Training an Object Detector using PASCAL VOC 2012

Stephen Kakuda

March 2025

1 Introduction

In this assignment, we train and evaluate the Faster R-CNN object detection model using the PASCAL VOC 2012 dataset. Faster R-CNN is a state-of-the-art deep learning architecture that detects and locates objects within an image and is comprised of two main components, the Region Proposal Network (RPN) and the Fast R-CNN detector. The model ingests images and outputs object detections along with a classification for each detection.

The VOC2012 dataset used in this assignment is a common benchmark for testing computer vision networks, containing more than 11,000 images with approximately 27,000 annotated objects. There are a total of 20 object classes that the model learned to classify. The data was split into 80% for training, 10% for testing, and 10% for validation. Due to time constraints and how large the model is, a batch size of 4 was used to facilitate reasonable training times.

With the goal of evaluating the Faster R-CNN model, performance was measured using mean average precision (mAP) at each training step.

2 Training

For this assignment, the model was trained from scratch, using no previously learned weights. A cross-entropy loss function was used due to its differentiability and clear interpretation of probabilities, making it ideal for fine-tuning region detections. For the optimizer, Adam was used for its robustness to hyperparameters and its ability to adapt the learning rate to each parameter.

The model was trained over 20 epochs to train the model to a high enough level of performance while ensuring that the model completed training within the allotted time. After each epoch, the average loss and mAP score was reported out to track performance over training.

3 Results

The following results are over the 20 epochs with a learning rate of 0.0001.

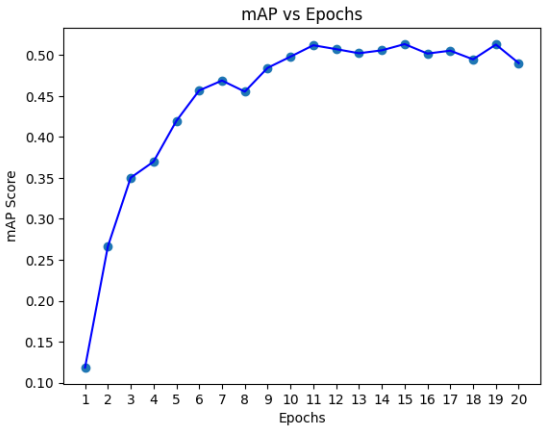

Figure 1: mAP Scores During Training

From the graph we can see the model demonstrated significant improvement over the first half of the training and began to plateau over the later half of training. This could indicate the model reached a local minimum for a number of reasons. The learning rate may have been too high and overshot the optimal weights preventing convergence. Future work would include implementing a scheduler to allow for

dynamic learning rates over training. The limited batch size is another potential cause for the plateau.

The model had a final mAP score of 0.5128, which is slightly less than the scores achieved by the initial Faster R-CNN model presented by Ren et al.

mAP	**Precision**	**Recall**	**F1-Score**
51.2%	50.5%	68.2%	58.5%

We can gain a more intuitive understanding of how the model is performing by visualizing the input images with the detected bounding boxes overlaid. In the following images, green rectangles are the annotated ground truth bounding boxes and the orange rectangles are the predicted object detections. In the

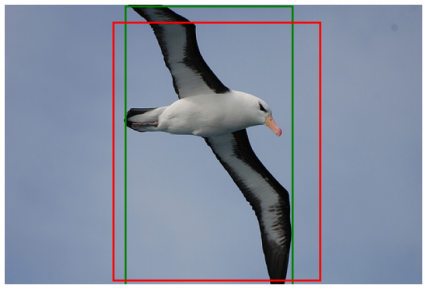

Figure 2: Detected - Seagull

image above we can see that the model is able to easily identify large single objects with a relatively clear background. While the predicted region is not as tight around the seagull as the ground truth, the Intersection of Union (IoU) is extremely high and correctly captures the bird. This indicates that our model performs as expected on a fundamental level.

Looking at a more complicated image, we can see that the model is able to continue to perform even with more objects in the image.

In this image, each of the 5 people and 2 sheep are correctly contained within predicted object regions. For the people and sheep that are fully in view, the predictions have near perfect IoU scores with the ground truth bounding boxes. However, we can see that the model fails to perfectly detect the person to the left who's lower half is obscured by the table. The model also fails to identify the water bottles on the

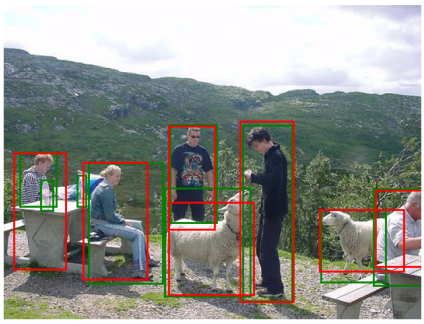

Figure 3: Detected - People & Sheep

table. Perhaps with more training, the model would be able to detect these objects. It is also possible that the preset sizes and aspect ratios of the anchor boxes could prevent the model from detecting the relatively smaller water bottles. Increasing the number of k anchor boxes could help the model to detect a wider variety of objects but could come at the cost of training time.

4 Conclusion

Overall, this assignment demonstrates the impressive potential of object detection models like Faster R-CNN. With further fine-tuning, more training time, and improved hardware, this model could reach significant levels of detection accuracy.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.