TinyVit: A Small Vision Transformer

Master's Degree in Computer Science

Lucian Dorin Crainic (1938430)

Academic Year 2024/2025

SAPIENZA
UNIVERSITÀ DI ROMA

This project is based on the well known **Vision Transformer** (ViT) architecture, a deep learning model that applies self attention mechanisms to image processing. We will start by introducing the core principles of ViT, explaining how it differs from traditional CNNs.

Then, we will explore the **TinyViT** a smaller implementation of the ViT architecture. Finally, we will present benchmark results demonstrating its performance across various tasks.

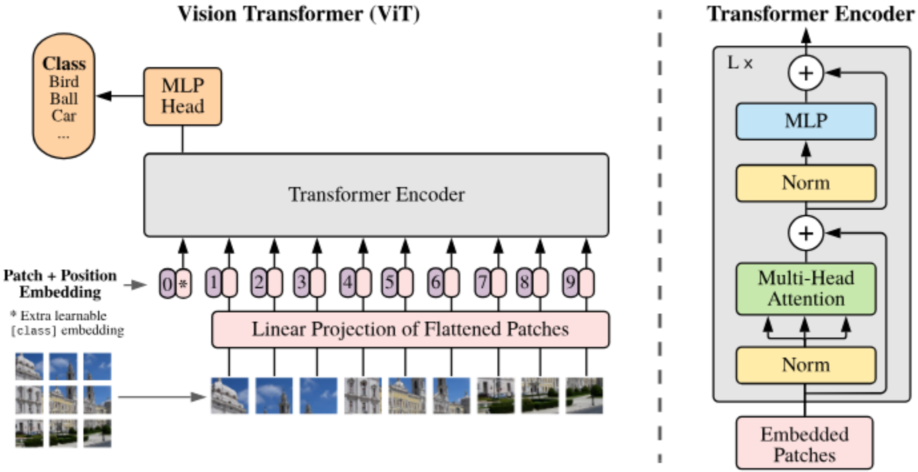
Table of Contents

Introduction

- Vision Transformers (ViTs) have revolutionized image recognition.
- Traditional ViTs require large datasets and heavy computation.
- TinyViT is a simplified alternative, retaining core transformer components while being computationally efficient.

Vision Transformer Architecture

1 Introduction



Vision Transformer Overview

- Inspired by NLP transformers (Vaswani et al., 2017).
- Treats images as sequences of patches instead of processing pixels directly.
- Uses a stack of transformer encoder blocks for feature extraction.

Step 1
Patch Embedding

Step 2
Positional Encoding

Step 3
Transformer Encoder

Step 4
Classification Head

Table of Contents

2 Vision Transformer

Mathematical Foundations

TODO

Vision Transformer Architecture

Patch Embedding and Positional Econding

Vision Transformer Architecture

Transformer Encoder and Classification Head

- CNNs use convolutional layers with local receptive fields.
- ViTs process images globally using self-attention mechanisms.
- CNNs have built-in spatial hierarchies, whereas ViTs rely on attention.
- ViTs typically need more data to perform well but can model long-range dependencies.

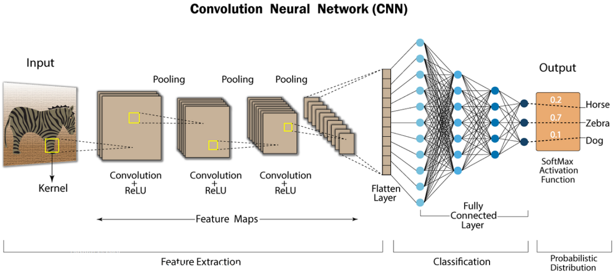


Table of Contents

Side-Picture Slides
3 TinyViT

- Opened with
  `\begin{sidepic}{<image>}{<title>}`
- Otherwise, `sidepic` works just like `frame`

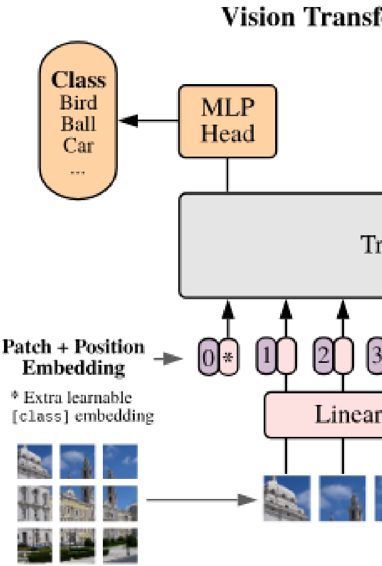


Table: Parameters of the TinyViT Model for CIFAR-10

Parameter	Value
Number of Classes	10
Embedding Dimension	128
Image Size	32
Patch Size	4
Input Channels	3
Number of Attention Heads	8
Number of Transformer Layers	6
MLP Hidden Dimension	512

Table of Contents

CIFAR 10 - CIFAR 100

STL10

Table of Contents

5 Training Setup

Models Training Setup

Compared to PowerPoint, using LaTeX is better because:

- **Optimizer** : AdamW
- **Learning Rate** : AdamW
- **Weight Decay** : AdamW
- **Loss Function** : AdamW
- **Epochs** : AdamW
- **Batch Size** : AdamW

Table of Contents

6 Results

Results CIFAR 10 - CIFAR 100

TinyViT outperforms CNNs on both datasets. Handles complex class distributions better than CNNs.

Dataset	Model	Accuracy	F1	Recall	MCC	Precision
CIFAR-10	**Tiny ViT**	**82.59**	**82.43**	**82.59**	**80.70**	**82.76**
	CNN	80.67	80.55	80.67	78.53	80.56
CIFAR-100	**Tiny ViT**	**59.40**	**59.04**	**59.40**	**59.00**	**60.00**
	CNN	46.13	44.57	46.13	45.61	45.61

Note: All values are percentages (%). Bold indicates best performance in category.

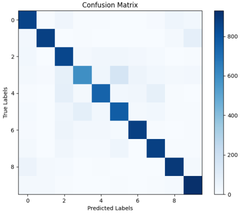
Results STL-10

CNN performs better on STL-10, likely due to the higher image resolution. TinyViT may struggle with lower-resolution images in datasets with fewer samples.

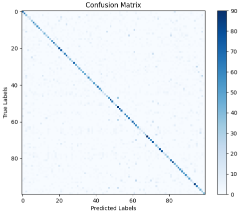
Dataset	Model	Accuracy	F1	Recall	MCC	Precision
STL-10	Tiny ViT	64.27	64.30	64.27	60.47	66.13
	CNN	**68.36**	**68.10**	**68.36**	**64.95**	**68.78**

Note: All values are percentages (%). Bold indicates best performance in category.

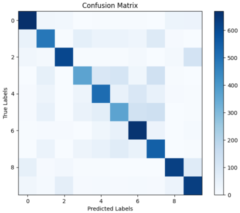
Confusion Matrices
6 Results



CIFAR-10 dataset using Tiny ViT



CIFAR-100 dataset using Tiny ViT



STL-10 dataset using CNN

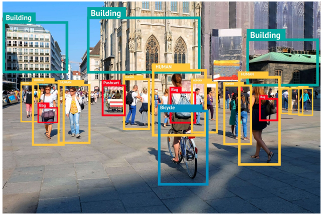
Table of Contents

Conclusion

- TinyViT proves that **transformer based models** can be efficient with fewer resources.
- Outperforms CNNs on **smaller datasets** like CIFAR-10 and CIFAR-100.
- **Requires improvements** for larger images like STL-10.

Future Work

- Implement TinyViT for **Object Detection** (DETR) and **Segmentation** (Segmenter).
- Experiment with different **hyperparameters** (layers, embedding size, attention heads).
- Explore pretraining on **larger datasets** to improve performance.



TinyVit: A Small Vision Transformer

Thank you for listening!
Any questions?