Multiple Object Detection Using Single Shot Multi-Box with MobileNet in Real-Time.

Md.Samin Rahman[1],AsifAhammad Miazee[1], Md. Mamun Ahmed[1], Md.Aktarujjaman[1]

[1]Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology, Cumilla Cantonment, Cumilla, Bangladesh.

*asifahammad7@gmail.com

Abstract: In the field of artificial intelligence, one of the challenging tasks is detecting real-time objects because it needs faster computation power to recognize the content at that moment. Here, a novel object recognition approach is proposed, which combines Single Shot Multi-Box Detection (SSD) with a lightweight network model known as MobileNet. SSD speeds up the classification of sub-windows by formulating the problem as a sequential decision process. Additionally, MobileNet provides better multi-scale handling to detect objects of all sizes without rescaling the input image. This speed-up builds upon the scale invariance property of image statistics in natural images that offers a powerful relationship for approximating feature responses of adjacent scales. Experimental results showed that this combination of MobileNet with the SSD template, which is the proposed novelty of the research, improves the level of validity when recognizing real-time household objects.

Keywords: *Object recognition, SSD, MobileNet*

1. Introduction

Object recognition is a common application for the position and labelling of objects within a video series by computer vision techniques. It is very important and difficult to detect and track moving objects or targets in real-time video. As there are many approaches in the process of detection that have so far been observed, the rate of precision remains uncertain. The neural network method has therefore been developed to detect objects in the sequence of video. One of these systems is the deep neural network and it makes the hidden layers more precise to detect the object in the video. In 2014, the R-CNN was initially used for the detection process based on a deep convolution neural network model. Then there were more improved methods such as simple R-CNN, Faster RCNN, faster R-FCN, and SPP-CNN. It cannot be used for multiple different types of recognition due to its complex networking structure. For that a single network and better output are important. Therefore, the single-shot multi-box detector depends on VGG and that has some extra layers since the reduction of the feature was specially designed for real-time object detection [17]. Few of the variation in multilevel property and the existing box idea apply to reclaim the drop in precision SSD.

Because SSD uses the filter of convolution to detect the altitude of the images it loses its precision when it gets the low-resolution frame [3]. Here, we proposed combining MobileNet because it harnesses the quality of separable and profound convolution, which minimizes the length of parameters dramatically compared to standard coalescent filters of the same length. Thus, the lightweight of the deep neural network is obtained which eventually makes the whole system faster..

2. Previous Works

The greatest challenge in creating a stable object detector is the volume of picture variance. Important factors leading to this variability include; the location, distance, or direction of the target about the frame, significant in-class object class variations, background noise, color discrepancies, shifts in lighting, and (partial) occlusions [14]. Target recognition algorithms strive to identify artefacts under certain circumstances and to be as resilient to such differences as possible [15]. A significant parameter influencing both speed and detection rate is the defined sliding-window phase size at which the detector scans through the image [16].

The research by S. Kanimozhi et al [12] inspires most of the research presented in this paper. Similar to our research the authors proposed a very fast object detector; however focusing on the problem of pedestrian detection. In their detection framework, the core contribution for reaching high detection speeds is a ground place estimation technique for reducing the search space of images [13]. The authors report detection speeds of 100 fps, using the ground plane estimation technique requiring a stereo-camera. Using the stereo-camera setup and the ground plane estimation technique the authors can reduce the search space to 640 × 60 pixels regions over 10 scales; hence this allows them to run the detector at such high speeds. In the research presented here, rather than considering ground plane estimation, speeding up the classification by manipulating the sub-windows is studied.

3. Single Shot Multibox MobileNet (SSD)

It implemented an updated version of MobileNet called Mobile-Det is used with the combination of Single Shot MultiBox Detector (SSD) form. We have developed this joint version to measure the favor of using this joint version and do a feasible review with other models like YOLO and VGG-based SSD. The SSD component of this project is enormous and far-reaching, so it will have a small prolog on how it works in similar pieces. Generally, SSD used different types of layers as classifiers, using the combination of various type ratios in the type of existing boxes at a specific location to evaluate each feature map in a coevolutionary manner [3]. The classifier also predicts the scores of class according to ranking on the containers. The accuracy in measuring the default boxes is only performed into account at the time of training if its Jaccard overlaps with the ground [7].

Figure-1 demonstrates the Mobile-Net design with the same frame arrangement as the SSD-VGG-300 [10]. But we are using MobileNet as a basis instead of using VGG in our research [11]. In addition, the deeply separable convolution form is restored in our approach instead of regular convolution.

Finally, it is obvious that the implementation of the SSD framework works well in the complete training of the figure instead of hanging on the frame of reference. Consequently, also, in theory, the transient data is increasingly reliable. But the main problem is it turns very slow as more convolutions are added in this model.

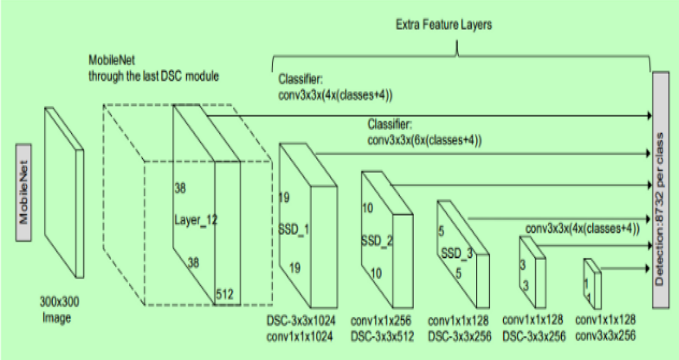

Fig.1. *Mobile Net SSD.*

For the extraction of feature charts, SSD uses VGG16. Then it uses the Conv4_3 layer to detect artefacts. For illustration, we draw the Conv4_3 to be 8 × 8 spatially (it should be 38 × 38).For each cell, it makes 4 object predictions.

4. Methodology

Object detection methods usually fall into either machine-based learning approaches or deep learning-based approaches. It is a type of application that is done by the mechanism of deep learning technique. Deep learning techniques can detect end-to-end artefacts without specific characteristics, which are usually based on convolution neural networks (CNN).

The proposed MobileNet SSD approach can be useful in-depth structure to recognize the real-time artefacts more efficiently. For generating the actual outcome whole methodology is divided into three steps:

Step 1 explains how the API generated for object detection in Tensor flow is used to detect artefacts.
Step 2 explains the best SSD process for the detection of artefacts in real-time.
Step 3 addresses how MobileNet contributes to making the whole system more lightweight.

Step 1: In expressing and implementing machine learning algorithms, it is an interface for the detection of the object. The object detection Tensor flow Application Programming Interface is a platform developed over Tensor Flow that makes it easy to train and deploys different object models. To detect and monitor this object, it used this Tensor Flow API. Design of a learning system It is still a difficult computer vision task to identify and correctly detect multiple objects in a single frame. A calculation communicated using Tensor Flow is conducted with almost no modification in a large range of heterogeneous systems [9], ranging from mobile phones and tablets to large scale distributed frameworks with multiple instruments and a large number of the computer device.

Step 2: Mobile Nets operates based on the layer of Deep Separable Convolution (DSC). It used some initial type of model to minimize layer computing costs for two reasons: The standardized method of convolution was replaced by the Depth-wise model. Spatial convolution was conveyed independently over the input of each channel in-depth type approach. Another reason is point wise convolution [6]. A simple convolution layer is used for the information of channel projection from depth into the extra space of the channel. Since deep separable Convolution has fewer parameters than regular layers of convolution, it also needed to calculate only less operation [8]. That's why it's cheaper and faster.

Step 3: This section explains the structure model that we have used to implement this project. Figure-1 shows the network model consisting of different DSC modules. Rectified Linear Unit, normalization of the batch, point, and depth-wise operations are the layers included in the DSC section. The module's first layer is used for regular convolution and while the end layer is used to minimize the spatial type of resolution.co-author's ORCID ID there is a look-up option included in Manuscript Central.

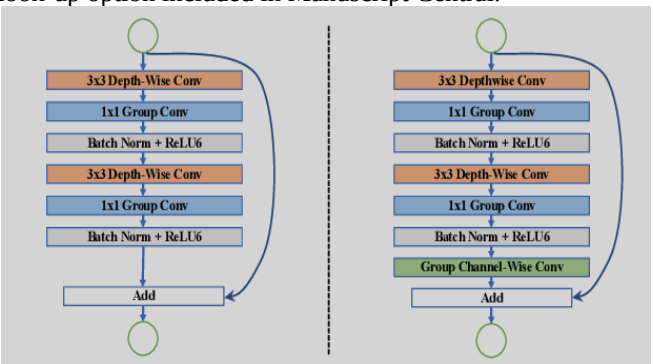

Fig.2. *Structure of a Separable Convolution according to Depth-wise typical module.*

5. Environment

The implementation environment was an Acer SF314-55 laptop computer fitted with an Intel Core i7-4210U 1.70 GHz Processor, 8 GB RAM, and 840 M GPU NVIDIA GeForce. Windows 10 is the operating system. Anaconda was the principal software tool. The framework for object detection and its associated methods was implemented as a combination of pre-existing and self-programmed with Anaconda tools.

6. Data-Set

Our image data is taken from our custom dataset which is similar to the COCO dataset. We have created a dataset with images used in today's life, such as cell phones, cups, mouse, keyboard, scissors, etc. Between them, 60% of the data is used during the training process and 40% of the data is used in the testing/validation and test sets in the test process with the distribution of images and objects.

7. Results

Fig.3. *Different types of detection in our Experimentation.*

It is the complete output figure of our thesis. Here cell phone, cup, mouse, scissors, and keyboard are detected. There is box which covered the sample. It denotes the accuracy percentage of each sample. It has some major advantages. It can detect the entire sample together with minimal light.

8. Loss and Accuracy

In fig.4. X axis denotes total trained steps and Y axis denotes the loss scale.Here we count the smooth total loss value. We trained our machine about 35,000 steps. After a sequence of 5,000 steps up to 35,000 steps, we showed smooth values. The main concern of train steps is to keep the smooth under 0.05 for better accuracy. Already it focused on it.

There have many methods of object detection like R-CNN, fast Yolo, Yolo (vgg16), SSD, and so on. At first, we try to make a comparison between these methods by three attributes. These are mAP, FPS, and the number of boxes that create during the implementation of the method. Here, mAP means mean average precision, FPS means frames per second. This comparison is made to pick the best object detection method and make it more efficient. SSD has been used to implement this project. It can easily notice that SSD has high mAP with low FPS. And it can divide a picture into a huge number of boxes which is around 25 thousand. And this deviation helps the machine to detect an object more correctly.

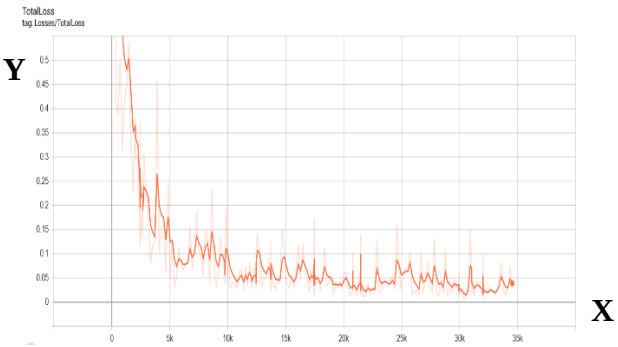

Fig.4.*Total Loss.*

Table 1 Trained steps VS Smooth value

Trained Steps	Smooth Value
5000	0.1248
10,000	0.0964
15,000	0.0834
20,000	0.0598
25,000	0.0254
30,000	0.0234
35,000	0.0225

9. Discussion

Many CNN detection methods for R-CNN [1], for example, start from the recommendation to the classifiers of objects of the different sites and scales in a test picture and provide the resulting classifiers of the raised region to detect an artefact.

Upon identification, the bounding boxes are corrected and the boxes are re-scoring using other items in this frame upon storage. The invented RCNN version is then given, such as F-RCNN [2] and Faster-RCNN [3], which is used several mechanisms to reduce regional proposal manipulation. Faster RCNN is useful to detect artefacts via the KITTI [4] dataset traffic data system. However, this method does not provide a new detection rate for real-time data which directly clarify that much more research is still needed to improve the inference velocity for the data of real-life. In the YOLO system [5] the problem with increasing real-life data inference speed was solved by combining the area plan with the classification approach including a tangible problem of regression straight from pixel picture to

the bordering of the box with group probabilities. As a whole pipeline discovery is a special network, it may improve the performance of direct detection.

10. Conclusion

We tried to identify the image we display in front of a webcam in this document. The model developed was tested and trained using a framework provided by Google's Tensor Flow Object Detection API. Studying a structure from a web camera generates many problems, so well structures per the second approach are needed to minimize problems with input/output [11]. So it concentrated on the techniques of threading which greatly increases structure per second to improve the processing time for each object. Although the software correctly identifies each object in front of the camera, moving the detected object box over the subsequent object in that video takes about 3-5 seconds. Using the mentioned function, it can be detected and focus the object in the area of sports so that the machine can prepare profoundly.

By using the Public Surveillance Camera to track the emergency Vehicle in traffic, it will monitor the traffic signals. Through monitoring the abnormal behavior of the people, it can also identify the public place occurred crime in the. Even monitoring their movement on the border of our country, we can introduce this application in the satellite to avoid terrorism activities. The plan will be useful in identifying and monitoring objects and making life easier.

11. References

[1]T. D. R. Girshick, J. Donahue, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". Computer Vision and Pattern Recognition (CVPR) Conference,2014.

[2]R. Girshick. "Fast R-CNN", In Proceedings of the IEEE International Conference on Computer Vision, pages 1440–1448, 2015.

[3]S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks". In Advances in neural information processing systems, pages 91–99, 2015.

[4]A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. "Vision meets robotics: The kitti dataset". International Journal of Robotics Research (IJRR), 2013.

[5]J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You only look once: Unified, real-time object detection", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788, 2016.

[6]K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang, "Object detection in videos with tubelet proposal networks," in CVPR, 2017.

[7]J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in ACM MM, 2014.

[8]Chu, W.T., Cheng, W.C.: Manga-specific features and latent style model for manga style analysis. In: International Conference on Acoustics, Speech and Signal Processing, pp. 1332–1336. IEEE (2016).

[9]C. Wojek, P. Dollar, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, p. 743, 2012.

[10]D. Ribeiro, A. Mateus, J. C. Nascimento, and P. Miraldo, "A real-time pedestrian detector using deep learning for human-aware navigation," arXiv:1607.04441, 2016.

[11]W.T. Chu, W.W. Li, "Manga Face Net: Face Detection in Manga based on Deep Neural Network," Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 412-415, June 2017.

[12]S. Kanimozhi, G. Gayathri and T. Mala, "Multiple Real-time object identification using Single-Shot Multi-Box detection," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-5.

[13]J. Redmon, S. Divvala, R. Girshick, and A. Farhadi.—You only look once: Unified, real-time object detection‖, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788, 2016.

[14]LubomirBourdev and Jonathan Brandt. Robust object detection via soft cascade. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 236–243. IEEE, 2005.

[15]Roman Juránek. Detection of dogs in video using statistical classifiers. In Computer Vision and Graphics, pages 249–259. Springer, 2009.

[16]K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang, —Object detection in videos with tubelet proposal networks,‖ in CVPR, 2017.

[17]P. Doll´ar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," PAMI, 2014.