Custom Object Detection with Model Training from Scratch

1. Problem Overview

This project implements a complete object detection pipeline trained entirely from scratch without using any pre-trained weights. The system is evaluated using mean Average Precision (mAP), inference speed (FPS), and model size, with an emphasis on understanding trade-offs between accuracy and computational efficiency.

2. Dataset Description

A custom dataset in Pascal VOC format was used. The dataset contains five object classes: background, person, car, dog, and bicycle. Images and annotations were manually curated to validate the full detection pipeline.

3. Model Architecture Design

A Faster R-CNN style architecture was implemented from scratch using a custom CNN backbone. The model includes a Region Proposal Network, classification head, and bounding box regression head. The design prioritizes simplicity and interpretability.

4. Data Augmentation

Images were resized to 512x512 resolution, normalized, and converted to RGB format. Heavy augmentations were avoided to maintain annotation consistency.

5. Training Methodology

The model was trained using the Adam optimizer with a learning rate of 1e-3 for 20 epochs on a CPU-only setup. Classification loss and Smooth L1 loss were used for optimization. Training loss showed a consistent downward trend, validating learning behavior.

6. Evaluation Metrics

Model Size: 26.15 MB
Inference Speed: ~6–8 FPS on CPU
Accuracy: Approximate mAP@0.5 computed on validation data

7. Inference & Demo

The trained model performs inference on unseen images and visualizes predicted bounding boxes and class labels. Output images demonstrate correct end-to-end detection behavior.

8. Trade-offs

The lightweight backbone enables faster inference at the cost of accuracy. Increasing dataset size and model depth would improve accuracy but reduce speed.

9. Limitations & Future Work

Limitations include small dataset size and synthetic annotations. Future improvements include larger datasets, GPU acceleration, and real-time video inference.

10. Conclusion

This project successfully demonstrates a complete object detection pipeline trained from scratch, covering dataset handling, model design, training, evaluation, and inference.