Training a custom pigeon detector with YOLOv7

Joule Voelz, Dominik Wielath

March 2023

Abstract

In this project, we train and test the YOLOv7 model on a custom dataset of pigeons and monk parrots. We experiment by training the model on randomized subsamples of the full dataset to assess the model's performance in terms of precision and recall on different sample sizes. We find that precision and recall improve rapidly with training sample size and start to plateau near a sample size of 160. In general, the model is more successful at identifying monk parrots than pigeons. The model has difficulty in distinguishing pigeons from other dark objects in the foreground when pigeons are small relative to the other objects. This reveals a bias in our training set that can be corrected in future experimentation.

1 Introduction

YOLO (You Only Look Once) is a real-time object detection model first developed in Redmon et al (2015). [1] In contrast to previous approaches to object detection which identify "proposal regions" where objects may be found, YOLO looks at the whole image with one convolutional neural network and gives class predictions for bounding boxes within the image. Many subsequent versions of the model have been developed and continue to be state-of-the-art in object detection. In this project, we fine-tune the YOLOv7 model to detect our arch-nemeses: pigeons.

The idea for our project arose from our experiences living in a place with many pigeons: the center of Barcelona. Often, they occupy balconies or ceilings and leave their excrement behind. This is not just annoying for humans wanting to enjoy breakfast or barbecue outside, but also threatens public health, as pigeons can transmit diseases such as Cryptococcosis, Histoplasmosis, and Psittacosis. Further, pigeon excrement increases the erosion of buildings, which can be particularly an issue in cities with many old structures and monuments.

As the number of people exposed to this issue is high, many products on the market are advertised as solutions to the problem. From our own experience, placing a plastic owl or glitter pompoms on a balcony might have a short-term effect by shooing away pigeons. However, after a brief period, the pesky creatures get accustomed to these static solutions and return to occupy their favored places. However, while pigeons get used to stationary devices, they remain afraid of objects moving in their direction. Shooing them away with a broom remains an effective option, but requires one to constantly observe the balcony with a broom in reach to be prepared for the case of a winged interloper landing on one's property.

These issues made us realize there is a need for a more advanced solution leveraging the capabilities of deep learning techniques and modern computational power. In particular, we want to use computer vision to detect

(a) Monk parrot

(b) Pigeon

Figure 1: We collected and hand-annotated a custom dataset.

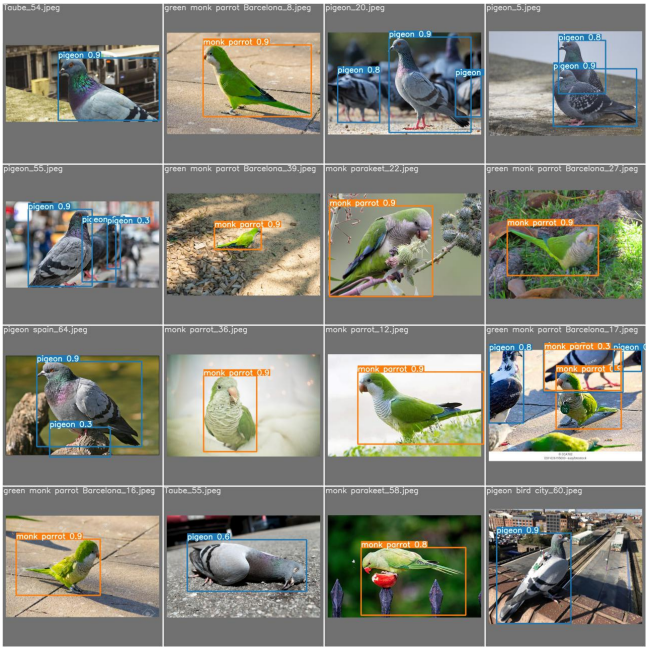

Figure 2: The model achieved high precision and recall after 100 epochs at an initial learning rate of 0.001.

pigeons and transmit their location to a robotic device executing a rapid movement of a broom in the direction of the threat. To create such a mechanism, we must overcome several challenges to make it effective and safe. It needs to identify and locate pigeons in real-time while not proposing a threat to other animals or humans using the balcony.

The YOLO algorithm is the perfect instrument to resolve some main challenges. It lets us train the model to identify pigeons in real-time and distinguish them from other animals and humans. In this project, we fine-tune the YOLOv7 model to distinguish between pigeons and one of Barcelona's other iconic and more aesthetically pleasing species: the monk parrot. To test the model's ability to perform with little training, we experiment by training the model on random samples of different sizes and recording precision, recall, and mean average precision at two different thresholds.

2 Data

In order to train our model to detect pigeons and monk parrots, we collected and hand-annotated over 240 pictures from the Internet (Figure 1) using the open-source tool LabelImg. The dataset was nearly balanced, with 113 photos featuring mostly pigeons and 131 photos featuring mostly monk parrots. In only a few training images did the two types of birds appear together. Pictures were a variety of sizes and aspect ratios.

3 Methods

For our custom bird detector, we chose to train the YOLOv7 model from the official repository of YOLOv7 (Wang et al 2022).[2] Before running experiments, we attempted to fine-tune the pre-trained model on the entire dataset with a 80/20 test-train split. Through experimentation, we noticed that the default learning of rate of 0.01 was too big and that the model was losing precision during training. Thus we changed the learning rate to 0.001, which resulted in slower but much more stable learning. Within 100 epochs, the model achieved a precision over 0.9 and recall over 0.87 on the test set (Figure 2). This result made us confident that our model could work, and thus we proceeded with our experiments.

We wanted to test model performance on 4 different sample sizes: 40 pictures, 80 pictures, 160 pictures, and 240 pictures. For each sample size, we drew 5 random samples with an 80/20 train-test split. Then for each sample, we trained the model for 100 epochs with an initial learning rate of 0.001 and recorded the results of running the trained model on the test set.

For each test run, we recorded four metrics: precision, recall, and mean average precision at 0.5 and 0.95 thresholds, respectively. Precision measures the accuracy of the model's predictions. It is the number of true positives divided by the number of true positives and false positives. In other words, when the network makes a prediction that a bird is present, is it predicting the right bird? If the network fails to predict a pigeon that is actually in the image, this does not enter into the precision.

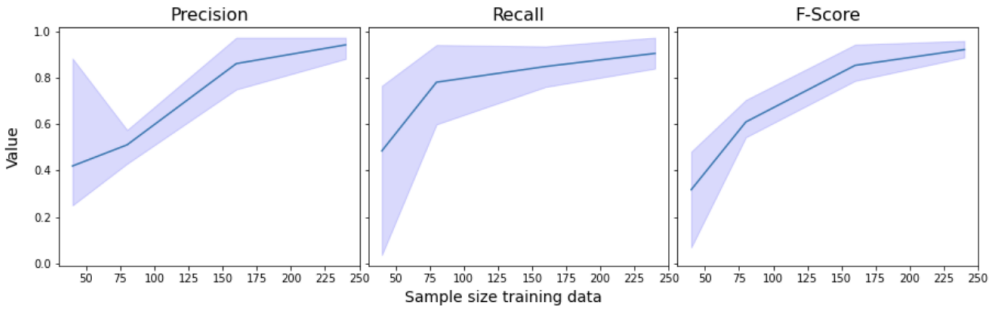

Figure 3: Experimental Results: Precision and Recall

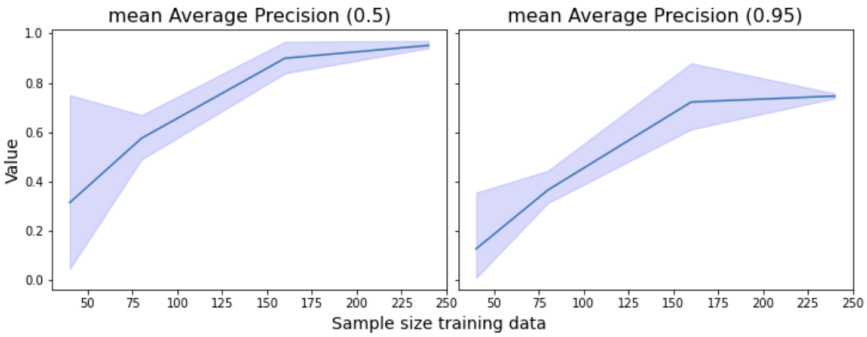

Figure 4: Experimental Results: Mean Average Precision

Recall measures how well the network can find all true positives. It is measured as the number of true positives divided by the number of true positives and false negatives. For example, if there are two parrots in an image but the network only predicts one parrot, it will have a precision of 1 but a recall of 0.5. The F1-score, which we also report, is the harmonic mean of the precision and the recall.

In object detection, precision and recall are calculated in terms of the intersection over union (IoU) of the predicted bounding box and the ground truth bounding box. The IoU is the area of these boxes' intersection divided by the total area of the boxes. Naturally, a higher IoU means that the predicted bounding box overlaps more with the ground truth box. We classify the network's predictions as true positives and false positives with respect to an IoU threshold. For example, if the if IoU threshold is 0.4, and the IoU value for a prediction is 0.6, then we classify the prediction as true positive. On the other hand, if IoU is 0.3, we classify it as a false positive. In our implementation of YOLOv7, the IoU threshold is 0.2.

By default, YOLOv7 reports the mean average precision (mAP) at thresholds of 0.5 and 0.95. That is, it reports the average precision for when the IoU threshold is set to 0.5 and 0.95. Clearly, 0.95 is a very high threshold and a high mAP0.95 means that the network is performing very well.

4 Results

The results of our experiments were as expected. In general, precision was higher for monk parrots than for pigeons. This makes sense because monk parrots, with their distinctive green feathers, stand out much more from a city background than the dowdy gray pigeon. As sample size increased, so did precision and recall (Figure 3). The model's ability to recognize and accurately label pigeons and birds was low with a sample size of 40 (and thus only 32 training pictures). However, its performance greatly improved and leveled out at a sample size of 160 (128 training pictures). If the IoU threshold is set to 0.5, the model has nearly perfect mean average precision with sample size 240 (192 training pictures) (Figure 4). Training on a set this size for 100 epochs took approximately one hour. These results demonstrate that with a relatively small training set, one can train a fairly accurate pigeon detector that can differentiate between pigeons and monk parrots.

5 Discussion

In testing the model, we came across a few limitations of our training approach. Namely, that we should have included humans as a third class in our training. For example, when we showed the model images that included humans, the model predicted that the humans were pigeons. We believe this is because most of the training data were close-up photos of pigeons and parrots. The model was trained to find large pigeons at the center of the image. Thus when it is shown an image where humans are large and pigeons are small, it assumes that the humans are pigeons. This deficiency must be addressed before being put into production, lest the pigeon shooing device whack its own owner if he or she steps out onto their balcony for a pigeon-free smoke. This issue could be solved by training the model to detect humans as well, and showing it more images where pigeons are small relative to the whole image.

Given more time, we would like to add more birds to our training set to improve the utility of the model. We would also like to assess the speed of the model's predictions, which would be crucial in designing a device that recognizes and responds to real-time pigeon assaults. To be feasible for low-cost production, the algorithm would have to work with lower-resolution streaming video on a mobile device. Once we fine-tune our model for detection, we would have to optimize it for those conditions.

6 Conclusion

In this project, we trained and tested the YOLOv7 model on a custom dataset of pigeons and monk parrots. We experimented by training the model on randomized subsamples of the full dataset to assess the model's performance in terms of precision and recall. We found that precision and recall improved rapidly with training sample size and started to plateau near a sample size of 160. In general, the model was more successful at identifying monk parrots than pigeons. The model had difficulty in distinguishing pigeons from other dark objects in the foreground when pigeons were small relative to the other objects. This revealed a bias in our training set that can be corrected with a more diverse training set and further experimentation.

References

[1] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: (2015). URL: https://doi.org/10.48550/arXiv.1506.02640.

[2] Hong-Yuan Mark Liao Chien-Yao Wang Alexey Bochkovskiy. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". In: (2022). URL: https://doi.org/10.48550/arXiv.2207.02696.