ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0024

# Research on Target Detection Algorithm Based on Lightweight Mobilinet in Assistant Driving System

Xianglin Yan[1, a], Zhijiang Bai[1, b]

[1]Computer science and technology, Shanghai Maritime University, Shanghai, China

[a]yanxianglin@shmtu.edu.cn, [b]zjbai@shmtu.edu.cn

## Abstract

**Due to the large amount of network parameters and slow detection speed of SSD target detection network, it is not suitable for embedded platform to detect road emergencies in real time. On the basis of the original network, this paper uses mobile net network to replace SSD backbone network, and adds five layers of convolution neural network after it. All adopt deep separable convolution operation, and finally get a new network structure. Conv7, Conv10, Conv11, Conv12, Conv13 and Conv14 are used as effective feature layers to obtain the prediction results. In this paper, 10050 images are used to train the new network and detect pedestrians, vehicles and bicycles on the road. The results show that the parameters of the new network are 7.28 million, the detection speed is 45 FPS, and the average detection accuracy is 79.86%. Compared with the original network, the parameters of the new network are reduced by 70% and the detection speed is doubled. Comprehensive analysis can be applied to embedded platform for real-time detection of road emergencies.**

## Keywords

**Convolution neural network; Target detection; MobileNet; Deep learning.**

## 1. INTRODUCTION

In 2020, there will be more than 200000 road traffic accidents in China. The main cause of many accidents is the nonstandard driving of drivers. The emergence of the auxiliary driving system can remind the driver of the dangerous situation encountered in the process of driving. With the rapid development of artificial intelligence, target detection, as an important research direction of artificial intelligence, is also applied in many fields. In 2016, Liu W and others proposed SSD [4] algorithm based on Yolo [1] [2] [3].The algorithm has a good effect on the detection of small objects. But SSD algorithm has a large number of parameters, storage and speed are not suitable for lightweight embedded platform. Therefore, based on the original neural network, this paper optimizes the network by analyzing the calculation amount, real-time performance and model size of the neural network for target detection, so as to obtain a lightweight network, and design a lightweight target detection system for auxiliary driving system, Reduce the occurrence of traffic accidents. In this paper, vehicles, pedestrians, bicycles as the research object, remind the driver to respond in advance to reduce the occurrence of traffic accidents.

## 2. SSD TARGET DETECTION ALGORITHM

SSD algorithm is improved on VGG [5] network structure by replacing the original full connection layer fc6 and fc7 with convolution layer, and four groups of convolution layers are added for feature extraction, and then feature maps of different scales are fused to extract

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0024

features. Thus, the feature map of low level and high level has a large amount of target location information, which is conducive to target location, The high-level feature map has a lot of semantic information, which is conducive to target classification. Due to the feature extraction of SSD algorithm on multi-scale, it also has a good effect on small target detection. In the road, due to the distance between the camera and vehicles and pedestrians, there will be a large number of small targets, so SSD target detection algorithm can be applied to the auxiliary driving system.

Although SSD algorithm is better than the traditional target detection algorithm in accuracy and speed, SSD uses VGG as the backbone network, which shows that it has large parameters and complex calculation, and can not detect pictures in real time, so it is not suitable for portable embedded platform. Therefore, this paper will make a lightweight improvement on SSD target detection framework, The VGG backbone network with large parameters and complex calculation is replaced by the lightweight MobileNet [6], MobileNet structure is simple, and the use of depth separable convolution structure to achieve convolution operation, in the case of ensuring the accuracy of calculation, can greatly reduce the amount of calculation, improve the speed of SSD detection network.

## 3. LIGHTWEIGHT TARGET DETECTION ALGORITHM MODEL

### 3.1. MobileNet Model Structure

MobileNet adopts depth separable convolution, which is divided into point by point convolution and layer by layer convolution [6] Compared with the traditional standard convolution, the computational complexity is reduced. Suppose the input image size is $D_F \times D_F \times M$ The convolution kernel size is $D_K \times D_K \times M$, the number of channels is $N$ The standard convolution kernel is shown in Figure 1.

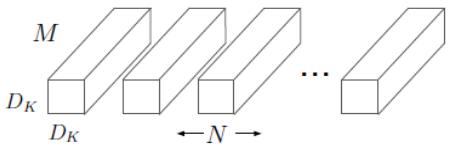



**Figure 1.** Standard convolution

Then the standard convolution computation is $D_F \times D_F \times D_K \times D_K \times M \times N$.

The convolution layer by layer is shown in Figure 2.

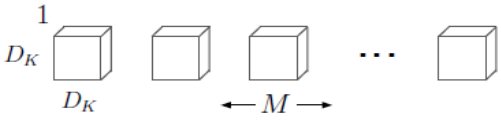



**Figure 2.** layer by layer convolution

The computation amount of convolution layer by layer is $D_F \times D_F \times D_K \times D_K \times M$

Point by point convolution is shown in Figure 3.

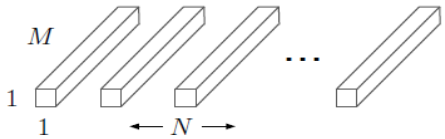



**Figure 3.** point by point convolution

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0024

The computational complexity of point by point convolution is as follows: $D_F \times D_F \times M \times N$ .

Then the total amount of depth separable convolution is as follows:

$$D_F \times D_F \times D_K \times D_K \times M + D_F \times D_F \times M \times N$$

The amount of computation reduced by depth separable convolution is as follows:

$$\frac{D_F \times D_F \times D_K \times D_K \times M + D_F \times D_F \times M \times N}{D_F \times D_F \times D_K \times D_K \times M \times N} = \frac{1}{N} + \frac{1}{D_K^2}$$

Therefore, the computational complexity of depth separable convolution is about the same as that of standard convolution $1/N$ Can achieve the purpose of reducing the network. In this paper, the depth separable convolution is applied to the later layers to further reduce the network parameters and calculation. The network structure is shown in Figure 4.

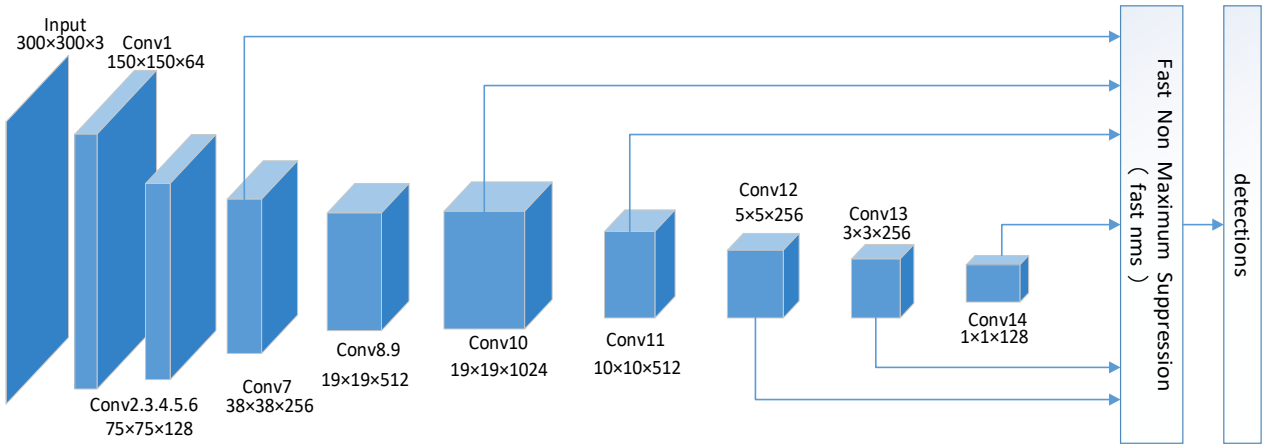



**Figure 4.** Improved network structure

The structure of convolution neural network designed in this paper is shown in Figure 4. The network is divided into 14 layers, the input picture is 300 × 300 × 3, the first layer uses 3 × 3 convolution kernel, the step size is 2, the padding adopts the same, then the output layer size is 150 × 150.In the second layer, 3 × 3 convolution kernel is used, step size is 2, and padding is the same. The third, fourth, fifth and sixth convolution kernels of 3 × 3 are used with step size of 1, and the same is used for padding. In the 14th layer, 3 × 3 convolution kernel is used, step size is 1 and padding is valid. All convolution layers in the network are activated by relu6 function. Among them, conv7, conv10, conv11, conv12, conv13 and conv14 convolution feature layers are used to predict the prediction results. The above six layers are convoluted twice, one for predicting the change of each prior frame, and the other for predicting the type of prediction frame on each grid point. The specific operation is shown in Figure 5 [7].

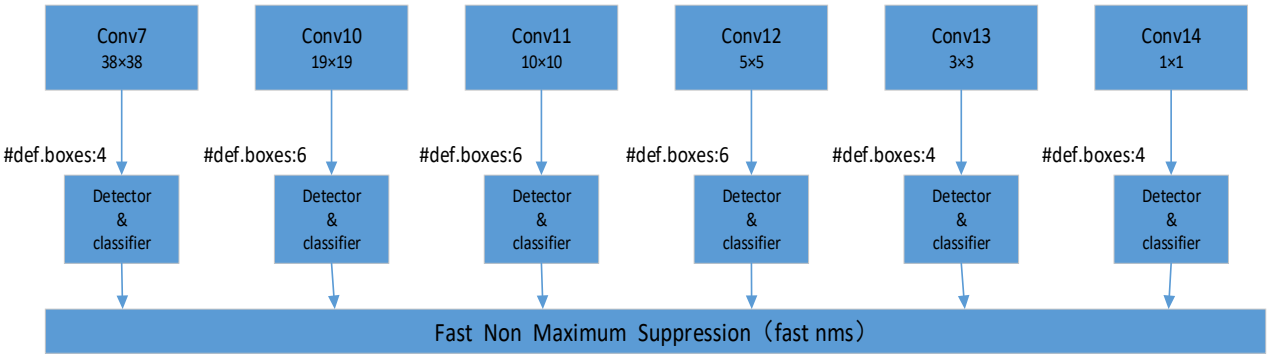



**Figure 5.** Feature extraction

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0024

### 3.2. Prior Box Selection

In this paper, the size of the effective characteristic layer is (38, 19, 10, 5, 3, 1), then the calculation formula of the size of prior frame is as follows: (1):

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1),\ k \in [1,m] \tag{1}$$

Among $s_k$ represents the size ratio of the input picture, $s_{\min}$ represents the minimum scale, which is set as 0.15 in this paper, $s_{\max}$ It is expressed as the maximum scale, which is set as 0.9 in this paper, $m$ In this paper, the ratio of the first effective feature layer is set to 0.07, and the other five feature layers participate in the operation. Then the ratio of the size of the basic prior frame of each feature layer to the input image is (0.07,0.15,0.33,0.51,0.69,0.87). If the size of the input image is 300 × 300, then the size of the basic prior frame of each feature layer is $b_i$ (21,45,99,153,207,261). The basic prior frame is regarded as the minimum square box, and then expanded from the minimum square box to other scale prior frame. In this paper, the maximum prior frame of the last feature layer is set as 315, and the ratio of length to width is set as $ar = \{\sqrt{2}, 1/\sqrt{2}, \sqrt{3}, 1/\sqrt{3}\}$.Then the specific calculation formula is as follows (2, 3):

Maximum square prior box:

$$b_i' = \sqrt{b_i \times b_{i+1}},\ i \in [1,6] \tag{2}$$

Rectangle prior box:

$$h_i = b_i \times ar,\ i \in [1,6] \tag{3}$$

The size of the prior frame corresponding to each grid point can be obtained by the above method.

### 3.3. Prediction and Classification

The effective feature layer extracted by Conv7 is divided into 38 × 38 grids, each grid contains 4 prior frames, a total of 5776 prior frames, but the 4 prior frames of each grid point can not determine any situation of the target, so it is necessary to adjust the prior frame. For the effective feature layer extracted by Conv7, two convolution operations are performed, one is used to predict the change of each prior frame, and the other is used to predict the type of prediction frame on each grid point. A convolution kernel of 3 × 3 × 4 is used to predict the change of prior frame, where 4 represents the offset of grid center (x, y) and the offset of grid height and width h, w, then the output shape is 38 × 38 × 16. The prediction box is obtained by adding the grid parameters with the offset after convolution. The type of prediction frame is convolution with convolution kernel of 3 × 3 × 6, where 6 represents the type of prediction. In this paper, bus, car, truck, person, bicycle and background are taken as prediction categories. The output shape is 38 × 38 × 24. Conv10, Conv11, Conv12, Conv13, Conv14 do the same operation, stack the adjusted grid and predicted types, and select the maximum IOU ratio by non maximum suppression method[8],At the same time, output its corresponding category, that is to complete the detection. The prediction results of all effective characteristic layers are shown in Table 1.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0024

**Table 1.** Forecast results

| Effective characteristic layer | Prior frame | Offset convolution shape | Type convolution shape |
|---|---|---|---|
| Conv7(38,38,512) | 4 | (38,38,16) | (38,38,24) |
| Conv10(19,19,1024) | 6 | (19,19,24) | (19,19,36) |
| Conv11(10,10,512) | 6 | (10,10,24) | (10,10,36) |
| Conv12(5,5,256) | 6 | (5,5,24) | (5,5,36) |
| Conv13(3,3,256) | 4 | (3,3,16) | (3,3,24) |
| Conv14(1,1,128) | 4 | (1,1,16) | (1,1,24) |

### 3.4. Loss Function Selection

In this paper, the loss function is divided into classified loss and prediction box position loss[10] ,Then the specific calculation formula is as follows (4, 5, 6):

$$L(x,c,l,g) = \frac{1}{N}\left(L_{conf}(x,c) + \alpha L_{loc}(x,l,g)\right) \tag{4}$$

$$L_{conf}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p}\log\left(\hat{c}_i^p\right) - \sum_{i\in Neg}\log\left(\hat{c}_i^0\right) \tag{5}$$

Among

$$\hat{c}_i^p = \frac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)}$$

$$L_{loc}(x,l,g) = \sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k} smooth_{L1}\left(l_i^m - \hat{g}_j^m\right) \tag{6}$$

$L(x,c,l,g)$ Total loss, $L_{conf}(x,c)$ Indicates the classification loss, $L_{loc}(x,l,g)$ Indicates the position loss of prediction box, $N$ Is the positive sample number of the prior frame, $c$ Is the predicted value of classification confidence; $l,g$ They are the position parameters of prediction box and real box respectively.

## 4. ANALYSIS OF EXPERIMENTAL RESULTS

### 4.1. Experimental Platform

The experimental environment is: Win10 system, NVIDIA GeForce GTX 1060 graphics card, i7-9750 processor. In this paper, Tensorflow 1.13 framework is used for training.

### 4.2. Road Data Collection and Model Training

In this paper, 1913 pictures were taken and downloaded from the network, and 8137 pictures were used in the voc07 + 12 data set, with a total of 10050 pictures. The labelimg tool was used to label the data as voc07 format, in which 10% of the pictures were used as the test set, and the rest were the training sets.

In the process of training, this paper uses two methods: rough training and fine training. The initial learning rate of rough training is 0.0005, and 16 pictures are input each time. If the loss does not decrease in two iterations, the learning rate is reduced to 1 / 2 of the original. If the loss does not decrease after six iterations, coarse training is stopped and fine training begins.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0024

The initial learning rate of fine training is 0.0001, and 8 pictures are transferred in each time, Same as rough training mode. The curve of loss value and learning rate in the process of improved network training is shown in Figure 6.It can be seen from the figure that the learning rate of the first 22 epochs in the rough training is 0.0005. When the loss value does not decrease, the learning rate will drop to half of the original value to continue training. The learning rate of 50-66 epochs is 0.0001, and then it is reduced to half of the original. Finally, the loss value of the model is reduced to 1.764, and the network stops training. During the training process, every time an epoch is carried out, the parameters with the lowest loss value are saved.

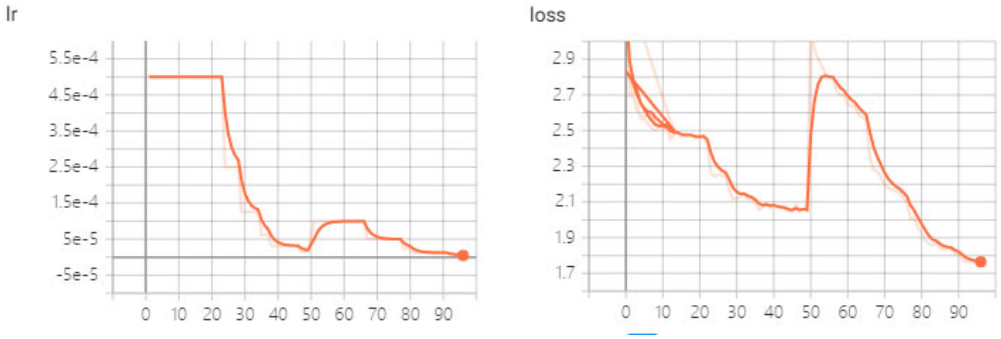



**Figure 6.** Loss vs. learning rate curve

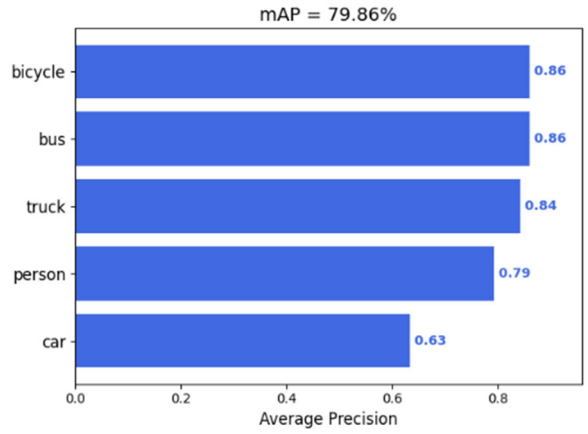
## 4.3. Evaluation Index

In this paper, the model size, FPS and map are used to evaluate the network performance. The size of the model is the size of the network parameters, FPS is the number of pictures that can be detected per second, map [9] Is the average detection accuracy, and its calculation formula is formula (7):

$$mAP = \frac{1}{|Q_R|}\sum_{q=1}^{Q_R} AP \tag{7}$$

## 4.4. Result Analysis

In this paper, three groups of experiments are compared. The parameters of SSD network model and the improved MobileNet model are compared, the detection accuracy mAP, and the number of frames per second of the detection picture FPS are compared. As shown in Figures 7 and 8.



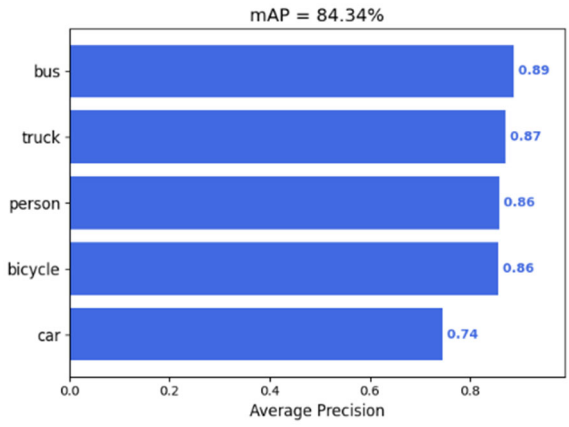




**Figure 7.** Improved mobile net network map

**Figure 8.** SSD network map

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0024

**Table 2.** Experimental comparison results

| network model | Parameter quantity | FPS (FPS) | mAP |
|---|---|---|---|
| SSD | 26,285,486 | 21 | 84.34% |
| Improved MobileNet | 7,826,990 | 45 | 79.86% |

It can be seen from the experimental comparison in Table 2 above that the parameter quantity of the original model is reduced by 70%, and the detection speed is nearly doubled, but the detection accuracy is slightly reduced. The specific test results are shown in Figure9 and Figure10.


**Figure 9.** Improved MobileNet prediction results


**Figure 10.** SSD prediction results

## 5. CONCLUSION

Through the research of SSD target detection algorithm, it is found that its parameters are large and the detection speed is slow, so it is not suitable for the light-weight vehicle terminal. Therefore, the depth separable convolution operation is used to replace the original standard convolution, which can greatly reduce the parameters while losing a little detection accuracy, and the FPS is nearly doubled. It can achieve the purpose of real-time, but the accuracy of the lightweight network in detecting small objects is low. Therefore, how to improve the detection accuracy while compressing the parameters needs further research.

## REFERENCES

[1] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection, Computer Vision & Pattern Recognition. IEEE, (2016),

[2] [Redmon J , Farhadi A . YOLO9000: Better, Faster, Stronger,IEEE Conference on Computer Vision & Pattern Recognition. IEEE, (2017),6517-6525.

[3] Redmon J, Farhadi A . YOLOv3: An Incremental Improvement. arXiv e-prints, (2018).

[4] Liu W , Anguelov D , Erhan D , et al. SSD: Single Shot MultiBox Detector[J]. Springer, Cham, (2016).

[5] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science, (2014).

[6] Howard A G , Zhu M , Chen B , et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (2017).

[7] Dong Yongchang, Shan Yugang, Yuan Jie. Pedestrian detection method based on improved SSD algorithm . Computer Engineering and Design, (2020) 2921-2926

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0024

[8] Jiang B , Luo R , Mao J , et al. Acquisition of Localization Confidence for Accurate Object Detection. (2018).

[9] Li Xiaowei. Research on lightweight deep learning target detection algorithm and system design. Anhui University, 2019.

[10] Lin T Y , Goyal P , Girshick R , et al. Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis & Machine Intelligence, (2017), 2999-3007.