Application of SSD_MobileNet Object Detection Algorithm in Autonomous Driving Technology

Cui Dongyan [1,2]

1 School of Artificial Intelligence, North China University of Science and Technology ,Tangshan, Hebei, China

2 Xingtian (Suzhou) Intelligent Control Technology Co., Ltd.Suzhou, Jiangsu, China

E-mail: *cdy_xxz@163.com*

Abstract

In autonomous driving environments, targets are extremely complex and there are many types of occluded or small targets. Therefore, enhancing accuracy and low latency in target detection of autonomous vehicles has become crucial to ensure the safety and efficiency of road drivers and pedestrians.

This article is based on the deep separable neural network model MobileNet as the SSD neural network skeleton network for feature extraction, which effectively solves this problem. It aims to utilize deep learning method to accelerate the target detection speed of autonomous vehicles, alleviate traffic congestion, and enhance traffic safety. The experimental results show that this model greatly reduces the number of parameters and computation compared to traditional object detection algorithms, improves the speed of object detection, and has certain practical application value.

Keywords: Autonomous Driving; Deep Learning; Target Detection; Convolutional Neural Network.

1. Introduction

In recent years, the rapid development of computer technology has driven the progress of artificial intelligence technology, and autonomous driving technology has begun to accelerate its innovation. At present, many well-known foreign car manufacturers are preparing to design and develop autonomous driving technology. Daimler will launch L3 level autonomous driving technology in 2021[1]; Tesla is an indispensable member of the driverless technology. Its Model S and ModelX models are equipped with a semi auto drive system. The owner can automatically change lanes by releasing the steering wheel [2]. German car manufacturers and research institutions are also actively researching autonomous vehicle technology. For example, the CARLA project at the Technical University of Munich and the autonomous vehicle project at Mercedes Benz are worth paying attention to. In 2020, China FAW Hongqi Automobile Company installed autonomous driving parking technology in its production vehicle category.

Object detection technology has become one of the key issues that urgently need to be addressed in autonomous driving technology, and many scholars are dedicated to the research of object detection technology. In 2005, Dalal et al. Proposed a combination of edge detection (HOG) that refers to the changes of local tissues in two gradient directions, and used information related to input and external features to detect changes in each local tissue in the image in both gradient directions. The specific method proposed can effectively handle changes in screen brightness scale and detect relatively stable [3]. Lowe proposed an External Invariant Scaling (SIFT) feature that uses Gaussian difference to register the maximum point as the critical point when displaying images of different sizes and input scales, and continues to determine the boundaries around the point and several external features of its region. SIFT has a strong ability to handle scaling and express language information related to texture effects [4]. In 2007, Felzenszwalb and other related personnel proposed the Folding Deformation Body Three Dimensional Diagram (DPM) [5].

Traditional target measurement methods are difficult to adapt to modern applications due to low accuracy and high latency. Deep learning based object detection is still constantly improving and has made significant breakthroughs compared to traditional methods. It has broad prospects for future development and is expected to replace humans in completing more accurate and efficient tasks. Regression based object detection techniques mainly include YOLO [6] and SSD [7-9]. Redmon and other researchers proposed YOLOv2, which combines Darknet-19 LAN connectivity and improves the transport layer on the original basis [10]. Redmon et al. proposed YOLOv3 in 2018, which can successfully further improve the quality and performance of software in detecting small moving objects [11-12].

This article aims to greatly improve the speed and accuracy of detection and enhance the real-time performance of object detection based on the application of deep learning in autonomous vehicles, in order to reduce the incidence of traffic accidents and congestion.

2. Simulation of Object Detection in Autonomous Driving Technology Based on SSD_MobileNet

Compared to traditional object detection algorithms, the initial SSD network has improved accuracy, but the detection speed and precision in autonomous driving technology are still difficult to meet the requirements. Therefore, in this article, we use MobileNet network instead of VGG in SSD network as the skeleton, and merge the two to construct SSD_MobileNet object detection algorithm, in order to meet the specific needs of object detection in autonomous vehicle scenarios.

2.1. SSD Network Structure

The SSD network can detect images of two sizes, 300 × 300 and 500 × 500, respectively. The network structure is shown in Figure 1.

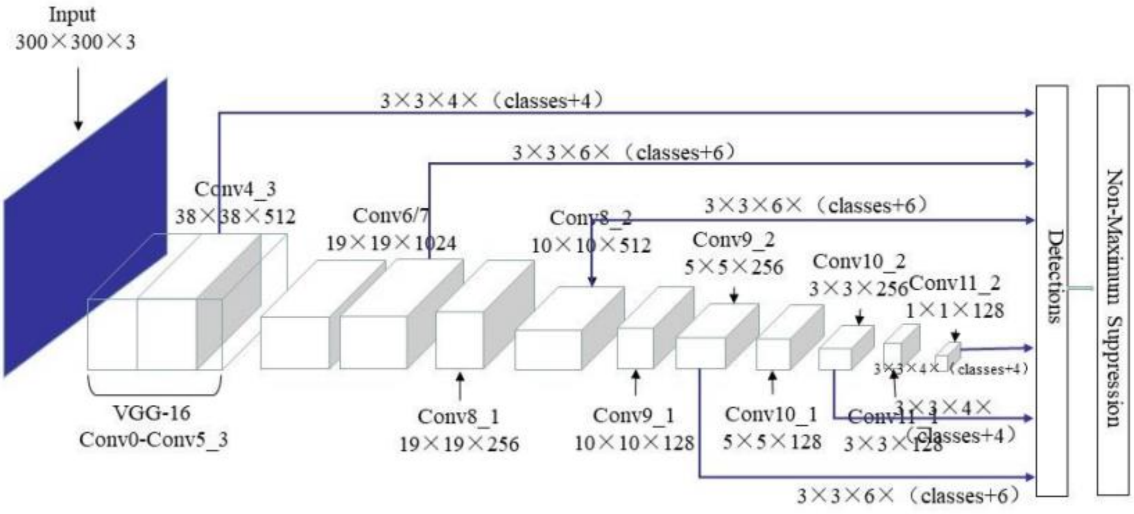

Fig.1. SSD Network Architecture

In the SSD network, feature extraction is performed using eleven different sized feature maps, which are then fed into the classification and localization network through a 3x3 convolution operation to ultimately form prediction box parameters. The specific parameters are shown in Table 1.

Table 1. SSD network parameters

Layers	Convolution operation	Input (w,h,c)	Output (w，h，c)
Conv6	1024，3×3，1，6	19×19×512	19×19×1024
Conv7	1024，1×1，1	19×19×1024	19×19×1024
Conv8_1	256，1×1，1	19×19×1024	19×19×256
Conv8_2	512，3×3，，2，1	19×19×256	10×10×512
Conv10_1	128，1×1，1	5×5×256	5×5×128
Conv10_2	256，3×3，1	5×5×128	3×3×256
Conv11_1	128，1×1，1	3×3×256	3×3×128
Conv11_2	256，3×3，1	3×3×128	1×1×256

The SSD network combines six sets of feature maps into a pyramid structure to detect targets of different sizes through information transmission.

2.2. MobileNet Network Architecture

The convolutional structure of depthwise separable convolution consists of two parts: depthwise convolution and 1x1 convolution. By utilizing the separate convolution of two blocks, a significant amount of computation can be saved. In this network, first perform a 3x3 convolution, then add ReLU and BN layers, followed by a 1x1 convolution, and then add ReLU and BN layers again. The combination of deep convolution and 1x1 convolution significantly reduces the computational complexity of the network.

In depthwise separable convolution, 1x1 convolution accounts for 90% of the computation and 70% of the parameters. The calculations and parameters of different units vary in Table 2.

Table 2. The parameter and computational complexity of different structures

Type	Parameter quantity	Computation
Conv1×1	74.59%	94.86%
Conv DW 3×3	1.06%	3.06%
Conv3×3	0.02%	1.19%
Fully Connected	24.33%	0.18%

2.3. SSD_SobileNet Network Structure

This article improves the SSD network by replacing the original VGGNet with a MobileNet network as the backbone network, and integrating it with SSD to form the SSD_MobileNet network. The specific structure is shown in Figure 2.

The Softmax layer and fully connected layer in the SSD network have been removed from the new model network, and four convolutional layers have been added for better feature extraction. The specific parameters are shown in Table 3

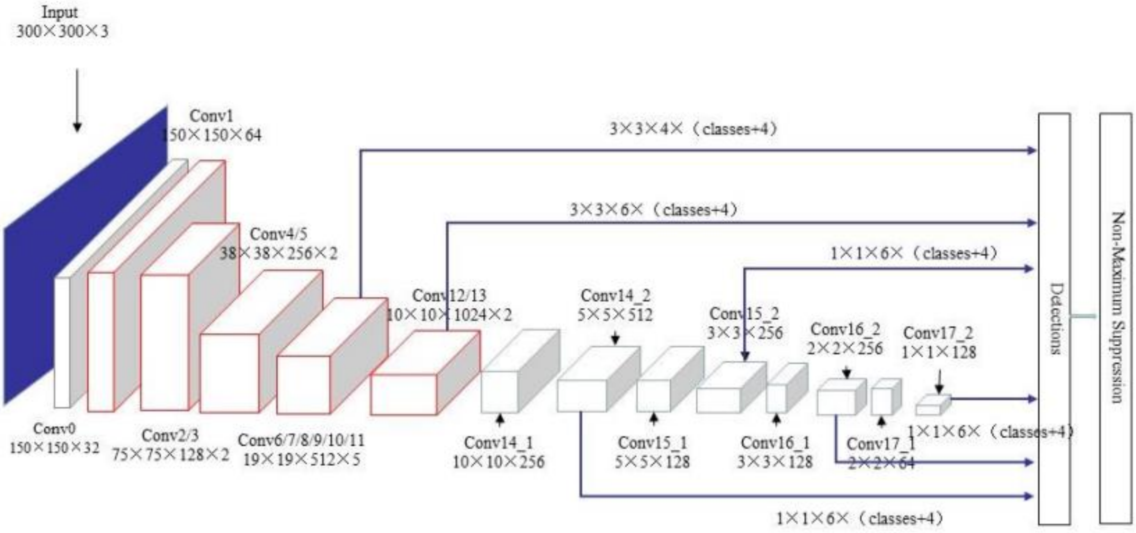

Fig.2. SSD_Mobile Network Architecture

Table 3. Anchor boxes for each prediction layer

Prediction layer	Prediction layer size	Number of anchor boxes	Total number of anchor boxes
Conv11	19×19	3	1083
Conv13	10×10	6	600
Conv14_2	5×5	6	150
Conv15_2	3×3	6	54
Conv16_2	2×2	6	24
Conv17_2	1×1	6	6

3. Experimental simulation and result analysis

3.1. Environmental Configuration

The experimental environment is as follows: The hardware environment consists of an Inter (R) CORE (TM) i7-8750H processor and an NVIDIA GTX 1050 graphics card. The software is Windows 11 operating system and Pytorch framework.

3.2. Training Process

All deep learning models used in this study were VOC2007 and VOC2012 deep learning models. The VOC2007 database dataset contains 9963 images, with a total of 24640 moving objects labeled. VOC2012 deep learning model image annotation 11530 times. The VOC deep learning model includes specific classifications of 20 moving objects, including: airplanes, bicycles, birds, bottles, buses, cars, and other specific classifications.

To achieve such an effect, it is necessary to evaluate the overall standardization of SSDnMobile, strengthen the VOC deep learning model through training, select 15k images from the database dataset for reinforcement training, and test 2k images.

The parameters set for the new model during the training phase are shown in Table 4.

Table 4. Training parameters

Parameter	Value
Batch	16
Decay	0.005
Learning_rate	0.001
Max_batches	20000

During the experiment, 8 images were used for each training session. The learning response rate of the improved SSD_MobileNet model was 0.001, and the iteration time interval for the minimum product was 80000 times. Referring to the loss value, continuously improve the parameters until the obtained loss value gradually disappears and tends to 0, or reaches the minimum iteration number for reinforcement training. Stop training and the training loss curve is shown in Figure 3.

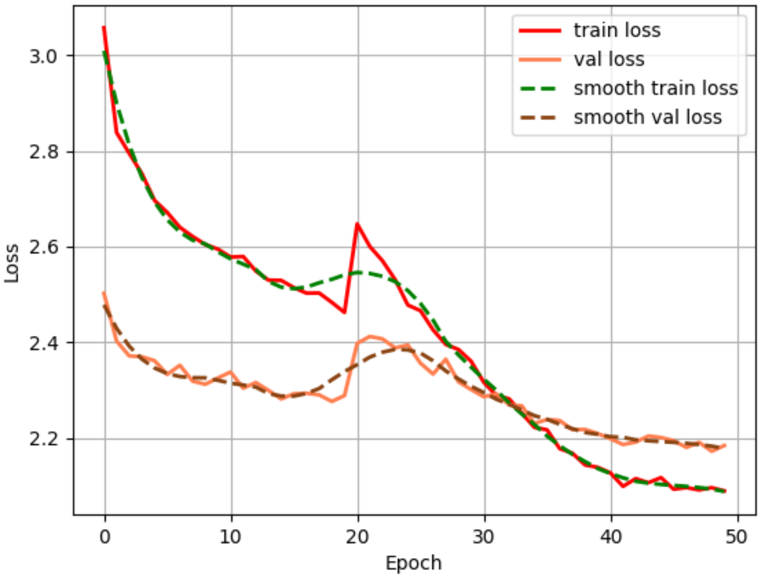

Fig.3. Training loss curve graph

It can be seen that the loss value in the early stage of training did not decrease significantly, but there were some small spikes in the middle. However, after a certain stage of learning, the curve loss gradually returned to normal, and the changes were not as obvious as at the beginning, indicating that the learning rate was relatively appropriate.

3.3. Experimental Results and Analysis

The results of testing 15 categories on the test set using the improved SSDnMobileNet network are shown in Table 5.

SSD_MobileNet has higher accuracy than SSD in many categories. The performance comparison of commonly used algorithms is shown in Table 6. The mAP of the improved SSD-MobileNet is 2.3% higher than that of SSD. It can also be seen that the proposed modified SSD-MobileNet has a significant improvement in detection accuracy compared to other algorithms.

Table 5. Comparison of target detection accuracy

Serial Number	Category	SSD(%)	SSD_Mobilenet(%)
1	Areo	73.2	80.2
2	Bicycle	79.1	82.2
3	Bird	78.3	82.1
4	Boat	82.3	83.4
5	Bottle	82.1	84.2
6	Bus	81.1	83.6
7	Car	80.5	81.2
8	Cat	76.3	77.3
9	Chair	77.8	79.8
10	Cow	75.4	79.6
11	Table	78.2	81.2
12	Dog	82.5	85.8
13	Horse	80.1	79.7
14	Motorbike	76.4	80.2
15	Person	81.2	82.8

Table 6. Comparison of Algorithm Detection Performance

Network	Backbone network	mAP（%）
YOLO	CNN	70.3
R-CNN	CNN	76.6
YOLOv2	Darknet-19	77.4
SSD	VGG-16	79.2
SSD_MobileNet	MobileNet	81.5

Table 7 shows the comparison of detection speed between SSD and SSD_maobile Net, measured in FPS.

Table 7. The Comparison of Detection Speed（GTX 1050）

Algorithmic network	FPS
SSD	65
SSD_MobileNet	78

From Table 7, it can be seen that SSDnMobileNet is much faster than SSD in terms of detection speed. Figure 4 shows the detection performance of the improved SSDnMobileNet network.

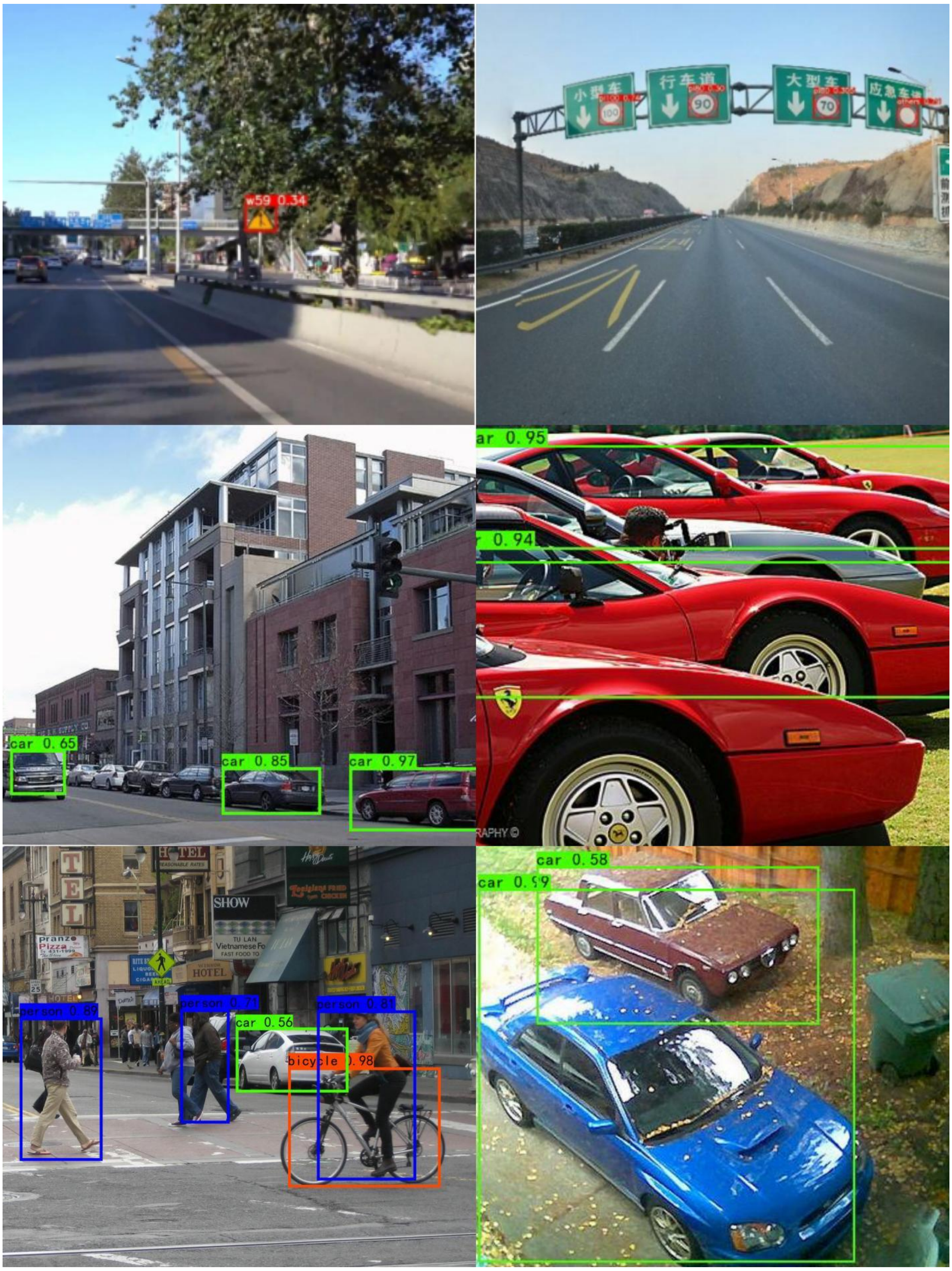

Fig.4. Actual measurement effect diagram

In Figure 4, the improved algorithm detects vehicles, pedestrians, and traffic safety signs in the image. It can be seen that the correct box selection and prediction are made for small targets located in the middle of the image. For nearby overlapping targets, it can also be correctly recognized, but the detection effect is lacking in overly distant and complex environments. Many small overlapping targets are not detected because they are located in the corners of the detection box.

4. Conclusion

Since the advent of autonomous vehicle, target detection technology has been used to identify targets quickly and accurately, so that vehicles can cooperate with each other and avoid road traffic congestion and key problems of road traffic safety quickly and effectively. This article replaces VGG-16 with MobileNet as the new network structure for SSD, and constructs the SSD-MobileNet network. This structure reduces the number of parameters and computation, which not only improves detection speed but also enhances detection accuracy. Finally, data comparison was obtained through Matlab simulation, and the conclusions were validated using the VOC dataset.

The algorithm in this study has demonstrated certain advantages in object detection, but there are some flaws. In the experiment, some small-sized or occluded targets were not accurately detected, and the detection classification was still not diverse enough. We hope to optimize algorithms in the future to overcome these problems.

References

[1] YAN Lixin, QIN Lingqiao, XIONG Yubing, et al. A Safe Evaluation on Intelligent Vehicles with Multi-mode Cooperative Driving. Journal of Transportation Information and Safety, 2018, 36(003):1-7,26.

[2] Liu Wen. Analysis of Intelligent Vehicle Assisted Driving Technology. Automobile Applied Technology,2021, 46(02):35-37.

[3] WANG Wenguang, LI Qiang, LIN Maosong. Efficient target detection method based on improved SSD. Computer engineering and applications, 2019, 55(13): 28-35.

[4] KIM H, PARK J, KIM H , et al. Robust facial landmark extraction scheme using multiple convolutional neural networks. Multimedia Tools and Applications, 2019, 78(3):3221-3238.

[5] MENG Zhe, SUN Xiaoyu, ZHAO Bin. Railway signboard recognition method based on convolutional neural network. Acta automatica Sinica, 2020, 46(03):518-530.

[6] REDMON J, DIVVALA S K, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection. Computer Vision and Pattern Recognition, 2016: 2(08)779-788.

[7] XIANG W, ZHANG D, ATHITSOS V, et al. Context-Aware Single-Shot Detector. Computer Vision and Pattern Recognition, 2017, 32(4):325-329.

[8] HUANG G, LIU Z, MAATEN L, et al. Densely Connected Convolutional Networks. CVPR. IEEE Computer Society, 2017 6(9):463-467.

[9] ARGEN Z, LIU Yun. YOLO9000: Better, faster, stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2019.

[10] HOWARD A, ZHU M, CHEN B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Computer Vision and Pattern Recognition, 2017, 37(4):1-4.

[11] KIM H, PARK J, KIM H , et al. Robust facial landmark extraction scheme using multiple convolutional neural networks. Multimedia Tools and Applications, 2019, 78(3):3221-3238.

[12] LI J, LIANG X, SHEN S, et al. Scale-Aware Fast R-CNN for Pedestrian Detection. IEEE Transactions on Multimedia, 2017, 12(9):44-50.