BoostDiff: Double the Inference Speed of Diffusion Models with Parameter-Efficient Distillation

Anonymous submission

Abstract

In this work, we aim to accelerate inference speed in top-performing models, such as stable diffusion, with limited training data and computational budget. Contrary to existing work primarily aimed at reducing diffusion steps, we introduce an orthogonal method that distills the conditional and unconditional models from classifier-free guidance into a single model, effectively doubling the inference speed. For the first time, we demonstrate that distillation in a parameter-efficient way can yield results comparable to the original diffusion model. Empirically, we double the inference speed of stable diffusion v2.1 with parameter-efficient methods such as LoRA/BitFit using 1.66M/0.34M learnable parameters, which is only 0.2%/0.04% of the origin model. Moreover, our method is compatible with existing acceleration methods, allowing for additional speed enhancements when combined.

Introduction

Diffusion Models have emerged as a powerful method for image generation with high quality. However, the practical application is often impeded by their computational inefficiency during inference. Recently guided diffusion models(Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2021; Saharia et al. 2022) have been widely recognized and extensively used due to excellent quality and explicit semantics of the generated images. Iterative process(Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Song et al. 2020) and classifier-free guidance(Ho 2022) are critical for sample quality. However, these two mechanisms are also responsible for low sample efficiency.

Contrary to existing works which focus on reducing diffusion steps(Lu et al. 2022; Zhao et al. 2023; Salimans and Ho 2022; Meng et al. 2023; Song et al. 2023; Berthelot et al. 2023; Shao et al. 2023), we aim to accelerate the classifier-free guidance process by distilling the conditional and unconditional models into a single model, effectively doubling the inference speed.

To achieve effective distillation, our key observation is that the conditional and unconditional model (i.e. two teacher models) shares the same model architecture and weights, but with different model inputs. Therefore, the student model for distillation can be initialized with the same model architecture and weights, and adapted in a parameter-efficient way. We quantitatively and qualitatively demonstrate that through LoRA(Hu et al. 2021)/BitFit(Ben-Zaken, Ravfogel, and Goldberg 2021) tuning, the distilled single student model can achieve comparable results to combined results of two teacher models from stable diffusion v2.1, with 1.66M/0.34M learnable parameters, which is only 0.2%/0.04% of the origin model.

Our contributions include: a) proposing an effective and efficient distillation method that doubles the inference speed of diffusion models with limited training data and computational budget; b) demonstrating that parameter-efficient methods are effective for distillation of diffusion models; c) demonstrating that the proposed method is orthogonal to existing methods and can further boost the inference speed when combined.

Related Work

Fast Sampling In general, fast sampling of diffusion models can be categorized into training-free and training-based methods. Training-free methods are mainly designed to reduce sampling time through quantification(Li et al. 2023; Wang et al. 2023), pruning(Fang, Ma, and Wang 2023), and optimized samplers(Lu et al. 2022; Zhao et al. 2023). While for training-based methods, fast sampling is principally accomplished via distilling a few-step student model(Salimans and Ho 2022; Meng et al. 2023; Song et al. 2023; Berthelot et al. 2023; Shao et al. 2023).

Parameter-Efficient Tuning Parameter-efficient tuning is of great importance for low-cost tuning of large-scale pre-trained models, both in the area of natural language processing(Houlsby et al. 2019; He et al. 2021) and computer vision(Jia et al. 2022; Xiang et al. 2023). Some methods prepend a few learnable vectors to activations or inputs, such as prefix-tuning(Li and Liang 2021) and prompt-tuning(Lester, Al-Rfou, and Constant 2021). The other methods introduce a small amount of learnable parameters, for examples, Adapter-tuning(Pfeiffer et al. 2020), LoRA-tuning(Hu et al. 2021), and BitFit-tuning(Ben-Zaken, Ravfogel, and Goldberg 2021). Extensive study has qualitatively and quantitatively demonstrated that parameter-efficient tuning can achieve considerable results in various domains.

Background

Stable Diffusion Stable Diffusion is a pretrained text-guided diffusion model. Given samples x from a data distribution $p_{data}(x)$, noise scheduling functions α_t and σ_t, a text-guided diffusion model $\hat{x}_\theta$ is trained via minimizing the weighted mean squared error

$$E_{t\sim U[0,1],x\sim p_{data}(x)}[\omega(\lambda_t)\,\|\hat{x}_\theta(z_t, e_{text}) - x\|_2^2], \quad (1)$$

where θ is the parameters of text-guided diffusion model, $\lambda_t = log(\alpha_t^2/\sigma_t^2)$ is a signal-to-noise ratio, $\omega(\lambda_t)$ is a prespecified weighting function, $z_t \sim q(z_t|x)$, and e_{text} is the embedding of caption or empty string extracted from a pretrained text encoder.

LoRA LoRA adds trainable pairs of rank decomposition matrices δW to existing weight matrices W_{origin}

$$W = W_{origin} + \delta W = W_{origin} + A * B, \quad (2)$$

where $W_{origin} \in \mathbb{R}^{d_{in}\times d_{out}}$, $A \in \mathbb{R}^{d_{in}\times r}$, $B \in \mathbb{R}^{r\times d_{out}}$, and r is the rank.

BitFit BitFit is a parameter-efficient approach that tunes only the bias of each linear projection.

The Proposed Method

As illustrated in Figure 1, we propose to distill a student model to match the combined output of conditional and unconditional teacher models in a parameter-efficient approach. The combined output is modulated via a guidance scale variable w range from w_{min} to w_{max}, seeking a trade-off between sample quality and diversity, as the following equation

$$\hat{x}_\theta^w(z_t, e_{text}) = \hat{x}_\theta(z_t, e_{empty}) + w * (\hat{x}_\theta(z_t, e_{caption}) - \hat{x}_\theta(z_t, e_{empty})). \quad (3)$$

For simplicity, we select one guidance scale w_{sel} to calculate the combined output and optimize the student model using the following objective

$$E_{t\sim U[0,1],x\sim p_{data}(x)}[\omega(\lambda_t)||\hat{x}_{\theta+\triangle\theta}(z_t, e_{caption}) - \hat{x}_\theta^{w_{sel}}(z_t, e_{text})||_2^2], \quad (4)$$

where θ is the parameters of teacher model, $\triangle\theta$ is the learnable parameters of student model, $z_t \sim q(z_t|x)$, and $\omega(\lambda_t)$ is the loss weight function. Various parameter-efficient methods can be applied, for examples, prefix-tuning, prompt-tuning, adapter-tuning, LoRA-tuning, and BitFit-tuning. In this paper, we take two cases of LoRA-tuning and BitFit-tuning to demonstrate effectiveness of the proposed method. For LoRA-tuning, the learnable low rank matrices are attached to weights from attention modules of guided diffusion models. For BitFit-tuning, we simply finetune all the bias parameters of guided diffusion models.

Experiment

Implemented Details We focus on the text-guided image generation diffusion model Stable Diffusion v2.1 base pretrained on subset of LAION-5B(Schuhmann et al. 2022).

Table 1: **FID performance of parameter-efficient distillation for Stable Diffusion v2.1 base model at 512x512 resolution.** The results are evaluated on 10,000 captions from the COCO2017(Lin et al. 2014) validation set.

Method	Sampler	Guidance Scale	FID-20step	FID-50step
-	DDIM	3.0	15.0	14.4
-	DDIM	1.0	34.7	28.1
LoRA-R8	DDIM	1.0	15.6	14.5
BitFit	DDIM	1.0	15.4	14.5
-	DPM++	3.0	14.2	14.1
-	DPM++	1.0	28.6	28.4
LoRA-R8	DPM++	1.0	14.5	14.5
BitFit	DPM++	1.0	14.5	14.4

Similar to the pretraining stage, we use a subset of LAION-5B as our training data, which has 474,741 image-caption pairs. We train the student model for 1600 gradient updates with constant loss(Salimans and Ho 2022) in a parameter-efficient approach. Specifically, for LoRA-tuning, the learnable parameters are the low rank matrices attached to the attention modules of the student model and the rank is 8. While for BitFit-tuning, the learnable parameters are all the bias parameters of the student model. During training, learning rate is 5e-5 and batch size is 512.

Quantitative Results As shown in Table 1, FID performance rapidly increases without classifier-free guidance, for both DDIM sampler and DPM++ sampler. After LoRA/BitFit distilling, FID performance of a student model is decreased back to the value that comparable to the combined output of two teacher models.

Qualitative Results As illustrated in Figure 2 and Figure 3, after LoRA/BitFit distilling, text-guided samples of the single model are visually comparable to those generated by two teacher models.

Conclusion

In this paper, we propose a parameter-efficient distillation approach for fast sampling of guided diffusion models. We quantitatively and qualitatively show that a student model can be distilled to match the combined output of two teacher models via LoRA/BitFit tuning. With 1.66M/0.34M parameters updated, which is only 0.2%/0.04% of the origin model, the student model can achieve comparable results to the combined results of two teacher models. In the future work, we plan to experiment on parameter-efficient progressive distillation for fast sampling of guided diffusion models.

References

Ben-Zaken, E.; Ravfogel, S.; and Goldberg, Y. 2021. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. *ArXiv*, abs/2106.10199.

Berthelot, D.; Autef, A.; Lin, J.; Yap, D. A.; Zhai, S.; Hu, S.; Zheng, D.; Talbot, W.; and Gu, E. 2023. TRACT: Denoising Diffusion Models with Transitive Closure Time-Distillation. *ArXiv*, abs/2303.04248.

Fang, G.; Ma, X.; and Wang, X. 2023. Structural Pruning for Diffusion Models. *ArXiv*, abs/2305.10924.

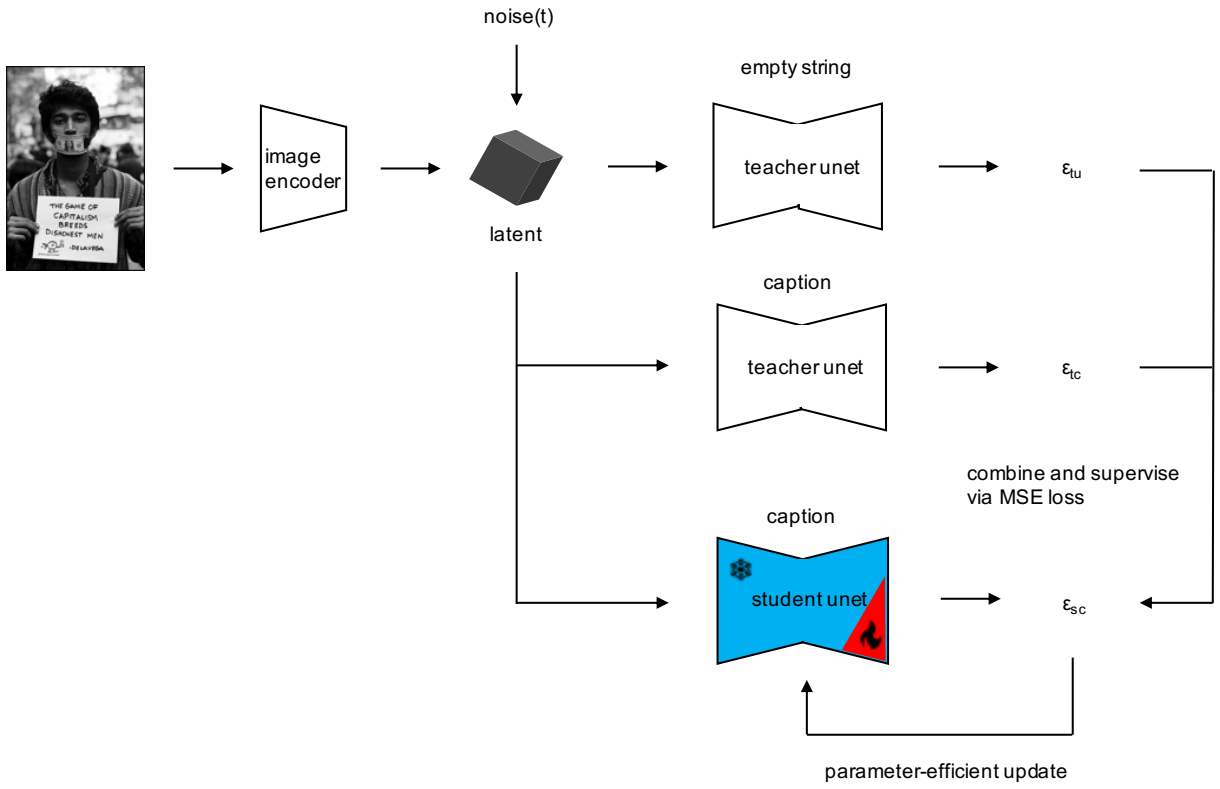

Figure 1: **The framework of BoostDiff, the proposed parameter-efficient approach to distill guided diffusion models.**

He, J.; Zhou, C.; Ma, X.; Berg-Kirkpatrick, T.; and Neubig, G. 2021. Towards a Unified View of Parameter-Efficient Transfer Learning. *ArXiv*, abs/2110.04366.

Ho, J. 2022. Classifier-Free Diffusion Guidance. *ArXiv*, abs/2207.12598.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In *International Conference on Machine Learning*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*, abs/2106.09685.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, 709–727. Springer.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. *ArXiv*, abs/2104.08691.

Li, X.; Lian, L.; Liu, Y.; Yang, H.; Dong, Z.; Kang, D.; Zhang, S.; and Keutzer, K. 2023. Q-Diffusion: Quantizing Diffusion Models. *ArXiv*, abs/2302.04304.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, abs/2101.00190.

Lin, T.-Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *ArXiv*, abs/2211.01095.

Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; and Salimans, T. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14297–14306.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*.

Pfeiffer, J.; Kamath, A.; Rücklé, A.; Cho, K.; and Gurevych, I. 2020. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. *ArXiv*, abs/2005.00247.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen,

M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, abs/2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Salimans, T.; and Ho, J. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. *ArXiv*, abs/2202.00512.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402.

Shao, S.; Dai, X.; Yin, S.; Li, L.; Chen, H.; and Hu, Y. 2023. Catch-Up Distillation: You Only Need to Train Once for Accelerating Sampling. *ArXiv*, abs/2305.10769.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *ArXiv*, abs/2010.02502.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency Models. *ArXiv*, abs/2303.01469.

Song, Y.; Sohl-Dickstein, J. N.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. *ArXiv*, abs/2011.13456.

Wang, C.; Wang, Z.; Xu, X.; Tang, Y.; Zhou, J.; and Lu, J. 2023. Towards Accurate Data-free Quantization for Diffusion Models. *ArXiv*, abs/2305.18723.

Xiang, C.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023. A Closer Look at Parameter-Efficient Tuning in Diffusion Models. *ArXiv*, abs/2303.18181.

Zhao, W.; Bai, L.; Rao, Y.; Zhou, J.; and Lu, J. 2023. UniPC: A Unified Predictor-Corrector Framework for Fast Sampling of Diffusion Models. *ArXiv*, abs/2302.04867.

Figure 2: **Qualitative results of random 512x512 text-guided samples with DDIM sampler in 50 steps.** From up to bottom, the texts are respectively ”a glass of orange juice”, ”a mountain”, ”a penguin standing on a sidewalk”, and ”beautiful lake and trees, professional photography”. From left to right, the models are respectively conditional and unconditional Stable Diffusion v2.1 base with guidance scale of 3.0, conditional Stable Diffusion v2.1 base, conditional Stable Diffusion v2.1 base after LoRA distillation, and conditional Stable Diffusion v2.1 base after BitFit distillation.

Figure 3: **Qualitative results of random 512x512 text-guided samples with DPM++ sampler in 20 steps.** From up to bottom, the texts are respectively ”a glass of orange juice”, ”a mountain”, ”a penguin standing on a sidewalk”, and ”beautiful lake and trees, professional photography”. From left to right, the models are respectively conditional and unconditional Stable Diffusion v2.1 base with guidance scale of 3.0, conditional Stable Diffusion v2.1 base, conditional Stable Diffusion v2.1 base after LoRA distillation, and conditional Stable Diffusion v2.1 base after BitFit distillation.