Review of “You only look once: Unified, real-time object detection”

Riyad Bin Rafiq

1. Paper summary

This paper [1] introduced a fast new approach called YOLO (you only look once) to object detection. The paper presented a single network that predicts bounding boxes and class probabilities directly from input images. YOLO outperformed other detection networks in real-time testing as the base model processes images at 45 frames per second. Moreover, it can learn a very general representation of the objects.

2. Contribution

2.1 Single network

This research solved the multi-stage training pipeline of the previous object detection methods. YOLO is a single convolutional neural network that can simultaneously predict multiple bounding boxes and class probabilities for those boxes. So this model can train on full images with direct optimization. That’s why, YOLO is fast in detecting objects. A unified single network is the most significant contribution of the research.

2.2 Limitations

This research explained the limitations of the approach which is a great contribution in my opinion. For example, the authors said that the model is not good in localizing the small objects. They also provided an explanation why this happens to the model. Other research works on object detection didn’t mention and explain such cases.

2.3 Comparison

The research provided a comparison to other detection systems. This kind of discussion helps to understand the uniqueness of the proposed approach. Moreover, the paper also compared the method to other real-time systems and provided mAP and FPS.

2.4 Generalization

In the real-world, the image varies from what the system has seen before. So YOLO was applied to the Picasso Dataset and the People-Art Dataset. Moreover, the comparative performance was listed and YOLO outperformed all other detection methods. Therefore, it can be said that the approach has good generalizability.

3. Critique

3.1 Grid

The unified model divides the image into an SxS grid. In the research, the authors chose S=7, but it was not defined why they chose that particular number. What would happen if a higher or lower number was chosen? This issue was cryptic in the research.

3.2 Why Leaky ReLU?

The authors used leaky rectified linear activation for training the convolutional layers. One research [2] demonstrated how ReLU learns faster and efficiently in CNN as learning has a great influence on the performance of the model. But it was not explained why the authors chose leaky ReLU. How did it perform with ReLU? I think this is one of the biggest gaps in the research.

3.3 Why not combine Faster R-CNN with YOLO?

The research combined Fast R-CNN with YOLO to boost the performance of object detection. But a question arises as to why the authors didn't apply Faster R-CNN with YOLO and if there is any specific reason for that. Because Faster R-CNN is the improved version of Fast R-CNN and so, the result can be better than that. In my opinion, this should be cleared in the research.

Reference

1. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 779–788.

2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html