Traffic Sign Detection with YOLOv1: A Theoretical Exploration

Introduction

This project implements the **YOLOv1 (You Only Look Once)** model for traffic sign detection. YOLOv1 is a **single-stage object detection** model that simultaneously predicts bounding boxes and class probabilities for objects in an image. The model was trained on a dataset of traffic sign images, and the project includes code for data preprocessing, model training, and evaluation. The goal of this project was to gain a deeper understanding of the YOLO algorithm and its application to object detection.

YOLOv1 Architecture

Grid-Based Detection

YOLOv1 divides an image into an $S \times S$ grid. Each grid cell is responsible for predicting the bounding boxes and class probabilities for objects whose center falls within the cell. Let $S = 7$ (the original YOLOv1 setting), and let each cell predict:

- B bounding boxes (e.g., $B = 2$).
- The confidence score for each bounding box.
- The class probabilities for C classes.

Predictions

Each grid cell predicts a tensor of size:

$$S \times S \times (B \times 5 + C)$$

where each bounding box prediction consists of:

- (x, y): The coordinates of the bounding box center relative to the grid cell.

- (w, h): The width and height of the bounding box relative to the entire image.
- c: The confidence score, representing the probability that an object is present and the accuracy of the bounding box.

The confidence score is defined as:

$$c = P(\text{object}) \times \text{IoU}_{\text{pred, truth}}$$

where $\text{IoU}_{\text{pred, truth}}$ is the Intersection over Union (IoU) between the predicted and ground-truth bounding boxes.

Bounding Box Prediction

The bounding box parameters (x, y, w, h) are predicted relative to the grid cell:

$$x = \sigma(t_x) + i, \quad y = \sigma(t_y) + j, \quad w = e^{t_w} p_w, \quad h = e^{t_h} p_h$$

where:

- i, j are the grid cell indices.
- $\sigma(t_x)$ and $\sigma(t_y)$ ensure x, y are within the cell.
- p_w, p_h are the anchor box dimensions.

Loss Function

YOLOv1 uses a custom loss function to balance localization and classification errors. The loss L is a combination of three components:

$$L = \lambda_{\text{coord}} L_{\text{bbox}} + L_{\text{obj}} + \lambda_{\text{noobj}} L_{\text{noobj}} + L_{\text{class}}$$

where:

- λ_{coord} (e.g., 5) emphasizes localization accuracy.
- λ_{noobj} (e.g., 0.5) reduces the impact of background predictions.

The terms are defined as:

1. **Localization Loss** (for bounding box coordinates):

$$L_{\text{bbox}} = \sum_{i=1}^{S^2} \sum_{j=1}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2 \right]$$

2. **Objectness Loss** (for confidence scores):

$$L_{\text{obj}} = \sum_{i=1}^{S^2} \sum_{j=1}^{B} \nVdash_{ij}^{\text{obj}} (c_i - \hat{c}_i)^2$$

3. **No-Object Loss**:

$$L_{\text{noobj}} = \sum_{i=1}^{S^2} \sum_{j=1}^{B} \nVdash_{ij}^{\text{noobj}} (c_i - \hat{c}_i)^2$$

4. **Classification Loss** (for class probabilities):

$$L_{\text{class}} = \sum_{i=1}^{S^2} \nVdash_{i}^{\text{obj}} \sum_{c \in C} (p_i(c) - \hat{p}_i(c))^2$$

Data Preprocessing

- **Resizing**: All images resized to 448×448 pixels.
- **Normalization**: Pixel values scaled to $[0, 1]$.
- **Augmentation**: Random rotations, flips, and brightness adjustments to enhance robustness.

Training

- **Optimizer**: Stochastic Gradient Descent (SGD).
- **Learning Rate**: $\eta = 0.001$, with a learning rate decay.
- **Batch Size**: 16.
- **Epochs**: 100.

Evaluation Metrics

- **Mean Average Precision (mAP)**: Measures the average precision across all classes.
- **Intersection over Union (IoU)**: Quantifies the overlap between predicted and ground-truth bounding boxes:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Insights

- YOLOv1 offers a fast, single-pass approach to object detection, making it suitable for real-time applications.
- The model performs well for traffic sign detection but can struggle with small objects due to the coarse 7×7 grid.
- The theoretical understanding gained from this project highlights the trade-offs between detection speed and accuracy.

Conclusion

This project provided a practical implementation of the YOLOv1 model for traffic sign detection and deepened the theoretical understanding of the YOLO algorithm. YOLOv1's single-stage approach offers a balance of speed and accuracy, making it a powerful tool for real-time object detection tasks.