



## Normalized Vision Transformers

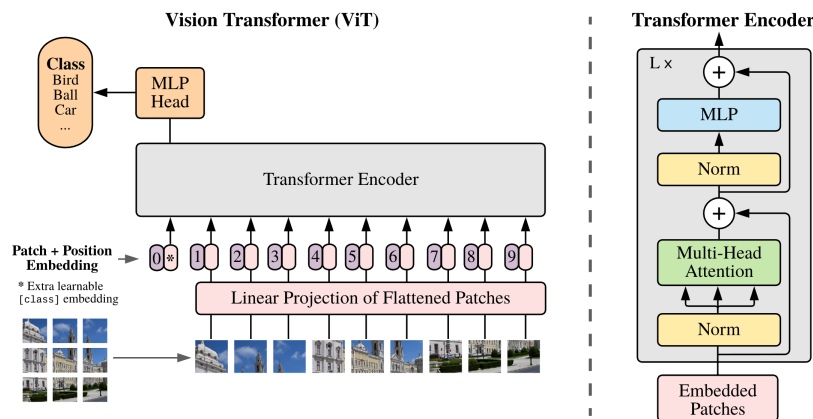


Figure 1: The original vision transformer architecture

Vision Transformers (ViTs) and language models share identical architectures, yet recent breakthroughs in efficient transformer training remain unexplored for vision. NVIDIA's nGPT achieves 4-20x faster training by normalizing all representations to a hypersphere, a geometric constraint that improves conditioning and stability [1].

This project investigates whether hypersphere learning transfers to visual domains. You will implement normalized ViT variants, replacing standard normalization with unit-norm constraints on patch embeddings, attention matrices, and MLP blocks. The goal is to benchmark training efficiency on ImageNet/CIFAR and analyze whether the conditioning improvements observed in language models apply to image patches.

The project will be done fully remote. We would have weekly meetings on Zoom to go over results, discuss open questions, and resolve any potential problems. You will have a lot of possibilities to shape the project in the directions you find the most interesting.

## Requirements

Solid programming skills in Python (PyTorch preferred) and knowledge of transformer architectures and computer vision evaluation are required. Knowledge of linear algebra is beneficial.

## Contact

In a few short sentences, please describe your interest in this project and any relevant coding experience or background (e.g., projects or coursework).

- Andreas Plesner: [aplesner@ethz.ch](mailto:aplesner@ethz.ch), ETZ G95
- Till Aczel: [taczal@ethz.ch](mailto:taczal@ethz.ch), ETZ G60.1

## References

- [1] Ilya Loshchilov et al. *nGPT: Normalized Transformer with Representation Learning on the Hypersphere*. 2025. arXiv: [2410.01131](https://arxiv.org/abs/2410.01131) [cs.LG]. URL: <https://arxiv.org/abs/2410.01131>.