Name: Aditi Tiwari
Paper:

Latent Consistency Models: Synthesizing High-Resolution Images With Few-Step Inference

Authors: Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, Hang Zhao (Institute for Interdisciplinary Information Sciences, Tsinghua University)

While **Latent Diffusion Models (LDMs)** like Stable Diffusion have revolutionized text-to-image synthesis, their **real-time deployment** remains constrained by **high memory requirements** (up to 780GB for large models) and **computational overhead.** The computational bottleneck caused by lengthy iterative sampling processes requiring 20-50 steps for inference is a critical challenge in high-resolution image generation using LDMs. This challenge has been addressed in the paper. The authors introduce **Latent Consistency Models (LCMs),** achieving high-fidelity image generation in just 2-4 steps while requiring less than 48GB of memory. This **breakthrough is particularly significant as it enables real-time applications of high-resolution generative models, requiring only 32 A100 GPU hours for training compared to previous methods needing 45+ A100 GPU days.** The practical impact is demonstrated through significant speedups: **3.25x on A100 GPUs** and **4.5x on more cost-effective A6000 GPUs.**

The **technical innovation** of LCMs centers around three carefully designed components: (a) a novel one-stage guided distillation approach that efficiently solves an augmented Probability Flow ODE (PF-ODE), incorporating classifier-free guidance (CFG) directly into the model, (b) Latent Consistency Fine-tuning (LCF) for adapting pre-trained LCMs to customized datasets while preserving few-step inference capabilities, and (c) the "skipping-step" technique with k=20 that accelerates convergence by ensuring consistency between non-adjacent timesteps. The augmented PF-ODE enables direct integration of CFG, eliminating the need for separate guided distillation stages and reducing training complexity. Moreover, their double quantization technique saves approximately 0.37 bits per parameter (around 3GB for a 65B model), optimizing memory usage without compromising performance.

The empirical **results** are compelling, particularly in high-resolution image generation. On the LAION-5B-Aesthetics dataset at 768×768 resolution, LCMs achieve remarkable performance with **4-step inference**, obtaining a Fréchet Inception Distance **(FID) score of 13.53** compared to 20.08 for DPM++ and 24.28 for DDIM. The effectiveness of their guidance approach is demonstrated through CLIP scores, where **4-step LCM achieves 28.60**, nearly matching DPM++'s 8-step score of 29.84. The impact of the guidance scale (ω) is significant, with scores improving from 27.83 to 28.69 as ω increases from 2 to 8. Even single-step inference achieves an FID of 34.22, significantly outperforming baselines that score above 120. Ablation studies show that their skipping-step technique with k=20 achieves an optimal balance between convergence speed and stability, with the DDIM solver performing better at this value compared to smaller or larger k values.

One of the **strong points** of the paper is **flexibility in task-specific adaptation**. The LCM framework's adaptability to customized datasets through LCF represents a notable strength. This fine-tuning technique allows the model to maintain its few-step inference efficiency even when trained on specialized data, which could benefit niche applications like medical imaging or scientific visualization where general-purpose models may lack domain specificity. The flexibility in task adaptation could encourage further adoption and customization of the model across diverse industries. One more noteworthy aspect is LCMs' ability to perform high-fidelity synthesis at 768x768 resolution with minimal steps is a key strength, as it effectively

demonstrates that such models can produce complex, high-quality visuals without requiring extensive computational resources.

One of the **limitations/weakness** of the paper is that while the LCM framework performs well with shorter generation sequences, the paper provides limited analysis on the consistency and quality of outputs when sequences are extended beyond the 2-4 step inference regime. In applications that may require progressive refinement or iterative generation (e.g., video synthesis or stepwise image editing), LCMs' efficacy in maintaining coherence over extended sequences remains unclear and warrants further investigation. Additionally, although LCMs demonstrate strong performance on specific datasets like LAION-5B-Aesthetics, the paper does not extensively explore how well the model handles varied or nuanced text prompts, particularly for languages or cultural references that are less common in the training data. This could be a limitation in real-world applications where users might input diverse prompts that require more sophisticated linguistic or cultural understanding, potentially impacting the model's ability to generate semantically accurate and contextually rich images.

Future work could investigate a multi-scale guidance approach to improve model fidelity across a broader range of tasks. By incorporating hierarchical guidance that adapts CFG levels based on task-specific details (e.g., content complexity or spatial scale within an image), LCMs might achieve better control over image detail and diversity, balancing fine-grained textures with larger structural features. This approach could enhance model flexibility and quality, particularly for tasks like scene generation or architectural rendering. Another promising direction could involve adding self-supervised feedback mechanisms to improve single-step and few-step inference. For instance, the model could be adapted to self-assess and iteratively refine its outputs by feeding generated samples back into the model to enhance details or correct inconsistencies. This feedback loop could allow LCMs to achieve even higher fidelity in fewer steps, making the model more robust to different input conditions and enhancing its real-time generation capabilities across more complex scenarios.