Object detection using deep neural networks

Comănac Dragoș-Mihail

***Abstract*—We try to explore the recent milestone methods based on deep neural networks, which pushed object detection further than before and helped object detection to be an important use case in day to day applications. We hope that this paper could serve as an introductory guide in the vast and complex field of object detection.**

I. Introduction

Computer vision is a broad scientific field that deals with the extraction of meaningful information from visual data such as images or videos. Given the potential of deep neural networks in solving the problem of object detection, the purpose of this paper is to study some recent architectural trends in building object detectors using deep convolutional neural networks (CNN).

II. Placement in the general field

The first object detection techniques heavily relied on hand-crafted features. These methods were popular before 2014 and they collectively compose what is now the so called ”traditional object detection period”. The main innovation that took place in 2014 was that the deep neural networks reached a maturity such that they became a viable solution in general in the field of object detection. Broadly speaking, in the context of deep neural networks, the object detection problem mainly branches out in two-stage object detection and one-stage object detection.

III. Survey of recent object detectors based on deep convolutional networks

In this section we will review some of the more recent object detection techniques.

A. General architecture

The usual practice is to build one-stage object detectors using a single deep CNN, which has three components: backbone, neck and head. The main advantage of this approach is that it allows the object detection system to learn end-to-end the complex patterns directly from the data in order to predict the bounding boxes directly from the entire image.

B. You Only Look Once

You Only Look once (YOLO) is one of the first one-stage object detectors. The object detection problem is treated as a regression problem. There is only one neural network that predicts the bounding boxes and class probabilities from an image. This way, the network can benefit from using end-to-end learning, and the inference time is greatly reduced, thus achieving real-time performance.

C. Region based convolutional neural networks

This type of method falls into the category of two-stage object detectors. Traditionally, there was a clean separation in terms of performance between this type of object detectors and the one stage class of object detectors, but with time, both types of methods became better in what they lacked by bringing several optimizations over the original architecture. Two stage methods gained speed and one stage methods gained accuracy for example.

D. RetinaNet

RetinaNet is an important milestone for one-stage object detectors because the authors introduce a novel way of computing the loss, namely the focal loss. This approach proposes to bridge the precision gap between one-stage and two-stage object detectors. After this point, one-stage methods seem to dominate the literature.

E. Fully convolutional One-Stage Object Detection

Fully Convolutional One-Stage Object Detection (FCOS) is one of the first anchor-free solutions that do not require complicated post-processing and has a good recall. Before FCOS, object detection was one of the few computer vision tasks that deviated from the ”fully convolutional per-pixel prediction” due to the dependency on anchors. As such, this is exactly what FCOS tries to do, to solve object detection in a similar fashion to semantic segmentation and it manages to be better than anchor-based detectors in terms of performance.

F. Task-aligned One-stage Object Detection

Object detection is basically a composed task, consisting of the simpler classification and localization. The general approach in solving this composed task, is to solve the individual simpler tasks separately. The main disadvantage is that the two tasks do not work collaboratively, thus bounding boxes with good localization might be discarded during non-maximum suppression only because of the bad score. As such Task-aligned One-stage Object Detection (TOOD) proposes to actively align the two tasks.

IV. Conclusions

In conclusion, object detection has come a long way and several important advances have been made, such as end-to-end learning, focal loss, anchor-free methods or the alignment of tasks that have pushed further and further the performance of object detectors. Looking towards the future, we believe that many of these ideas can be integrated together to create other object detection methodologies that surpass existing ones.