ML Lab Week 14: CNN Image Classification Report

Student: PES2UG23CS176

---

1. Introduction

This lab focused on designing, building, and training a Convolutional Neural Network (CNN) using PyTorch to classify hand gesture images into three categories: rock, paper, and scissors. The objective was to develop a model capable of accurately recognizing these gestures from the Rock Paper Scissors dataset, which contains over 2,000 images organized into class-specific folders.

---

2. Model Architecture

The CNN architecture implemented consists of two main components: a convolutional feature extractor and a fully-connected classifier.

Convolutional Feature Extractor: The feature extraction block contains three convolutional layers with progressively increasing channel depths:

- **Block 1:** Convolves from 3 input channels (RGB) to 16 output channels using a 3×3 kernel with padding=1, followed by ReLU activation and 2×2 max pooling
- **Block 2:** Convolves from 16 to 32 channels using the same kernel configuration, followed by ReLU and max pooling
- **Block 3:** Convolves from 32 to 64 channels, followed by ReLU and max pooling

Each max pooling layer reduces spatial dimensions by half, transforming the input from 128×128 to 64×64, then 32×32, and finally 16×16.

Fully-Connected Classifier: After three pooling operations, the feature maps are flattened from 64×16×16 dimensions (16,384 features) into a 1D vector. The classifier consists of:

- A linear layer mapping 16,384 features to 256 hidden units
- ReLU activation
- Dropout layer with p=0.3 for regularization
- Final linear layer mapping 256 units to 3 output classes

---

3. Training and Performance

Hyperparameters:

- **Optimizer:** Adam optimizer
- **Loss Function:** CrossEntropyLoss
- **Learning Rate:** 0.001
- **Number of Epochs:** 10
- **Batch Size:** 32

Training Results: The model demonstrated excellent convergence during training. The loss decreased consistently from 0.5934 in epoch 1 to 0.0021 by epoch 10, indicating effective learning without significant overfitting.

Test Accuracy: The model achieved a final test accuracy of **98.40%** on the held-out test set of 438 images, demonstrating strong generalization capability.

---

4. Conclusion and Analysis

The model performed exceptionally well, achieving near-perfect classification accuracy on the test set. The high accuracy (98.40%) indicates that the CNN architecture successfully learned discriminative features to distinguish between rock, paper, and scissors hand gestures.

Challenges: The primary challenge was determining the appropriate architecture depth and fully-connected layer dimensions. Calculating the flattened feature map size ($64 \times 16 \times 16 = 16,384$) required careful tracking of spatial dimension changes through multiple pooling layers.

Potential Improvements:

1. **Data Augmentation:** Implementing random rotations, flips, and brightness adjustments could improve model robustness to variations in hand orientation and lighting conditions.
2. **Learning Rate Scheduling:** Using a learning rate scheduler (e.g., ReduceLROnPlateau) could potentially achieve even faster convergence and possibly marginally higher accuracy by fine-tuning in later epochs.