A Comprehensive Evaluation of YOLOv7 as a Single-Stage Object Detector in Comparison to R-CNN and Vision Transformer

Yoobin Cheong, Yeong Koh and **Yoon Tae Park**
Center for Data Science
New York University
60 5th Avenue, New York, NY
{yc5206, yk2678, yp2201}@nyu.edu

Abstract

Object detection is a fundamental task in computer vision, which involves identifying and localizing objects within an image or video. Over the years, several object detection algorithms have been developed, including single-stage detectors such as YOLOv7 (Wang et al., 2022), two-stage detectors such as R-CNN (Ren et al., 2016), and recent transformer-based detectors like Vision Transformer (Dosovitskiy et al., 2021). In this research project, we propose a comprehensive evaluation of YOLOv7 as a single-stage object detector in comparison to R-CNN and Vision Transformer. We provide a detailed overview of each algorithm, including their architecture, training methodology, and inference procedure. We then compare MAP (Mean Average Precision) on benchmark datasets, such as COCO, Pascal VOC, and a custom dataset, to assess their performance. As a potential extension, we consider the trade-off between speed and accuracy and compare the detection speed of each algorithm. This aspect is particularly important in real-time applications, where the detection speed can directly impact the system's performance. Overall, this project aims to provide a comprehensive analysis of state-of-the-art object detectors and help researchers and practitioners select the most appropriate algorithm for their specific use case.

1 Introduction

Object detection is a crucial task in computer vision, involving the identification and localization of objects within images or videos. Over the years, many object detection algorithms have been developed, each with its unique advantages and disadvantages. Among these, YOLOv7 (You Only Look Once) is a popular single-stage detector that achieves real-time performance by predicting object bounding boxes directly from the image.

As with any machine learning model, it is essential to ensure that YOLOv7's performance is both reproducible and generalizable to new datasets. Therefore, in this paper, we aim to conduct a comprehensive evaluation of the YOLOv7 model's reproducibility, generalization and performance.

2 Method/Approach

To evaluate the performance of the YOLOv7 model, we set up three different tests. First, we validate that our implementation of the YOLOv7 model reproduces the results reported in the original paper. Then, we evaluate the model's ability to detect objects in new datasets not used during training, including those with varying object sizes, occlusion, and lighting conditions. Finally, we compare the performance of YOLOv7 with R-CNN and Vision Transformer on benchmark datasets such as COCO, Pascal VOC, and a custom dataset. We assess the MAP score of each algorithm and consider the trade-off between speed and accuracy. Through this evaluation, we aim to provide researchers and practitioners with a detailed analysis of the YOLOv7 model and its performance compared to other state-of-the-art object detection algorithms.

3 Experiments

In the initial experiment, we test the reproducibility of the YOLOv7 model by training it on the COCO train dataset and comparing the results to those reported in the original paper. Due to limited resources, we run the experiment for 20 epochs, but we plan to increase the number of epochs to at least 100, ideally 300, to ensure a fair comparison. Next, we evaluate the YOLOv7 model's generalization performance by investigating how well it adapts to new datasets using concepts of transfer learning. We first examine the model's performance on a dataset it has not been trained on, namely Pascal-VOC, using a pre-trained YOLOv7-v7 model. We also test how well the YOLOv7-v7 model can gen-

eralize to new datasets without using pre-trained weights by training it on Pascal-VOC. For all experiments, we use a batch size of 16 and plan to increase the number of epochs for the reproducibility experiment. In addition to the above experiments, we plan to explore and compare the performances of R-CNN and Vision Transformer with that of YOLOv7.

4 Results and Conclusions

The results of our initial experiments so far can be summarized as follows:

Reproducibility We found that YOLOv7 exhibits good reproducibility, as we were able to achieve the same level of performance on the COCO test dataset using pretrained weights, and the inference on the given example image worked well. Using a non-pretrained model to train on the COCO dataset took a relatively long time, but we are seeing good results as we increase the number of epochs. We will check the final results after completing the training.

Generalization In terms of generalization, we conducted two tests. The first test, which involved using pretrained weights to train the Pascal VOC 2012 dataset, showed good results, with a mean average precision (MAP) of 0.685 at 0.5 intersection-over-union (IoU) and 0.492 at IoU of 0.5 to 0.95. The second test, which involved training a non-pretrained model on the same dataset, resulted in relatively low performance with a MAP of 0.419 at 0.5 IoU and 0.243 at IoU of 0.5 to 0.95. This was expected, as we used a small number of epochs for training, and the model was underfitted to the dataset. We expect to see better results by increasing the number of epochs.

Performance We are planning to measure the performance of the R-CNN and Vision Transformer models on the COCO and Pascal VOC 2012 datasets. We will use similar parameters to ensure that these models are in a similar environment to YOLOv7, and we will compare their performance with that of the YOLOv7 model.

5 Challenges

One of the main challenges we have faced so far was the limited resources we had available for training the models. The experiments we conducted used a reduced number of epochs and batch sizes, which may have resulted in outcomes that are inconclusive and yet to be verified using more epochs for training.

6 Future Scope

As future work, we plan to further investigate the trade-off between speed and accuracy in object detection algorithms. We also plan to further explore the use of transfer learning techniques for improving the generalization performance of object detection models by exploring how the performance of these models can be improved by incorporating additional data.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2022. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.