

Vision Transformer on CIFAR-10 - Report (Based on Submitted Notebook)

Name: Yadnesh Sutar

Roll No: 241072

Dataset: CIFAR-10

Framework: PyTorch

Training Platform: Google Colab

1. Overview

This implementation builds a Vision Transformer (ViT) from scratch using PyTorch modules and evaluates it on the CIFAR-10 dataset. Unlike pre-trained ViT models, this approach focuses on architectural understanding and controlled experimentation suitable for academic evaluation.

2. Data Preprocessing and Regularization

2.1 Preprocessing Techniques

- Random Crop (32x32, padding=4)
- Random Horizontal Flip
- Normalization (Mean: 0.4914, 0.4822, 0.4465; Std: 0.2023, 0.1994, 0.2010)

Effect: These steps improve translation invariance and stabilize gradients, acting as implicit regularization.

2.2 Regularization Methods Used

- Dropout (p=0.1): Applied in Transformer encoder layers.
- Weight Decay: Value of 0.05 using AdamW optimizer.
- Early Stopping: Training stops if validation accuracy does not improve for 3 epochs.

3. Vision Transformer Architecture and Mechanism

3.1 Patch Embedding

Input image (32x32) is divided into 64 non-overlapping patches of size 4x4. A convolution layer projects patches into a 128-dimensional embedding space.

3.2 Class Token and Positional Embedding

A learnable [CLS] token is appended to the patch sequence. Learnable positional embeddings preserve spatial information.

3.3 Transformer Encoder

Each encoder block contains Multi-Head Self Attention (8 heads), Feed Forward Network (MLP ratio = 4.0), Layer Normalization, and Residual Connections. Depth = 8 layers.

4. Hyperparameter Tuning

Patch Size	4
Embedding Dimension	128

Vision Transformer on CIFAR-10 - Report (Based on Submitted Notebook)

Transformer Depth	8
Attention Heads	8
Optimizer	AdamW
Learning Rate	3e-4
Weight Decay	0.05
Batch Size	64 (train), 128 (test)
Scheduler	Cosine Annealing
Epochs	Up to 15

Observation: Cosine learning rate scheduling improved convergence stability and final accuracy.

5. Results

- Best Test Accuracy: 74.53%
- Training Strategy: Early stopping used to prevent overfitting.

6. Strengths and Limitations

Strengths:

- Full ViT implemented from scratch.
- Proper regularization and training control.

Limitations:

- Dataset size is small for ViT models.
- No pre-training leads to lower accuracy than standard CNNs.

7. Conclusion

This implementation demonstrates a solid conceptual understanding of Vision Transformers and applies appropriate regularization techniques. While limited by dataset size, the model is well-suited for academic evaluation.