Abstract

Efficient and accurate object detection has been an important topic in the advancement of computer vision systems. With the advent of machine learning and deep learning techniques, the accuracy for object detection has increased drastically. The project aims to incorporate state-of-the-art technique for object detection with the goal of achieving high accuracy with a real-time performance. In this project, we use a completely machine learning with opencv and deep learning based approach to solve the problem of object detection in an end-to-end fashion. The network is trained on the most challenging publicly available dataset, on which a object detection challenge is conducted annually. The resulting system is fast and accurate, thus aiding those applications which require object detection.

Chapter 1

Introduction

1.1 Problem Statement

As we move towards more complete image understanding, having more precise and detailed object recognition becomes crucial. In this context, one cares not only about classifying images, but also about precisely estimating the class and location of objects contained within the images, a problem known as object detection. 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.

It has the following objectives:

- Object detection, segmentation, location, and recognition
- Object tracking
- Different perspectives on the same scene or object based image retrieval

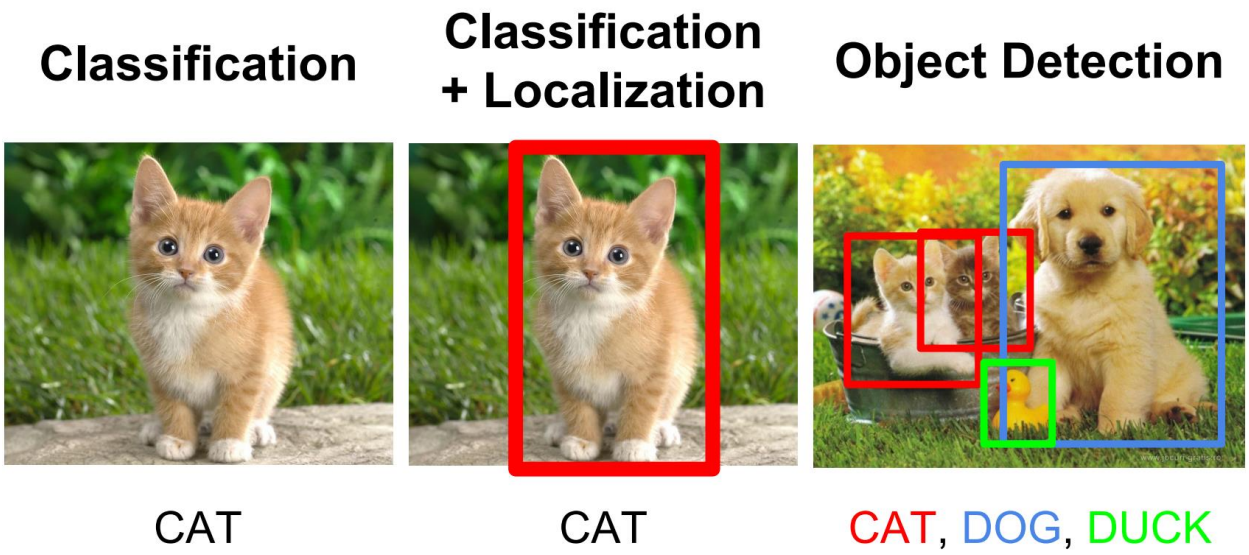


Figure 1.1 . Process of Image Detection in Computer Vision

Figure 1.1 shows the process of image detection in computer vision. In the figure, the first step is video input. The next steps involve the application of image processing techniques and extraction of the object. Following this, image recognition takes place. Finally, a description

of the object is obtained. In this process, image processing is a very important step. It involves converting the original image data to digital data for the use of the computer.

1.2 Application

A well known application of object detection is face detection, that 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. Also it has a important role to play in surveillance systems. These systems can be integrated with other tasks such as pose estimation where the _rst stage in the pipeline is to detect the object, and then the second stage will be to estimate pose in the detected region. It can be used for tracking objects and thus can be used in robotics and medical applications. Thus this problem serves a multitude of applications.

Chapter 2

Literature Survey

In past few years, the detection of Objects in real time and Image processing has become an active area of research and several new approaches have been proposed.

Several researchers have conducted many studies about Object detection-

- S.V. Viraktamath, Mukund Katti, Aditya Khatawkar & Pavan Kulkarni has conducted a study of openCV and also have published an IEEE paper for Face Detection and Tracking using OpenCV. Their work is related with converting web cam captured 2D Images and convert them into 3D Images related to human faces by constructing 3D Geometry data outputs.
- A.Rodriguez used HSV (Hue, Saturation, Value) model for more accurate processing of image. For training and recognizing image ANN (Artificial Neural Network) algorithm is used. To separate foreground and background from image sobel edge detection is used.
- J. A. Hesch, S.I. Roumeliotis also provide an idea to implement ORB(Oriented FAST and Rotated BRIEF(Binary Robust Independent Elementary Features) algorithm to increase the execution speed by utilizing the reconfigurable nature and pipeline. ORB algorithm builds on the well-known FAST key point detector and the recently developed BRIEF descriptor.

Chapter 3

Related Work

There has been a lot of work in object detection using traditional computer vision techniques. Deep architectures for object detection and parsing have been motivated by part-based models and traditionally are called compositional models, where the object is expressed as layered composition of image primitives. Further examples of compositional models for detection are based on segments as primitives, focus on shape, use Gabor filters, or larger HOG filters. These approaches are traditionally challenged by the difficulty of training and use specially designed learning procedures. Moreover, at inference time they combine bottom-up and top-down processes.

Neural networks (NNs) can be considered as compositional models where the nodes are more generic and less interpretable than the above models. Applications of NNs to vision problems are decades old, with Convolutional NNs being the most prominent example [16]. It was not until recently than these models emerged as highly successful on large-scale image classification tasks in the form of DNNs. Their application to detection, however, is limited. Scene parsing, as a more detailed form of detection, has been attempted using multi-layer Convolutional NNs. Segmentation of medical imagery has been addressed using DNNs. Both approaches, however, use the NNs as local or semi-local classifiers either over super pixels or at each pixel location. Our approach, however, uses the full image as an input and performs localization through regression. As such, it is a more efficient application of NNs.

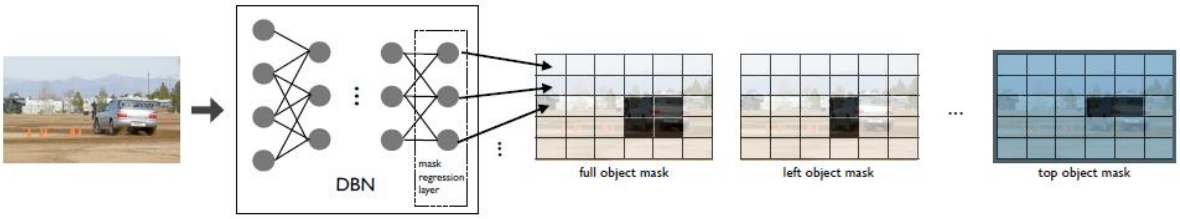


Figure 3.1: A schematic view of object detection as DNN-based regression.

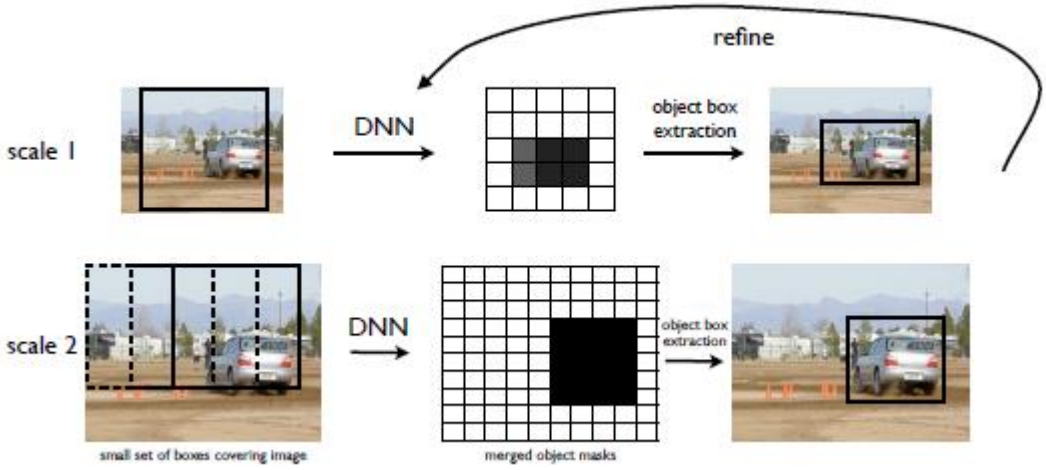


Figure 3.2: After regressing to object masks across several scales and large image boxes, we perform

object box extraction. The obtained boxes are refined by repeating the same procedure on the sub images, cropped via the current object boxes. For brevity, we display only the full object mask, however, we use all five object masks.

ORB also uses learning method for de-correlating BRIEF features under rotational invariance, leading to better performance in nearest-neighbor applications. The recognize method for object recognition is Scale invariant feature transform (SIFT), which is popular for its invariance to scaling, rotation and illumination, is computationally complex due to its heavy workload required in local feature extraction and matching operation. Thus computer vision kind of applications demands high performance and low complexity solution and ORB provides better solution to it.

3.1 Unified Method

The difference here is that instead of producing proposals, pre-de_ne a set of boxes to look 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 o_sets. The broad idea is depicted in Fig. 6. 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 layers.
3. 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.

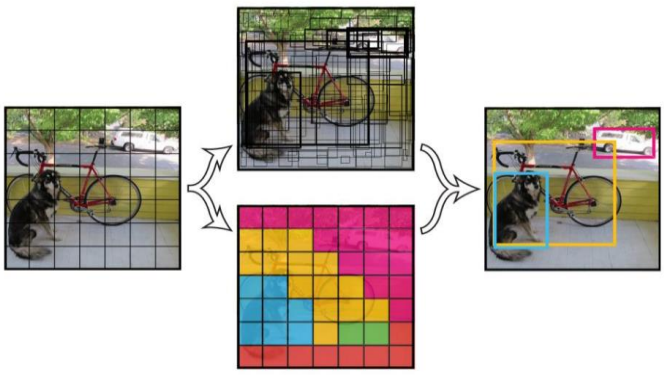


Figure 3.3: Unified Method

Chapter 4

Methodology

Neural Network Deep learning used by the network has been constantly improving, in addition to the changes in the network structure, the more is to do some tune based on the original network or apply some trick to make the network performance to enhance. The more well-known algorithms of object detection are a series of algorithms based on R-CNN, mainly in the following.

R-CNN

Paper which the R-CNN (Regions with Convolutional Neural Network) is in has been the state-of-art papers in field of object detection in 2014 years. The idea of this paper has changed the general idea of object detection. Later, algorithms in many literatures on deep learning of object detection basically inherited this idea which is the core algorithm for object detection with deep learning. One of the most noteworthy points of this paper is that the CNN is applied to the candidate box to extract the feature vector, and the second is to propose a way to effectively train large CNNs.

SSD

The network used in this project is based on Single shot detection (SSD).

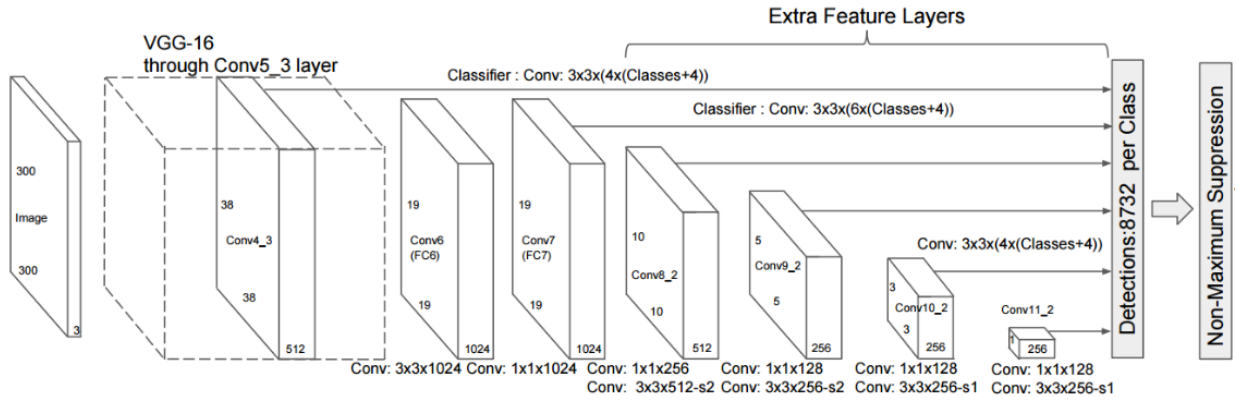
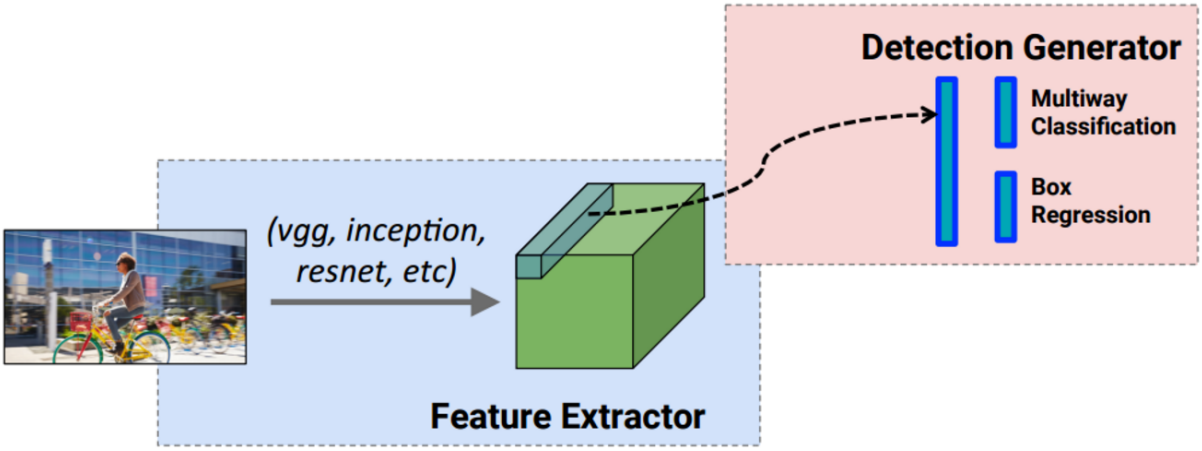


Figure 4.1: SSD Architecture

The SSD normally starts with a VGG 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 (conv4 3). The added layers produce 19x19, 10x10, 5x5, 3x3, 1x1 feature maps. All these feature maps are used for predicting bounding boxes at various scales (later layers responsible for larger objects). Thus the overall

idea of SSD is shown in Fig. 8. Some of the activations are passed to the sub-network that acts as a classifier and a localizer.

Figure 4.2: 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. 4.3.A model is trained to make two predictions for each anchor:

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

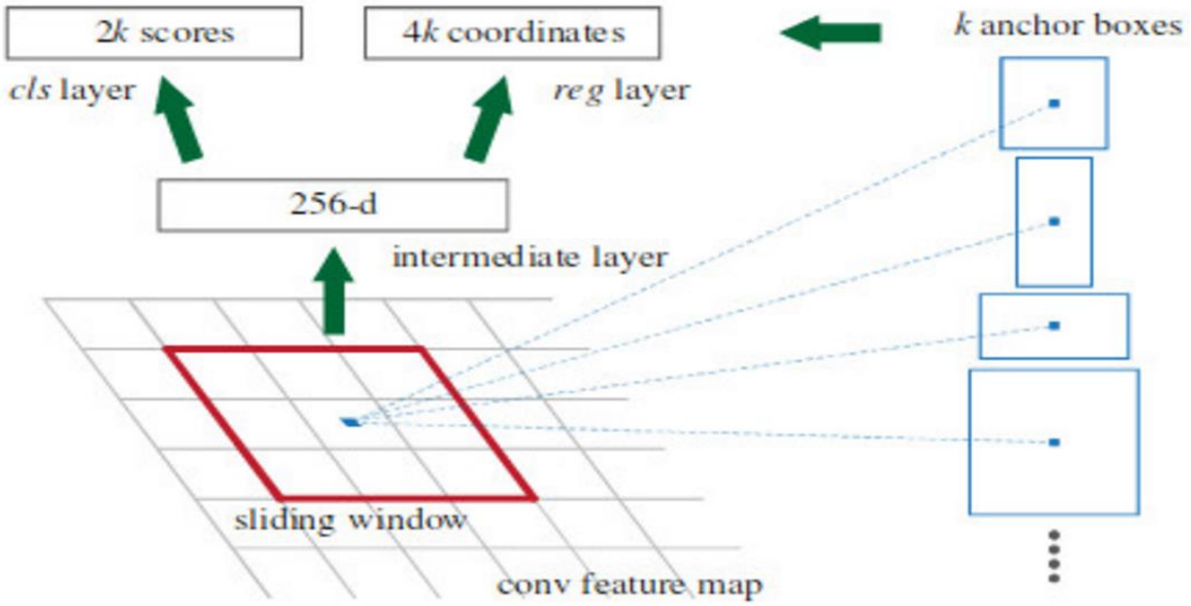
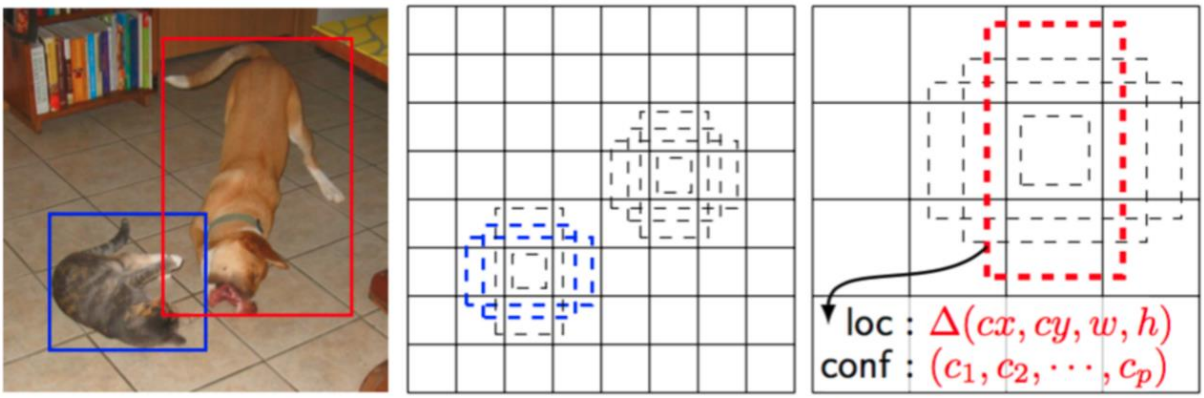


Figure 4.3: Anchors

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. 4.4, where the cat has two anchors matched and the dog has one anchor matched. Note that both have been matched on different feature maps.



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

Figure 4.4: Matches

Chapter 5

Experiments and Results

5.1 Dataset: We evaluate the performance of the proposed approach on the test set of the Pascal Visual Object Challenge (VOC) 2007. The dataset contains approx. 5000 test images over 20 classes. Since our approach has large number of parameters, we train on the VOC2012 training and validation set which has approx. 11K images. At test time an algorithm produces for an image a set of detections, defined bounding boxes and their class labels. We use precision-recall curves and average precision (AP) per class to measure the performance of the algorithm.

5.2 Implementation Details:

The project is implemented in python 3. Tensorflow, sklearn 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 i5-7700 3.60 GHz, RAM - 32 Gb, GPU - Nvidia Titan Xp.

5.2.1 Network

The entire network architecture is shown in Fig. 13. The model consists of the base network derived from VGG net and then the modi_ed convolutional layers for _ne-tuning and then the classi_er and localizer networks. This creates a deep network which is trained end-to-end on the dataset.

5.2.2 Qualitative Analysis

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

Input Ground Truth Prediction

Table 5.1: Detection results on PASCAL VOC dataset.

5.2.3 Quantitative Analysis

The evaluation metric used is mean average precision (mAP). For a given class, precision recall 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 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) con_dence 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.

The average precision for all the object categories are reported in Table 5.2.

Class	Average Precision
Motorbike	0.724
Bottle	0.272
Bird	0.635
Cat	0.909
Aeroplane	0.727
Chair	0.360
Person	0.542
Diningtable	0.534
Boat	0.544
Train	0.909
Sofa	0.710
Bicycle	0.636
Bus	0.726
Horse	0.726
Tvmonitor	0.633
Cow	0.632
Pottedplant	0.359
Car	0.634
Dog	0.818
Sheep	0.633

Table 5.2: Average precision for all classes

Chapter 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. Custom dataset was created using labelling and the evaluation was consistent. These results come at some computational cost at training time, one needs to train a network per object type and mask type.

REFERENCES

[1] D. G. Lowe, International Journal of Computer Vision, vol. 60, no. 2, (2004), pp. 91.

[2] J.-H. Park and T. S. Sidhu, Journal of Information and Communication Convergence Engineering, vol. 2, no. 3, (2004), pp. 187.

[3] 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.

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

[5] A. Rodriguez, Assisting the Visually Impaired: Obstacle Detection and Warning System by Acoustic Feedback, 17476 – 17496, Sensors 2012.

[6] Y. Wu J. Lim, and M.-H. Yang, "Online object tracking: A benchmark,"in Proc. IEEE Conf. Compute. Vis. Pattern Recognit., Jun. 2013.

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

[8] OpenCV (Open Source Computer Vision), http://opencv.org.

[9] Computer Vision, http://www.aistudy.com/physiology/vision/computer_vision.htm.