Convolutional Neural Networks: Spatial Intelligence, Robustness, and Interpretability

Course: CS-671 Deep Learning and Applications

Assignment: CNN Coding Assignment

Objective: Understanding the spatial intelligence of CNNs

---

1 Problem Definition & Methodology

Convolutional Neural Networks (CNNs) are designed to extract spatially meaningful representations from image data through localized receptive fields and weight sharing. The objective of this experiment is to interpret how CNNs learn visual features by analyzing learned filters and activation maps.

The experiment was conducted using the MNIST dataset. A CNN model was trained and later used for:

- Extracting first-layer convolution filters
- Visualizing activation maps at different depths
- Understanding hierarchical feature learning

The convolution operation is mathematically defined as:

$$A^{(l)} = f(W^{(l)} * X^{(l-1)} + b^{(l)}) \tag{1}$$

where $W^{(l)}$ represents convolutional kernels and $f(\cdot)$ is the ReLU activation function. Pooling reduces dimensionality:

$$X_{pooled} = \max(X_{region}) \tag{2}$$

This allows efficient hierarchical feature abstraction.

2 Hyperparameter Tuning & Architecture Design

2.1 CNN Architecture

- Conv Layer 1: $1 \rightarrow 32$ filters, kernel 3×3, padding=1
- MaxPool: 2×2

- Conv Layer 2: $32 \rightarrow 64$ filters, kernel 3×3, padding=1
- MaxPool: 2×2
- Fully Connected Layer: $64 \times 7 \times 7 \rightarrow 128$
- Output Layer: 10 classes

Flatten dimension:

$$64 \times 7 \times 7 = 3136$$

2.2 Hyperparameters

- Optimizer: Adam
- Learning rate: 0.001
- Batch size: 64
- Epochs: 3
- Loss function: CrossEntropyLoss
- Activation: ReLU

Training loss decreased:

- Epoch 1: 0.1700
- Epoch 2: 0.0484
- Epoch 3: 0.0335

Final test accuracy:

$$\mathbf{98.9\%}$$

3 Results, Visualizations, and Interpretations

3.1 Question 3.1: Filter Gallery

The weights of the first convolutional layer were extracted and visualized to understand the primitive visual features learned by the CNN.

$$W \in R^{32 \times 1 \times 3 \times 3}$$

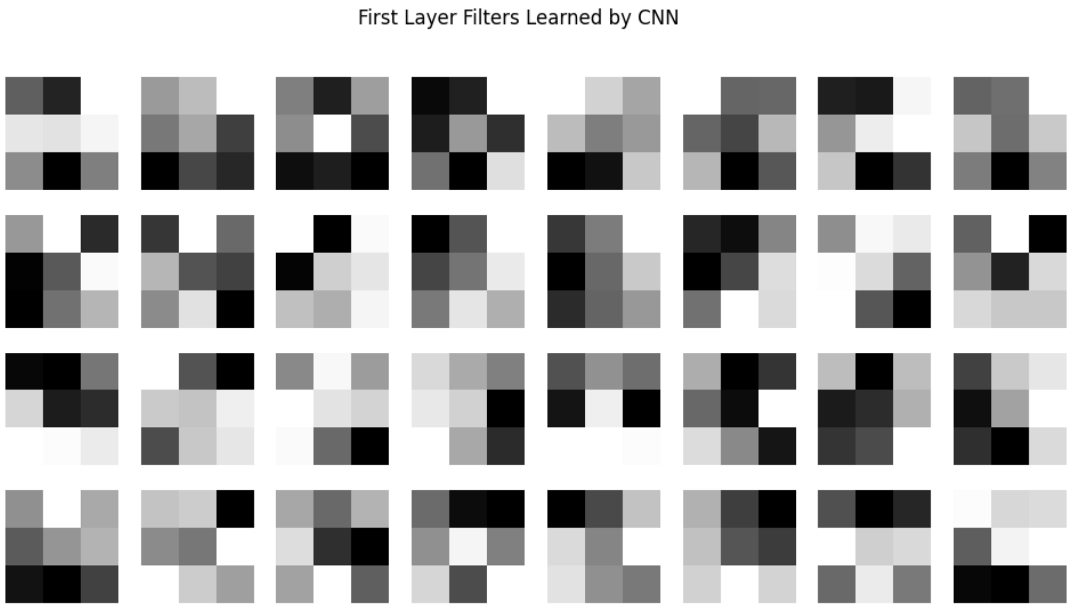

Figure 1: First-layer convolutional filters learned by the CNN

Interpretation:
The filters display structured patterns resembling:

- Edge detectors
- Gradient transitions
- Stroke orientation detectors

These filters act as fundamental building blocks for higher-level feature extraction. Similar behavior has been observed in biological vision systems, where early neurons respond to simple visual primitives.

3.2 Question 3.2: Receptive Field Experiment

A sample MNIST digit image was passed through the CNN, and activation maps were visualized at different convolutional depths to analyze spatial feature evolution.

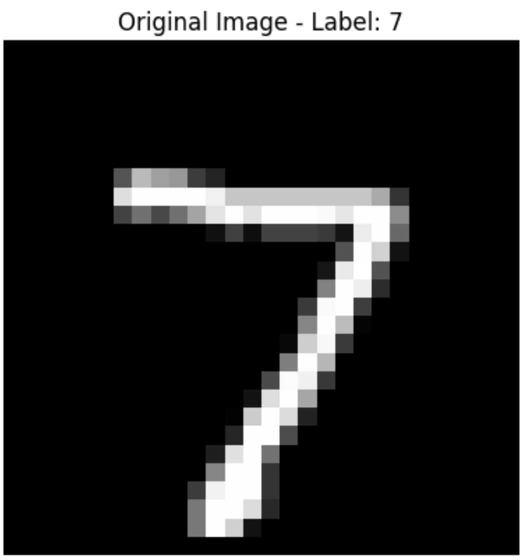

Figure 2: Original MNIST input image

3.2.1 Activation Maps After Conv1

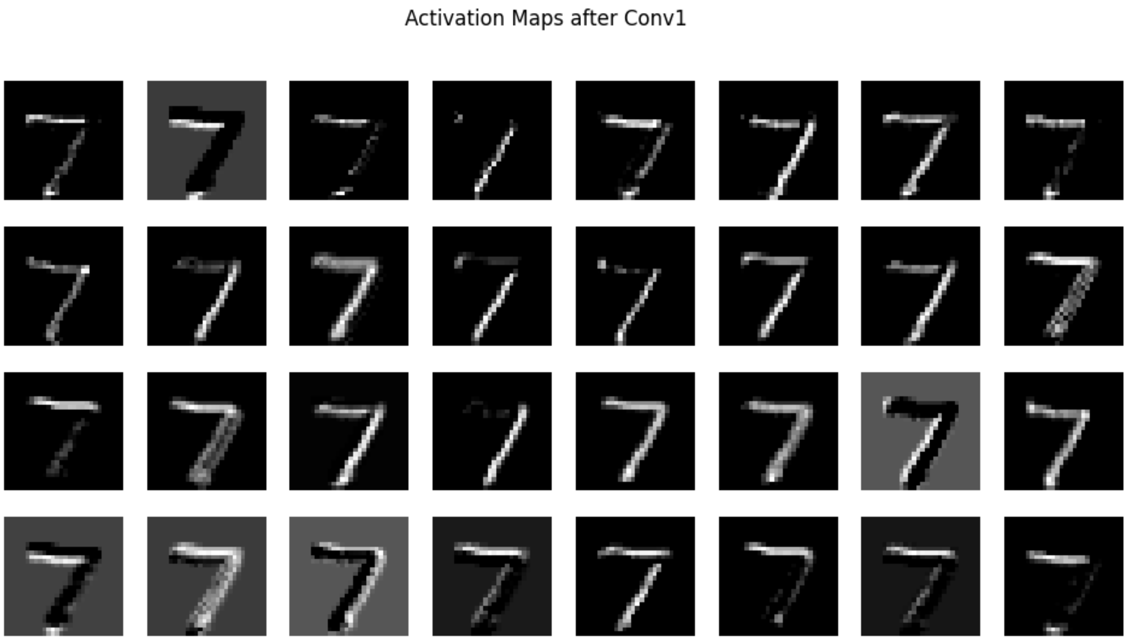

Figure 3: Activation maps after the first convolution layer

Observations:

- Strong responses at digit edges and stroke boundaries.
- Localized feature detection.
- High sensitivity to pixel intensity changes.

This layer captures primitive spatial information and operates with a small receptive field.

3.2.2 Activation Maps After Conv2

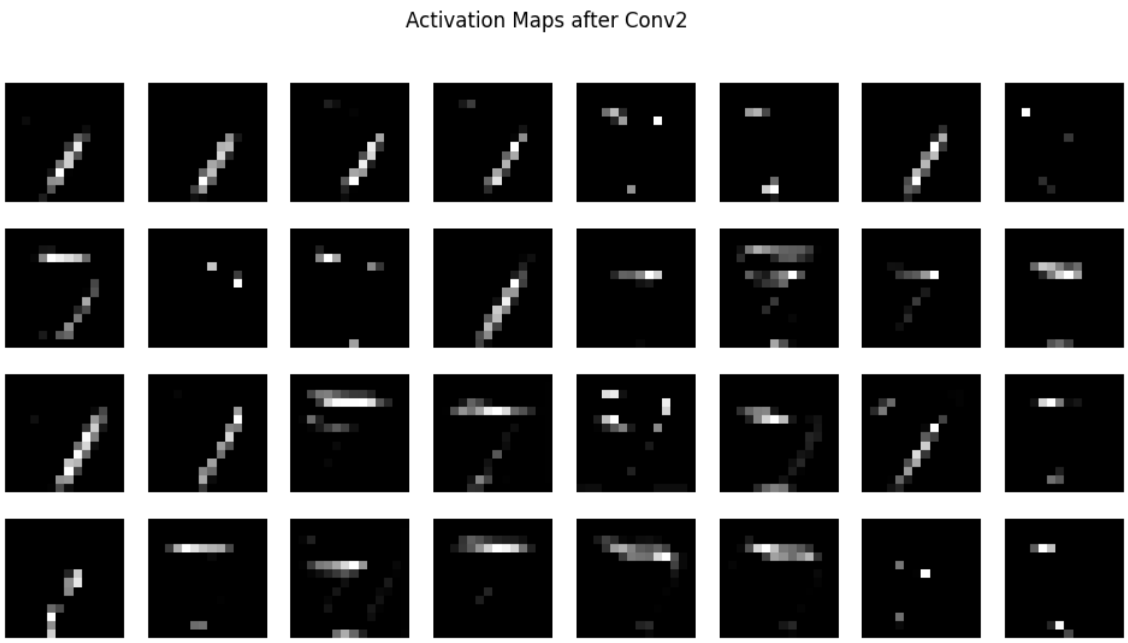

Figure 4: Activation maps after the second convolution layer

Observations:

- Larger spatial activation regions.
- Combination of multiple local features.
- Emergence of digit-specific structural representations.

This indicates increased receptive field size and deeper contextual understanding.

$$RF = k + (k - 1)(L - 1)$$

where k is kernel size and L is number of layers.

3.3 Interpretation of Hierarchical Feature Learning

The CNN progressively transforms visual information:

$$\text{Pixels} \rightarrow \text{Edges} \rightarrow \text{Shapes} \rightarrow \text{Digit Representation}$$

- Early layers focus on local spatial patterns.

- Intermediate layers combine features into shapes.
- Deeper layers capture semantic digit structure.

This confirms that CNNs learn hierarchical representations rather than memorizing pixel patterns.

4 Key Findings and Explanations

- CNNs automatically learn interpretable visual filters.
- First-layer filters behave as edge and gradient detectors.
- Activation maps demonstrate spatial feature progression.
- Receptive fields expand with network depth.
- Hierarchical learning enables robust digit recognition.

Major Insight:

CNNs transform raw pixels into semantic understanding through layered spatial abstract

The visualization results demonstrate that convolutional neural networks progressively learn structured internal representations, beginning with low-level edge features and advancing toward high-level semantic understanding of digit structures.

Conclusion:

Feature visualization confirms that CNNs are not black-box models. Instead, they perform hierarchical feature learning in which early layers detect primitive visual patterns, while deeper layers integrate these patterns into meaningful object-level representations.

This ability to expand receptive fields and capture spatial dependencies enables CNNs to achieve robust and interpretable image classification performance.