TinyViT: A Small Vision Transformer for Image Classification

Lucian Dorin Crainic
Department of Computer Science
crainic.1938430@studenti.uniroma1.it

La Sapienza University of Rome

Abstract

Vision Transformers (ViTs) excel in image classification but require large datasets and complex architectures. We introduce TinyViT, a minimalist ViT that retains core components—patch embedding and attention—while drastically reducing scale. By simplifying token processing and prioritizing parameter efficiency, we test whether a tiny ViT can achieve good accuracy despite limited complexity. Experiments on standard benchmarks show TinyViT delivers acceptable results even with reduced data requirements, challenging the assumption that ViTs inherently demand heavy resources. This work demonstrates that minimalist transformer architectures can learn meaningful representations, offering a pathway to simpler, more accessible models without sacrificing core functionality.

1 Introduction

Transformers, introduced by [Vaswani, 2017] for natural language processing (NLP), have become the dominant architecture for sequence modeling due to their scalability and self-attention mechanisms. Inspired by their success in NLP, [Alexey, 2020] pioneered the Vision Transformer (ViT), demonstrating that transformers can achieve state-of-the-art results in image recognition by treating images as sequences of patch tokens. By splitting an image into fixed-size patches, linearly embedding them, and processing the sequence with a standard transformer encoder, ViT outperformed convolutional neural networks (CNNs) [He et al., 2016] on large-scale datasets like ImageNet when pretrained on massive datasets (e.g., JFT-300M). However, ViT's strong performance comes at a cost: it requires extensive computational resources and large pretraining datasets, raising practical barriers for adoption in settings where such infrastructure is unavailable.

In this work, we aim to (1) elucidate the foundational mechanics of Vision Transformers and (2) present TinyViT, a minimalist implementation designed to test the viability of ViTs in simplified, resource-efficient settings. Unlike the original ViT, which emphasizes scaling to massive datasets, TinyViT reduces architectural complexity—employing fewer transformer layers, smaller embedding dimensions, and streamlined attention mechanisms—while retaining the core principles of patch-based processing and self-attention. We evaluate TinyViT on widely adopted benchmarks like CIFAR-10 and CIFAR-100 [Krizhevsky et al., 2009], and STL-10 [Coates et al., 2011], datasets that reflect real-world scenarios where data and computational resources are often constrained.

2 Architecture

References

[Alexey, 2020] Alexey, D. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

[Coates et al., 2011] Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.

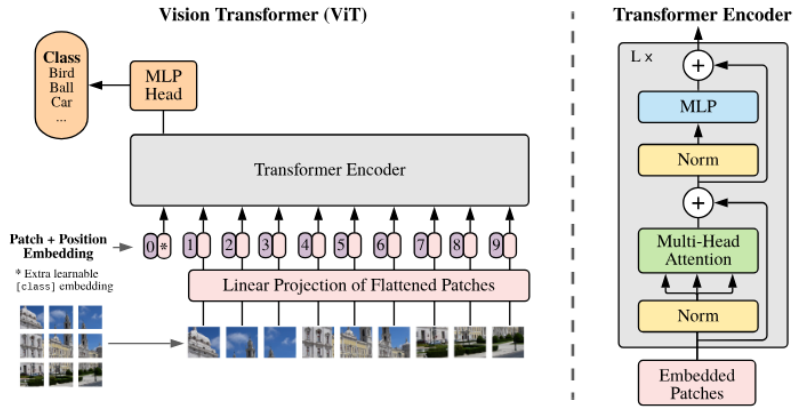

Figure 1: A figure spanning both columns but only 70% of the page width.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Krizhevsky et al., 2009] Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

[Vaswani, 2017] Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.