International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Recent Trends on Real Time Object Detection using Single Shot Multibox Detector

Ritika Dhiman[1], Dr. Jaswanti[2]

[1, 2]Chandigarh College of Engineering & Technology, Chandigarh, India

Abstract: ***This research paper investigates the running of object detection algorithm on low-end devices to detect different kinds of objects in images. Deep convolutional neural networks (CNNs) which are used in SSD have recently proven extremely capable of performing object detection in single-frame images. Single shot multi-box object detectors have been recently shown to achieve state-of-the-art performance on object detection tasks. The implementation can be done using Pytorch object detection library, and COCO (Common Objects and Context) or Pascal VOC (Visual Object Classes) dataset to detect the common objects around a person like cars, dogs, laptops, etc. The main advantage of using SSD is that, unlike other methods it can be used in laptops and other personal devices.***

Keywords: Computer Vision, Image Procession, Object Detection, CNNs (Convolutional Neural Networks), Low-end devices, COCO (Common Objects And Context), Pascal VOC (Visual Object Classes), Python, Open CV and ROIs (Region Of Interest)

1. Introduction

Computer vision is the theoretical and technological concern that arises when building an artificial system capable of obtaining information from images or multi-dimensional data [1]. Object detection is a process widely used in computer vision, and image processing, to detect semantic objects of a particular class (e.g. cat, fire hydrant, human) in digital images or video. State of the art computer vision systems have involved a range of object detection models that use convolutional neural networks in their working [2,3]. Google photos has deployed models like SSD Mobile Net which is known for its speed and isn't memory intensive, here performance isn't the most important factor, but memory efficiency is. In case of self-driving cars though, the requirement for an extremely accurate model is a priority, as these real time systems are performing tasks which can lead to a life and death situation in their surroundings.

If we look closely at the body of modern cars, we can spot a lot of different sensors, which were not built in into older cars [4]. These sensors are essential for implementing an artificial intelligence. Just like the human being, before planning and reacting to the environment, we need some kind of perception. Humans gather information with our "human sensors" like the nose, ears, and eyes. The car has several different ways to perceive the surroundings, for example, laser or lidar sensors to scan point clouds of the surroundings. Another more human-like approach to incorporate the environment is a camera. Just like the eye, a camera is able to catch the light and transfer this information. Nevertheless, only gathering images about the outer world is not enough. Vision only begins with the eyes, but truly takes place in the brain [5]. Just like a person needs the brain to process the visual input, the system needs to derive crucial information out of the camera output.

Computer Vision can be used not only to locate objects in pictures but also to classify them and determine their pose in the environment [6]. However, recent research has brought up a different approach. Deep learning has shown to surpass former state-of-the-art technology in object detection. A lot of different deep learning object detection architectures have been published. Each of them uses a convolutional feature extractor as its basis. In the last years, a lot of deep neural feature extractor hit the stage, too. This leads to a huge number of different combinations of feature extractors and object detectors [7,8].

This paper presents the running of object detection algorithm on low-end devices to detect different kinds of objects in images. Deep convolutional neural networks (CNNs) which are used in SSD have recently proven extremely capable of performing object detection in single-frame images. The implementation can be done using Pytorch object detection library, and COCO (Common Objects and Context) or Pascal VOC (Visual Object Classes) dataset. The evaluations carried out demonstrate that low-end devices have sufficient computing power to run certain object detection algorithms with a real-time video feed.

2. Methodology

The main advantage of using SSD is that, unlike other methods it can be used in laptops and other personal devices. The real-life objects and living beings captured using a normal webcam can be detected by interfacing python in Spyder IDE and Open CV library. It combines high detection accuracy with real-time speed. However, it is widely recognized that SSD is less accurate in detecting small objects compared to large objects, because it ignores the context from outside the proposal boxes. The SSD algorithm uses semantic segmentation to create the bounding boxes of varying colors based on the different classes present in an image as shown in Figure 1. The objects are identified by comparing with pre-trained objects present in the COCO or Pascal VOC dataset as shown in Figure 2. Along with the name of each object, their confidence score is also obtained which can be used to compare the accuracy of detection.

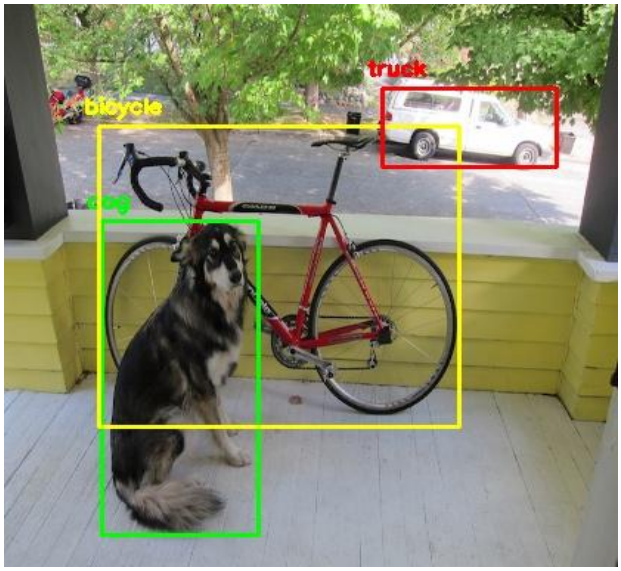

Figure 1: Semantic Segmentation on the different classes.

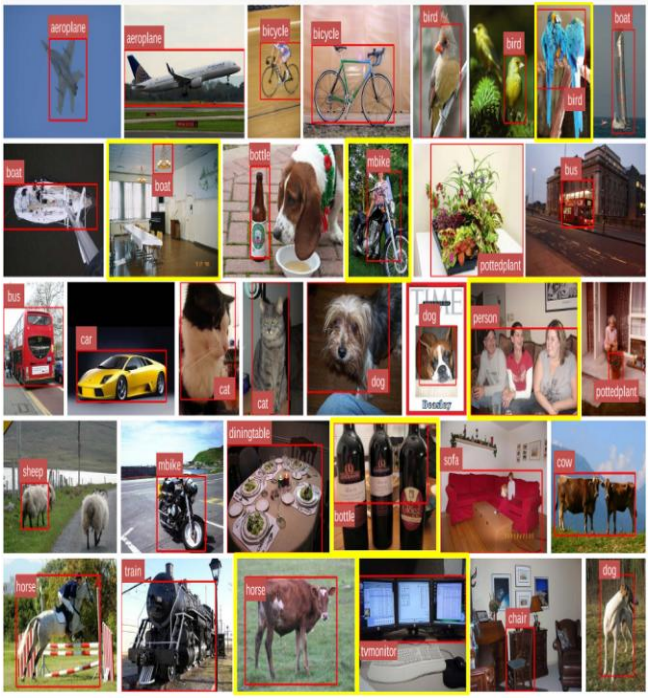

Figure 2: COCO or Pascal VOC dataset on objects

3. Results and Discussion

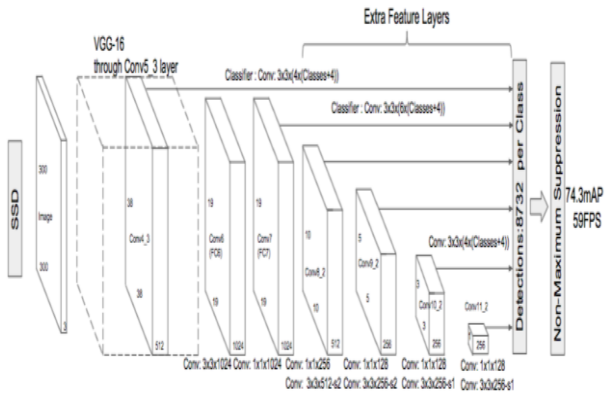

Figure 3: Single Shot Multibox Detector Architecture

Here, Figure 3 represents the SSD Architecture. The VGG-16 is used as a base network because of its strong performance in high quality image classification tasks. A set of auxiliary convolutional layers from conv6 onwards are added thus enabling to extract features at multiple scales and progressively decrease the size of the input to each subsequent layer and each of the convolutional layers are used to classify objects.

It deploys a technique to pick ROIs (Region Of Interests). ROIs is basically the object which we want to detect e.g. a face or a tree. End-to-end training is performed to predict a class and compute the boundary shift for a particular ROI. Like Faster RCNN, SSD works on the same concept of using anchor boxes to create ROIs. All this is done by using the feature maps of the last layer of shared convolution layer. In a layer of such feature maps, k anchor boxes which have varying aspect ratios are picked to be around each pixel. So if a feature map has a resolution of m x n then, that amounts to m x n x k ROIs for that layer. To deal with detecting objects of various sizes, SSD uses many such feature layers that have different dimensions to generate the ROIs. Earlier layers in a deep convolutional network capture only low-level features; it utilizes features maps from after certain levels and layers. After going through some specific transformation, if an ROI matches with the Ground Truth for a class and has 0.5 or more Jaccard overlap value then it is labelled as positive. This is also called the multi box concepts as shown by Figure 4 and Figure 5. In Figure 4, it shows how the image is divided into different segments by points having boxes of different kinds surrounding it. On the other hand, Figure 5 shows the boxes which detected complete or part of object in the image i.e. the boat in this case. Different resolutions of feature maps and varying aspect ratios of each box make the task more challenging. To handle that, SSD performs scaling and works with different aspect ratios to be close to the ground truth dimensions. This facilitates the calculation of Jaccard overlap for default boxes for each pixel at a particular resolution. SSD uses a small kernel of dimensions 3*3*p (p channels) to predict the value of four offsets (cx, cy, h, w) for each candidate ROI from the Ground Truth box with a confidence score for every class.

So, for each box out of k at a given pixel location, c class scores and four offset values relative to actual default box shape are computed. ROIs are generated by using multiple feature maps with varying dimensions. The ROIs are labeled as positive or negative based on the value of the Jaccard overlap after the ground box has scaled appropriately to deal with the differences in the resolutions of the input image and feature map.

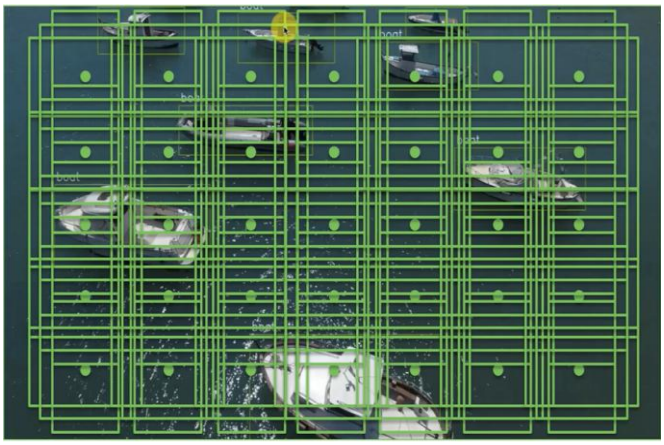

Figure 4: Image is divided in different segment by points

Figure 5: Detector to detect complete or part of object in the image

4. Conclusion

This paper introduces SSD, a fast single-shot object detector for multiple categories. A key feature of our model is the use of multi-scale convolutional bounding box outputs attached to multiple feature maps at the top of the network. This representation allows us to efficiently model the space of possible box shapes. SSD model provides a useful building block for systems that employ an object detecting component. The evaluations carried out demonstrate that low-end devices have sufficient computing power to run certain object detection algorithms with a real-time video feed. It can be used as a part of system using recurrent neural networks to detect and track objects in videos simultaneously by considering a video as a set of frames or images taken one at a time at a very fast pace. The overall performance of the proposed SSD method is relatively poor. The reason for this seems to be partially due to low resolution and partially due to the fact that the original method was designed to capture objects of various aspect ratios instead of objects of various scales. When increasing the resolution we see that performance increases for the more frequent occurring samples in our dataset (hard and faces of 50 < pixels). Here we see that inference/training on a resolution of 700x700 is the best resolution for the WIDER dataset. The future enhancement of this project can be used for facial recognition in order to take the attendance in the school to avoid mal-practices, to detect the movement of the eyes while driving in order to ensure that the driver is not sleeping to prevent the accidents by alerting the driver by vibrating or by using some sound alerts, it can be used to calculate the distance between two vehicles for parking the vehicle by speech alerts, and so on.

References

[1] A. Bonnaccorsi, “On the Relationship between Firm Size and Export Intensity,” Journal of International Business Studies, XXIII (4), pp. 605-635, 1992. (journal style)

[2] Yann LeCun, Leon Bottou, Yoshua Bengio, Patrick Haffner, Gradient-Based Learning Applied to Document Recognition, *IEEE Conference*, 1998

[3] Szegedy,C.,Liu,W.,Jia,Y.,Sermanet,P.,Reed,S.E.,Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. *In: CVPR, IEEE Computer Society*, pp. 1–9.

[4] Torii, A., Sivic, J., Okutomi, M. and Pajdla, T., 2015. Visual place recognition with repetitive structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(11), pp. 2346– 2359.

[5] Zepeda, J. and Perez, P., 2015. Exemplar svms as visual feature encoders. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3052–3060.

[6] O. Hahm,E. Baccelli,H. Petersen,N. Tsiftes, Operating Systems for Low-End Devices in the Internet of Things: A Survey, *IEEE Internet of Things Journal*, 2016

[7] Ms.Godlin Jasil S.P,Shaik Asif Moinuddin, Shaik Bab Ibrahim, M.Sakthivel,B.Sakthi Arjun, Home security alert system using moving object detection in video surveillance system, *ARPN Journal of Engineering and Applied Sciences*, 2016

[8] Rajat Agarwal, Bhushan Gajare, Omkar Kute3 Pravin Wattamwar, Shobha S. Raskar, Review of Security System based on P.I.R Sensor using Face Recognition Concept, IJSRD - *International Journal for Scientific Research Development*, 2017

[9] Shen, Z., Liu, Z., Li, J., Jiang, Y., Chen, Y. and Xue, X., DSOD:learning deeply supervised object detectors from scratch. In: *IEEE International Conference on Computer Vision*, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 1937–1945.