Object Detection using Single Shot Detector for a Self-Driving Car

Mrs. Swetha M. S.
Assistant Professor
Department of Information Science & Engineering
BMS Institute of Technology & Management, Yelahanka, Bangalore -560064, Karnataka, India

Ms. Vallae Haritha
Assistant Professor
Department of Information Science & Engineering
BMS Institute of Technology & Management, Yelahanka, Bangalore -560064, Karnataka, India

Abstract

Object detection is a computer technology related to computer vision and image processing that deals with detecting, in digital images and videos, instances of semantic objects of a certain class, such as humans, buildings, cars, etc. Single Shot Multi Box Detector (SSD) is a deep learning method and one of the fastest algorithms which uses a single convolutional neural network to detect the object and also classify the stationary candidate objects in an image. SSD has the advantages of fast speed and high accuracy but is less accurate in detecting small objects compared to large objects. The proposed system is to use the SSD algorithm for a video simulation of a self-driving car to detect person and some suspected type of objects which include backpack, handbag, etc. quickly along with the curvature of the road and also to be able to use in a real time video.

Keywords: Single Shot Multi Box Detector (SSD)

I. INTRODUCTION

Deep learning achieves great success in image classification, object detection, semantic segmentation and natural language processing. A few years ago, by exploiting some of the leaps made possible in computer vision via CNNs, researchers developed R-CNNs to deal with the tasks of object detection, localization and classification. A R-CNN is a special type of CNN that is able to locate and detect objects in images the output is generally a set of bounding boxes that closely match each of the detected objects, as well as a class output for each detected object. Many improved methods based on the R-CNN, such as fast R-CNN, faster R-CNN emerged in the object detection area. These methods achieved high accuracies, but their network structures are relatively complex.

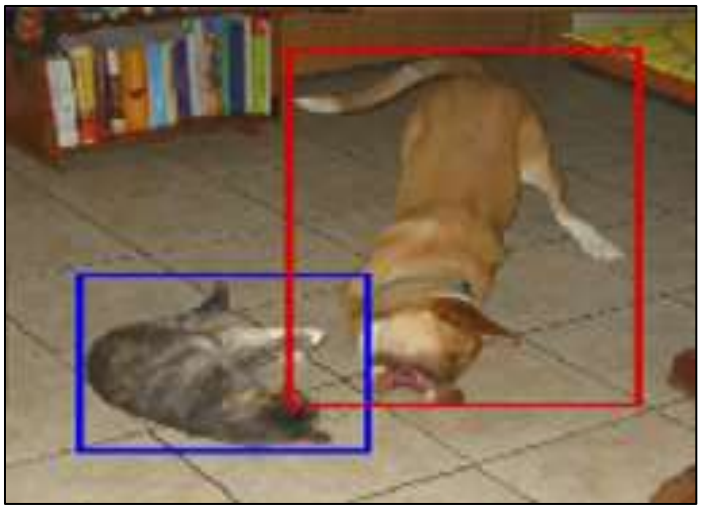


Fig. 1: Object Detection using SSD

SSD is also a part of the family of networks which predict the bounding boxes of objects in a given image. It is a simple, end to end single network, removing many steps involved in other networks which tries to achieve the same task, at the time of its publishing. It works better than the state of the art Faster-RCNN in cases of higher dimensional images.

The fundamental concept of SSD is mostly based on the feed forward convolution network. It is discretized the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. It then generates scores forth presence of each object class in each default box and produces adjustments to better match object shape.

The SSD model is comprised of mainly two structures: Base network and Auxiliary network. The Base network is the early part of the model which is based on standard architecture used for high quality image classification. The Auxiliary network has features

mainly focused for objects with different scales or aspect ratios. SSD has two components in its structure. The first component, called a base network, is used for classification of images. The second component has three useful features including multi-scale features maps for detection, convolutional predictors for detection, and six aspect ratios of detection boxes at the end of November 2016 and reached new records in terms of performance and precision for object detection tasks, scoring over 74% mAP (mean Average Precision) at 59 frames per second on standard datasets such as PascalVOC and COCO. To better understand SSD, let's start by explaining where the name of this architecture comes from:

– Single Shot

This means that the tasks of object localization and classification are done in a single forward pass of the network

– Multi-Box

This is the name of a technique for bounding box regression developed by Szegedy et al. (we will briefly cover it shortly)

– Detector

The network is an object detector that also classifies those detected object.

In Multi-Box, the bounding box regression technique of SSD is inspired by Szegedy's work on MultiBox, a method for fast class-agnostic bounding box coordinate proposals. Interestingly, in the work done on Multi-Box an Inception-style convolutional network is used. 1x1 convolutions that help dimensionality reduction since the number of dimensions will go down (but "width" and "height" will remain the same). Such applications demand a protocol that can hide the node identities and geographical positions as well as their traffic information. The protocols that ensure this type of routing are known as anonymous routing protocols. Anonymous routing protocols in MANETs are playing a crucial role to offer secure communication. They provide sec

Some of the fields where SSD is used combining with Transfer Learning for Ship Detection Using Chinese Gaofen-3 Images and SAR Target Detection Based on SSD with Data Augmentation.

II. LITERATURE SURVEY

The automatic detection of objects in real time video surveillance videos proposes a robust, scalable framework for automatic detection of abandoned, stationary objects in real time surveillance videos that can pose a security threat. The background modelling method to generate a long-term and a short-term background model to extract foreground objects is used. Subsequently, a pixel-based FSM detects stationary candidate objects based on the temporal transition of code patterns. In order to classify the stationary candidate objects, the deep learning method (SSD) is used to suppressant false alarm and also to remove other stationary candidate objects other than the suspected stationary objects and also check if there is no person near by the suspected detected objects for a particular time. Accordingly, another method -The Inception block to replace the extra layers in SSD, and call this method Inception SSD (I-SSD) helps to catch more information without increasing the complexity. Usage of the batch-normalization (BN) and the residual structure in this I-SSD network architecture. The proposed I-SSD algorithm achieves 78.6% mAP on the Pascal VOC2007 test and an Outdoor Object Detection (OOD) dataset to testify the effectiveness of the proposed I-SSD on the platform of unmanned vehicles.

Autonomous driving has gained attention both from industrial and academic research facilities but a detailed knowledge of the current state of a self-driving car and the surroundings is a crucial problem that can be best addressed with a set of different sensors, such as LiDAR, RADAR and RGB cameras. LiDAR provides very accurate information about the distance of a 3D point, it yields a sparse representation of the scene and only little semantic information is encoded but LiDAR sensors are relatively expensive. In contrast, RGB cameras provide a very dense representation and hence can be used to obtain rich semantic information. As they are quite cheap, most consumer vehicles are already equipped with at least one RGB camera. Finally, combining the data of these sensors allows for a complete scene representation of the car's surrounding environment to obtain information from each sensor individually to achieve best possible performance and to focuses on the data acquisition of single monocular images.

SAR images are used for ship detection to ensure marine transportation and security. SSD is applied to ship detection in SAR images. Transfer learning is adopted because it performs well even in small training dataset. Two types of SSD models integrated with transfer learning, namely SSD-300 and SSD-512 and are applied to ship detection.

To evaluate the approaches three SAR images acquired by Chinese Gaofen-3 satellite are used. First, SAR images are cut into training, validation, and test with respect to machine learning routines. Both training and validation dataset are resized to feed to SSD with fine tuningVGG16 to train the model. Finally, test images are used to evaluate Experimental results reveal that compared to SSD-300, SSD-512achieves more than 0.02 in probability of ship detection, whereas 0.05 worse in false alarm.

CSSD–shorthand for context-aware single-shot multi box object detector is built on top of SSD, with additional layers modeling multi-scale contexts. It describes two variants of CSSD, which differ in their context layers, using dilated convolution layers (DiCSSD) and deconvolution layers (DeCSSD) respectively. The experimental results show that the multi-scale context modeling significantly improves the detection accuracy and SSD coupled with context layers achieves better detection results especially for small objects (+3.2%AP@0.5 on MSCOCO compared to the newest SSD).

III. SCOPE OF PRESENT WORK

There has been a lot of work in object detection using traditional computer vision techniques (sliding windows, deformable part models). However, they lack the accuracy of deep learning based techniques and some of them consist drawbacks too however efficient the algorithms are. To overcome those drawbacks we have to face challenges and find solutions. There have been various

approaches proposed to improve the performance of SSD by enhancing its capabilities to collect more features. The performance of SSD is always very low for small objects no matter how much deep feature extractor we use. It was also found that the performance of SSD degrades more when the training dataset contains small, complex and deforming objects. Its robustness for real life applications is a point of concern. Motivated by these limitations and SSD's popularity because of its capability to operate in real-time, we wish to find the possible improvements to SSD which can allow us to augment its performance for small and complex object detection By having an adaptive default box generation algorithm which creates a set of default boxes based on the distribution of aspect ratios across the training data, we can optimize the selection of the default boxes. Consequently, this creates better localization and classification. & route anonymity protection together. Detection and then the developed system plays an important role in surveillance systems. These systems can be integrated with other tasks such as pose estimation where the first 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, automated cars and even in medical applications.

IV. PROPOSED SYSTEM

Object detection applications are easier to develop than ever before. Besides significant performance improvements, these techniques have also been leveraging massive image datasets to reduce the need for large datasets. In addition, with current approaches focusing on full end-to-end pipelines, performance has also improved significantly, enabling real-time use cases

SSD is widely used for different types of applications. Some applications need accuracy and speed at the same time in order to achieve the main objective. This can be done by getting more data, inventing/creating more data, re-scaling the data, transforming the data or by feature selection. Further, these models will be optimized by tuning the algorithms. Finally, based on the prior research either one of the model or an ensemble (i.e. Combining the models/combining the views/ stacking) of the two will be used to implement a simulated Self-Driving Car. This car will be able to detect the lane curvature and the deviation of the car simultaneously. In addition to this it will be able to detect objects in the pathway.

The above mentioned process is done while maintaining real-time speed, in addition to producing promising detection results on small objects. The evaluation code and trained models are publicly available and also customized as per the requirement.
The work focus on following is the objectives:

1) To find to create new datasets which consists real time images and videos and train the model developed using SSD for object detection.
2) To try our best to improve the accuracy when it comes to the detecting small objects.
3) To maintain the speed while maintain the accuracy.
4) To use in video simulation of self-driving car to detect pedestrian, road curvature and alert the driver regarding the turns.

V. METHODOLOGY

A. Architecture

The network used in this project is based on Single shot detection (SSD). The architecture is shown in Fig.2.The SSD normally starts with a VGG model, which is converted to a fully convolutional network.
Then we attach

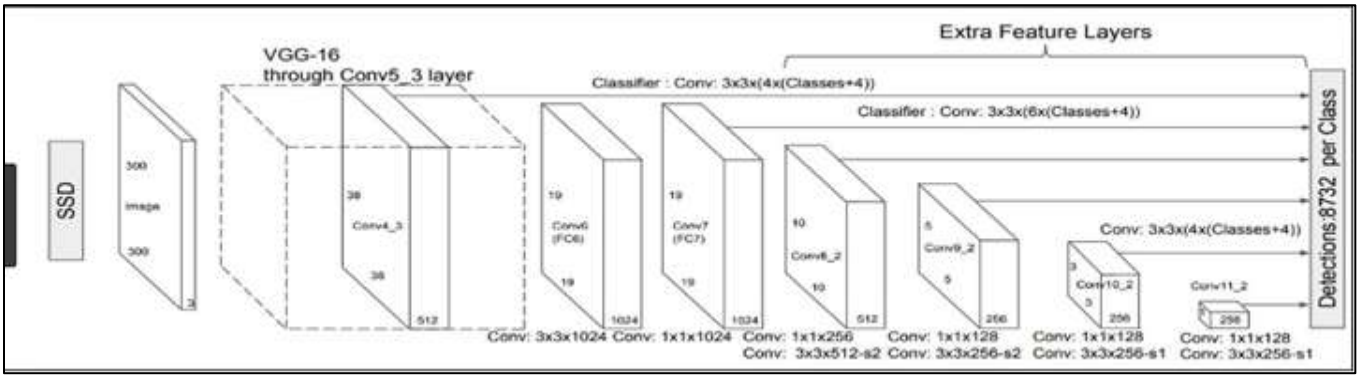


Fig. 2: SSD Architecture

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. 3. Some of the activations are passed to the sub-network that acts as a classifier and a localizer.

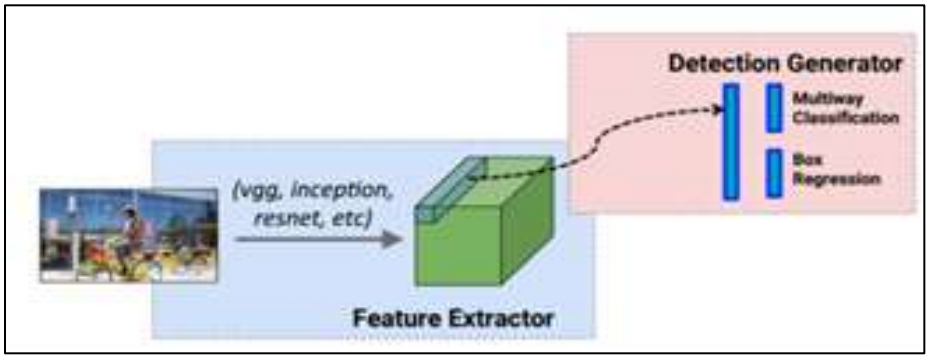


Fig. 3: SSD Overall Idea

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. 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

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

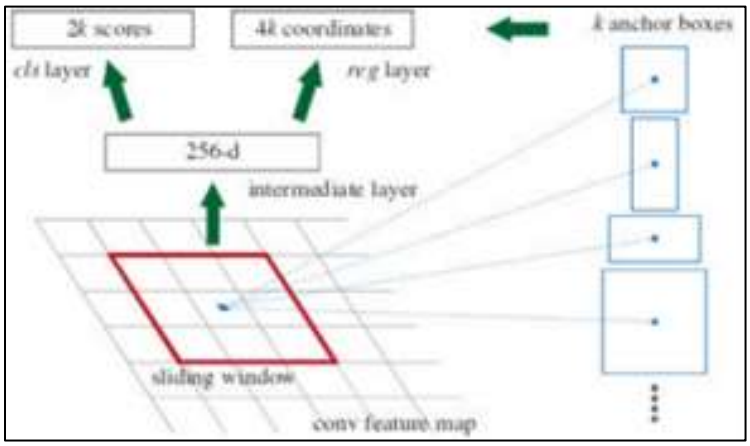
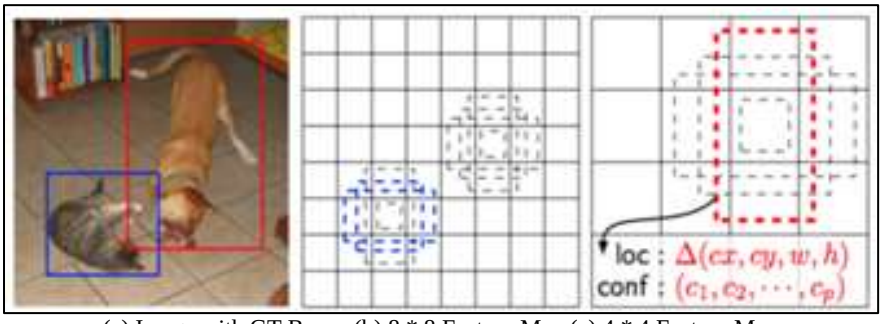


Fig. 4: Anchors

The loss function used is the multi-box classification and regression loss. The classification loss used is the soft max 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.



(a) Image with GT Boxes (b) 8 * 8 Feature Map (c) 4 * 4 Feature Map

Fig. 5: Matches

B. Image Classification

While classification is about predicting label of the object present in an image, detection goes further than that and finds locations of those objects too. In classification, it is assumed that object occupies a significant portion of the image.So the images where multiple objects with different scales/sizes are present at different locations, detection becomes more relevant. So it is about finding all the objects present in an image, predicting their labels/classes and assigning a bounding box around those objects. In image classification, we predict the probabilities of each class, while in object detection, we also predict a bounding box containing the object of that class. So, the output of the network should be:

1) Class probabilities (like classification)
2) Bounding box coordinates. We denote these by cx(x coordinate of center), cy(y coordinate of center), h (height of object), w(width of object).

Class probabilities should also include one additional label representing background because a lot of locations in the image do not correspond to any object.

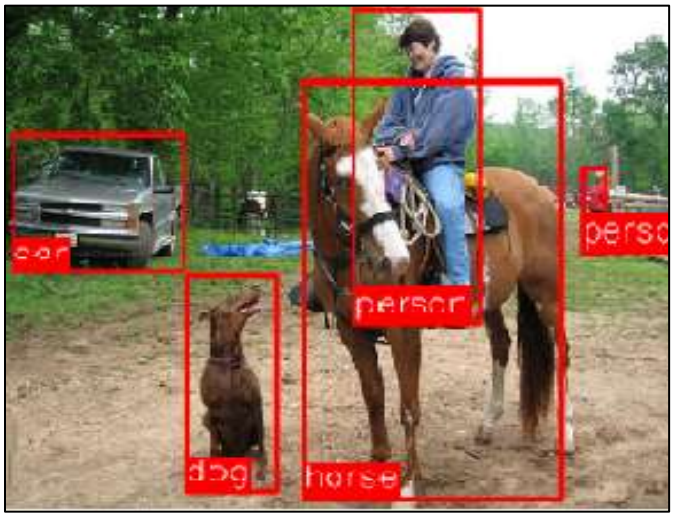


Fig. 6: Image Classification in SSD

C. *Training Datasets*

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. 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.

Fig. 7: detection Results on PASCAL VOCS Datasets

Fig. 8: detection Results on PASCAL VOCS Datasets

Once the model is trained with the available datasets that is PASCAL VOC then the model is trained my using custom datasets which are given by us. The cvustom datasets coinsusts of real time exa, mples and images where the images are taken by us and fed to the recogisation. Some of the examples are given below in the Fig 9 and fig 10.

Fig. 9: Person's Image

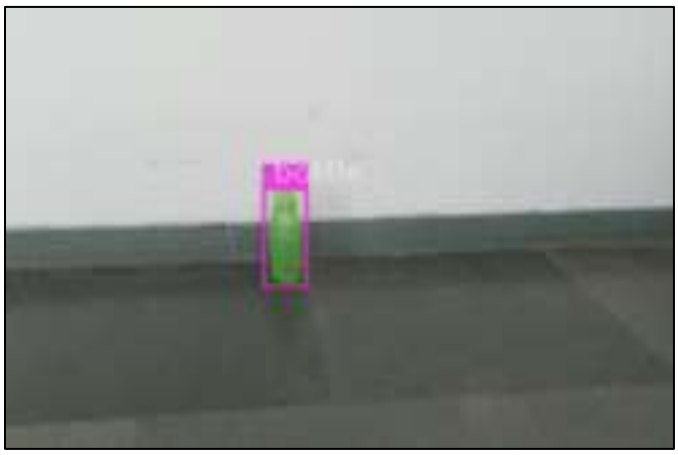


Fig. 10: Example of Tiny Object (Water Bottle)

D. Performance 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 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. T 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.

VI. 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 real time images 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 self-driving

car to keep track of the road curvature, pavement and also the object detection of humans and other candidate objects including the Small objects in the larger frame like backpacks, hangbags etc.

REFERENCES

[1] Nils Gahlert , Marina Mayer, Lukas Schneider, Uwe Franke and Joachim Denzler " MB-Net: MergeBoxes for real-Time 3D Vehicles Detection," IEEE Intelligent vehicles symposium (IV) Changshu, Suzhou, China, June 26-30,2018.
[2] Wei Xiang,Dong-Qing Zhang, Heather Yu, Vassilis Athitsos," Context-Aware Single-Shot Detector",2018 IEEE Winter Conference on Applications of Computer vision.
[3] Yuanyuan Wang,Chao Wang , Hong Zhang"Combining single Shot Multibox Detector with Transfer Learning for ship Detection using Sentinel-1 images", IEEE, 2017.
[4] Devadeep Shyam, Alex Kot, Chinmayee Athalye"Abandoned objectDetection using Pixel-Based Finite State machine and single Shot Multibox Detector"
[5] Takuya Fukagai, Kyosuke Maeda, Satoshi Tanabe, koichi Shirahata,Yasumoto Tomita,Atsushi Ike,Akira nakagawa," Speed-Up of Object Detection Neural Network with GPU", (ICIP), 2018.
[6] Zhaocheng Wang , Lan Du , Senior Member, IEEE, Jiashun mao, Bin Liu, and Dongwen Yang," SAR Target Detection based on SSD With Data Augmentation and Transfer learning," IEEE Geoscience and Remote Sensing letters,2018.
[7] Viral Thakar, Walid Ahmed, Mohammad M Soltani Jia Yuan Yu," Ensemble-based Adaptive Single-shot Multi boxDetector", IEEE,2018.
[8] Swetha M S,Dr.Thungamani M,Ankita Mishra, Enhancement of Performance Analysis in Anonymity MANET through Trust-Aware Routing Protocol.In the proceedings of International Journal of Advance Research in Computer Science and Management Studies, Volume 5, Issue 5, May 2017
[9] Swetha M.S, Dr. Thungamani M , Pooja Reddy, Richa Singh, A Survey on Privacy-Preserving and Authenticated Routing in Mistrustful Mobile Ad-Hoc Networks.In the proceedings of International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2016
[10] Mrs.Swetha M S, Dr.Thungamani M, Enhanced Anonymity in Hierarchical Routing Protocol for MANETs. 978-1-5386-4304-4/18/$31.00 ©2018 IEEE.