

Comparative Study of Feature Extraction vs. End-to-End Deep Learning for Object Localization

Max Ma

Yanbo Wang

Introduction to Machine Learning
NYU Tandon School of Engineering

Abstract

Object localization is a fundamental computer vision task that requires both accurate spatial reasoning and robust feature representations. In this project, we study the trade-offs between classical feature extraction pipelines and end-to-end deep learning approaches for object localization. Using a synthetic dataset derived from MNIST digits placed on a larger canvas, we formulate localization as a regression problem that predicts normalized bounding box coordinates. We compare four models: PCA with linear regression, PCA with decision tree regression, PCA with a shallow multi-layer perceptron (MLP), and a convolutional neural network (CNN) trained end-to-end. Results demonstrate a clear performance progression from classical baselines to deep learning models, highlighting the importance of learned spatial feature representations for localization tasks.

1 Introduction

Object localization aims to identify the spatial extent of an object within an image, typically represented by a bounding box. Unlike image classification, localization requires preserving spatial information and accurately estimating object size and position. Traditional machine learning approaches often rely on hand-crafted or pre-computed features, while modern deep learning methods learn hierarchical spatial features directly from raw pixels.

The goal of this project is to empirically compare these two paradigms in a controlled setting. Specifically, we investigate whether dimensionality reduction via Principal Component Analysis (PCA) combined with classical or shallow learning models can compete with an end-to-end convolutional neural network (CNN) for object localization. This comparison is conducted within the scope of an introductory machine learning course, emphasizing clarity, interpretability, and experimental rigor over architectural complexity.

2 Project Milestones

- **Milestone 1: Dataset Construction and Verification**

Target date: Week 1

Load the MNIST dataset and generate a synthetic MNIST-on-Canvas dataset by placing single digits on a larger canvas. Generate normalized bounding box labels (x_c, y_c, w, h) for each sample. Verify data correctness through visualization and define the pool/test data split.

- **Milestone 2: Baseline Model Implementation (PCA + MLP)**

Target date: Week 2

Implement the classical machine learning pipeline including image flattening, feature standardization, and PCA-based dimensionality reduction. Design and train an MLP regressor to predict bounding box parameters. Evaluate baseline performance using Mean Squared Error (MSE) and Intersection-over-Union (IoU).

- **Milestone 3: PCA Hyperparameter Selection via Cross-Validation**

Target date: Week 3

Perform K -fold cross-validation on the training pool to select the optimal number of PCA components. Ensure that preprocessing steps (standardization and PCA) are fit only on training folds to avoid data leakage. Select the PCA dimension that maximizes mean cross-validation IoU.

- **Milestone 4: End-to-End CNN Model Design and Training**

Target date: Week 4

Design an end-to-end convolutional neural network for bounding box regression using raw images as input. Train the CNN model using the same data split and loss function for fair comparison with the PCA+MLP baseline. Inspect intermediate tensor shapes to verify architectural correctness.

- **Milestone 5: Model Evaluation and Comparison**

Target date: Week 5

Evaluate both models on a held-out test set using MSE and IoU metrics. Perform quantitative comparison between the PCA+MLP and CNN approaches. Visualize predictions qualitatively by overlaying predicted and ground-truth bounding boxes on test images.

- **Milestone 6: Analysis, Reporting, and Final Presentation**

Target date: Week 6

Analyze experimental results and discuss trade-offs between classical machine learning and deep learning approaches. Summarize findings, limitations, and potential extensions of the project. Prepare the final report, presentation slides, and code repository for submission.

3 Problem Formulation

We formulate object localization as a supervised regression problem. Given a grayscale image $X \in \mathbb{R}^{64 \times 64}$ containing a single object, the task is to predict a normalized bounding box:

$$\mathbf{y} = (x_c, y_c, w, h) \in [0, 1]^4,$$

where (x_c, y_c) denote the center coordinates and (w, h) denote the width and height of the bounding box, normalized by image dimensions.

Model performance is evaluated using two metrics:

- Mean Squared Error (MSE) between predicted and ground-truth bounding box parameters.
- Mean Intersection-over-Union (IoU), measuring spatial overlap between predicted and ground-truth bounding boxes.

4 Dataset Construction

To avoid manual annotation while maintaining full ground-truth access, we construct a synthetic localization dataset using MNIST. Each sample is generated by randomly placing a scaled MNIST digit onto a 64×64 blank canvas. The bounding box is computed analytically based on the digit’s placement.

Digit scaling is restricted to ensure that the resized digit does not exceed its original 28×28 resolution, preventing shape inconsistencies during data generation. A total of 8,000 samples are generated and split into training and test sets with an 80/20 ratio.

5 Models

The models evaluated in this project are organized into three categories:

5.1 Shallow Learning Model: PCA + MLP

A multi-layer perceptron (MLP) is trained on PCA features to evaluate whether a trainable non-linear model improves performance over classical regression methods. The network consists of two hidden layers with ReLU activations and dropout regularization.

5.2 Deep Learning Model: End-to-End CNN

A convolutional neural network processes raw image pixels directly. Convolutional layers preserve spatial locality and learn hierarchical features, followed by fully connected layers that regress bounding box parameters. This model represents an end-to-end learning approach commonly used in modern vision systems.

6 Experimental Setup

We evaluate two regression-based localization models on the MNIST-on-canvas dataset. The dataset contains $N = 8000$ samples and is split into 80% training and 20% held-out test data. For the PCA+MLP baseline, each image is flattened to a 4096-dimensional vector, standardized using a training-set `StandardScaler`, and reduced using PCA with a candidate number of components k ; an MLP then predicts the 4 box parameters with a sigmoid output layer. The optimal k is selected via 5-fold cross-validation on the training set, using mean IoU as the primary selection metric (with MSE also reported). For the CNN regressor, the raw $1 \times 64 \times 64$ images are fed directly into a three-layer convolutional network followed by a small fully connected head. Both models are trained using MSE loss with the Adam optimizer (learning rate 10^{-3} , batch size 128) for 10 epochs, and final performance is reported on the test set using MSE and mean IoU, complemented by qualitative visualizations of predicted versus ground-truth bounding boxes.

7 Results

Table 1 summarizes the final test-set performance.

Table 1: Test-set performance comparison for object localization

Model	MSE ↓	Mean IoU ↑
PCA + MLP	0.000968	0.7778
CNN (End-to-End)	0.000682	0.8100

Overall, performance improves consistently from classical baselines to shallow neural networks and finally to the CNN model.

8 Qualitative Analysis

Qualitative visualizations reveal several common failure modes. PCA-based models often struggle with small objects and boundary placement due to the loss of spatial detail during feature compression. While center coordinates are generally predicted accurately, small errors in width and height significantly reduce IoU. In contrast, the CNN produces tighter and more precise bounding boxes, particularly near image boundaries, due to its ability to learn spatially localized features.

9 Conclusion

This project presents a controlled comparison between feature extraction pipelines and end-to-end deep learning for object localization. Results demonstrate that while PCA-based classical and shallow models provide reasonable performance, convolutional neural networks consistently achieve superior localization accuracy. These findings highlight the importance of learned spatial representations for vision tasks and reinforce the advantages of end-to-end deep learning approaches for object localization.