Object Detection

Halil Eralp Koçaş

January 19, 2020

1 Object Detection

Object detection

1.1 Motivation

Various models are developed through years in the field of object detection. All these models are developed using different methods. Although all of these models are based on convolutional neural networks, their way of identifying objects in given images or frames are varies. Also, their backbone networks, scanning methods, multi-scale object detection, error functions are varies. These varieties affect the performance of models in different aspects such as small and large object detections, speed of the model, etc.

The aim in this article is to investigate, analyze, and compare the performances of various state-of-the-art object detectors through the time trained on video data. Then, promising detectors will be selected and these selected detectors will be analyzed on Video Object Detection dataset on both mean average precision and frame rate per second basis.

1.2 Literature in Static Object Detection

Static object detection refers to object detection in single image. In this context, object detectors do not use temporal information about objects in given image or frame sequences. Each given images are considered separately. The aim of investigating static object detection is to make a generalization from single frame to multiple frames. Understanding the pros and cons of each object detectors in various static object detection cases are important to make a generalization and creating a baseline for video object detection which requires more features to be considered for a well- performing detector.

TBC!!

1.3 Literature in Video Object Detection

Literature in Video Object Detection

2 Detector Features

As mentioned in Motivation section, in this article, the aim is to analyze object detectors based on various detector features. In the following five subsections, the following features will be examined:

- Backbone Networks
- Scanning Methods
- Multi-scale Handling
- Loss Functions
- Bells-and-whistles

These listed features are the essential features for detectors. When one studies the object detectors, one can see that major increases on the performances of object detectors are caused by changes on these features. These features will be analyzed and shown in the following subsections and section.

2.1 Backbone Networks

Backbone networks are the initial part of a detector's architecture. The given frame is processed for the first time in backbone networks. These backbone networks are implemented as Fully Convolutional Neural Networks. The used networks are the ones that are already proven to be perform well in images. The most frequent used ones are as following:

- VGG
- ResNet
- DarkNet
- Feature Pyramid Network
- Hourglass

The main function of these networks is to compute convolutional feature map over given frame, so that, these extracted feature maps are used to detect and localize objects in given frames.

2.2 Scanning Methods

There are many methods developed to scan given images through years. The following is a list for most frequent used scanning methods in object detectors:

- Sliding windows
- Region Proposals
- Grid Cells
- Anchor Boxes

All these methods have different advantages against each other. Also, Anchor Boxes are used together with other methods.

2.3 Multi-scale Handling

Multi-scale handling refers to detect objects that have different sizes in given frame. Multi-scale handling is a crucial feature for having a well performing detector since most of the objects in images have different sizes. There are various ways to handle this problem:

One of the solution is to run detection multiple times that in each run, the resolution of input frame has to be changed. Then, all the detected objects has to be combined after all iterations are completed. In case of multiple detection for same object, a suppression has to be performed to reduce single detection. Although this method works well, its runtime is slow.

Another solution is to use single feature map. In this method, single feature map is extracted from given frame and this feature map is passed through multiple convolutional layers to obtain a final feature map with more fine-grained features. Then, this feature map is used to predict the objects in given frame. This method is used to obtain fast detection but its performance is relatively worse than other methods.

Another solution is to use pyramidal feature hierarchy in which multiple feature maps are used to make prediction. In this method, a feature map is extracted from given frame and as it is in second solution, this feature map is passed through multiple convolutional layers. However, in each layer, the extracted feature map is used to make prediction. This method is relatively works slower than second solution but it performs better.

Another solution is to use feature pyramid network but this method will be examined detaily in section 3.2.8.

2.4 Loss Functions

2.5 Bells-And-Whistles

3 Static Detector Types

3.1 Two-Stage Detectors

3.1.1 Faster R-CNN

3.1.2 Mask R-CNN

3.2 One-Stage Detectors

3.2.1 You Only Look Once: Unified, Real-Time Object Detection

3.2.2 YOLO9000: Better, Faster, Stronger

3.2.3 YOLOv3: An Incremental Improvement

3.2.4 SSD: Single Shot MultiBox Detector

3.2.5 DSSD: Deconvolutional Single Shot Detector

3.2.6 CornerNet: Detecting Objects as Paired Keypoints

3.2.7 RetinaNet

3.2.8 FPN

3.2.9 EfficientDet: Scalable and Efficient Object Detection

3.2.10 FreeAnchor: Learning to Match Anchors for Visual Object Detection

4 Datasets and Metrics

5 Performance Comparison

5.1 Accuracy and Real-Time Applicable

5.2 On COCO

5.3 In Video

5.3.1 MOT17Det

5.4 Quantitative Analysis

5.4.1 Based on Size

5.4.2 Based on Location on Image

5.4.3 Based on Visibility Ratio

5.4.4 Based on Confidence Scores

5.4.5 Based on Combinations