Vision Transformer Implemented from Scratch

Gourav Bajwa
240407

February 7, 2026

1 Introduction

This report describes the design and training of a Vision Transformer (ViT) model implemented entirely from scratch using PyTorch and trained on the CIFAR-10 dataset. In contrast to conventional convolutional neural networks (CNNs), which extract features through convolution and pooling operations, Vision Transformers treat images as sequences of fixed-size patches and process them using the Transformer architecture. This formulation enables the model to capture global context and long-range dependencies across the entire image.

2 Data Preprocessing and Regularization

2.1 Dataset

The CIFAR-10 dataset contains:

- 60,000 color images
- Image dimensions of $32 \times 32 \times 3$
- 10 object categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck

2.2 Preprocessing Steps

The following data augmentation and preprocessing techniques were applied during training:

1. **Random Horizontal Flip**
   Images are randomly flipped along the horizontal axis. *Effect:* Enhances generalization by encouraging invariance to left-right orientation.

2. **Random Crop with Padding**
   Images are padded and then randomly cropped back to a size of 32×32. *Effect:* Improves robustness to minor spatial shifts and translations.

3. **ToTensor**
   Converts images into PyTorch tensors and rescales pixel intensities to the range $[0, 1]$. *Effect:* Ensures compatibility with neural network computations.

4. **Normalization**
   Images are normalized using the dataset-specific mean and standard deviation. *Effect:* Stabilizes training dynamics and speeds up convergence.

2.3 Regularization Techniques

To reduce overfitting and improve training stability, the following regularization strategies were employed:

- **Dropout:** Randomly deactivates neurons during training to improve generalization.
- **Weight Decay (L2 Regularization):** Discourages excessively large weights and promotes smoother optimization.
- **Gradient Clipping:** Prevents exploding gradients in deep Transformer models.

3 Hyperparameter Tuning

3.1 Key Hyperparameters

Parameter	Tested Values	Selected Value
Patch Size	2, 4, 8	4
Embedding Dimension	256, 384, 768	364
Attention Heads	4, 8	8
Transformer Layers	4, 6, 8	6
Dropout Rate	0.1, 0.2	0.1
Batch Size	64, 128	128
Learning Rate	10^{-3}, 3×10^{-4}, 10^{-4}	3×10^{-4}

3.2 Observations

- Smaller patch sizes yielded higher accuracy but increased computational overhead.
- Increasing the number of Transformer layers improved performance up to a point, after which gains saturated.
- Higher learning rates resulted in unstable training behavior.

4 Vision Transformer Architecture

The Vision Transformer architecture represents images as sequences of tokens, enabling image classification using Transformer-based processing.

4.1 Patch Creation

Input images are divided into non-overlapping patches, each of which is treated as an independent token.

4.2 Patch Embedding

Each flattened patch is linearly projected into a fixed-dimensional embedding space, converting raw pixel values into meaningful feature vectors.

4.3 Positional Embedding

To compensate for the lack of inherent spatial structure in Transformers, positional embeddings are added to the patch embeddings to encode spatial information.

4.4 Class Token

A learnable class token is prepended to the sequence of patch embeddings. This token aggregates global information and is used for final classification.

4.5 Multi-Head Self-Attention

Through self-attention, each patch can attend to all other patches, allowing the model to capture global relationships rather than relying solely on local features.

4.6 Feed-Forward Network (MLP Block)

This component applies non-linear transformations to each token embedding, enhancing the expressive power of the model.

4.7 Classification Head

The final representation of the class token is passed through a linear layer to generate class scores.

5 Results

After training for 50 epochs, the model achieved the following performance:

- **Training Accuracy:** 95.03%
- **Best Test Accuracy:** 79.30%

The results indicate that the model successfully learned meaningful visual representations. The comparatively lower test accuracy can be attributed to the limited size of the CIFAR-10 dataset, as Vision Transformers typically benefit from large-scale pretraining.