End-Evaluation Report: Vision Transformer

Aayushman Tripathi
240025

Vision Transformer based Image Classification on CIFAR-10:

1. Introduction

Image classification has traditionally been dominated by Convolutional Neural Networks (CNNs). However, Vision Transformers (ViTs) have recently emerged as a powerful alternative by applying the transformer architecture, originally designed for Natural Language Processing, to image data.

In this project, a Vision Transformer model was implemented and trained on the CIFAR-10 dataset using PyTorch in Google Colab. The objective was to study the working mechanism of ViT, apply preprocessing and regularization techniques, perform hyperparameter tuning, and evaluate the classification performance.

2. Dataset Description

The CIFAR-10 dataset consists of 60,000 colored images of size 32×32 divided into 10 classes:

- Airplane
- Automobile
- Bird
- Cat
- Deer
- Dog
- Frog
- Horse
- Ship
- Truck

Split:

- Training images: 50,000
- Testing images: 10,000

3. Preprocessing Techniques Used

1. **Random Horizontal Flip**
    - Randomly flips images
    - Helps model generalize better
2. **Random Cropping (with padding)**
    - Adds variation in object position
    - Prevents overfitting
3. **Normalization**
    - Scales pixel values to a standard range
    - Stabilizes training

Effect of Preprocessing

Technique	Effect
Horizontal Flip	Makes model invariant to direction
Random Crop	Improves robustness to spatial variations
Normalization	Faster and more stable convergence

Regularization Techniques Used:

Regularization was necessary because Vision Transformers have a large number of parameters.

Methods Applied

1. **Dropout (inside Transformer layers)**
    - Prevents neurons from co-adapting
    - Reduces overfitting
2. **AdamW Optimizer (Weight Decay)**
    - Penalizes large weights
    - Improves generalization
3. **Data Augmentation**
    - Acts as implicit regularization

Effect of Regularization

- Reduced training–testing accuracy gap
- More stable loss curves
- Better generalization on test data

Mechanism of Vision Transformer

The Vision Transformer processes images similarly to how transformers process sentences.

Step 1: Image → Patches

A 32×32 image is divided into smaller patches.

Example:

- Patch size = 4×4
- Total patches = 64

Each patch acts like a "word token".

Step 2: Patch Embedding

Each patch is:

- Flattened into a vector
- Passed through a linear projection
- Converted into an embedding

Step 3: Positional Encoding

- Since transformers do not understand spatial position naturally, positional embeddings are added to each patch embedding to preserve location information.

Step 4: CLS Token

- A special learnable token is added:
- Stores global information
- Used for final classification

Step 5: Transformer Encoder

- Each encoder block contains:
- Multi-Head Self Attention
- Feed-Forward Network
- Layer Normalization
- Residual Connections

Step 6: Classification Head

- The CLS token output is passed through a linear layer to predict the image class.

6. Hyperparameter Tuning

Different configurations were tested to observe performance changes.

Parameters Tuned

Parameter	**Values Tried**
Patch Size	4
Embedding Dimension	128
Number of Heads	4
Transformer Layers	4
Learning Rate	1e-4, 3e-4

Observations

- Smaller patch size → better accuracy
  (more spatial detail captured)
- Too many layers → overfitting on CIFAR-10
- Learning rate 1e-4 gave better accuracy than 3e-4

7. Training Details

- Optimizer: AdamW
- Loss Function: Cross-Entropy Loss
- Epochs: 20
- Batch Size: 128

Observations

- ViT performs well but requires large datasets.
- CIFAR-10 is relatively small, which limits performance.
- CNNs often outperform ViTs on small datasets.

Model	Dataset	Accuracy
CNN (ResNet-like)	CIFAR-10	~85–90%
Our ViT	CIFAR-10	**73.92%**

Conclusion:

A Vision Transformer was successfully implemented and trained on the CIFAR-10 dataset. The project demonstrated:

- How images can be processed as sequences
- The role of self-attention in visual learning
- The importance of preprocessing and regularization
- The impact of hyperparameter tuning

Although the ViT did not outperform CNNs on CIFAR-10, it showed promising results and highlighted the potential of transformer-based architectures in computer vision tasks.