International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Accurate Real-Time Object Detection using SSD

Kanishk Wadhwa[1], Jay Kumar Behera[2]

[1]*Department of Information Technology, SRM Institute of Science and Technology, Chennai -603203, T.N., India*
[2]*Department of Information Technology, SRM Institute of Science and Technology, Chennai -603203, T.N., India*

Abstract - *In the earlier days, object detection has been a big challenge in the field of machine learning. But with advanced technology and various object detection algorithms it has become very easy to detect objects accurately. The main subject of our paper is Real Time Object Detection Using SSD. SSD also known as Single Shot Detector is a single convolutional neural network. It works in corporation of the extracted features and bounding boxes. SSD allows more aspect ratios for generating default bounding boxes around the objects detected. SSD boxes can wrap around the objects in a tighter and more accurate manner. With the help of this paper we will present the analysis and implementation of Real-Time Object detection using SSD which is one of the fastest object detection algorithms. The goal of this paper is to analyze different models and their knowledge in this domain.*

Key Words: open computer vision; real-time object detection; single shot multi-box detector (SSD); region proposal network; COCO dataset, MobilNet

1. INTRODUCTION

A few years ago, Hardware and software Image Processing Systems were limited to the development of User Interface (UI) which was mainly developed by the programmers in the firm. However, with the development of operating systems such as Windows Operating System and MacOS, when the developers started solving the problems of image processing itself but this has not led to the significant progress in solving tasks of recognizing faces, car numbers, road signs, analyzing remote and medical images and many more. Each of these problems is solved by trial and error by the efforts of various groups of the engineers and scientists. The task of automating the software creation tools for complex problems is formulated and intensively solved. In the field of object detection, the required tool kit should support the analysis and image recognition of various unknown content by ensuring the effective development of applications by ordinary and efficient programmers.

There is an exponentially increasing amount of image data in the world. However, most of these images are stored on cloud on present on the Internet. To effectively manage such huge data, we need to have some information about its contents. Automated image processing is useful for a large number of tasks that require image captioning or image detection. Objects contained in image files can be located and identified automatically. This is called object detection and is one of the most common problem of computer vision that needs to be solved for future purposes.

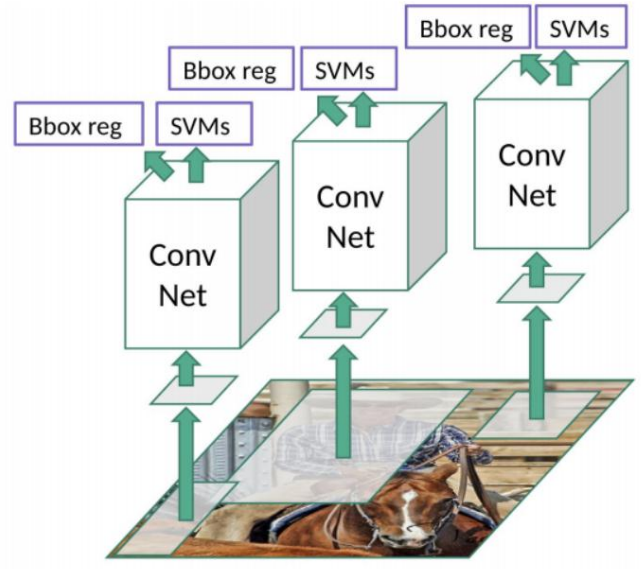
Modern object detectors use feature extractor and feature classifier as traditional object detectors. These two parts can be understood independently in the traditional object detectors whereas they are considered to be united in the modern object detectors. The feature classifier is usually a linear Support Virtual Machine (SVM) which is a non-linear boosted classifier, or considered as an additive kernel SVM. With the rise of its applications such as autonomous vehicles, smart video surveillance, facial detection and various people counting applications, fast and accurate real-time object detection systems are facing a high rise in demand. Besides recognizing and classifying every object in an image, these systems localize each object one by one by drawing the bounding box around that object. This makes Real-Time object detection a harder procedure than the traditional computer vision detectors image recognition. SSD also known as Single Shot Detectors is the fastest of the other models as it detects the object accurately. However, it is not the most accurate model than the other models present but its speed makes it significant to be used in applications which require fast detection such as counting number of people, autonomous vehicles and so on.

2. LITERATURE STUDY

Object Detection is applied in various fields to detect different types of objects accurately. There has been a lot of research in some of the very famous object detection algorithms like RCNN, Resnet, R-FCN and YOLO

2.1 RCNN

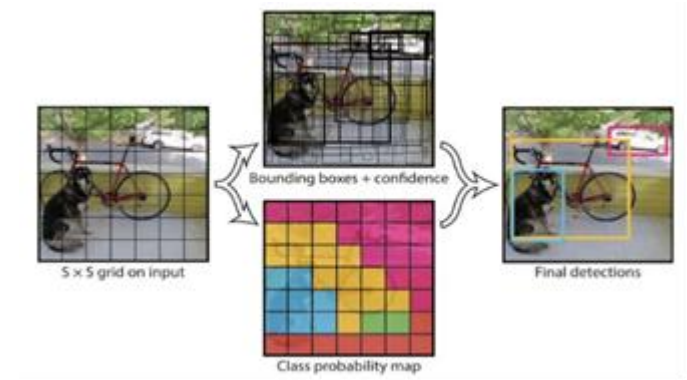
Ross Girshick et al. proposed a method where we use selective search to extract just 2000 regions from the image and he called them region proposals. Instead of trying to classify a huge number of regions, the algorithm proposed to work with just 2000 regions. These 2000 region proposals are bent/twisted in the form of a square and then fed into a convolutional neural network that produces a 4096-dimensional feature vector as output. The CNN acts as a feature extractor and the output dense layer consists of the different features extracted from the input image. Then, the extracted features are fed into an SVM to classify the presence of the object within the candidate region proposal.



The problems with R-CNN are that it still takes a huge amount of time to train the network as it has to classify 2000 region proposals per image, it cannot be implemented real time. Therefore, no learning is happening at that stage. This could lead to the generation of bad region proposals.

2.2 YOLO (You Only Look Once)

YOLO also known as You Only Look Once has the latest version as Yolo-v3 which is considered as custom CNN architecture, which is called DarkNet-53. The initial and first version, Yolo v1 architecture was inspired by Google Net which performed down sampling of the image detected and produced final predictions from a tensor. This tensor is obtained as in the Region of Interest pooling layer of the Faster R-CNN network. The next-generation Yolo v2 architecture used a 30-layer architecture, 19 layers from Darknet-19 and an additional 11 layers used significantly for the detection of objects. This new architecture provided a more accurate and faster object detection mechanism but it struggles with the detection of small objects in the region of interest detecting them incorrectly which can be a problem.



YOLO works by taking an image as input and splitting it into an SxS grid, it takes m bounding boxes within each grid. For every bounding box, the network gives an output 'a' class probability and offset values for each bounding box formed. The bounding box having the class probability above a threshold value is selected and used to further locate the object within the image. YOLO is faster by order of magnitudes (45 frames per second) than other object detection algorithms present. The limitation and disadvantage of YOLO algorithm is that it faces problem with the small objects within the image, for example, it might have difficulties in identifying a bird in the image. This is due to the spatial constraints of the YOLO algorithm.

2.3 ResNet

ResNet helps to achieve excellent performance by training hundreds or thousands of layers. The performance of many Computer Vision applications other than image classification has been weakened such as face recognition and object detection by taking advantage of its powerful representational ability. Deep neural networks are difficult to train because of the vanishing gradient problem — as the gradient is back-propagated to previous layers and repeated multiplication may make the gradient infinitively small. Therefore, as the network goes deeper, the performance starts degrading rapidly or get saturated. The core idea of ResNet is introducing a "identity shortcut connection" which means to skip one or more layers.

These feature layers gradually reduces in size that allowed prediction of the detection on multiple scales. When the input size is given as 300 and 320, it is experimentally known that it replaces the SSD's underlying convolution network with a residual network reducing its accuracy rather than increasing it.

2.4 R-FCN

R-FCN also has to obtain region proposals but in R-FCN the fully convolutional layers after ROI pooling are removed. The process FC layers after pooling does not share the ROI and takes time which makes RPN slow. And the FC layers increase number of connections which increase complexity. All the region proposals will be using the same set of score maps to perform average voting, which is simple calculation. These score maps are convolutional feature maps that have been trained to recognize certain parts of each object. As a result, RFC is faster than Faster RCNN with competitive mAP.

2.5 Fast RCNN and Faster RCNN

Fast RCNN and Faster R have made further evolution in the field of object detection. They use convolutional layers which are initialized with pretraining for ImageNet classification to extract region-independent features and multi-layer perceptron (MLP) for classification. Fast RCNN has improved detection speed over RCNN because it performs feature extraction over the image before it processes regions. It replaces the SVM with a softmax layer which helps to extend a neural network instead of creating a

new model. However, in Faster RCNN Selective Search method is replaced by Region Proposal Network (RPN) which aims to learn the proposal of an object with the help of feature maps. The feature maps extracted from CNN are passed to RPN for proposing the regions. K number of anchor boxes are to generate such region proposals for each location of feature maps. Further, Region of Interest (ROI) pooling operation is performed as it is done in the second stage of Fast RCNN. As in Fast RCNN, ROI feature vector is obtained from the fully connected layers and it is classified by softmax to determine which category it belongs to.

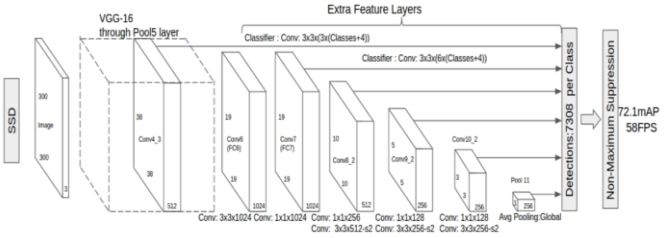
3. IMPLEMENTATION AND TRAINING

In this work, all the experiments were performed with the Tensorflow and Keras open-source deep learning framework, which is developed by the Google research team. We have used SSD know as Single Shot Multi-Box Detector with MobilNet v1 which is fastest model to detect objects. Using multi-box detector allows us to detect multiple objects at the same time by creating m bounding boxes around each object present in the image. We have implemented this model using a low-cost GPU using the pre-trained network with the COCO dataset. Moreover, we have tuned some parameters and extended our training set to improve the performance as much as possible.

MobileNet V1 uses convolutional layers, that are essential to computer vision tasks but are quite expensive to compute, can be replaced by depth wise separable convolutions. The job of the convolution layer is split into two subtasks, first there is a depth wise convolution layer that filters the input. Then this 1X1 convolution combine these values to create the new features. Then, the depth wise and pointwise convolutions form a depth wise separable convolution block. It does approximately the same thing as traditional convolution but is much faster.

3.1 SSD

The task of object detection is to identify "what" objects are inside of the given input image and "where" they are. By giving an input image, the algorithm outputs a list of objects, each associated with a class label and location (with bounding box coordinates). Object detection has been a central problem in computer vision and pattern recognition. It does not only inherit the major challenges from image classification, such as robustness to noise, transformations, occlusions but also introduces new challenges, such as detecting multiple objects, overlapping images identifying their locations in the image.



The two important main tasks in object detection: identify what objects in the image (classification) and where they are (localization). SSD is a multi-scale sliding window detector that leverages deep CNNs for both these tasks.

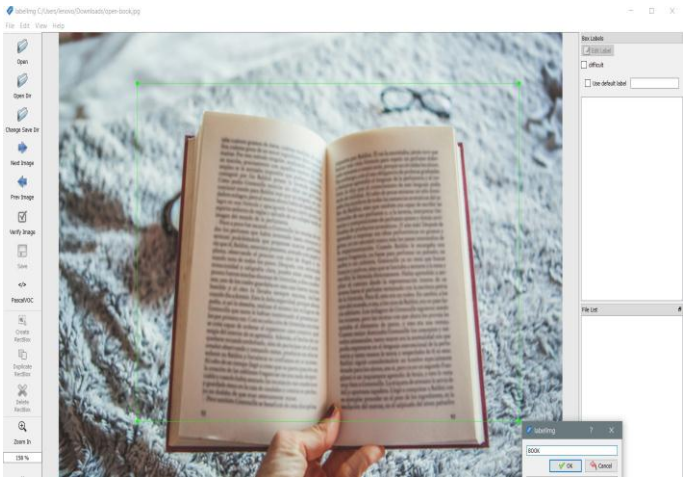
The sliding window detection, as the name suggests, slides a local window across the image and identifies at each location whether the window contains any object of interests or not. Multi-scale increases the robustness of the detection by considering windows of different sizes.

SSD makes the detection relatively very easy. Summarizing the rationale with a few observations:

- Deep convolutional neural networks can classify object strongly against the realting transformation, due to the cascade of pooling operations and non-linear activation. SSD does sliding window detection where the receptive field acts as the local search window just like all other sliding window methods, SSD's search also has a finite resolution, decided by the stride of the convolution and the pooling operation. SSD will get poorly sampled information – where the receptive field is off the target. But even though because of deep features, this doesn't break SSD's classification performance.

- Deep convolutional neural networks can predict not only an object's class but also its precise location. SSD can map the same pixels to a vector of four floating numbers, representing the bounding box. The detection is now free from unwanted shapes; hence it gets much more accurate detection with very less computation.

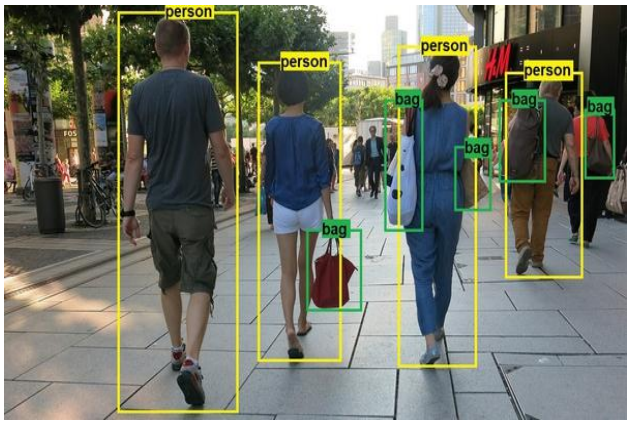
- SSD allows feature sharing between the classification task and the localization task. It is only the very last layer is different between these two tasks. This significantly reduces the computation cost and allows the network to learn the features.



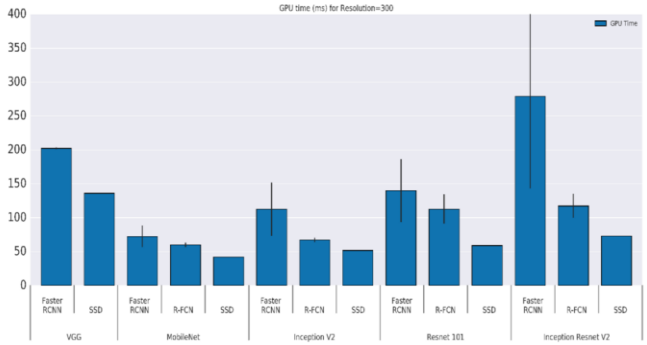
The algorithm used for object detection and tracking is SSD. The whole implementation is done in Python. The custom dataset consisting for 100 images having different objects in the images and was used to train the SSD model which was pre-trained for MS COCO dataset. The images were annotated using LabelImg tool. The Images are separated as training and testing data Images were labelled in (.jpg) format then they are converted to (.xml) format using the XML converter. After all this information is passed inside the code of the main file. Then the training of SSD model with the custom dataset begins with the number of epoch decided depending on the size of the dataset and trying to achieve maximum accuracy. An Input video is passed through the system and then at first total number of frames are extracted and forwarded to object detector which is SSD. Being an object detector SSD generated bounding boxes with class ID and confidence for each bounding box.

4. RESULTS AND ANALYSIS

The proposed system is tested several times with many objects and it is able to detect and identify the objects quite accurately. The project is python-based and evaluated several times through the web cam. It runs with a decent FPS. Input video is broken down into total number of frames and passes each image to out custom-made object detector and once detection is done the bounding box information is passed to SSD algorithm and object detection is performed. Below are the output images from the live video tested on our proposed system which gives the output as bounding boxes with the class name and confidence score.



From the given image, it can be seen that SSD model with MobilNet v1 has the lowest GPU time in milliseconds when compared to other object detection models. However, Faster RCNN performed slowest in terms of GPU time. This was performed for image resolution of 300. This advantage makes SSD useful in applications such as counting object or people as it is the fastest model to detect objects by taking minimum time possible. Moreover, SSD is also the fastest Object-Detection Model in detecting smaller objects with accuracy whereas other models such as RCNN are unable to detect smaller objects accurately.



GPU time (milliseconds) for each model, for image resolution of 300.

5. CONCLUSION AND FUTURE WORK

By using SSD (single Shot Detector) in this experiment, we are able to detect objects in the fastest manner. It helps to identify individual objects in the image in the fastest way. By using Single Shot Multi-Box Detector, we are able to detect multiple objects at the same time in real-time. We have presented how the prevailing fully convolutional neural networks can be used in Real-Time Object Detection systems. The networks were trained on a low-cost GPU using MS-COCO dataset by adding few more examples to the dataset to increase the accuracy of the objects detected by increasing the probability of the objects detected, the performance of the model was then calculated. Moreover, SSD performed better in object localization than Fast RCNN and Faster RCNN. Finally, we are able to perform Real-Time Object Detection using SSD with MobilNet v1 which is faster and efficient than traditional methods of object detection.

The objective of our experiment is to develop an Object Recognition system to identify the 2-D and 3-D objects in the image. Moreover, the Object Detection systems can be further improved by increasing the global or local features so that efficiency of these systems can be increased. The proposed Object Detection system uses grey-scale image and discards the color information from the images. The color information from the images can be used for recognition of objects more accurately as it plays a vital role in Robotics. Night vision mode can also be made available as n in-built feature in tracking devices and CCTV cameras. To make the systems fully automatic and to overcome above limitations, multi-view tracking can be implemented using multiple cameras because of its wide coverage with different viewing angles for the objects to be tracked.

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] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C; SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, the Netherlands, 11–14 October 2016.

[3] Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV 104, 2013

[4] Ugur Alganci, Mehmet Soydas , Elif Sertel ; Comparative Research on Deep Learning Approaches for Airplane Detection from Very High-Resolution Satellite Images. Feb, 2020.

[5] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)

[6] Yun Ren, Changren Zhu, Shunping Xiao ; Object Detection Based on Fast/Faster RCNN Employing Fully Convolutional Architectures.2018,January.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik; Region-Based Convolutional Networks for Accurate Object Detection and Segmentation IEEE Transactions on Pattern Analysis And Machine Intelligence. 2016.

[8] K. Simonyan ,A. Zisserman;Very deep convolutional networks for large-scale image recognition Computer Vision and Pattern Recognition, 2014.

[9] Lin, T.Y., Marie, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.; Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, September 2014.

[10] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, A. W. M. Smeulders ;Selective search for object recognition, International Journal of Computer Vision".2013.

[11] Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: CVPR (2014)

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna;Rethinking the inception architecture for computer vision in Proceedings of the 2016 IEEE Conference on ComputerVision and Pattern Recognition (CVPR '16), July 2016.