Object Detection Using Tensor Flow

Author
SarikAnwar

**SCHOOLOFCOMPUTERSCIENCEANDENGINEERING
GALGOTIASUNIVERSITY,GREATERNOIDA**

Abstract

Efficient and accurate object detection has been a crucial topic within the advancement of computer vision systems. With the arrival of deep learning techniques, the accuracy for object detection has increased drastically. The paper aims to include state-of-the-art technique for object detection with the goal of achieving high accuracy with a real-time performance. A major challenge in many of the thing detection systems is that the dependency on other computer vision techniques for helping the deep learning based approach, which results in slow and non-optimal performance. In this paper, we use a totally deep learning based approach to unravel the matter of object detection in an end-to-end fashion. The network is trained on the foremost challenging publicly available dataset (PASCAL VOC), on which a object detection challenge is conducted annually. The resulting system is fast and accurate, thus aiding those applications which require object detection. Here we are proposing an application which can be used to identify different types of objects like human objects present in a picture consisting of different other objects. We will apply supervised learning to make the system learn how a human object is recognized by teaching it with some examples. This model is going to work on data sets. The data sets have some patterns that are combined to form a result pattern and resultant pattern is analysis with the input and provide results. Our Model is going to more accurate with more balanced data sets. Fueled by the steady doubling rate of computing power every 13 months, object detection and recognition has transcended from an esoteric to a well-liked area of research in computer vision and one among the higher and successful applications of image analysis and algorithm based understanding. Because of the intrinsic nature of the problem, computer vision is not only a computer science area of research, but also the object of neuro-scientific and psychological studies, mainly because of the general opinion that advances in computer image processing and understanding research will provide insights into how our brains work and vice- versa

1. Introduction

The goal of this article is to provide an easier human-machine interaction routine when user authentication is needed through object detection and recognition. With the aid of a regular web camera, a machine is able to detect and recognize a person's object; a custom login screen with the ability to filter user access based on the users' object features will be developed. The objectives of this thesis are to provide a set of detection algorithms that can be packaged in an easily portable framework among the different processor architectures we see in machines (computers) today. These algorithms must provide at least a 95% successful recognition rate, out of which less than 3% of the detected objects are false positive is processed to crop and extract the person's object for easier recognition. Object Recognition where that detected and processed object is compared to a database of known objects, to decide who that person is. Since 2002, object detection can be performed fairly easily and reliably with Intel's open source framework called OpenCV . This framework has an inbuilt Object Detector that works in roughly 90-95% of clear photos of a person looking forward at the camera. However, detecting a person's object when that person is viewed from an angle is usually harder, sometimes requiring 3D Head Pose Estimation. Also, lack of proper brightness of an image can greatly increase the difficulty of detecting a object, or increased contrast in shadows on the object, or maybe the picture is blurry, or the person is wearing glasses, etc. Object recognition however is much less reliable than object detection, with an accuracy of 30-70% in general. Object recognition has been a strong field of research since the 1990s, but is still a far way away from a reliable method of user authentication. More and more techniques are being developed each year.The Eigenobject technique is considered the simplest method of accurate visual perception , but many other (much more complicated) methods or combinations of multiple methods are slightly more accurate. OpenCV was started at Intel in 1999 by Gary Bradski for the needs of accelerating research in and commercial applications of computer vision within the world and, for Intel, creating a requirement for ever more powerful computers by such applications. Vadim Pisarevsky joined Gary to manage Intel's Russian software OpenCV team. Over time the OpenCV team moved on to other companies and other Research. Several of the first team eventually ended up working in robotics and located their thanks to Willow Garage. In 2008, Willow Garage saw the need to rapidly advance robotic perception capabilities in an open way that leverages the entire research Eigenobjects is considered the simplest method of accurate object recognition, but many other (much more complicated) methods or combinations of multiple methods are slightly more accurate.Most resources on visual perception are for basic Neural Networks, which usually don't work as well as Eigenobjects does. And unfortunately there are only some basic explanations for better type of object recognition than Eigenobjects, such as recognition from video and other techniques at the Object Recognition Homepage or 3D Object Recognition Wikipedia page and Active Appearance Models page But for other techniques.

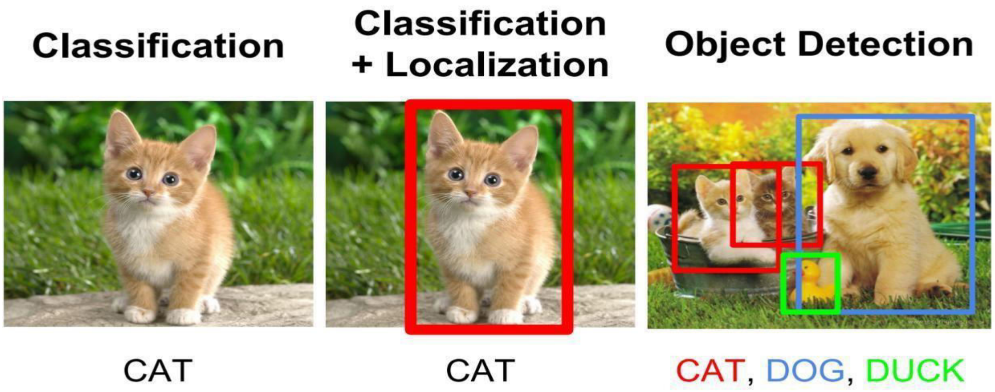
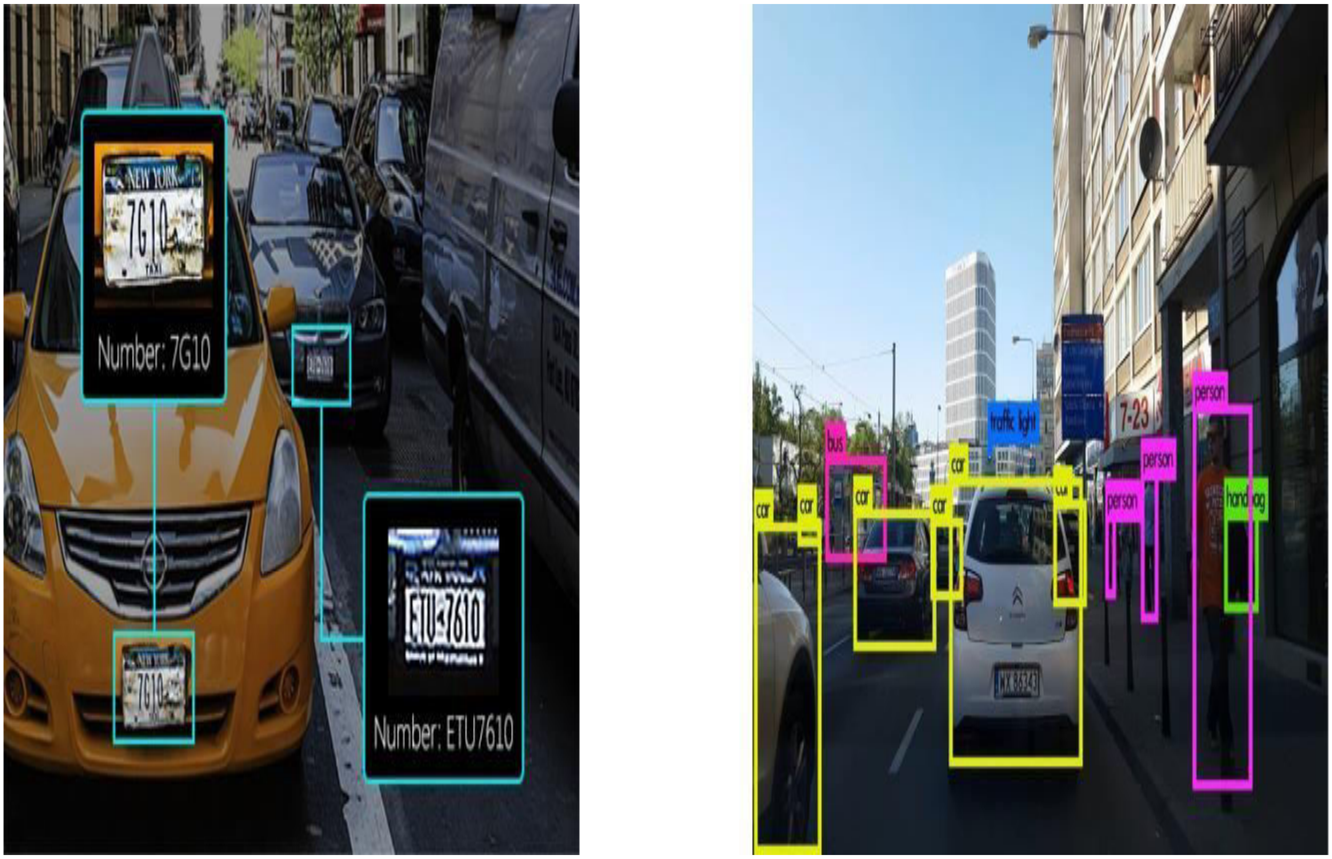


Figure 1: Computer Vision Tasks

• Applications

A documented application of object detection is face detection, that's utilized in most the mobile cameras. A more generalized (multi-class) application are often utilized in autonomous driving where a spread of objects got to be detected. Also it's a important role to play in surveillance systems. These systems are often integrated with other tasks like pose estimation where the primary stage within the pipeline is to detect the thing , then the second stage are going to be to estimate pose within the detected region. It are often used for tracking objects and thus are often utilized in robotics and medical applications. Thus this problem serves a mess of applications



(a) Surveillance

(b) Autonomous vehicles

Figure 2: Applications of object detections

- **Challenges**

The major challenge during this problem is that of the variable dimension of the output which is caused thanks to the variable number of objects which will be present in any given input image. Any general machine learning task requires a hard and fast dimension of input and output for the model to be trained. Another important obstacle for widespread adoption of object detection systems is that the requirement of real-time (>30fps) while being accurate in detection. The more complex the model is, the longer it requires for inference; and therefore the less complex the model is, the less is the accuracy. This trade-o between accuracy and performance must be chosen as per the appliance . the matter involves classification also as regression, leading the model to be learnt simultaneously. This adds to the complexity of the matter.

- **Problem Statement**

Many problems in computer vision were saturating on their accuracy before a decade. How-ever, with the increase of deep learning techniques, the accuracy of those problems drastically improved. one among the main problem was that of image classification, which is defined as predicting the category of the image. a rather complicated problem is that of image localiza-tion, where the image contains one object and therefore the system should predict the category of the situation of the thing within the image (a bounding box round the object). The more compli-cated problem (this paper), of object detection involves both classification and localization. during this case, the input to the system are going to be a image, and therefore the output are going to be a bounding box like all the objects within the image, along side the category of object in each box. an summary of of these problems is depicted.

2 Related Work

There has been tons of labor in object detection using traditional computer vision techniques (sliding windows, deformable part models). However, they lack the accuracy of deep learning based techniques. Among the deep learning based techniques, two broad class of methods are prevalent: two stage detection (RCNN [1], Fast RCNN [2], Faster RCNN [3]) and unified detection (Yolo [4], SSD [5]). the main concepts involved in these techniques are explained below.

- **Bounding Box**

The bounding box may be a rectangle drawn on the image which tightly fits the thing within the image. A bounding box exists for each instance of each object within the image. For the box, 4 numbers (center x, center y, width, height) are predicted. this will be trained employing a distance measure between predicted and ground truth bounding box. the space measure may be a jaccard distance which computes intersection over union between the anticipated and ground truth boxes as shown in Fig. 3.

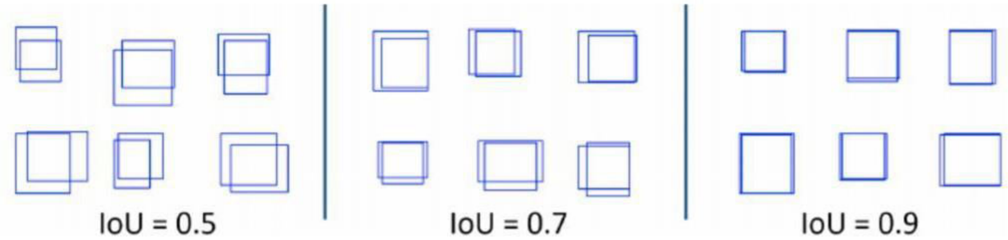


Figure 3: Jaccard distance

- **Classification+Regression**

The bounding box is predicted using regression and the class within the bounding box is predicted using classification. The overview of the architecture is shown in Fig. 4

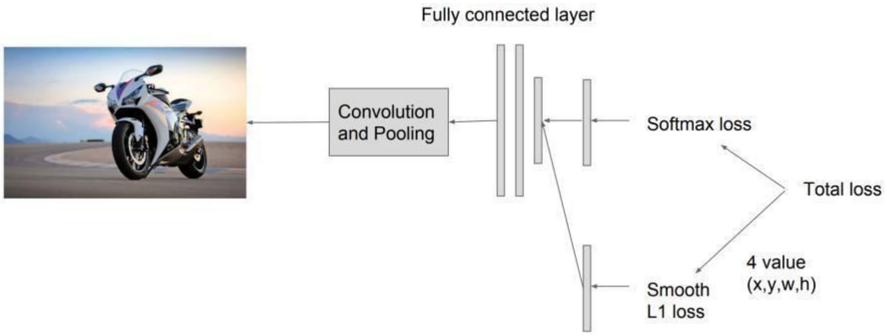
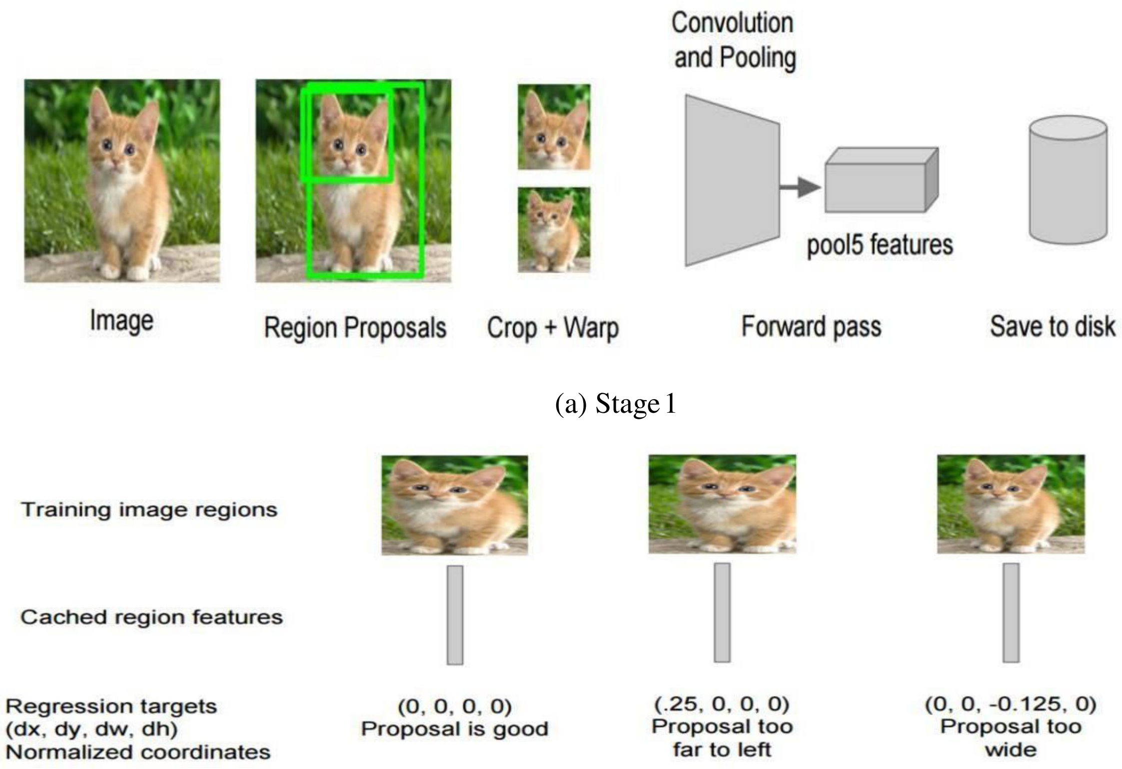


Figure 4: Architecture overview

- **Two-stage Method**

In this case, the proposals are extracted using another computer vision technique then resized to fixed input for the classification network, which acts as a feature extractor. Then an SVM is trained to classify between object and background (one SVM for every class). Also a bounding box regressor is trained that outputs some correction (offsets) for proposal boxes. the general idea is shown in Fig. 5 These methods are very accurate but are computationally intensive (low fps).


Convolution and Pooling
pool5 features
Image
Region Proposals
Crop + Warp
Forward pass
Save to disk
Training image regions
Cached region features
Regression targets (dx, dy, dw, dh) Normalized coordinates
(0, 0, 0, 0) Proposal is good
(.25, 0, 0, 0) Proposal too far to left
(0, 0, -0.125, 0) Proposal too wide

(a) Stage 1

(b) Stage 2

Figure 5: Two stage method

3 Approach

The network used in this paper is based on Single shot detection (SSD) [5]. The architecture is shown in Fig. 6.

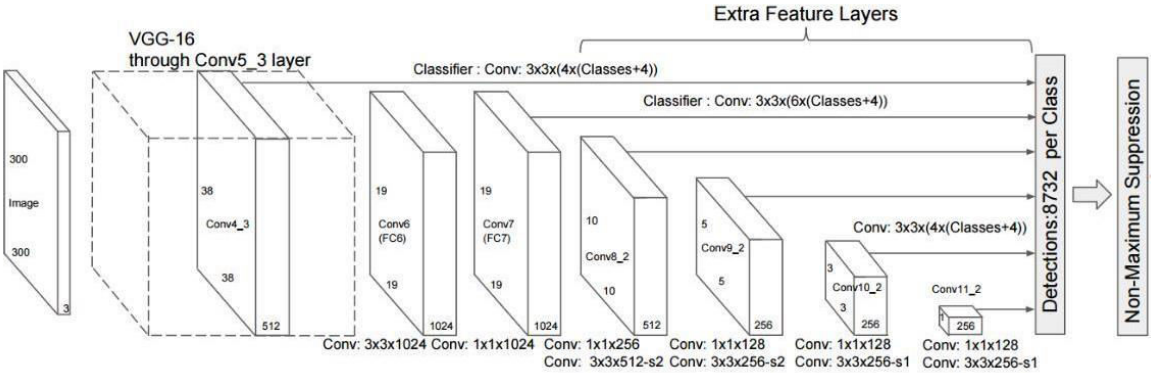


Figure 6: SSD Architecture

The SSD normally starts with a VGG [6] model, which is converted to a totally convolu-tional network. Then we attach some extra convolution layers that help to handle bigger objects. The output at the VGG network may be a 38x38 feature map (conv4 3). The added layers produce 19x19, 10x10, 5x5, 3x3, 1x1 feature maps. of these feature maps are used for predicting bounding boxes at various scales (later layers liable for larger objects).

Thus the general idea of SSD is shown in Fig. 7. a number of the activations are passed to the sub-network that acts as a classifier and a localizer.

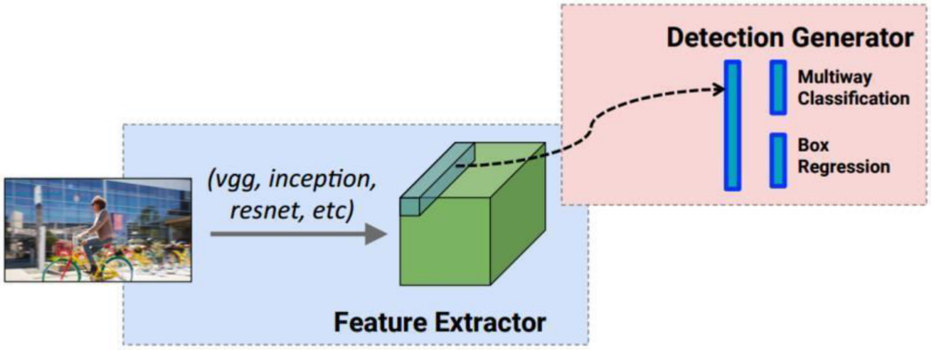


Figure 7: SSD Overall Idea

Anchors (collection of boxes overlaid on image at different spatial locations, scales and aspect ratios) act as reference points on ground truth images as shown in Fig. 8.

A model is trained to form two predictions for every anchor:

A discrete class

A continuous offset by which the anchor must be shifted to fit the ground-truth bounding box

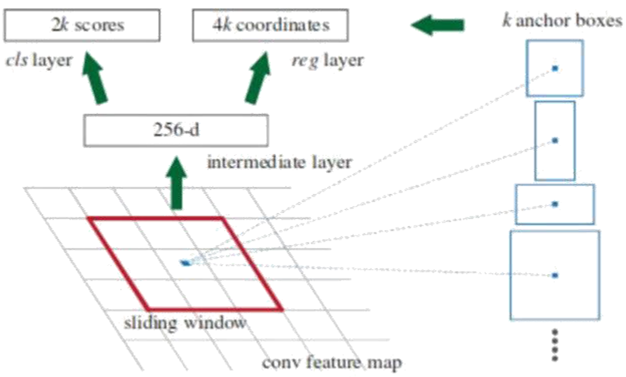


Figure 8: Anchors

During training SSD matches ground truth annotations with anchors. Each element of the feature map (cell) features a number of anchors related to it. Any anchor with an IoU (jacquard distance) greater than 0.5 is taken into account a match. Consider the case as shown in Fig. 10, where the cat has two anchors matched and therefore the dog has one anchor matched. Note that both are matched on different feature maps.

4 Experimental Results

• Dataset

For the aim of this paper, the publicly available PASCAL VOC dataset are going to 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 employed within the PASCAL VOC Challenge which runs per annum since 2006.

Visual Object Classes Challenge 2012 (VOC2012)
PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Figure 9: Dataset

• Implementation Details

The paper is implemented in python 3. Tensorflow was used for training the deep network and OpenCV was used for image pre-processing.
The system specifications on which the model is trained and evaluated are mentioned as follows: CPU - Intel Core i7-7700 3.60 GHz, RAM - 16 Gb, GPU - Nvidia 1050TI.

- **Pre-processing**

The annotated data is provided in xml format, which is read and stored into a pickle file along side the pictures in order that reading are often faster. Also the pictures are resized to a fixed size**.**

- **Network**

The entire specification is shown in Fig. 10. The model consists of the bottom network derived from VGG net then the modified convolutional layers for fine-tuning then the classifier and localizer networks. This creates a deep network which is trained end-to-end on the dataset.

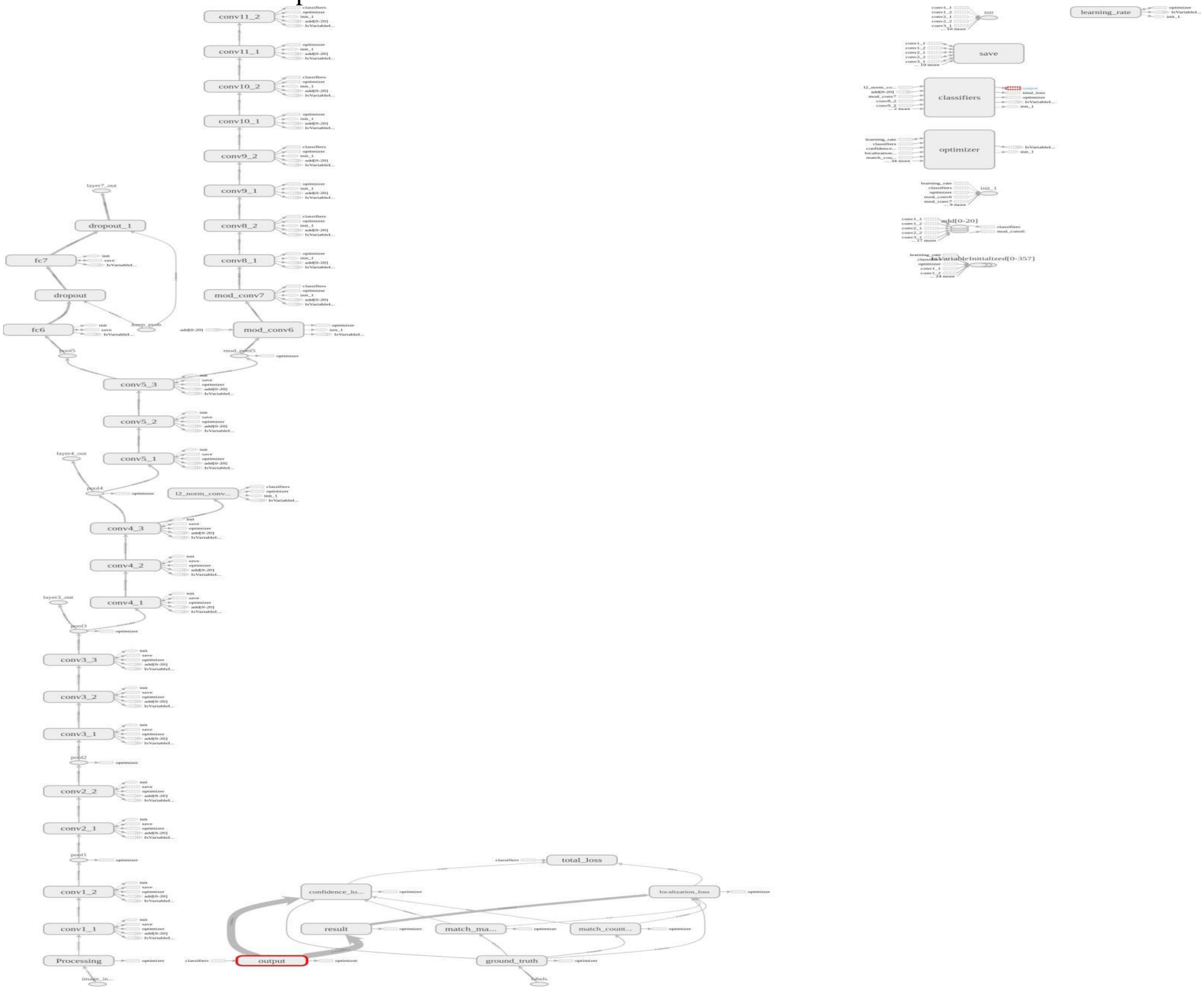


Figure 10: Network in Tensorboard

- **Qualitative Analysis**

The results from the PASCAL VOC dataset are shown in Table 1.

put Ground Truth Prediction

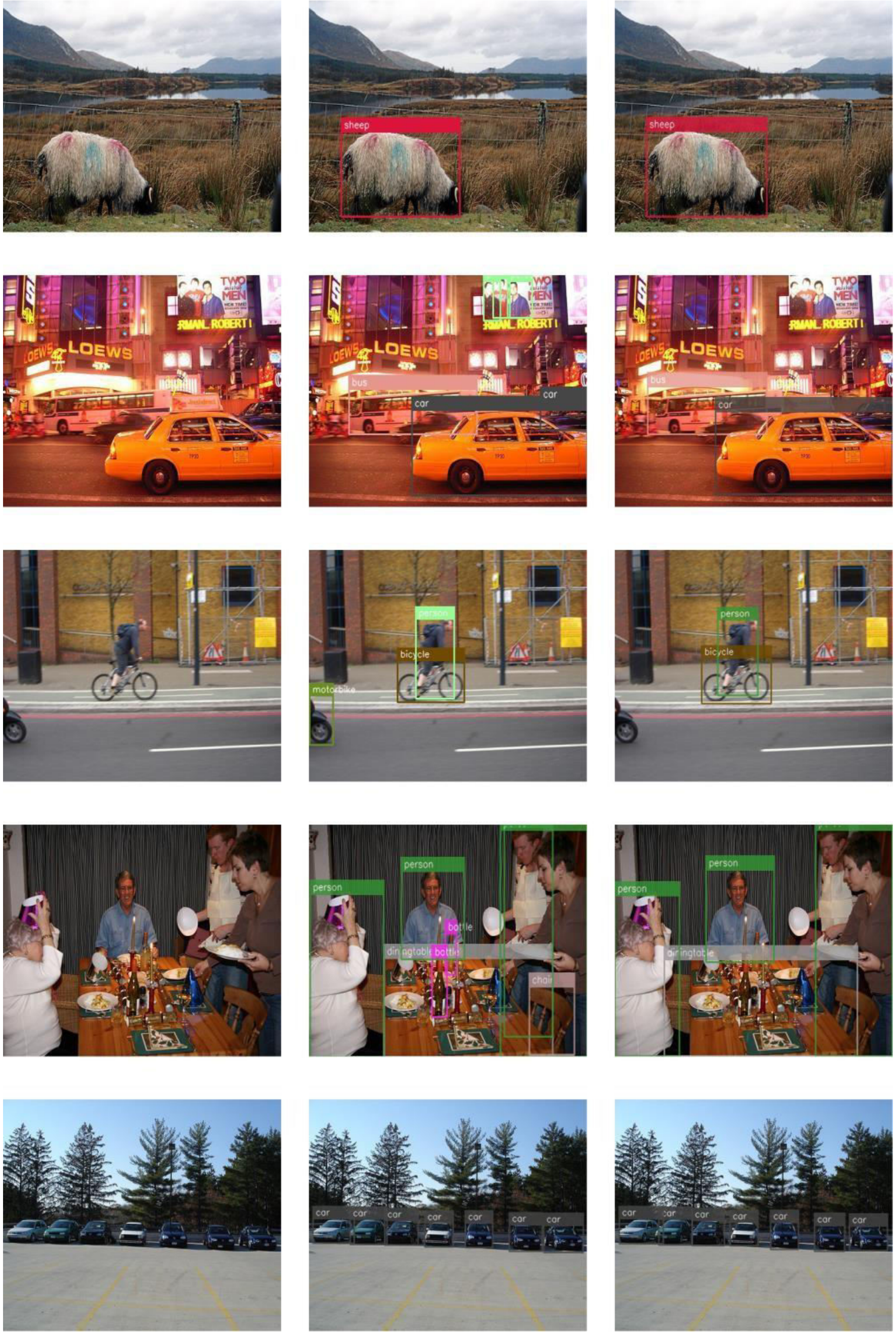


Table 1: Detection results on PASCAL VOC dataset.

The results on custom dataset are shown in Table 2.

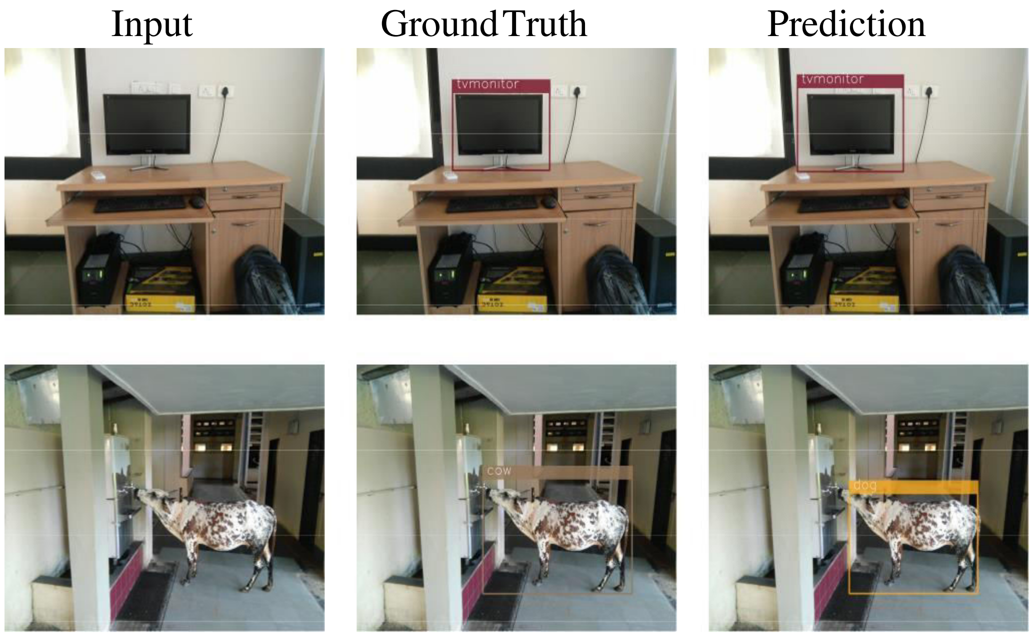
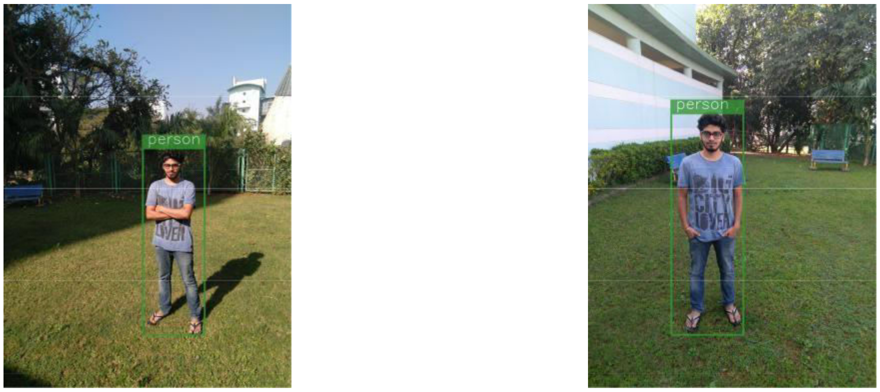


Table 2: Detection results on custom dataset.

The system handles illumination variations thus providing a strong detection. In Fig. 14 an equivalent person is standing within the shade then within the sunny environment.



(a) High illumination (b) Low illumination

Figure 11: Detection robust to illumination variation

However, occlusion creates a drag for detection. As shown in Fig. 15, the occluded birds aren't detected correctly.

Also larger object dominated when present along side small objects as found in Fig.

16. this might be the rationale for the typical precision of smaller objects to be less in comparison to larger objects. This has been reported within the next section.

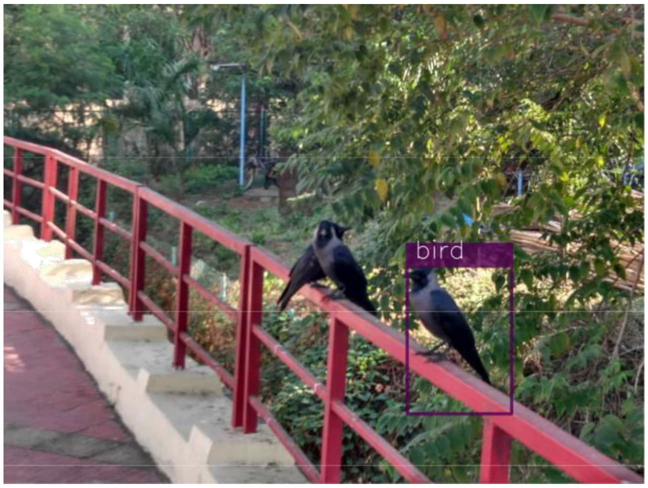
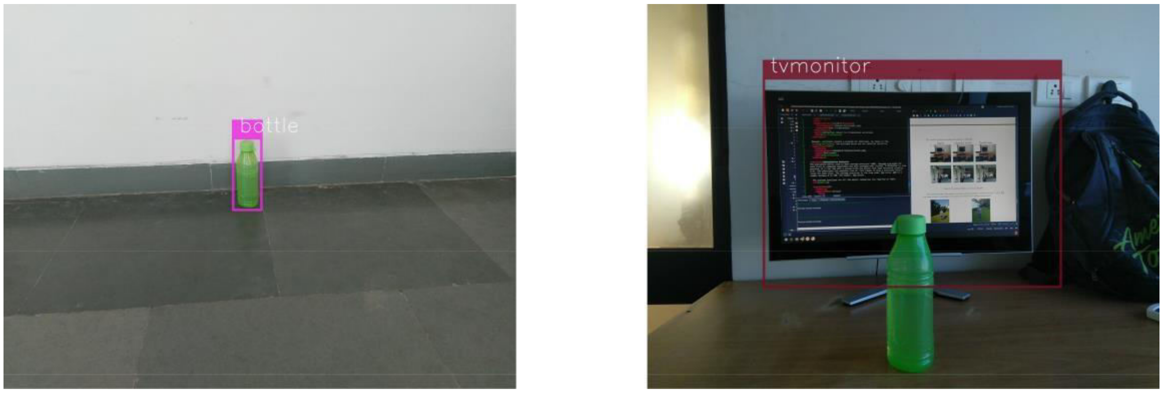


Figure 12: Occlusion



(a) Only small object in image (b) Small and large object in image

Figure 13: Domination of larger object in detection

Conclusion

An accurate and efficient object detection system has been developed which achieves compa- rable metrics with the prevailing state-of-the-art system. This paper uses recent techniques within the field of computer vision and deep learning. Custom dataset was created using labeling and therefore the evaluation was consistent. this will be utilized in real-time applications which require object detection for pre-processing in their pipeline.

An important scope would be to coach 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 Girhick, Je Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[2] Ross Girhick. Fast R-CNN. In Internatonal Conference on Computer Vision (ICCV), 2015.

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

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

[5] 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.

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