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

Cheryl Ngai, Irene Chang, Karamjeet Gulati

UCDAVIS

Introduction

- The pursuit of high-quality text-to-image generation with efficient inference remains a key challenge in the field of deep learning.
- While **Stable Diffusion** has demonstrated remarkable capabilities, achieving both visual fidelity and computational efficiency requires sophisticated approaches.
- This research presents a novel methodology that combines:
  - **Latent Consistency Model (LCM) training**
  - **LoRA (Low-Rank Adaptation) distillation**
  - **Knowledge distillation**
- **Outcomes:**
  - Generates visually appealing images with minimal computational cost.
  - Rigorous evaluation using benchmark datasets.
  - Significant performance gains, emphasizing image quality and computational efficiency.
- **Impact:** Provides valuable insights for optimizing text-to-image generation techniques, especially in resource-constrained environments.

Dataset

Flickr Datasets

Flickr30k

- a popular benchmark for sentence-based picture portrayal
- comprising 31,783 images
- capture people engaged in everyday activities and events
- each image has 5 descriptive captions provided by human annotators

Flickr8k: a smaller version of Flickr30k, contains 8,000 images

Methodology

Phase 1: Pre-trained Stable Diffusion Model

- We leverage the pre-trained Stable Diffusion v1-5 model from RunwayML, trained extensively on a diverse set of text-image pairs.

Phase 2: Latent Consistency Model (LCM) Training

- The LCM approach emphasizes consistent latent representations across various diffusion timesteps, employing a specialized loss function.
- Using our custom LCM algorithm, we process text-image pairs to generate and refine latent representations guided by the consistency loss function.

Methodology

Phase 3: LoRA Distillation Cont'd

- LoRA Network applies low-rank updates to the pre-trained model's weights, enhancing computational efficiency.
- The LoRA network is trained to mimic the LCM model by minimizing discrepancies in latent representations.
- A combination of mean squared error (MSE) and KL divergence guides the optimization.

Phase 4: Evaluation

- Baselines: Original Stable Diffusion model and normal LoRA fine-tuned Stable Diffusion
- Metrics: Frechet Inception Distance (FID) and inference speed. Lower FID scores indicate better image quality, while faster inference speed denote improved efficiency.

Results

FID Score:

- **Base Model Excels:** The original Stable Diffusion v1.5 significantly outperforms all fine-tuned LoRA models in terms of FID scores, indicating superior image quality.
- **Fine-Tuning Impact Limited:** While LCM LoRA shows slight improvement with more data, the overall impact of fine-tuning on FID scores is minimal, even for normally trained LoRA models.

Inference Speed:

- **Base Model Slower:** The base Stable Diffusion v1.5 model takes noticeably longer to generate images (2.07 seconds on average) compared to the fine-tuned LoRA models.
- **LORA Efficiency:** Both LCM LoRA models demonstrate significantly faster inference speeds (0.86 seconds) compared to the base model, regardless of the dataset size used for fine-tuning.
-

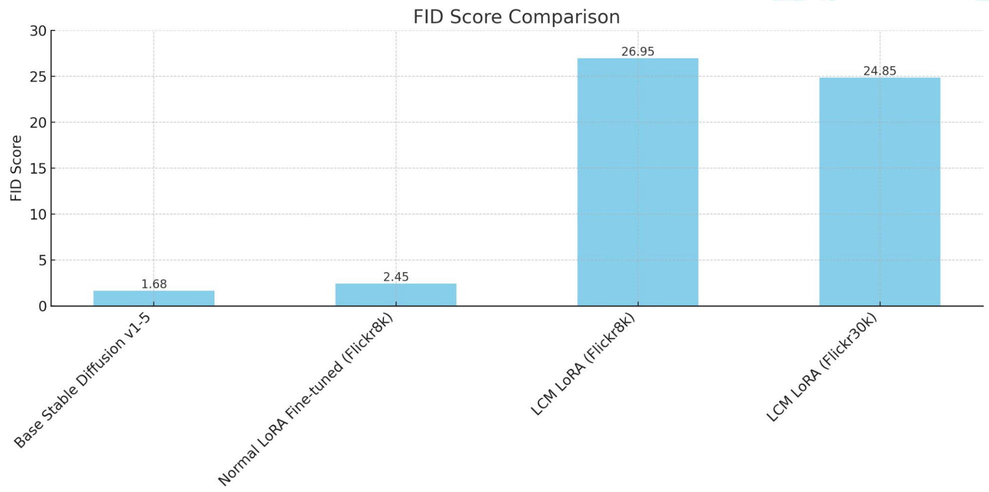

Figure 1: FID Score Comparison

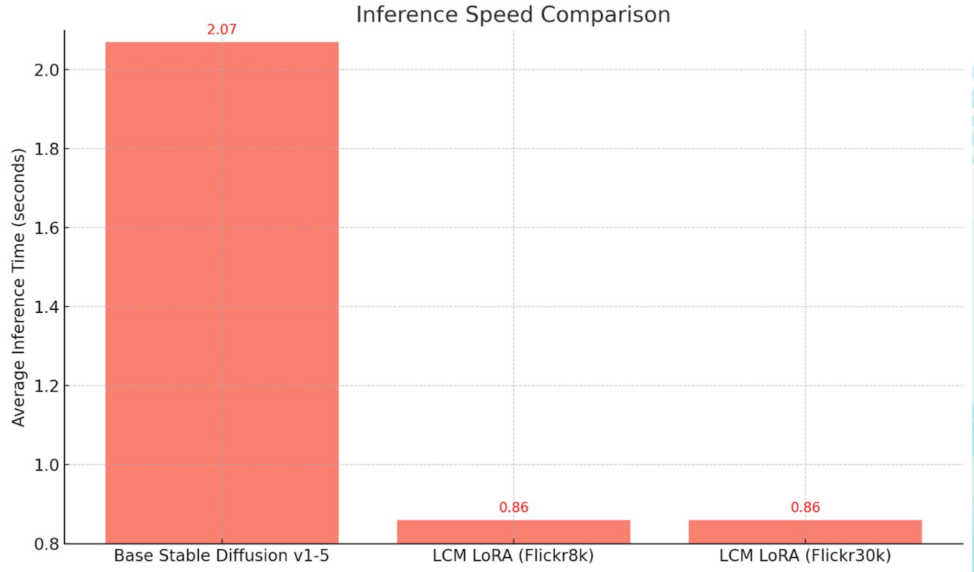

Figure 2: Inference Speed Comparison

Result

Fig 3a : Base model

Fig 3B : LoRA fine tuned model

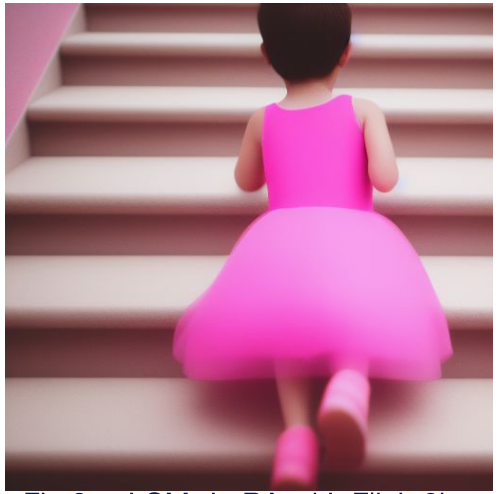

Fig 3c : LCM- LoRA with Flickr8k

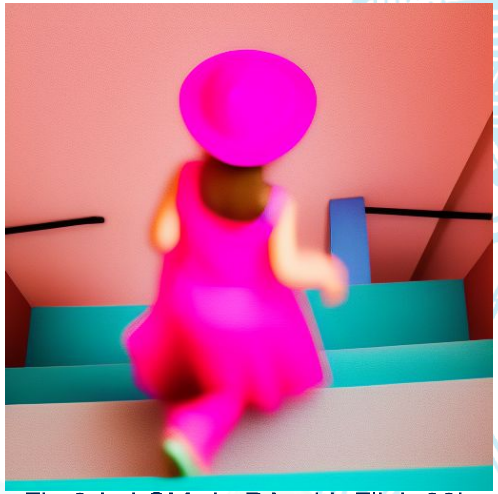

Fig 3d : LCM- LoRA with Flickr30k

Fig 3: Generated Image with a caption of "*A child in a pink dress is climbing up a set of stairs in an entry way*"

Optimal Performance: Base Stable Diffusion exhibits the highest realism with the lowest FID score of 1.68, highlighting exceptional image quality.

Enhanced Fine-tuning: Normal LoRA Fine-tuned model shows improved realism with a moderate FID score of 2.45, striking a balance between quality and performance.

Comparative Analysis: While LCM LoRA models have higher FID scores (26.95 and 24.85), indicating reduced fidelity, selecting models like Base Stable Diffusion or Normal LoRA Fine-tuned is advisable for tasks requiring high visual

Conclusion

- **Scalability and Efficiency:** Increasing dataset size from 8k to 30k improves FID scores and reduces inferencing time, highlighting the LCM model's efficiency.
- **Quality Improvement:** Larger datasets enhance image quality, showcasing the potential of the LCM model in scalable environments.

Limitations

- **Quality vs. Base Models:** Despite improvements, LCM models still lag behind base models in visual fidelity.
- **Resource Intensity:** Larger data improves performance but demands significant computational resources, potentially limiting deployment flexibility.