{:}\,T) = \mathrm{p}(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}\,|\mathbf{x}_t)$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = N\big(\mathbf{x}_{t-1};\mu_\theta(\mathbf{x}_t,t),\sigma_\theta(\mathbf{x}_t,t)\big) \quad (2)$$

Song et. al. [15] Introduce a stochastic differential equation (SDE) capable of achieving a seamless transformation from a complex data distribution to a known prior distribution through the gradual introduction of noise. Additionally, develop a corresponding reverse-time SDE, which smoothly reverts the prior distribution back into the original data distribution by progressively eliminating the injected noise.

$$\mathrm{d}x_t = f(t)x_t dt + g(t)d\omega_t, \quad x_0{\sim}q(x_0) \quad (3)$$

The Fréchet inception distance (FID) [16] is used to measure the quality of image generation. The FID compares the distribution of generated images with the distribution of a set of real images, it is also a good measure for visual acuity.

3. ACCELERATION TECHNIQUES

Despite their high-fidelity generation capabilities, diffusion models suffer from slow sampling speeds, limiting their practical applicability. To address this issue, various advanced techniques fall into four categories: Training based, training schedule enhancement, training-free acceleration, and parallel processing. In this section, we provide a concise overview of these methods.

3.1 Training based

Training based methods are knowledge distilation in which efficient smaller networks are created by transferring knowledge from larger, complex teacher models to simpler student models [18, 19]. In the context of diffusion models, distillation techniques aim to achieve faster sample generation or use smaller networks. Similar to traditional distillation methods applied to large pre-trained models, distillation in diffusion models is based on the concept of alignment, seeking to minimize differences between generated samples and their corresponding original samples. Generally, the distillation processes in diffusion models involve modifying trajectories to optimize transportation costs between different distributions. These optimal mappings result in shorter and more efficient paths, leading to reduced generation costs and controllable generation.

DDPM normally uses a *1,000* step discretization of the SDE (Stochastic Differential Equation). These de-noising steps are computed sequentially and the sampling can be very slow due to a full pass through the neural network $p(\theta)$ each step. As a result, popular works such as DDIM and DPMSolver [5] used coarser discretization to reduce de-noising steps with the price of trading quality with speed. Other people try to use ODE (Ordinary Differential Equation) trajectory method pass the teacher's model to student's model using ODE to map prior distribution to target distribution along an optimum

path [8]. Salman [8] presented a two-step approach, one is a new parameterizations which offer increased stability using few sampling steps. Second, a method to distill a trained deterministic diffusion sampler, and generate a new diffusion model that takes half of sampling steps.–Luman [9] created a new connection between knowledge distilling and image generation, which can reduce multiple sampling steps into single step.

Song [10] proposed a one-step mapping from noise to image called consistency model. It can achieve the new state-of the-art FID of 3.55 on CIFAR-10 and 6.20 on ImageNet 64 x 64 for one-step generation.

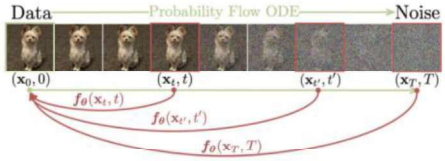


Figure , Consistency map: one step mapping, source:[10]

3.2 Training-free Sampling

Training-free methods aim to expedite the sampling process of pre-trained diffusion models by utilizing more advanced samplers, thus eliminating the necessity for model re-training. This section classifies these methods into several aspects, including the acceleration of the diffusion ODE and SDE samplers.
Song [4] proposed a De-noising Diffusion Implicit Model (DDIM) where the model can be trained with an arbitrary forward steps but only sample from few of them in the generative process. ODE sampling can accelerate the process, but in the cost of quality scarification. On the contrary, SDE always present superior quality. Jolicoeur-Martineau [18] designed a more efficient SDE solver. The new solver is equipped with an adaptive step size. This approach generates data in 2x to 10x speed up.

3.3 Parallel Computing

Instead of reducing the de-noising steps which will compromise the synthesis quality, Shih et. al. [1] proposed to parallelize sampling via Picard iterations, this approach speculates the future solution of de-noising steps and refine it iteratively until convergence. Several de-noising steps are carried in parallel. The essence of this approach hinges on improving the sample latency reduce the time for generate simple sample by solving the de-noising steps in parallel. They improved the sampling speed 2 to 4 times. The time for 100-steps diffusion policy can be dropped down to 0.2s, and 1,000-steps are cut down to 16s. And there is no obvious drop of the FID and CLIP score.

The Picard's method, primarily employed for approximating solutions to differential equations, operates iteratively. It starts with initial guess of y_0, employs successive approximation, gradually refining the numerical results with each iteration. The process generates a sequence of approximations $x_1(v), x_2(v), \ldots, x_T(v)$, to the solution of the differential equations. Each subsequent approximation is derived from one or more previous approximations, leading to increasing accuracy in the results. An ODE is defined by a drifting function $f(x, v)$

$$x_t^{k+1} = x_0 + \int_0^t f(x, v)dv \quad (1)$$

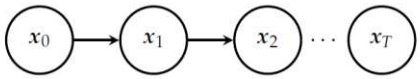


Figure: Sequencial graph, source [1]

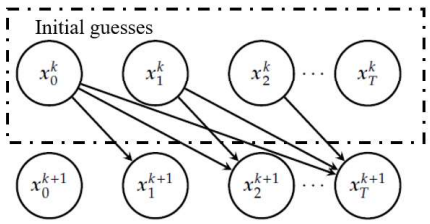


Figure Picard Iteration, source [1]

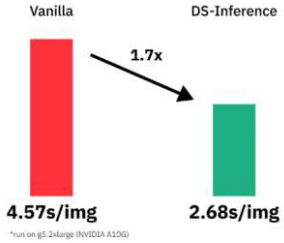
In order to compute Picard iteration numerically, equation (1) can be re-written in a discrete form, equation (2).

$$x_t^{k+1} = x_0^k + \frac{1}{T}\sum_{i=0}^{t-1} f(x_i^k, i/T) \quad (2)$$

In equation (2), time t update depends on all previous time steps, this way illustrates skip dependencies in figure, and allows for rapid propagation of information along the chain and enhances the speed of sampling.
Picard iteration requires to store the complete array of *x(0:T)* which can be prohibitively to fit to the GPU memory. Instead, they suggest a batch processing which open perform Picard iteration only in a window size of *p*, *x(t, t+p)*. However, Picard iteration has its own limitation, the parallelization require iterations till convergence, the total number of model evaluation increase substantially.

The other parallel methodology is DeepSpeed (DS)[17] framework developed by Microsoft. DS offers an inference extension for workload speed-up. The

extension, inference-v2, 2023combine model parallelism technology such as tensor, pipeline-parallelism, with custom optimized CUDA kernels. DS provides a seamless inference mode for compatible transformer based models trained using DeepSpeed, Megatron, and HuggingFace



Stable diffusion optimized by deep-speed can decrease model latency from 4.57s to 2.68s, a 1.7x speedup.

6. CONCLUSION

In conclusion, this survey serves as a comprehensive guide to researchers and practitioners seeking to accelerate diffusion models. By exploring a wide range of acceleration techniques and their implications, we aim to promote the widespread adoption of diffusion-based methods in real-world applications, where real-time, high-quality image synthesis from textual descriptions is crucial.

REFERENCES

[1] Shih, Andy and Belkhale, Suneel and Ermon, Stefano and Sadigh, Dorsa and Anari, Nima, "Parallel Sampling of Diffusion Models ," *arXiv:2305.16317,* 2023.

[2] H. Cao, C. Tan, Z. Gao, Y. Xu, G.Chen, P. Heng, Z. Li, "A Survey on Generative Diffusion Model," *arXiv:2209.02646*, 2023.

[4] J. Song, C. Meng, and S. Ermon. "Denoising Diffusion Implicit Models". *In 9th International Conference on Learning Representations*, ICLR, 2021.

[5]. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. "Dpm-solver: Afast ode solver for diffusion probabilistic model sampling in around 10 steps". *In: arXiv preprint arXiv:2206.00927*, 2022.

[6] Deep Speed acceleration on Diffusion Models *https://www.philschmid.de/stable-diffusion-deepspeed-inference*

[7] J. Siddiqui "Diffusion Made Easy," *https://towardsdatascience.com/diffusion-models-made-easy-8414298ce4da,* access 2023.

[8] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv:2202.00512*, 2022.

[9] E. Luhman and T. Luhman, "Knowledge distillation in iterative generative models for improved sampling speed, *arXiv:2101.02388*, 2021.

[10] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," *arXiv: 2303.01469,* 2023.

[11] W. Sun, D. Chen, C.Wang, D. Ye, Y. Feng, and C. Chen, "Accelerating diffusion sampling with classifier-based feature distillation," *arXiv:2211.12039*, 2022.

[12] C. Villani, "Topics in optimal transportation," Graduate Studies in Mathematics, 2003.

[13] X. Liu, C. Gong et al., "Flow straight and fast: Learning to generate and transfer data with rectified flow," *in NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[14] S. Lee, B. Kim, and J. C. Ye, "Minimizing trajectory curvature of ode-based generative models, *arXiv:2301.12003,* 2023.

[15] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv:2011.13456*, 2020.

[16] 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", Advances in Neural Information Processing Systems, 2017. arXiv:1706.08500.

[17] R. Y. Aminabadi et al., "DeepSpeed- Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale," SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*,* 2022.

[18] A. Jolicoeur-Martineau, K. Li, R. Pich´e-Taillefer, T. Kachman, and I. Mitliagkas, "Gotta go fast when generating data with score-based models," arXiv:2105.14080, 2021