Comparitive analysis of the state-of-the-art Architectures for Object Detection

Yash Dharmadhikari
Virginia Tech
yashd@vt.edu

Abstract

This project explores the state-of-the-art Neural Network architectures - Trasnfomer-based Vision Transformer(ViT) and CNN-style YOLOv7 algorithm for object detection. We aim to compare the performance of both of these architectures on similar object detection tasks and shed more light on the future trend of Object Detection and other areas of Computer Vision.

1. Introduction

Object Detection has been a field of extensive research since the rise of Deep Learning approaches. Many algorithms and architectures have been proposed to solve the problem of object detection effectively, ranging from traditional machine learning methods wherein features of interest were handcrafted from images for classification to modern Deep Learning frameworks, which perform end-to-end object detection. Some of the most influential object detection algorithms in the past were - R-CNN, Fast-R-CNN, Faster-RCNN, YOLO, SSD, Retinanet, etc. The YOLO series [3] has been one of the most popular algorithms for this task, and multiple versions of this series have been proposed, with the latest one being the YOLOv7 [4] algorithm. There has been a revolution in the field of Natural Language Processing as well with the introduction of the Transformer Architecture in 2017 [2]. Following the remarkable achievements of Transformers in NLP, transformer-based architecture has also been developed in Computer Vision called the Vision Transformer(ViT) [1]

Transformer-based architectures are relatively new in the field of Computer Vision. Despite their novelty, they have shown promising results in most of the major areas of Computer Vision. The Vision Transformer architecture [1] was introduced in 2021 and has significantly impacted the trend in Computer Vision; the authors of this paper successfully proved that ViT models could outperform the current state-of-the-art architectures like ResNets on multiple image recognition benchmarks given a huge amount of dataset.ViT models, when trained on huge datasets like ImageNet-21k and JFT-300M, approach or beat ResNets, getting an accuracy of 88.55% on ImageNet, 90.72% on ImageNet-Real and 94.55% on CIFAR-100[1]. Initially, three main ViT architectures were proposed in the ViT paper[1], namely - ViT Base consisting of 86M parameters, ViT Large consisting of 307M parameters, and ViT huge of 632M parameters[5]. It was shown that the larger models - Vit-L and ViT-H outperform the smaller ViT-B model as well as the ResNet152x2(BiT) model with a growing dataset size[1].

These benefits of ViT are due to the fact the Transformer based models have a much less image-specific inductive bias than their CNN counterparts. Resolution Adjustment and patch extraction are the only points at which an inductive bias about the 2D structure of the images is manually injected into the Vision Transformer. The self-attention method allows ViT to integrate information across the complete the image even in the lowest layers giving it a remarkable performance.

In comparison, The YOLO family [3] of networks has become very popular due to its inference speed and the ability to integrate drawing bounding and labeling into a single-stage end-to-end differentiable network. The first version of the network was built on the Darknet framework.The YOLO network consists of 3 main sections: The Backbone, the Neck, and the head. The head is the part of the network that makes the bounding boxes and class predictions. It is guided by 3 types of loss functions for YOLO class, objectness, and box. The YOLO Neck combines and mixes the ConvNet layer representations before passing them on to the prediction head. The Backbone is a CNN pools image pixels to form features of different granularities and is typically pretrained on a classification dataset [3].

YoloV7 was claimed to be the most accurate and fastest real-time object detection model for computer vision tasks and was released in July'2022 [4]. The code has been made open source, with everything released in a GitHub repository. The Architecture is built upon the PyTorch framework, which brings curiosity about its inference speed when parallelism is employed. The Yolov7 network is integrated with BlendMask allowing it to perform instance segmenta-

tion and its integration with YOLO-Pose allows it also to do pose estimation.

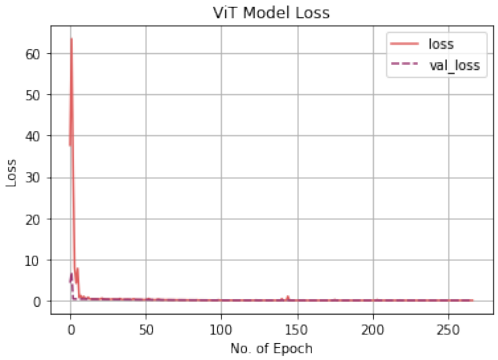

Figure 1. Training and Validation Loss

2. Methods and Experimental Setup

For my part of this project, I have trained the ViT model from scratch [5] to understand better how it works. I have used extensive fine-tuning and hyperparameter selection to get the best performance(a high IoU score) to my ability on the model. After trying multiple combinations of hyperparameters, I have included the top 3 high IoU score models, which can be summarised in the following paragraphs. Note that early stopping was implemented in all of the following configurations. All the configurations were trained on the Calteh-101 dataset. The image size was 224x224 for training. The total parameters(all trainable) for this model were 28.7M.

Configuration 1 : In this configuration my hyperparameters were - learning rate =0.001,weight decay = 0.0001, batch size = 32, number of epochs = 100, patch size = 32x32. The mean IoU score was 0.49482441867530425

Configuration 2 : In this configuration my hyperparameters were - learning rate =0.0015,weight decay = 0.0002, batch size = 32, number of epochs = 500, patch size = 16x16.The mean IoU score was 0.5593205934110492 (with early stopping at the $195/500^{th}$ epoch.)

Configuration 3 : In this configuration my hyperparameters were - learning rate =0.0017,weight decay = 0.00012, batch size = 32, number of epochs = 500, patch size = 16x16.The mean IoU score was 0.7826509197584113 (with early stopping at $267/500^{th}$ epoch)

I picked Configuration 3 for further testing against YOLOv7 since it had the highest mean IoU score. The training and validation loss is shown in Figure 1. For YOLOv7, I use pretrained model[4]for testing.

Figure 2. YOLOv7 Object Detector

Figure 3. ViT Object Detector

3. Experimental Results

I tested ViT trained from scratch and the pre-trained(and fine-tuned) YOLOv7 model to perform object detection on cars. Figure 2 shows the result of deploying the ViT object detector on cars. Figure 3 shows the results I got from deploying the YOLOv7 model on the same set of images.

Upon seeing the results, it is clear that the YOLOv7 model outperforms the ViT model. IoUs for the YOLOv7 model are higher than for ViT for all the images shown. It is still worthwhile noting that our ViT model has comparable performance to YOLOv7, having been trained on a smaller dataset.

4. Discussions and Comparison

On comparing the results, it is evident that the YOLOv7 model outperforms the ViT model. This is because the YOLOv7 was pretrained on the large MS-COCO dataset, and the ViT one was trained on the Caltch-101 dataset, which is a very small dataset as compared to MS-COCO. We couldn't train ViT on a dataset like MS-COCO because

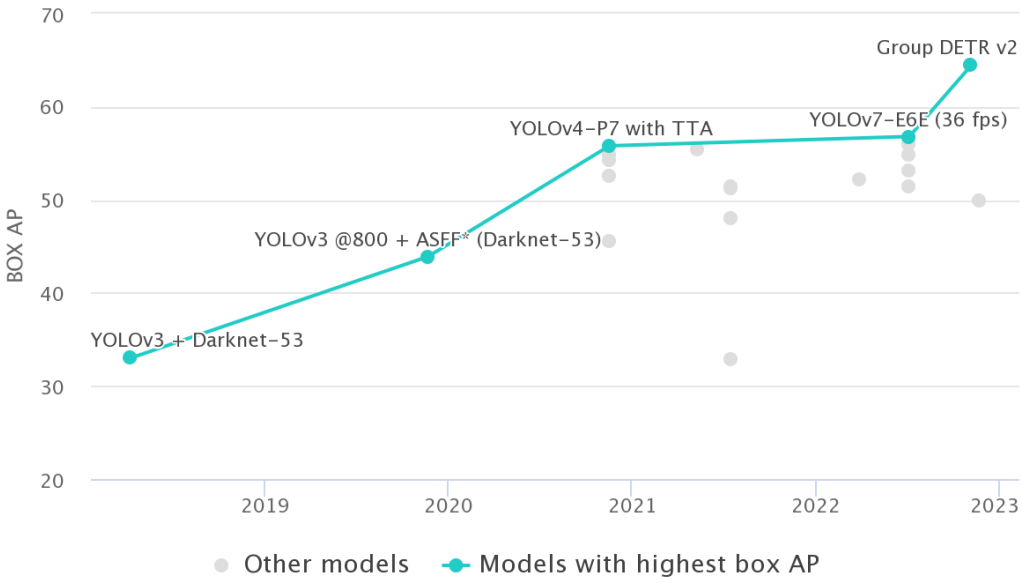

Figure 4. ViT variants Vs. YOLOv7(Source: Papers With Code)

of the computational constraints that the training presents. Even with similar training situations, YOLOv7 would still outperform ViT- due to the inductive bias present in the CNN-style architecture of YOLOv7 -on the Small to Large dataset range. It is only when a very large model like the ViT-H/16 trained on huge datasets like JFT-300M that ViT performance overtakes YOLOv7.

In addition, the current state-of-the-art(as of 2022) is a ViT-H based Group DETRv2 vision transformer which outperforms YOLOv7(YOLOv7-E6E). Group DETRv2[6] has a Box AP of 64.5, while the YOLOv7-E6E has 56.8, as shown in Figure 4. This further supports the hypothesis that ViT-based models seem to be the future of Computer Vision. In the future, it will be very interesting to see a Neural Network architecture with no inductive bias; transformers are a stepping stone to that.

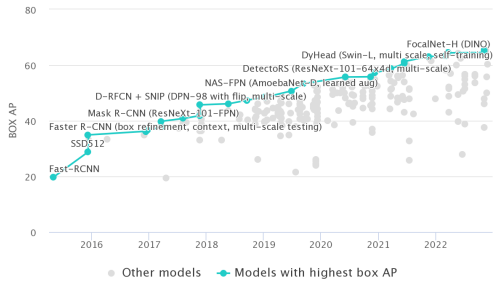

Figure 5. Object Detection trends till date(source:papers with code)

5. Conclusion

In conclusion, our comparison sheds light on the current state of art high-performance object detection architectures. We compared ViT to YOLOv7 and concluded that for small to large datasets, YOLOv7 is the top-performing architecture primarily because of the inductive bias present in the CNN-style architectures. Once we get to very large datasets(14M-300M images), we can see that ViT models dominate. As shown in Figure 5, Vision-based transformers and their variants are increasingly becoming popular because of the large amounts of data and high-performance infrastructure to train these models.

6. Future work

Transformer Neural networks have revolutionized Deep Learning, and we can already see their power in the most popular Deep Learning models like DALL-E and GPT-3, transformer-based language models.

7. Contributions

In this project, I worked on training and testing the Vision Transformer from scratch based on the code available on the Keras forum[5]. I further fine-tuned YOLOv7 for ViT vs. YOLOv7 object detection comparison.

References

[1] A. K. D. W. Alexey Dosovitskiy, Lucas Beyer. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[2] N. P. J. U. Ashish Vaswani, Noam Shazeer. Attention is all you need. *NIPS*, 2017.

[3] P. Azevedo. Object detection state of the art, 2022.

[4] A. B. Chien-Yao Wang and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2022.

[5] K. V. Dave. Object detection with vision transformers, 2022.

[6] C. H. S. Z. Z. L. X. C. J. C. Qiang Chen, Jian Wang. Group detr v2: Strong object detector with encoder-decoder pretraining. *ICLR*, 2022.