International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Object Detection and Recognition Using Single Shot Multi-Box Detector

Shailesh Wagh[1], Sumeet Shukla[2], Harshal Shah[3], Ajay Yadav[4], Shirish Sabnis Prof[5]

[1,2,3,4]*Shailesh Wagh, Address: P9, Pragati Complex, Fatherwadi, Vasai.*
[5]*Shirish Sabnis Prof, Dept. of Information Technology, RGIT Mumbai, Maharashtra, India*

Abstract - *Efficient and accurate object detection has been an important topic in the advancement of computer vision systems. With the advent of deep learning techniques, the accuracy for object detection has increased drastically. We present a method for detecting objects in images using a single deep neural network. Our approach, named SSD, discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. SSD is simple relative to methods that require object proposals because it completely eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network. This makes SSD easy to train and straightforward to integrate into systems that require a detection component. Experimental results on the PASCAL VOC, COCO, and ILSVRC datasets confirm that SSD has competitive accuracy to methods that utilize an additional object proposal step and is much faster, while providing a unified framework for both training and inference. For 300×300 input, SSD achieves 74.3% mAP on VOC2007 and for 512 × 512 input, SSD achieves 76.9% mAP, outperforming a comparable state-of-the-art Faster R-CNN model. The resulting system is fast and accurate, thus aiding those applications which require object detection.*

Key Words: **Object Detection, Video Detection, SSD, Localization**

1. INTRODUCTION

ThisCurrent state-of-the-art object detection systems are variants of the following approach: hypothesize bounding boxes, re-sample pixels or features for each box, and apply a high quality classifier. This pipeline has prevailed on detection benchmarks since the Selective Search work through the current leading results on PASCAL VOC, COCO, and ILSVRC detection all based on Faster R-CNN albeit with deeper feature. While accurate, these approaches have been too computationally intensive for embedded systems and, even with high-end hardware, too slow for real-time applications.

Often detection speed for these approaches is measured in seconds per frame (SPF) and even the fastest high-accuracy detector, Faster R-CNN, operates at only 7 frames per second (FPS). There have been many attempts to build faster detectors by attacking each stage of the detection pipeline, but so far, significantly increased speed comes only at the cost of significantly decreased detection accuracy.

This paper presents the first deep network-based object detector that does not resample pixels or features for bounding box hypotheses and is as accurate as approaches that do. This results in a significant improvement in speed for high-accuracy detection (59 FPS with mAP 74.3% on VOC2007 test, vs. Faster R-CNN 7 FPS with mAP 73.2% or YOLO 45 FPS with mAP 63.4%). The fundamental improvement in speed comes from eliminating bounding box proposals and the subsequent pixel or feature resampling stage. We are not the first to do this (cf [4,5]), but by adding a series of improvements, we manage to increase the accuracy significantly over previous attempts. Our improvements include using a small convolutional filter to predict object categories and offsets in bounding box locations, using separate predictors (filters) for different aspect ratio detections, and applying these filters to multiple feature maps from the later stages of a network in order to perform detection at multiple scales.

2. EXISTING SYSTEM

Today, there is a plethora of pre-trained models for object detection like (YOLO, CNN)). CNN: CNNs are the basic building blocks for most of the computer vision tasks in deep learning era.

What do we want? We want some algorithm that looks at an image, sees the pattern in the image and tells what type of object is there in the image.

How can we teach computers to learn to recognize the object in the image? By making computers learn the patterns like vertical edges, horizontal edges, round shapes and maybe plenty of other patterns unknown to humans.

2.1 DESCRIPTION OF COMPONENTS

A. IMAGE CLASSIFICATION

To classify the objects on a given input image into there respective classes.

B. OBJECT LOCALIZATION

To find the fix position of the detected object in the given inputs images.

C. ARTIFICIAL NEURAL NETWORK

Inspired by the functional aspects it is a model which makes use of mathematics and computation power. Using

the connectionist approach of computation, artificial neurons are present in groups and intertwined with each other. Based on the information internal and external that flows inside the network that is present, during the phase of learning so that the ANN will adapt to it and change accordingly.

D. *MACHINE LEARNING*

Machine learning helps to learn and improve experience without being programmed specifically. It is an Application of Artificial Intelligence (AI) that is provided by the system. The main aspect of machine learning is on the development of a computer program so that data can be accessed and it can them self-learn. This technique of machine learning is used which learns on its own on the given data set.

E. *TRAINING DATASET*

The collection of dataset i.e images of respective classes to train the method.

F. *BOUNDING BOX*

The bounding box is a rectangle drawn on the image which tightly fits the object in the image. A bounding box exists for every instance of every object in the image.

3. PROPOSED SYSTEM

A. Method

The difference here is that instead of producing proposals, pre-define a set of boxes for objects. Using convolutional feature maps from later layers of the network, run another network over these feature maps to predict class scores and bounding box offsets. The steps are mentioned below:

1. Train a CNN with regression and classification objective.
2. Gather activation from later layers to infer classification and location with a fully connected or convolutional layer.
3. During training, use Jaccard distance to relate predictions with the ground truth. During training, use Jaccard distance to relate predictions with the ground truth.
4. During inference, use non-maxima suppression to filter multiple boxes around the same object.

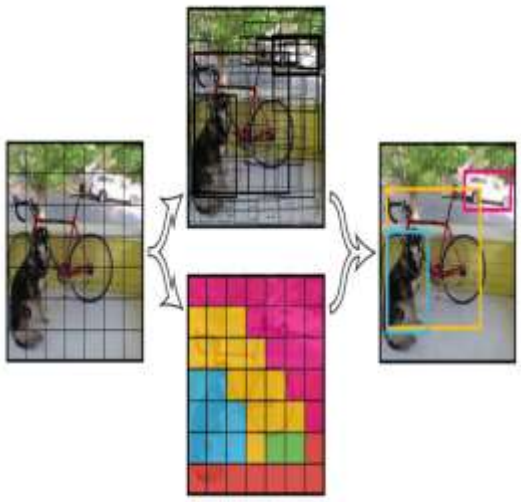

Figure 1 Unified Method.

B. Multi-scale feature maps for detection

We add convolutional feature layers to the end of the truncated base network. These layers decrease in size progressively and allow predictions of detections at multiple scales. The convolutional model for predicting detections is different for each feature layer (cf Overfeat[4] and YOLO[5] that operate on a single scale feature map).

C. Training

For the purpose of this project, the publicly available PASCAL VOC dataset will be used. It consists of 10k annotated images with 20 object classes with 25k object annotations (xml format). These images are downloaded from flickr. This dataset is used in the PASCAL VOC Challenge which runs every year since 2006.

D. Matching Strategy

During training we need to determine which default boxes correspond to a ground truth detection and train the network accordingly. For each ground truth box we are selecting from default boxes that vary over location, aspect ratio, and scale.

E. Method used for video detection

1) Detect function(frame, net, transform)

2)Get the height & width of they frame.

3)Convert Variables into torch Variables

4) Using method-Data to get following values(Batch no, no of occurence,Score)

5)for(i=0;i<=class.no;i++)

if(score>0.6)

choose

else
Reject
6)for(i=0;i<=frame.no;i++)
get frame
run Detect function on each frame.
Get output to compose a video.

4. EXPERIMENTAL RESULTS

The detections were assigned to ground truth objects and judged to be true/false positive by measuring bounding box overlap. To be considered a correct detection, the area of overlap between the predicted bounding box and ground truth bounding box must exceed a threshold.The average precision for all the object categories are reported in Table. The mAP for the PASCAL VOC dataset was found to be 0.633. The current state-of-the-art best mAP value is reported to be 0.739.

Class	Precision
Bike	0.712
Human	0.642
Cat	0.909
Table	0.534
Boat	0.564

Figure 2 detected image

Our video detected with the output can be found at: *link:https://drive.google.com/open?id=1656G1cCT06ZJxLuV2RasPzbHTvUN-5Xr*

5. FUTURE SCOPE

The given methodology of this paper can be further used by advanced features to enhance the working of the object Detection system. Live object detection, video surveillance concept can be implemented, such that they require less powerful hardware. This helps to implement object detection at even the basic level with similar accuracy.

6. CONCLUSION

An accurate and efficient object detection system has been developed which achieves comparable metrics with the existing state-of-the-art system. This project uses recent techniques in the field of computer vision and deep learning. This can be used in real-time applications which require object detection for pre-processing in their pipeline. An important scope would be to train the system on a video sequence for usage in tracking applications. Addition of a temporally consistent network would enable smooth detection and more optimal than per-frame detection.

REFERENCES

[1] Ross Girshick. Fast R-CNN. In International Conference on Computer Vision (ICCV),2015.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems (NIPS), 2015.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once:Unified, real-time object detection. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In ECCV, 2016.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.