Scalable Diffusion Models with Transformers

Author: Peebles et al.

created by paper2slides

Uploaded to arXiv: 2022-12-19

https://arxiv.org/abs/2212.09748

Executive Summary

- This paper introduces a new class of diffusion models named **Diffusion Transformers** (DiTs).
- The DiTs replace the traditional U-Net backbone with a transformer architecture.
- State-of-the-art image generation performance on ImageNet benchmarks at 512×512 and 256×256 resolutions.
- Empirical findings indicate that increasing transformer Gflops directly improves image quality.

Introduction

- Transformers have revolutionized NLP and vision tasks, but image-level generative modeling still predominantly uses convolution-based architectures like U-Nets.
- Diffusion models, particularly those based on U-Nets, have recently set new performance standards in image synthesis.
- The goal: Explore transformers as backbones for diffusion models due to their scalability and robustness.

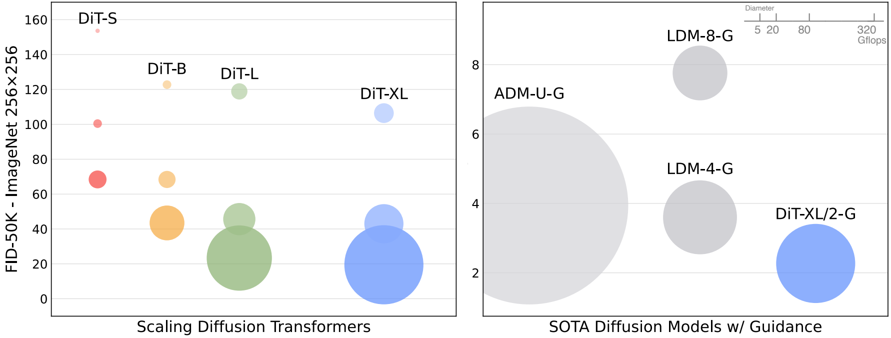

Figure: Diffusion models with transformer backbones achieving state-of-the-art image quality.

Proposed Method: DiT Architecture

- **Patchify**: Decompose the latent representation into patches.
- **Vision Transformer Backbone**: Sequence of transformer blocks processing the patches.
- **Adaptive Layer Norm (adaLN)**: Regresses normalization parameters conditioned on noise timesteps and class labels.
- **DiT Configurations**: Different sizes (S, B, L, XL) and patch sizes (2, 4, 8).

DiT Block Designs

- **In-context Conditioning**: Appends embeddings of conditional information as input tokens.
- **Cross-attention Block**: Introduces extra cross-attention layers for conditioning.
- **Adaptive Layer Norm (adaLN)**: Inspired by GANs, applies learned scaling parameters.
- **adaLN-Zero**: Initializes as identity function to improve stability and performance.

Key Hypotheses

- U-Net backbone is not crucial; transformers should provide similar or better performance.
- Increasing Gflops enhances model quality: larger DiTs should yield better results.
- Scaling properties of transformers should translate to improved diffusion models.

Implementation Details

- Training conducted on ImageNet for both 256×256 and 512×512 resolutions.
- Class-conditional models evaluated with Fréchet Inception Distance (FID).
- Maintained moving average of weights with decay of 0.9999.
- Models implemented using JAX on TPU-v3 pods.

Training Setup

- Optimizer: AdamW with learning rate 1×10^{-4} and no weight decay.
- Batch size: 256.
- Data augmentation: Only horizontal flips.
- Training iterations: Up to 7 million steps for top-performing models.

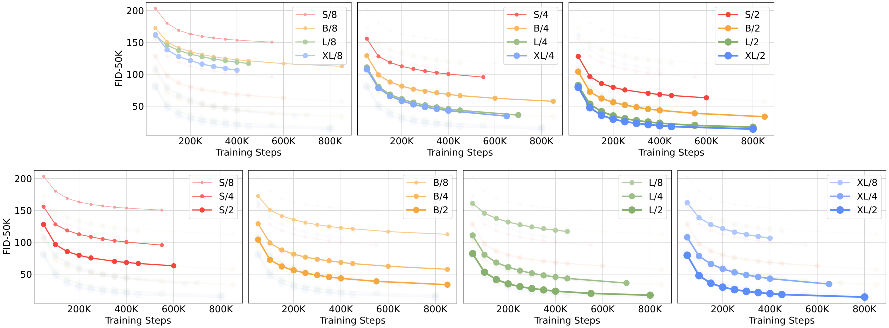

Figure: Scaling DiT model size and patch size consistently improves FID across all stages of training.

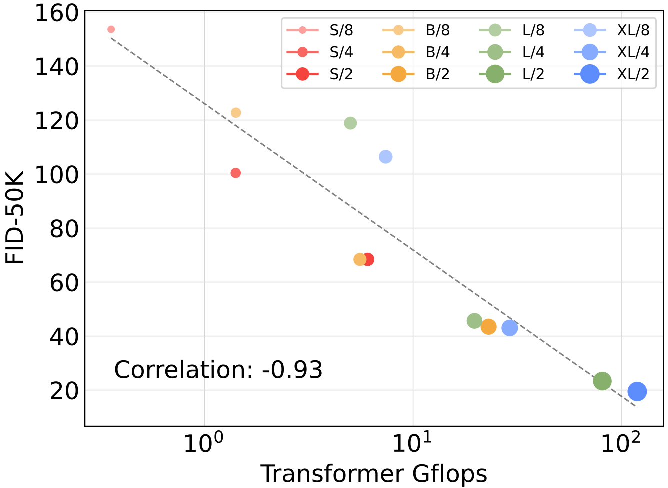

Figure: Clear inverse correlation between Gflops and FID, underscoring the importance of model compute.

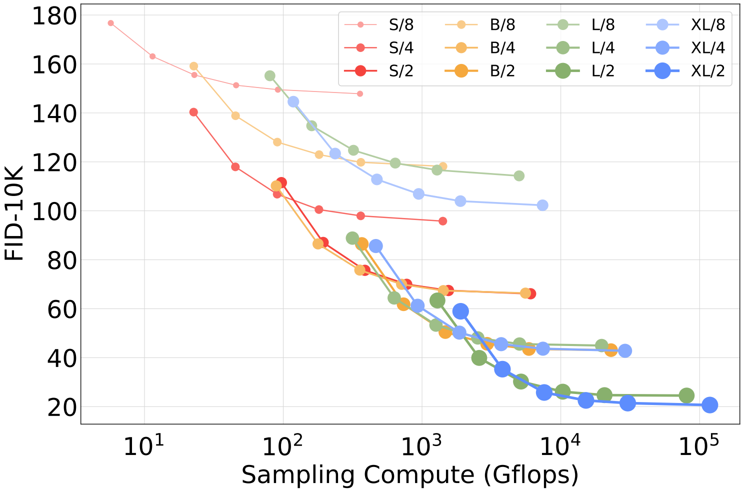

Figure: Scaling sampling steps improves FID, but cannot compensate for less model compute.

Results: 256×256 ImageNet

- **DiT-XL/2**: Achieved FID of 2.27, outperforming LDM-4 (3.60) and ADM-G (4.59).
- Higher recall values compared to prior latent-space diffusion models.
- Significant quality improvement due to more transformer Gflops.

Results: 512×512 ImageNet

- **DiT-XL/2**: Achieved FID of 3.04, setting a new state-of-the-art.
- More compute-efficient compared to U-Net models like ADM which used far more Gflops.
- Demonstrates the scalability and robustness of DiTs in high-resolution image generation.

Conclusion and Future Directions

- DiTs establish transformers as a competitive backbone for diffusion models.
- Significant headroom for scaling DiTs further, providing robust improvements in quality.
- Future work: Apply DiTs to text-to-image models and further scaling in both tokens and model size.