Single-stage Multi-scale Receptive Field Improvement Lightweight Object Detection Network Based on MobileNetV3

Zhenkai Tong[1], Yefu Wu[1], Yang Liu[2]
1. School of Computer and Artificial Intelligence ,Hubei Key Laboratory of Transportation Internet of Things, Wuhan University of Technology,Wuhan,China
2.Chongqing Guoyuan Port Co.Ltd, Chongqing, China
e-mails: zenkoton@outlook.com,wuyefu@whut.edu.cn, 274532014@qq.com

***Abstract*—The computing power of edge devices is difficult to keep up with the development of modern computer technology, and the computing power is not improved enough. In practical application environments, so there is the birth of lightweight models. Lightweight models specifically refer to some models with simple model architecture and low computational load. Although the lightweight model is fast and the model is relatively simple, the detection effect is not very good. This paper proposes a parallel convolution module, performs feature fusion through parallel processing of multi-scale convolution kernels, and then integrates the spatial channel attention mechanism into the module to implement a multi-scale target detection module on a single feature layer. Model fusion proposes anchor boxes to generate heads and become a single-stage object detection model.**

Key Words:target detection, small target detection, feature fusion, lightweight network, attention mechanism

I. INTRODUCTION

Lightweight networks trade off speed and accuracy. The SqueezeNet [1] network extensively uses 1x1 convolutional modules for compression and expansion, in order to reduce the number of parameters of the network. ShuffleNet [2] utilizes group convolution and channel shuffling operations to further reduce the amount of computation. ShiftNet [3] , it is proposed to replace the resource-intensive spatial convolution with gradually convolutional interleaving operations. MobileNetV1[4], the depth wise separable convolution is used to reduce the amount of parameters and improve the network efficiency. A resource-efficient module with inverse residuals and linear bottlenecks is introduced in MobileNetV2 [5], which improves the prediction speed of the network. In MobileNetV3[6], the hard swish function is used to replace the relu function to improve the network detection performance.

II. IMPROVED MODLE

A. Improved MobileNetV3 network

The overall structure of the detection network is shown in Figure 1.

The Fig.1 shows the improved MobileNetV3 network. In the network model, the convolutional layer can be divided into three different stages, namely shallow network, middle network, and deep network

The multi-scale separable convolution module borrows the idea of Inception [7] in structure. Through the parallel operation of different scale convolution kernels, the output feature map has multiple sizes of receptive field information. Improve the performance of the network on small object detection. Shallow network often contains

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00074

more original information. Therefore, for the information in the shallow network, convolution and fusion prediction help to improve the detection performance of the network for small targets.

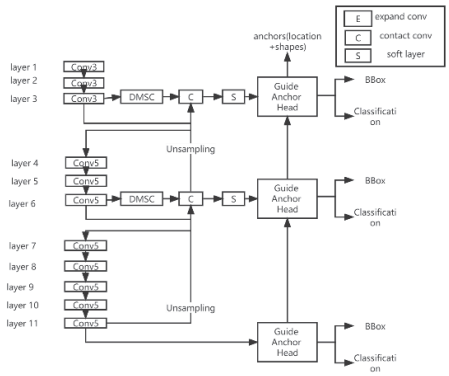

Fig. 1 The overall structure of the detection network

DMSC is multi-scale convolution module.In the multi-scale convolution module, the original feature map is firstly passed through three convolution kernels with different scales and sizes to adapt to small target information of different sizes. Each output feature map is connected to the CBAM attention module for processing. Convolution increases the weight of key information, and fuses and predicts each feature map.DMSC module struct is shown on Fig2

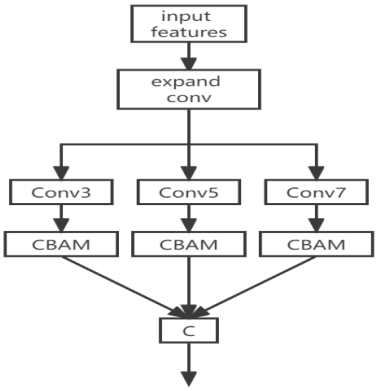

Fig. 2 Fusion attention mechanism multi-scale feature fusion module

Experiments show that adding receptive fields of different scales in the shallow network can greatly improve the prediction performance of the network for small targets.

B. Propose anchor box to generate head

The proposed anchor box generation head is different from the traditional two-part detection network, which is a hybrid detection head. The candidate anchor boxes are dynamically generated by predicting the position of the anchor point and the size of the anchor box by inputting the feature map. In the model using the feature pyramid, it is recommended that the anchor box generate the head, and the feature maps at different stages are predicted to improve the detection performance of the anchor box proposed to generate the head for small targets.

The structure diagram of the proposed anchor box to generate the head is shown in Figure 3

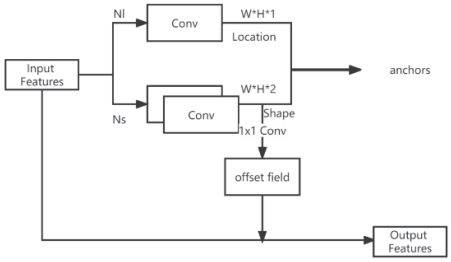

Fig. 3 Schematic diagram of the structure of the proposed anchor box to generate the head

C. Loss function

The loss of the model includes four main components: anchor point prediction loss, anchor box width and height prediction loss, classification loss, and regression loss. The total loss Loss is expressed as

$$Loss = \lambda_1 Loss_{locations} + \lambda_2 Loss_{shapes} + \lambda_3 Loss_{bbox} + \lambda_4 Loss_{classfication} \quad (1)$$

We hope to give a larger weight to the central area of the annotation, a smaller weight to the edge of the annotation, and a worse weight to

the part that is not within the annotation, so as to shrink the predicted position of the anchor point.

The loss function for when the anchor point is within the anchor box is defined as

$$Loss_{location} = \left(2 - 2Predication + \frac{(2x - x_1 - x_2)^2}{(x_2 - x_1)^2} + \frac{(2y - y1 - y_2)^2}{(y_2 - y_1)^2}\right)^2 \quad (2)$$

x_1, x_2, y_1, y_2 represent the horizontal and vertical coordinates of the four points of the real frame, respectively, and x, y represents the anchor point coordinates

The localization loss function for the non-anchor box

$$Losss_{location} = Predication * 2 \quad (3)$$

The prediction loss function Loss_shape of the anchor box in the image is not the same as the ordinary anchor box regression loss function. The predicted value of the anchor box height is not only determined by the information given by the feature map, but the prediction of width and height is limited by the anchor. The predicted value of the point, the width and height correspond to the prediction of the anchor point value one-to-one, so the loss function in the shape prediction is

$$Loss_{shape} = L_1\left(1 - \min\left(\frac{w}{w_g}, \frac{w_g}{w}\right)\right) + L_1\left(1 - \min\left(\frac{h}{h_g}, \frac{h_g}{h}\right)\right) \quad (4)$$

III. Experimental Result

The coincidence degree IOU between the predicted anchor frame and the real annotation is the limiting condition for determining whether the detection frame is the true value. This experiment refers to the detection index of the coco data set, and adds the limiting condition IOU as the limiting condition on AP and AR, and the IOU is 0.50: 0.95, starting from 0.50 and taking 0.5 steps to 0, 95 are taken as threshold calculation, the corresponding precision and recall rate under each threshold, and these values are averaged. Each model is trained for 30 epochs.

TABLE 1. ACCURACY AND REGRESSION ON ALL SCALES

Model Name	Average target recognition accuracy AP	Target average recall AR
MobilenetV2+ yolov3head	0.358	0.351
mobilenetV3+ yolov3head	0.383	0.415
Ours method	0.433	0.476

TABLE 2.ACCURACY AND REGRESSION ON SMALL SCALE

Model Name	Small target recognition accuracy AP	Small target recall AR
MobilenetV2+ yolov3head	0.060	0.137
mobilenetV3+ yolov3head	0.120	0.159
Ours method	0.148	0.213

The experimental results show that the model proposed in this paper can effectively improve the precision and recall rate of the network in small target detection and overall detection compared with the original mobilenetv3 network. Its precision and recall are significantly better than the other two models.

In order to show the regression of each model to the boundary, the larger the IOU threshold, the closer the regression margin predicted by the model is to the actual labeling margin. is shown in Figure 4

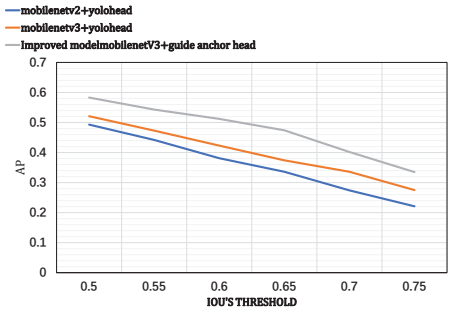

Fig. 4 Changes of AP with different IOU thresholds

Figure 4 shows that the model in this paper has a better regression effect on the boundary value, and the boundary value predicted by the model in this paper is closer to the actual standard value.

IV. CONCLUSION

In this paper, by adding a multi-scale convolution module to the mobilenetv3 model and integrating the spatial channel attention mechanism, the performance of the model in small target detection and boundary regression has been effectively improved. It is suggested that the addition of the head of the generated anchor frame is effective, reducing the problem of manual intervention of candidate frames and too many dense candidate frames in the image detection model. By regressing the image hotspot area, the detection performance of the model is effectively improved. By appropriately increasing the complexity of the MobilnetV3 model, the accuracy is improved without increasing too many model parameters, and the improved MobileNetv3 model can be better integrated into practical applications.

ACKONWLEDGEMENT

This work is support by National Natural Science Foundation of China under Grant U1764262.

REFERENCES

[1] Iandola, F.N., et al., SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016.

[2] Zhang, X., et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[3] Yan, Z., et al. Shift-net: Image inpainting via deep feature rearrangement. in Proceedings of the European conference on computer vision (ECCV). 2018.

[4] Howard A.G., et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[5] Sandler, M., et al. Mobilenetv2: Inverted residuals and linear bottlenecks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] Howard A , Sandler M , Chu G , et al. Searching for MobileNetV3[J]. 2019.

[7] Szegedy C , Ioffe S , Vanhoucke V , et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[C]// 2016.