OBJECT DETECTION AND RECOGNIZATION

Prof. Manikrao mulge ,Rishab lachuriye , Mohammed Faizan, Ajay Solanke

**Assistant Professor, Department of Computer Science & Engineering, Guru Nanak Dev engineering college, Bidar,
Department of Computer Science & Engineering, Bidar .**

Abstract- Computer Vision is the branch of the science of computers and software systems which can recognize as well as understand images and scenes. Computer Vision is consists of various aspects such as image recognition, object detection, image generation, image super-resolution and many more. Object detection is widely used for face detection, vehicle detection, pedestrian counting, web images, security systems and self-driving cars. In this project, we are using highly accurate object detection-algorithms and methods such as R-CNN, Fast-RCNN, Faster-RCNN, RetinaNet and fast yet highly accurate ones like SSD and YOLO. Using these methods and algorithms, based on deep learning which is also based on machine learning require lots of mathematical and deep learning frameworks understanding by using dependencies such as TensorFlow, OpenCV, imageai etc, we can detect each and every object in image by the area object in an highlighted rectangular boxes and identify each and every object and assign its tag to the object. This also includes the accuracy of each method for identifying objects.

KEY WORDS: road safety inspection, hazardous road location, accident prediction model.

I. INTRODUCTION

Many problems in computer vision were saturating on their accuracy before a decade. However, with the rise of deep learning techniques, the accuracy of these problems drastically improved. One of the major problem was that of image classification, which is defined as predicting the class of the image. A slightly complicated problem is that of image localization, where the image contains a single object and the system should predict the class of the location of the object in the image (a bounding box around the object). The more complicated problem (this project), of object detection involves both classification and localization. In this case, the input to the system will be a image, and the output will be a bounding box corresponding to all the objects in the image, along with the class of object in each box.

A well known application of object detection is face detection, which is used in almost all the mobile cameras. A more generalized (multi-class) application can be used in autonomous driving where a variety of objects need to be detected.

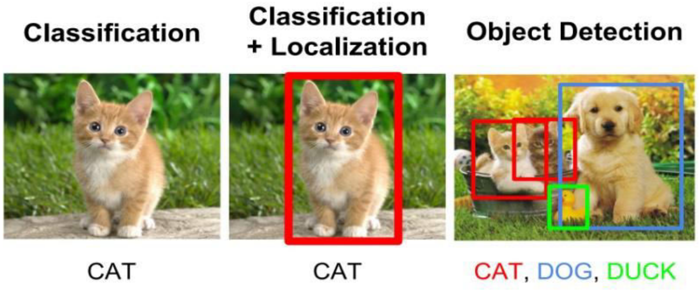


Figure 1: Computer Vision Task

II. APPROACH

The SSD normally starts with a VGG [6] model, which is converted to a fully convolutional network. Then we attach some extra convolutional layers, that help to handle bigger objects. The output at the VGG network is a 38x38 feature map . The added layers produce 19x19, 10x10, 5x5, 3x3, 1x1 feature maps. All these feature maps are used for predicting bounding boxes at various scales .

Thus the overall idea of SSD is shown in Fig. 2. Some of the activations are passed to the sub-network that acts as a classifier and a localizer.

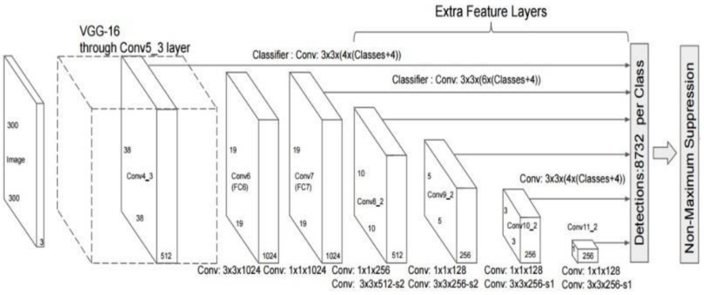


Figure 2: SSD Architecture

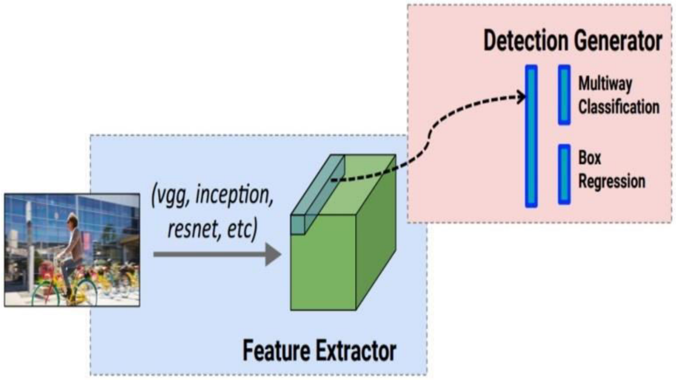
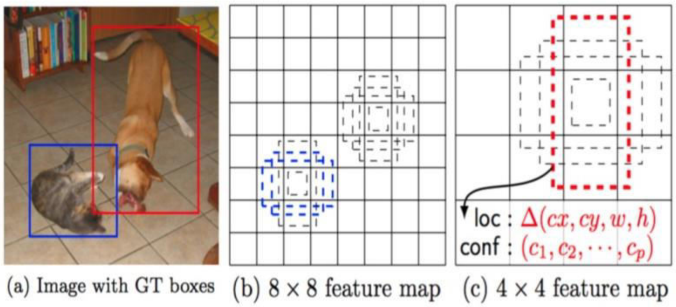


Figure 3: SSD Overall Idea

During training SSD matches ground truth annotations with anchors. Each element of the feature map (cell) has a number of anchors associated with it. Any anchor with an IoU (jaccard distance) greater than 0.5 is considered a match. Consider the case as shown in Fig. 4, where the cat has two anchors matched and the dog has one anchor matched. Note that both have been matched on different feature maps.


loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$

(a) Image with GT boxes (b) 8 × 8 feature map (c) 4 × 4 feature map

The loss function used is the multi-box classification and regression loss. The classification loss used is the softmax cross entropy and, for regression the smooth L1 loss is used.

During prediction, non-maxima suppression is used to filter multiple boxes per object that may be matched as shown in Fig. 4.

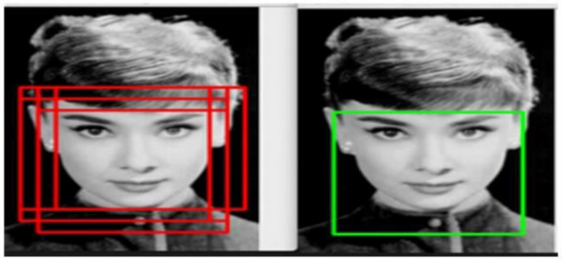


Figure 4: Non-maxima suppression

EXPERIMENTAL RESULTS

Dataset

For the purpose of this project, the publicly available PASCAL VOC dataset and COCO 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.

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

Figure 5: Dataset

IMPLEMENTATION

The project 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 - 32 Gb, GPU - Nvidia Titan Xp.

Pre-processing

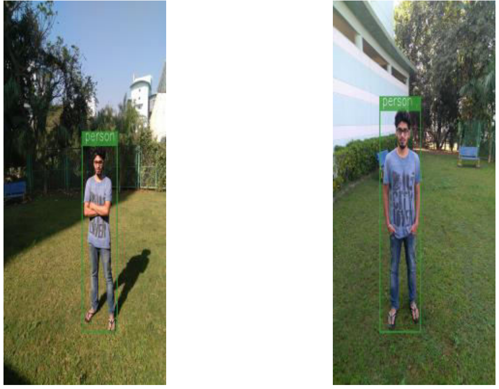
The annotated data is provided in xml format, which is read and stored into a pickle file along with the images so that reading can be faster. Also the images are resized to a fixed size.

Network

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

QUALITATIVE ANALYSIS

The system handles illumination variations thus providing a robust detection. In Fig. 6 the same person is standing in the shade and then in the sunny environment.



(a) High illumination (b) Low illumination

Figure 6: Detection robust to illumination variation

However, occlusion creates a problem for detection. Also larger object dominated when present along with small objects as found in Fig. 6. This could be the reason for the average precision of smaller objects to be less when compared to larger objects. This has been reported in the next section.

QUANTITATIVE ANALYSIS

The evaluation metric used is mean average precision (mAP). For a given class, precisionrecall curve is computed. Recall is defined as the proportion of all positive examples ranked above a given rank. Precision is the proportion of all examples above that rank which are from the positive class. The AP summarizes the shape of the precision-recall curve, and is defined as the mean precision at a set of eleven equally spaced recall levels [0, 0.1, ... 1]. Thus to obtain a high score, high precision is desired at all levels of recall. This measure is better than area under curve (AUC) because it gives importance to the sensitivity.

The detections were assigned to ground truth objects and judged to be true/false positives 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 output of the detections assigned to ground truth objects satisfying the overlap criterion were ranked in order of (decreasing) confidence output. Multiple detections of the same object in an image were considered false detections, i.e. 5 detections of a single object counted as 1 true positive and 4 false positives. If no prediction is made for an image then it is considered a false negative.

CONCLUSIONS

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. Custom dataset was created using labeling and the evaluation was consistent. 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, Jeff 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 Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*, 2015.

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

[4] 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.

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, ChengYang 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.