Masked Flexible Diffusion Transformers

Overview

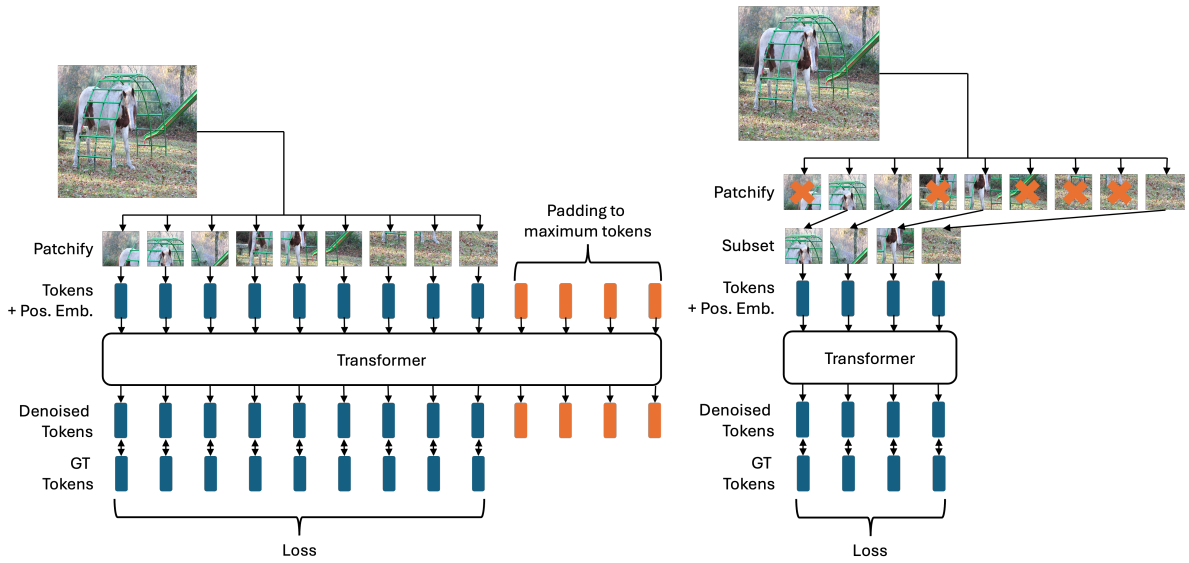
Diffusion models [5] have emerged as the preferred method for generating high-quality content such as images, video, and audio. However, most methods use a fixed aspect ratio and resolution and require expensive fine-tuning to generalize to other resolutions. Recently, Flexible Vision Transformers (FiT) [6] and NaViT [1] have introduced vision transformers [3] that are trained on different resolutions and aspect ratios. However, both methods waste computation during training by padding the input tokens to a fixed length.

In this project, we investigate how flexible diffusion models can be trained without the need for padding tokens. In particular, one approach would be to generate only a subset of the tokens during training. A similar idea for fixed size diffusion transformers has been successfully implemented in MDTv2 [4] and MaskDiT [8]. The resulting architecture doesn't waste computation on unused padding tokens, could be more robust due to less dependence on individual patches, and would reduce the computational and memory requirements during training.

The new model should be compared to FiT [6] on multiple resolutions using 400K training steps and the FiT-B/2 architecture with the same autoencoder[1] from stable diffusion [7]. The metrics can be computed using OpenAI's evaluation script[2] to compare the results with related work.

[1] `https://huggingface.co/stabilityai/sd-vae-ft-ema`
[2] `https://github.com/openai/guided-diffusion/tree/main/evaluations`

Figure 1: The goal of the project is to generate realistic images with different resolutions and aspect ratios without requiring padding tokens by subsampling tokens.



(a) The original FiT [6] architecture.

(b) Our architecture.

Figure 2: Instead of applying the transformer to the all tokens (and padding tokens), we only apply it to a subset of the tokens.

Goals

The goal of this project is to evaluate flexible diffusion transformer architectures by predicting only subsets of patches during training.

Tasks

1 Getting familiar with the related literature (~ 30h) and reproduce the baseline[3] (~ 40h);

2 Implementing different patch subsampling strategies (~ 80h);

3 Training and evaluating multiple models with patch subsampling on ImageNet [2] (~ 50h);

4 Ablate different model designs (~ 50h);

5 Writing a project report and preparing the presentations (~ 60h).

[3] `https://github.com/mindspore-lab/mindone/blob/master/mindone/models/fit.py`

Contact

Dominik Schnaus
Email: dominik.schnaus@tum.de

References

[1] Mostafa Dehghani et al. "Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution". In: *Advances in Neural Information Processing Systems* 36 (2024).

[2] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[3] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2020.

[4] Shanghua Gao et al. "Masked diffusion transformer is a strong image synthesizer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 23164–23173.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[6] Zeyu Lu et al. "Fit: Flexible vision transformer for diffusion model". In: *arXiv preprint arXiv:2402.12376* (2024).

[7] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[8] Hongkai Zheng et al. "Fast training of diffusion models with masked transformers". In: *arXiv preprint arXiv:2306.09305* (2023).