Enhancing Image Generation with LCM LoRA Distillation: A Deep Dive into Advanced Model Optimization Techniques

Cheryl Ngai
University of California, Davis
cyngai@ucdavis.edu

Irene Chang
University of California, Davis
seoch@ucdavis.edu

Karamjeet Gulati
University of California, Davis
ksgulati@ucdavis.edu

The link presentation video is here.
The link to the code is here.

1 Introduction & Motivations

The pursuit of high-quality text-to-image generation with efficient inference remains a key challenge in the field of deep learning. While models like Stable Diffusion (Rombach et al., 2022b) have demonstrated remarkable capabilities, achieving both visual fidelity and computational efficiency requires sophisticated approaches. This research presents a novel methodology that combines LoRA (Low-Rank Adaptation) (Hu et al., 2021), Latent Consistency Model (LCM) training, and knowledge distillation (Hinton et al., 2015) techniques to significantly enhance Stable Diffusion's performance. Our approach leverages the power of pre-trained models by focusing on consistent latent representations across diffusion timesteps. This is achieved through LCM training, which employs a specialized loss function to guide the learning process, emphasizing the coherence of latent codes across different diffusion stages. To further optimize performance and maintain computational efficiency, we introduce LoRA distillation. LoRA is a lightweight network that effectively adapts the pre-trained model's weights. Combining with knowledge distillation techniques, it allows the transfer of valuable knowledge from a larger, more complex model (LCM) to a smaller, more efficient model (LoRA). This knowledge transfer facilitates efficient knowledge acquisition and enables the LoRA network to mimic the LCM's performance with significantly reduced computational overhead. The resulting model demonstrates a remarkable ability to generate visually appealing images with minimal computational cost, even when trained on extensive datasets. This research provides valuable insights for developers and researchers seeking to optimize text-to-image generation techniques, particularly in resource-constrained environments.

2 Related Work

2.1 Image Synthesis using Stable Diffusion

Image synthesis, a prominent subfield of computer vision, entails the generation of images from textual instructions or existing image inputs. One of the foremost methodologies in this domain is diffusion models (Ho et al., 2020). Notably, recent advancements in diffusion models have yielded remarkable outcomes in image synthesis (Ho et al., 2020; Song et al., 2020), establishing themselves as the benchmark in class-conditional image generation (Dhariwal and Nichol, 2021; Ho et al., 2022). Stable diffusion models (Rombach et al., 2022a) stands out among these techniques, offering a streamlined approach to achieving state-of-the-art results in class-conditional image synthesis while markedly reducing computational overhead compared to conventional diffusion models.

2.2 Improving Stable Diffusion Model

In this project, we employ knowledge distillation (Hinton et al., 2015), LoRA (Hu et al., 2021), and LCM(Latent Consistency Model)-LoRA (Luo et al., 2023) to improve the existing stable diffusion model. Stable diffusion requires significant computational recourse for training and inference. Knowledge distillation uses a larger and more complex model called "teacher" to create a smaller and simpler model called "student" that replicate the behavior of the "teacher". This technique can improve a stable diffusion model in terms of efficiency and scalability without substantial loss in performance. In addition, LoRA creates a more lightweight and faster model by reducing the number of parameters from an existing model. By incorporating LoRA into a stable diffusion model, we can effectively reduce the resource usage while maintaining the quality of the output. LCM-LoRA extend this concept by inserting a small number LoRA layers and training them instead of the full model so it can learn the behavior of the original model without serious quality loss. This enables image generations with faster speed and less memory consumption.

3 Method

3.1 Problem Definition

Stable Diffusion has emerged as a powerful text-to-image generation model, enabling artists and content creators to produce high-quality visual content from textual descriptions. However, one of the challenges faced by users of fine-tuned Stable Diffusion models is the computational overhead associated with inference, which can result in significant latency and hinder real-time applications.

The primary objective of our project is to improve the inference speed of fine-tuned Stable Diffusion models without compromising their output quality. To achieve this goal, we propose a novel approach that combines Latent Consistency Model (LCM) training and Low-Rank Adaptation (LoRA) distillation techniques. By leveraging these methods, we aim to optimize Stable Diffusion for improved inference speed, while maintaining the image quality and consistency of the original model.

3.2 Method Design

This project aimed to enhance image generation capabilities by integrating multiple advanced model optimization techniques into a structured methodology. The approach was divided into three main phases: utilizing a pre-trained Stable Diffusion base model, fine-tuning with Low-Rank Adaptation (LoRA), and implementing Latent Consistency Model (LCM) for LoRA distillation.

3.2.1 Base Model: Pre-trained Stable Diffusion

The project began with the pre-trained Stable Diffusion v1-5 model from RunwayML. This model was selected due to its extensive training on a diverse set of text-image pairs, providing a robust foundation for subsequent enhancements.

3.2.2 Fine-tuning with Low-Rank Adaptation (LoRA)

The first enhancement phase involved fine-tuning the Stable Diffusion model using LoRA. LoRA applies low-rank updates to the model's weights, improving computational efficiency while maintaining performance.

The environment was set up with essential libraries such as `accelerate`, `datasets`, `torch`, and `transformers` to create a conducive training environment. The dataset used included the Flickr30k dataset, comprising 31,783 images with descriptive captions. The data was processed for compatibility, involving resizing, cropping, and normalization of images. The components of the Stable Diffusion model (tokenizer, text encoder, VAE, and UNet) were configured, and LoRA configurations were applied to the UNet, targeting specific modules for low-rank adaptations to enhance efficiency.

The training loop involved gradient accumulation, optimizer steps, and loss calculations. Latents were encoded using the VAE, noise was added through the noise scheduler, and predictions were generated using the UNet. The training focused on minimizing the loss between model predictions and the target noise, refining the latent representations.

3.2.3 Latent Consistency Model (LCM) for LoRA Distillation

The second enhancement phase introduced the Latent Consistency Model (LCM), aimed at achieving consistent latent representations across various diffusion timesteps. This phase utilized a custom algorithm and a specialized loss function to guide the training process.

The LCM algorithm processed text-image pairs, generating and refining latent representations guided by a consistency loss function. This ensured robustness and consistency across different timesteps. The LoRA network applied low-rank updates to the model's weights. The network was trained to mimic the LCM model by minimizing discrepancies in latent representations, guided by Huber loss. This combination helped maintain stability and reduced the sensitivity to outliers in the training data.

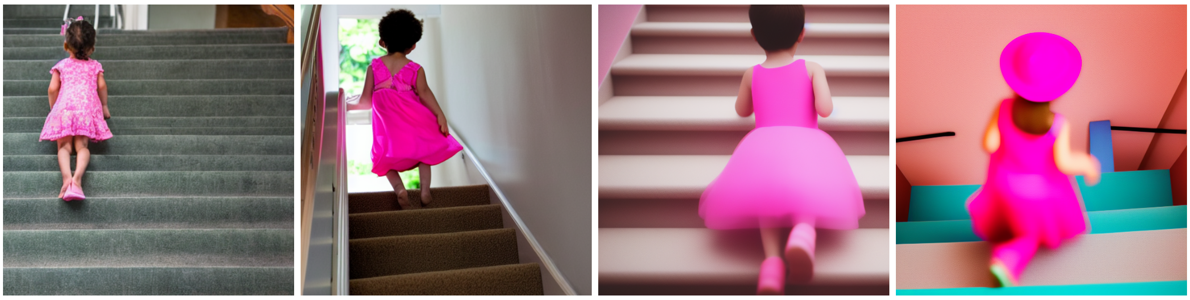
4 Experiments

4.1 Datasets

In this project, we utilize Flickr8k and Flickr30k dataset. The Flickr datasets are widely-used benchmark in the field of image captioning and multimodal learning. The Flickr8k dataset consists of 8,000 images sourced from the popular photo-sharing platform Flickr, each accompanied by five human-annotated captions describing the content of the image. The Flickr30k dataset is a larger version of the Flickr8k dataset, containing 31,783 images. These datasets are well-suited due to their

Model	FID Score	Inference Speed (seconds)
Base Stable Diffusion	1.68	2.07
Enhanced LoRA Fine-tuned	1.49	-
LCM LoRA (Flickr8k)	26.95	0.86
LCM LoRA (Flickr30k)	24.85	0.86

Table 1: The FID score and inference speed of the base stable diffusion, Enhanced LoRA Fine-tuned, LCM LoRA (Flickr8k), and LCM LoRA (Flickr30k) models.



(a) Base model (b) LoRA fine-tuned model (c) LCM- LoRA with Flickr 8k (d) LCM- LoRA with Flickr 30k

Figure 1: Generated Image with a caption of "A child in a pink dress is climbing up a set of stairs in an entryway"

diverse range of subjects and scenes, high-quality annotations, established use in multimodal learning, and manageable size for training. By preprocessing the images and captions appropriately, we can leverage Flickr8k's rich data to optimize Stable Diffusion's performance on multimodal tasks like image captioning and text-to-image generation.

4.2 Evaluation Protocol

To comprehensively assess the performance of the proposed LCM-LoRA and knowledge distillation strategies, we employ two key metrics: inference time and Fréchet Inception Distance (FID). We measure the time taken by each model (base model, LoRA fine-tuned, and LCM-LoRA fine-tuned) to generate images from text prompts during the inference phase. Lower inference time is desirable, as it translates to faster and more efficient text-to-image generation. The FID score is a metric that quantifies the quality of generated images by comparing the distribution of generated images to the distribution of real images. We compute the FID score by comparing the feature representations of the generated images from each model with the feature representations of real images from the Flickr8k dataset using 100 images. A lower FID score indicates that the generated images are more similar to real images, and thus, higher quality.

4.3 Results

The evaluation of the models involved comparing Frechet Inception Distance (FID) scores and inference speeds to assess the image quality and computational efficiency of different model configurations. The FID score measures the quality of generated images by comparing the distribution of generated images to real images, with lower FID scores indicating better image quality.

As shown in Table 1, the original Stable Diffusion v1-5 model demonstrated strong performance with an FID score of 1.68. The Enhanced LoRA fine-tuned model significantly outperformed the base model with the lowest FID score of 1.49, showcasing superior image quality. The LCM-enhanced LoRA models had higher FID scores of 26.95 and 24.85 for the Flickr8k and Flickr30k datasets, respectively, indicating a trade-off between fidelity and computational efficiency.

Inference speed measures the time taken to generate images, with faster speeds indicating improved computational efficiency. The base Stable Diffusion v1-5 model had a slower inference time of 2.07 seconds per image. In contrast, the Enhanced LoRA fine-tuned model achieved an inference speed of 0.91 seconds per image. Both LCM LoRA models demonstrated significantly faster inference speeds of 0.86 seconds per image, indicating enhanced efficiency.

These results highlight the efficiency gains from

LoRA fine-tuning and LCM distillation, albeit with a trade-off in image quality for the LCM models.

Visual examples were used to qualitatively assess the differences in image quality among the models. As shown in Figure 1, these examples included various scenes, such as "A child in a pink dress is climbing up a set of stairs in an entryway," showcasing the differences in generated image realism among the models.

5 Conclusion

The project successfully enhanced image generation capabilities by integrating advanced optimization techniques. The pre-trained Stable Diffusion model provided a strong foundation with competitive FID scores. Enhanced LoRA fine-tuning significantly improved image quality and performance, making it suitable for high-fidelity visual tasks. LCM LoRA distillation achieved substantial computational efficiency gains, making it ideal for resource-limited environments despite higher FID scores. This project demonstrated the potential of combining these techniques to optimize text-to-image generation, offering models tailored to various requirements.

6 Limitations

Several limitations impacted the findings. The trade-off between image quality and efficiency was evident, with LCM-enhanced LoRA models showing higher FID scores but faster inference speeds. This needs careful consideration for specific applications. Evaluations were based on Flickr30k and Flickr8k datasets, limiting generalizability. The resource-intensive training and fine-tuning processes could restrict deployment in constrained environments. Implementing LCM added complexity, requiring deep technical understanding. There is also a risk of overfitting to specific datasets. Future research should explore diverse datasets, optimize the quality-efficiency balance, and improve technique accessibility.

References

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Jonathan Ho, Ajay Jain, and P. Abbeel. 2020. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. 2023. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.