Image Classification using Vision Transformer and Convolutional Neural Network

Jay Vikrant
Email: jayvikarnt@gmail.com

***Abstract*—This project investigates two machine learning models for image classification: the Vision Transformer (ViT) and Convolutional Neural Network (CNN). We present their theoretical foundations, implementations, and evaluate their performance on the MNIST dataset (greyscale hand written numbers).**

I. Introduction

Image classification has seen significant advancements through various machine learning models. In this work, we focus on the Vision Transformer (ViT) and Convolutional Neural Network (CNN) models, and evaluate their performance on the MNIST dataset, which consists of grayscale images of handwritten digits.

II. Vision Transformer (ViT)

The Vision Transformer adapts principles from natural language processing for image data, treating image patches as tokens.

A. Model Architecture

The Vision Transformer model comprises the following components:

- **Patch Embedding:** The input image of size $H \times W \times C$ is divided into non-overlapping patches of size $P \times P$. Each patch is projected into a high-dimensional space using a convolutional layer, producing a sequence of patch embeddings $X \in R^{N \times D}$, where N is the number of patches and D is the embedding dimension.
- **Positional Encoding:** To retain spatial information, positional encodings $E_{pos} \in R^{(N+1) \times D}$ are added to the patch embeddings.
- **Self-Attention Mechanism:** The attention output A is computed as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V$$

  where Q, K, and V represent the queries, keys, and values obtained via linear projections of X.
- **Feed-Forward Network:** Each attention output is processed through a feed-forward network with two linear layers and a non-linear activation function (e.g., GELU).
- **Classification Head:** The final token (class token) is used to predict the class label:

$$\hat{y} = \text{softmax}(W_{cls} \cdot \text{Norm}(X_{cls}))$$

III. Convolutional Neural Network (CNN)

The Convolutional Neural Network is designed to perform image classification by leveraging hierarchical feature extraction.

A. Model Architecture

The CNN model consists of:

- **Convolutional Layers:** Convolutional layers extract hierarchical features from images. The convolution operation is defined as:

$$(I * K)(x, y) = \sum_{i,j} I(x + i, y + j) \cdot K(i, j)$$

  where K is the convolutional kernel.
- **Activation Function:** ReLU activation introduces non-linearity:

$$\text{ReLU}(x) = \max(0, x)$$

- **Pooling Layers:** Max pooling is used to downsample feature maps:

$$\text{MaxPool}(x) = \max(x)$$

- **Fully Connected Layers:** The output is passed through fully connected layers for classification:

$$\hat{y} = \text{softmax}(W_2 \cdot \text{dropout}(F_{\text{ReLU}}(W_1 \cdot \text{flatten}(x))))$$

IV. Results

A. Vision Transformer (ViT) Results

- **Training Loss:** [Insert Training Loss Data]
- **Test Loss:** [Insert Test Loss Data]
- **Test Accuracy:** [Insert Test Accuracy Data]

B. Convolutional Neural Network (CNN) Results

- **Training Loss:** [Insert Training Loss Data]
- **Test Loss:** [Insert Test Loss Data]
- **Test Accuracy:** [Insert Test Accuracy Data]

V. Conclusion

This paper demonstrates the application of Vision Transformers and Convolutional Neural Networks for image classification using the MNIST dataset. The Vision Transformer applies attention mechanisms and positional embeddings, achieving competitive results compared to the CNN, which relies on convolutional operations and pooling layers for feature extraction.