You Only Look Once: Arhitectural discussion Summary

Comănac Dragoș-Mihail

Abstract—**The main purpose of this paper is to discuss the arhitectural details of You Only Look Once method for solving object detection. Over time, this method has proven to be a robust way of performing object detection, given recent applications that use it and achieve competitive results on the standard Microsoft Common Objects in Context dataset for comparing object detectors.**

I. Introduction

You Only Look Once (YOLO) [5] is a widely used method for performing object detection. The importance of this subject comes from the need for fast and precise methods such as YOLO for finding objects in visual data such as images and videos. Some fields in which this method can be successfully implemented (among others) include the automotive industry with use cases such as autonomous driving, traffic monitoring, or parking management, and logistics with use cases such as inventory management.

II. Placement in the broader field

Computer vision is a scientific field that deals with the extraction of meaningful information from visual data such as images or videos. There are various methods that can be used in the field of computer vision, such as hand crafted features, but the most relevant methods are related to deep learning. Object detection is a computer vision problem, which consists of locating and classifying several objects in an image. Broadly speaking, in the context of neural networks, the object detection problem mainly branches out in two-stage object detection and one-stage object detection. YOLO is a one-stage object detection method.

III. Method description

In YOLO there is a single CNN with a classic backbone-neck-head or backbone-head, that predicts end-to-end the bounding boxes. The image is split in a grid in which each cell is responsible for detecting objects that appear inside it. The ground truth and the final CNN output must have the following structure: $C \times C \times B \times (5+N)$ where C is the size of the grid, B is the number of anchors and N is the number of classes. The bounding boxes are composed by the center of the object which is relative to the responsible cell and the width and height which are relative to the responsible anchor. Also, the probability that there is an object in the anchor is predicted, hence the $5+N$ term. C and B are important parameters because they control how dense is the detector, and that helps in controlling the recall.

The success of the YOLO method is given by the robustness it has shown through time. Over the years, several methodologies have been proposed that use YOLO as the key element. Each brings different optimizations over the original method, but the main aspects are still relevant. The fact that there is still an ongoing research interest into this method is another argument for its success and robustness.

IV. Related work

Single shot MultiBox Detector (SSD) [4] is another example of an object detection system that achieves real-time performance, encapsulating all operations in a single deep neural network. The main innovation that it brought are that predictions are made using multiple scale feature maps.

Region based convolutional neural networks (R-CNN) [2] fall into the category of two-stage object detectors. The first stage is to extract a lot of object proposals from the image using selective search or a region proposal network. Afterwards, in the second stage, these proposals are classified. This approach is traditionally more accurate and slower than one-stage methods, but the recent optimizations made both two-stage faster and one-stage more accurate.

V. Comparison with other methods in terms of performance

To measure an object detection system performance usually frames per second (FPS) and mean average precision (mAP) are the most used measures. We compare the performances of several models in terms of FPS and mAP on two standard object detection datasets PASCAL VOC [1] and COCO [3].

VI. Conclusions

In conclusion, YOLO is an important milestone for the object detection domain. The main innovations are that it managed to create a single viable and successful convolutional neural network architecture that is able to learn end to end to predict the bounding boxes from the raw data. Looking towards the future, we believe that the method is generic and versatile enough that it can be adapted to the newest advances in deep learning in general, as it was the case until the time of writing this paper. We also argue that this method is a good candidate for various practical applications in our modern society, given the low computational cost and relatively easy to understand and implement architecture.

References

[1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot MultiBox Detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.