Multi-Digit Recognition System Using YOLOv8

Pavan Karke

January 2025

Introduction

This report details the implementation of a digit recognition system using YOLOv8, focusing on the detection and classification of multiple digits in images. The project encompasses dataset generation, model training, and inference evaluation.

Dataset Generation

Methodology

- Created a custom dataset generator using MNIST digits as base elements.
- Generated 15,000 training samples with the following specifications:
  - Image dimensions: 168x40 pixels.
  - Fixed 4 digits per image.

Data Augmentation Techniques

- **Random Scaling:** Scale factor range from 0.9 to 1.2, enhancing model robustness to digit size variations.
- **Random Rotation:** Rotation range from -15° to +15°, improving model resilience to orientation changes.
- **Dynamic Positioning:** Random vertical placement within image bounds, maintaining aspect ratios while ensuring digits fit within boundaries.

Annotation Format

- Implemented YOLO-format annotations with normalized coordinates.
- Each annotation includes:
  - Class ID (digit value).
  - Normalized center coordinates (x, y).
  - Normalized width and height.
- Stored in individual text files corresponding to images.

Model Training

Configuration

- **Base Model:** YOLOv8x (pre-trained).
- **Training Parameters:**
  - Epochs: 15.
  - Image size: 640x640.
  - Batch size: 8.

Training Metrics

The following table summarizes the key metrics observed during training:

Epoch	Time (s)	Train Box Loss	Train Cls Loss	Train DFL Loss	Precision (B)
1	352.6	0.80706	0.78731	1.0783	0.97538
2	702.6	0.65652	0.49883	0.98337	0.98345
3	1050.1	0.60441	0.45789	0.95631	0.98677
4	1396.7	0.54996	0.40986	0.93135	0.98663
5	1742.3	0.5124	0.38496	0.91769	0.98781
6	2084.3	0.4883	0.31793	0.9154	0.98551
7	2428.9	0.4517	0.29326	0.89636	0.9913
8	2773.7	0.40987	0.2755	0.8779	0.99075
9	3118.1	0.37333	0.25092	0.86182	0.99296
10	3463.5	0.3446	0.23616	0.85241	0.99261
11	3809.2	0.32594	0.22356	0.84346	0.99512

12	4154.2	0.31391	0.21792	0.84139	0.99449
13	4495.4	0.2965	0.20365	0.8367	0.99385
14	4837.1	0.27463	0.18624	0.82535	0.99554
15	5176.6	0.26074	0.17844	0.82171	0.99574

Visualization

Figure 1: Image with ground truth labels on the validation set.

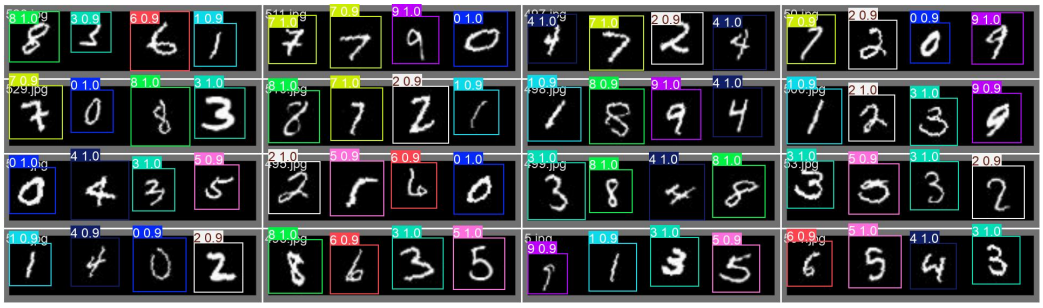

Figure 2: Image with predicted labels and probabilities for the same digits.

Comparison with ViT

A Vision Transformer (ViT) with an MLP head and MSE loss was also attempted. It achieved only 10% accuracy at 10 epochs and required significant time to converge. Therefore, YOLOv8 was chosen for its superior performance and efficiency.

Inference and Evaluation

Single Image Inference (Generated dataset):

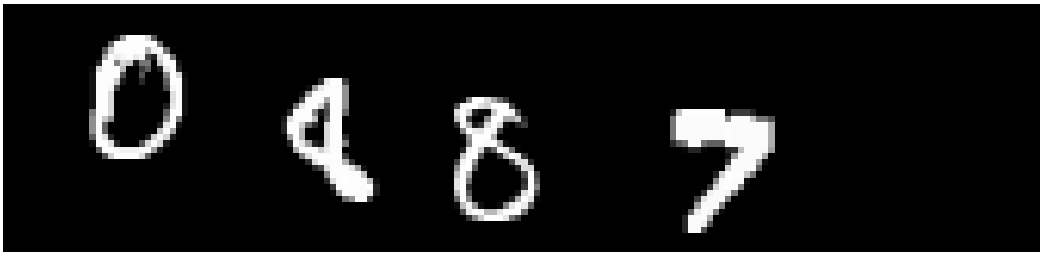

Figure 3: Test Image

- Prediction:
  - Class 0: 0.949 confidence.
  - Class 8: 0.928 confidence.
  - Class 8: 0.923 confidence.
  - Class 7: 0.921 confidence.
- High confidence scores (>0.92) for detected digits.

Dataset Evaluation (Given dataset)

- Evaluation on combined data from multiple .npy files.
- Test set accuracy: 69.33%.

Scope of improvement

- Further fine-tuning the model by adding additional layers and using transfer learning on dataset can be done.
- i.e. MLP can be added to predict sums instead of bounding boxes, which may further improve accuracy.