Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/62

Video Object Detection Using Densenet-Ssd

Durgaprasad Gangodkar[1], Vrince Vimal[2]

[1]Department of Computer Science & Engineering,Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002
[2]Department of Computer Science & Engineering,Graphic Era Hill University, Dehradun, Uttarakhand India, 248002

ABSTRACT

Despite the fact that there are well-established object detection methods that are based on images, the application of these approaches to video data on a frame-by-frame basis has two major drawbacks. One is inefficiency in terms of processing caused either by redundant information across image frames or by the absence of temporal and spatial correlation of features across image frames. And other is not being able to cope with the challenges of everyday utilization, such as motion blur and occlusion. In order to capture both temporal and spatial correlation of features, we proposed an improved SSD algorithm called DenseNet – SSD. From the experimental analysis, it is explained that our proposed object detection model gives better result.

Keywords: Object Detection, Deep Learning, Image Frames, Video Object Detection

INTRODUCTION

Many different computer vision tasks, including image classification, object recognition, semantic segmentation, human position estimation, and many others, have seen widespread applications of deep learning. Improvements in object detection on ImageNet and PASCAL VOC have been made possible by the evolution of deep convolutional neural networks during the past few years. Two of the most widely used computer vision benchmarks are ImageNet and PASCAL VOC. The most advanced techniques for object recognition are used to teach CNNs to identify proposed regions of an image as either background or one of the object classes. On the other hand, these methods concentrate on finding things in static images. It is becoming increasingly vital to automatically extract relevant information from video surveillance systems as more and more video surveillance systems are being deployed across a wide range of locations. However, the existing monitoring system does not provide the target identification capability, which is particularly important for the case that was presented. The purpose of this paper is to offer a framework based on deep learning that can detect the objects specified from video.

Targets may be relocated quickly and simply in the video surveillance system across situations with varying degrees of light and shade. When dealing with complicated scenarios, the standard approach of object detection often fails to accurately identify the targets, especially when there are multiple categories involved. As data mining techniques have advanced, so too have the benefits of deep learning in terms of how models are expressed. Target detection and behaviour identification are two areas where it shines. Deep learning's convolution neural network is particularly resilient to Displacement, scaling, and distortion because it relies on local perception, shared weights, and down sampling. The convolution neural network (CNN) has developed into a powerful tool for image recognition, with applications ranging from vehicle licence plate recognition to facial recognition. In this propose a DenseNet – SSD to capture temporal and spatial features from the video.

RELATED STUDY

The computer vision community has not yet solved the challenge of object detection. Both its intrinsic difficulty and its potential for broad application play a role in this. Although there has been significant development in the speed and accuracy of object detection frameworks for still images in recent years [1], less has been done in the related topic of object detection in the video domain. Compared to a still image, a video can tell you a lot more about the subject, including how they move and how deep the background is. Any or all of these data sets may be utilised to hone in on the target with greater precision. In this paper, we investigate a new approach to improving real-time object detectors by using contextual information extracted from videos [2]. In contrast, box-level methods utilise the output of object detectors applied independently across several sequential frames to perform operations on the resulting bounding boxes. There has been much greater focus on box-level approaches in the video object detection literature. For instance, Han et al. [3] developed a method that uses bounding boxes from numerous frames in place of the traditional Non-Maximal Suppression (NMS). Another box-level method for better object predictions is described by Tripathi et al. [4], which use a recurrent network to evaluate a series of outputs from object detectors.

Using the results of a single-frame detector, Kang et al. [5] create spatio-temporal "tubelets" that are then used to refine box predictions. To generate fresh bounding boxes and class probabilities, Lu et al. [6] feed feature maps from a single-frame detector into a recurrent network. All of these methods belong to a broader category of box-level methods called tracking-by-detection [7]. These methods are based on the fundamental idea of generating tracks by associating detections throughout the output of an object detector that has been separately applied to sequential single-frame images. These detections are then associated with one another to produce the track. After then, these tracks might be utilised to eliminate false positives and bring back any detections that had been missed. However, the following features of video data are ignored when object detection is performed independently on each image frame: (1) There are feature extraction redundancies between neighbouring frames due to the spatial and temporal correlations between image frames. Computational inefficiency is sacrificed when features are detected in each individual frame. (2) Some frames in a long video stream may have low quality because of factors such motion blur, video defocus, occlusion, and position changes. Object detection accuracies suffer when low quality frames are used. Strategies for video object detection aim to deal with the aforementioned problems. In order to address these problems, we proposed a DenseNet – SSD which captures both spatial and temporal features even in the low video streams.

PROPOSED METHOD

Object detection in single images has recently been shown to be well within the capabilities of deep convolutional neural networks (CNNs). We generalise a family of convolutional neural networks

(CNNs) originally developed for single-frame picture object detection to cover object detection in videos with multiple frames. To improve upon the Single Shot Detector, we developed the DenseNet Single Shot Detector (DenseNet - SSD), which uses contextual temporal information from video to identify shots. Our model enhances the accuracy of the object detector by adding a deep convolutional layer to an SSD architecture, allowing it to merge features across several frames and make use of the additional spatial and temporal information available in video data. To preserve the remarkable speed of SSDs, our technique employs a fully convolutional network architecture.

DenseNet – SSD

the SSD algorithm's framework, which consists primarily of two parts. One component is a front-end deep convolutional neural network that extracts basic features about the target using data classified by a VGGNet-16 network (i.e., a VGGNet with a 16-layer structure stripped of its classification layer). The other part is a multi-scale feature detection network, which is a cascade of convolutional neural networks (CNNs) that uses the feature layer generated by the front-end network to extract features under different scale situations. The fully connected layer in YOLO is removed by the SSD method, and in its place, multi-scale feature mappings are mapped onto the detection layer. This is done to account for the unavoidable reality that the input image contains targets of varied sizes. The SSD algorithm employs non-maximal suppression in the layer following the detection layer to achieve the best localised value. This is done so that the desired outcome can be attained. When the loss value stabilises at a lower level, this shows that the detection model has converged, and the loss function is used to assess the discrepancy between the expected and actual value of the object. This is how the best detection model is found through iterations of finding the best detection model. The loss function in the target detection field is composed of the positioning loss of the object as well as the classification loss. Both of these losses add up to the total amount of information that is lost. The formula can be divided into the following:

$$L(y,b,i,g) = \frac{1}{N}\left(L_{conf}(y,b) + \alpha L_{loc}(y,i,g)\right)$$

Where $\alpha L_{loc}(y,i,g)$ is the loss of object positioning.

The standard SSD approach is able to identify the item by fusing the conv4 3 layer of VGG16 in lieu of the conv 7 layer of the fully connected layer and appending the feature map acquired by the four convolutional layers. These three steps are followed by the addition of the feature map. The detection outcome is then obtained via the non-maximum value suppression method. By fusing the Dense Blocks of the DenseNet network and feeding the output of the three additional convolutional layers into the detection layer, the enhanced DenseNet-SSD algorithm was able to get the feature maps. This was done to meet the feature map size criteria of the feature mapping layer and improve the integration of high-level and low-level information. In addition, the improved algorithm was able to obtain the feature maps. The proposed model is shown in Figure 1.

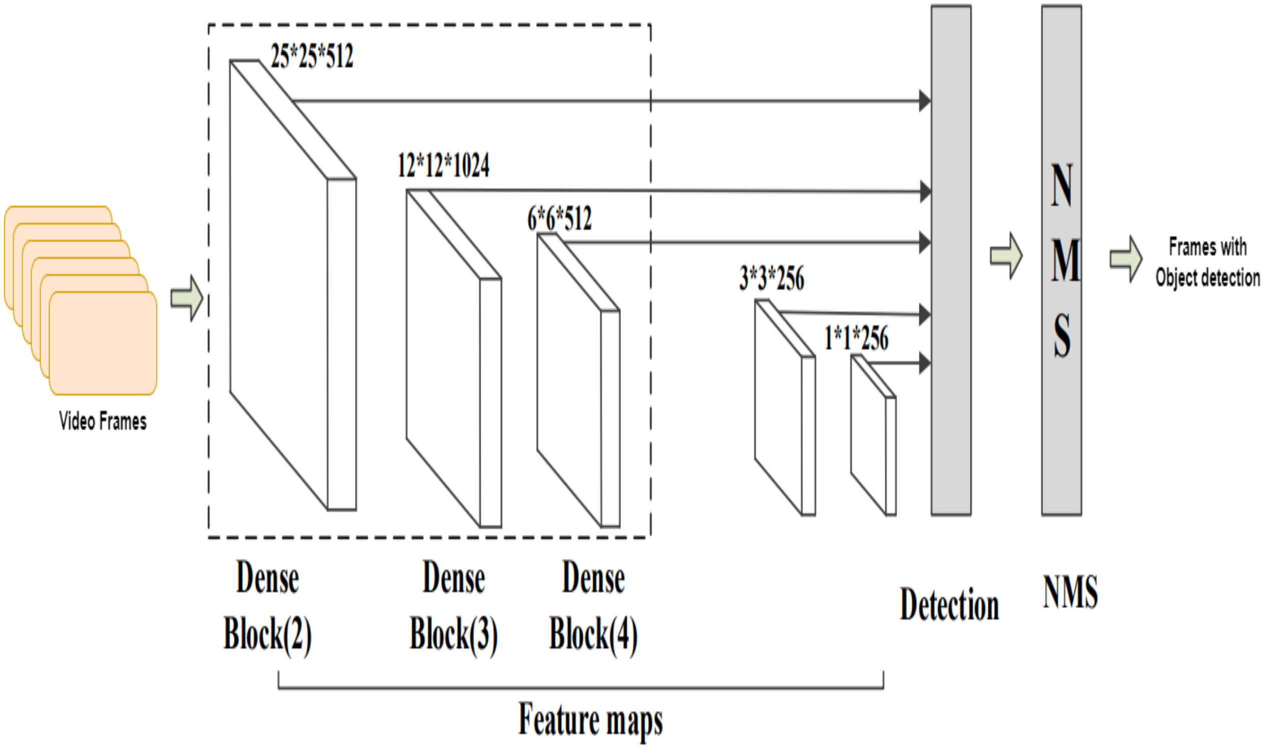


Figure 1: Proposed DenseNet – SSD

EXPERIMENTAL ANALYSIS

To begin, we conduct an analysis of how well our methodology performs on the KITTI detection dataset. We evaluate the efficiency of our proposed DenseNet-based Single Shot Detector in comparison to a baseline SSD that represents the state of the art. The KITTI autonomous driving dataset is comprised of variables that significantly from six hours of driving carried out under a wide range of environmental and traffic circumstances. The images were taken at a rate of 10 Hz and had a resolution of 1242 x 375. The dataset for 2D object detection contains 7,481 training photos and 7,518 testing images. A bounding box and a class label for one of three classes—cars, pedestrians, or cyclists—have been applied to each image in this collection. Unlike other frequently used video object recognition datasets, the vast majority of images in KITTI contain examples of numerous object types. Furthermore, KITTI is still a difficult dataset for object detection because to the large number of small items that are obscured, oversaturated, in shadow, or truncated. KITTI provides three additional unlabeled frames for each training image, allowing us to define the multi-frame video input, even if the dataset is optimised for single-frame object detection in its original form. In order to define train/val splits for our experiments on the KITTI detection dataset, we randomly divide the available training data in half. We do an evaluation of each model using the validation set, and we compare the results using a standard evaluation metric such as precision, recall and mAP. The training and validation accuracy of proposed model is shown in Figure 2.

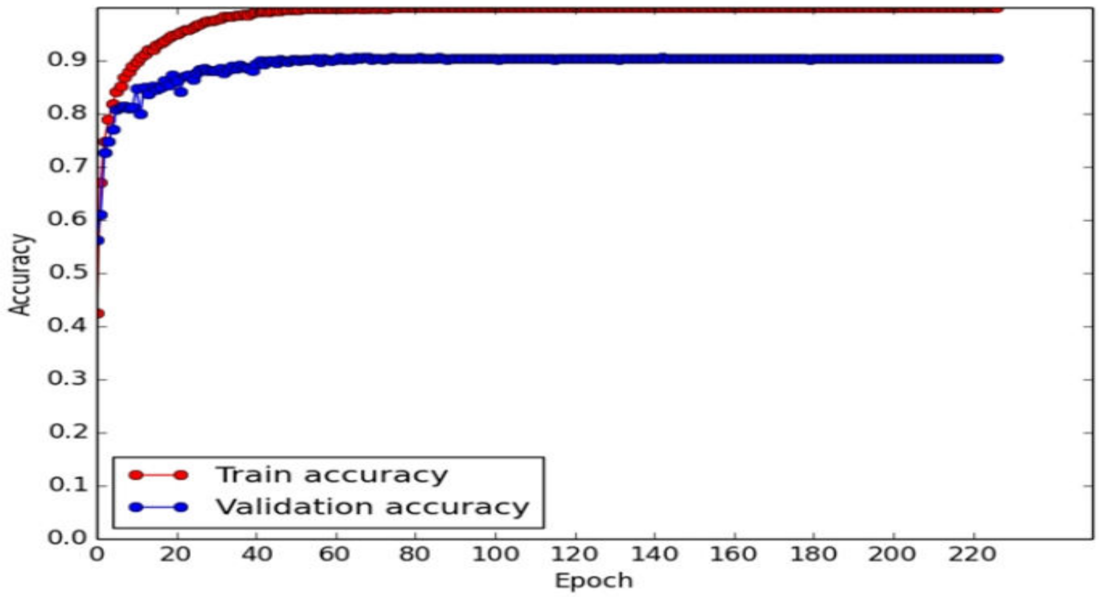
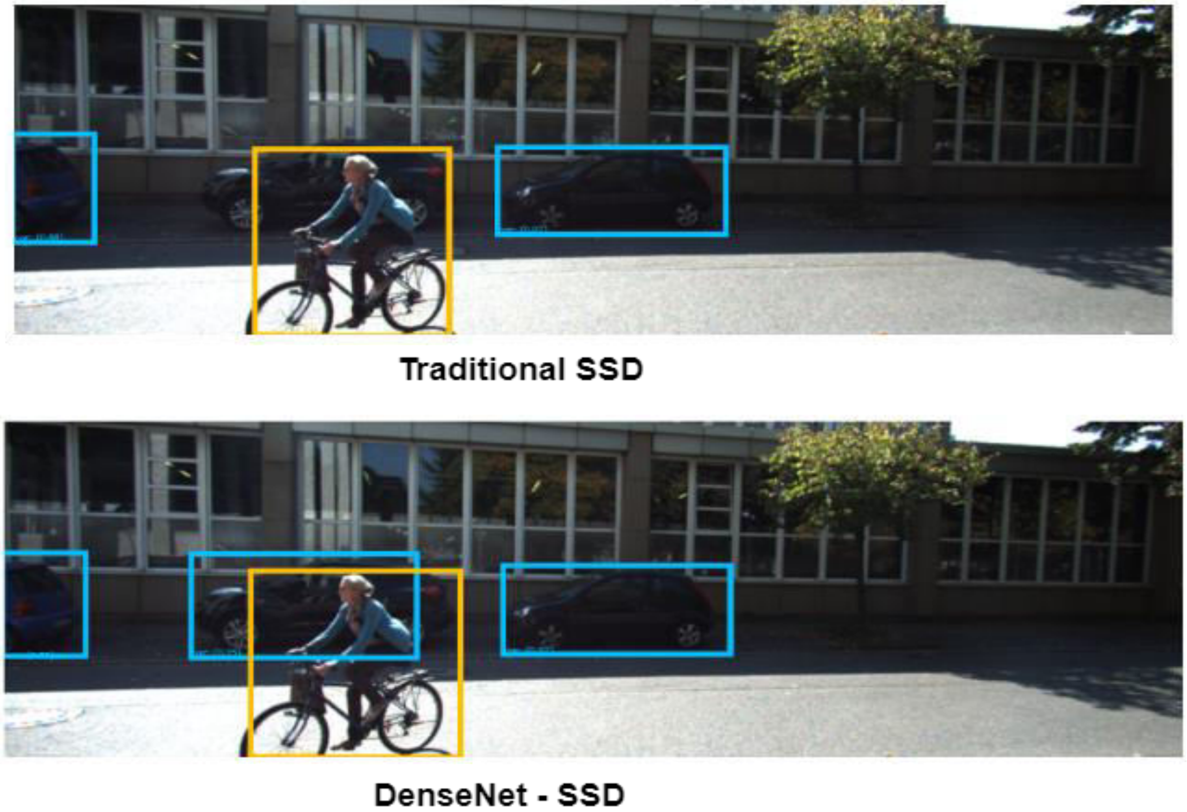


Figure 2: Accuracy of DenseNet – SSD



Traditional SSD

DenseNet - SSD

Figure 3: Sample Output

Model	Precision	Recall	mAP
SSD	90.5±0.2%	95%	0.9243
R-CNN	96.5±0.3%	96%	0.9645
YOLOV2	96.5±0.5%	93%	0.9345
DenseNet - SSD	98.5±0.1%	97%	0.9854

Table 1: Model Comparison

From Figure 3 and Table 1, it can be deduced that when compared to the traditional SSD methodology, the DenseNet-SSD method is able to achieve a good compromise between the recall and precision rates of various types of object detection. Therefore, the performance of the modified SSD algorithm in video object detection is superior to that of the old SSD method.

CONCLUSION

In this paper, we introduce the DenseNet Single Shot Detector. Notably, adopting DenseNet to take use of the additional spatio-temporal characteristics present in video data requires only a little adjustment to the standard Single Shot Detector design. A key consideration when dealing with video data is network speed, and the completely convolutional nature of the DenseNet-SSD architecture ensures that the detection network's runtime is extremely respectable. In addition, unlike some other methods, our method does not rely on computing optical flow, which can significantly slow down processing time. From the experimental analysis, it is evident that the proposed model obtained the better result.

REFERENCES

1. Bichen Wu, Forrest Iandola, Peter H. Jin, and Kurt Keutzer. SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving. In CVPR Workshops, 2017.
2. Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In ICCV. IEEE, 2017.
3. Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seqnms for video object detection. arXiv:1602.08465, 2016.
4. Subarna Tripathi, Zachary C Lipton, Serge Belongie, and Truong Nguyen. Context matters: Refining object detection in video with recurrent neural networks. arXiv:1607.04648, 2016.
5. Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In Computer Vision and Pattern Recognition, 2016.
6. Yongyi Lu, Cewu Lu, and Chi-Keung Tang. Online video object detection using association lstm. In ICCV, 2017.
7. Michael Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(9), 2010