YOLOv5s-MC: lightweight road target detection network

Yaohua Li[1], *Yuan* Yang[1,*], *Lei* Huang[1], *Chengyu* Yang[1], *Guoliang* Zhang[1], and *Qianye* Yang[2]

[1]College of Autimation and Information Engineering, Xi'an University of Technology Xi'an, China, 710048
[2]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China

Abstract. For the problem of large number of target detection algorithm parameters, a lightweight real-time detection algorithm YOLOv5s-MC based on improved YOLOv5s road scenes is proposed. firstly, CA attention is added to the model to improve the sensitivity of the network to detect targets; secondly, in the feature fusion network, add adaptive weight parameters using AS-Concat structure are added to better fuse the feature information of different layers and improve the detection accuracy of the algorithm ; adding a small target detection layer to improve the detection accuracy of tiny targets; finally introducing Mobilnetv2, a lightweight network, as the overall backbone layer to realize the lightweight requirement of the network; to verify the advantages of the proposed algorithm, experiments were conducted on the kitti dataset. The experimental results show that the proposed algorithm, compared with the original network, improves the average accuracy by 0.2% with 55.8% less parameters and 33.7% less computation, and the detection speed reaches 35 FPS, which meets the requirements of real-time detection and improves the ability of algorithm deployment in weak hardware computing power scenarios to a certain extent.

1 Introduction

Autonomous driving is an important direction of vehicle intelligence, and obtaining useful information from the external traffic environment is the basic link to realize autonomous driving[1]. Target detection is a key task in computer vision, and its main purpose is to detect the established target category and the location of the target that appears in the image. There are many application scenarios such as autonomous driving, target tracking, image retrieval, etc.[2] Target detection algorithm through the efforts of many scholars, detection accuracy and speed is gradually improving, and has a wide range of practical applications for the development of the economy and society[3]. Such structures then lead to models becoming larger and larger, and in some real-world applications, such as augmented reality, robotics, and autonomous driving scenarios, target detection tasks usually require real-time operations on platforms with limited computational power .As a result, the requirement for lightweight algorithms is gradually increasing.

* Corresponding author: yangyuan@xaut.edu.cn

In previous work, Xin[4] et al. introduced Swin-Transformer and CBAM attention mechanism into the feature fusion network and backbone of YOLOv5s, respectively, to improve the overall recognition accuracy of the model for infrared traffic more targets. Guo[5] et al. proposed a new AugFPN structure, which reduces the semantic information between different scale features in front of the feature fusion network and enhances the fusion effect. Raja SunKara[6] et al. improved the problem of information loss due to pooling layers in target detection networks by performing Focus operation on feature maps to reach the downsampling purpose and then changing the number of channels by convolution operation. In this paper, we design a lightweight target detection algorithm YOLOv5s-MC based on YOLOv5s.prune YOLOv5s and use the lightweight algorithm MobileNetv2 as the detection backbone. Optimize the feature fusion method and introduce adaptive learning parameters to make the network autonomously control the fusion weights of feature maps between different levels to increase the network aggregation capability. Improve the feature fusion network and add a small target detection layer to improve the detection accuracy of the model for small targets. the overall structure of YOLOv5s-MC is shown in Figure 1.

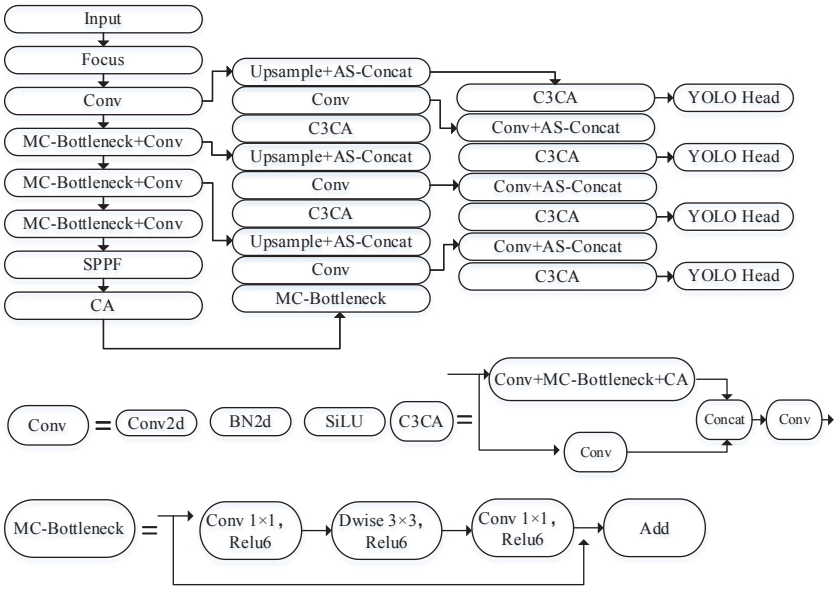

Fig. 1. YOLOv5s-MC network structure diagram.

2 Methodology

This section mainly introduces the network model of YOLOv5s-MC, the convolution module we used in the network and the AS-Concat mechanism.

2.1 MobileNet network model

MobileNet[7] network uses deeply separable convolution to construct lightweight convolutional neural networks, which greatly reduces the number of algorithm parameters and computational cost. For the input $X \in \mathbb{R}^{H\times W\times C}$ the output of a normal convolutional operation O can be expressed as Equation (1):

$$O = X \cdot f + b \tag{1}$$

$f \in \mathbb{R}^{k\times k\times C\times n}$denotes a convolution kernel of size$k \times k \times C$ of$n$ the convolution kernel, and b denotes the bias term. The depth-separable convolution is calculated as shown in Eqs. (2) and (3):

$$O' = X \cdot f_{dw} + b_1 \tag{2}$$

$$O = O' \cdot f_{pw} + b_2 \tag{3}$$

$f_{dw} \in \mathbb{R}^{k\times k\times 1\times C}$denotes a convolution kernel of size$k \times k \times 1$ of C depth convolution, and$f_{pw} \in \mathbb{R}^{1\times 1\times C\times n}$ denotes a convolution kernel of size$1 \times 1 \times C$ of n point convolution. Then the computational effort of the ordinary convolution and the depth-separable convolution is as in Equation (4):

$$\frac{P_{depthwise\ sepeable}}{P_{Conv}} = \frac{H \times W \times C \times (k \times k \times C + C \times n)}{H \times W \times C \times k \times k \times C \times n} = \frac{1}{k^2} + \frac{1}{n} \tag{4}$$

When the number of convolution kernels n the higher the number of convolution kernels, the closer the ratio is to$\frac{1}{k^2}$. i.e., the depth-separable convolution computation is reduced by k^2 times.

2.2 Adaptive feature fusion weights

In this paper, we propose the AS-Concat fusion method to address this problem. Adding learnable weights to feature fusion w_1, w_2 , allowing the network to adjust the weight parameters of different input features autonomously. This method makes the feature fusion network more flexible without increasing the parameters too much, and obtains different weight parameters when performing fusion for different levels of input feature maps, giving the network a stronger feature aggregation capability to achieve specific problem-specific analysis and make full use of the differentiation information between different levels, and its structure is shown in Figure 2 below.

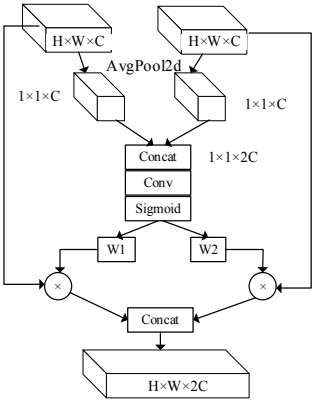

Fig. 2. AS-Concat structure diagram.

2.3 Improving multi-level feature fusion networks

In this paper, based on the original feature fusion network of YOLOv5s, we add a detection layer for small targets, multiplex and fuse the multilayer features of the network, retain more shallow features, and improve the recognition ability and robustness of the model for small targets. The specific operation is to upsample the feature maps after layer 17, after which the

features of layers 2 and 21 are fused by the Concat operation to obtain a 160×160 detection layer. Finally, the feature fusion network detection head sizes are 20×20, 40×20, 80×80, 160×160, and the corresponding sizes are set as (116×90, 156×198, 373×326), (30×61, 62×45, 59×119), (10×13, 16×30, 33×23), (5×6, 8×14, 15×11) The anchor frame size.

2.4 Attention mechanism

The CA attention [8] adds the feature map location information to the channel attention, and transforms the input feature map into two separate directions with dimensions of $1 \times W \times C$ and $H \times 1 \times C$ The two feature vectors with different directional feature information are encoded to obtain both channel information and location information, and each feature map has a long-range correlation of the input feature map along one direction. After that, the two obtained feature maps are stitched in the channel dimension, and the channel compression is performed by the convolution operation to get the dimension $C/r \times 1 \times (W + H)$ The feature map is then transformed into two feature maps by 1×1 Convolution transforms the feature map into two feature vectors in different directions, and using the Sigmoid function, the feature vectors are normalized to obtain the corresponding weight information of each channel. Finally, the new output feature map is obtained by multiplying the obtained weight vectors with the input feature map. CA attention not only captures the feature map cross-channel information, but also captures the location-sensitive information, which enables the model to better locate and focus on the target of interest, and the structure is shown in Figure 3.

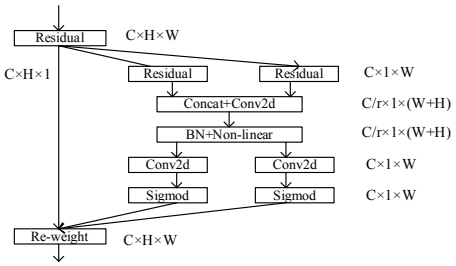

Fig. 3. Coordinate attention.

3 Experimental results

3.1 Experiment setup

All experiments in this paper were conducted under Ubantu 16.04 operating system with deep learning framework of pytorch1.7, cuda10.1, cudnn7.6.4, and hardware configuration GPU of Tesla K80 with 12GB of video memory.

3.2 Experiment results

To verify the performance of YOLOv5s-MC, We choose MAP@0.5 , the number of parameters, the amount of computation, and the accuracy rate are evaluated. Where the MAP@0.5 denotes the average accuracy when the IOU threshold is 0.5, precision indicates how many of the samples predicted to be positive are actually positive samples. Recall (recall) indicates how many of the positive samples are predicted to be positive. recall and accuracy are calculated as shown in equation (5) equation (6).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

where TN means the prediction is negative and the result is correct, TP means the prediction is positive and the result is correct, FN means the prediction is negative and the result is incorrect, FP means the prediction is positive and the result is incorrect, and the ablation comparison test as shown in Table 1.

Table 1. Ablation experiment.

Model	CA	Multi-head	AS-Concat	Parameter/M	Map50	FPS
YOLOv5s				7.06M	91.9%	40
YOLOv5sMobilenet				2.98M	88.9%	38
A	√			3.01M	89.7%	38
B	√	√		3.10M	91.5%	36
C	√	√	√	3.12M	92.1%	35

"√" indicates that the method is used in the experiments. It can be seen from the experiments that compared to the original algorithm the accuracy of the algorithm proposed in the paper is improved by 0.2% and the number of model parameters is drastically reduced by 55.8%. The improvement strategies proposed in this paper, as can be seen from the results in the table, have all improved the model performance, and the number of model parameters has not increased tremendously, Although the detection speed of the improved algorithm is reduced compared to the original algorithm, it still meets the real-time detection requirements. which proves that the improvement strategies proposed in the paper are effective.

4 Conclusion

In this paper, we propose a lightweight real-time target detection algorithm for road scenes based on the YOLOv5s algorithm. The lightweight MobileNetv2 network is used as the backbone of YOLOv5s. In addition, on the basis of lightweight, CA attention is introduced into the network, which is used to obtain the spatial and location information of targets and improve the overall feature extraction capability of the model; the detection head for small targets is added to improve the detection capability of tiny targets; the feature fusion network is optimized, and an adaptive fusion method AS-Concat, which allows the network to autonomously select the fusion weights between different levels and enhance the network's ability to fuse feature maps at different levels; through experiments on the KITTI dataset, the results show that the proposed algorithm reduces the amount of parameters by 55.8% compared with the original YOLOv5s and the detection accuracy is improved by 0.2% compared to the original YOLOv5s. The next work considers to improve the detection speed and accuracy of the algorithm based on the existing improved model through data enhancement and algorithm structure optimization, etc.

References

1. Liu Y, Cao S, Lasang P, et al. Modular lightweight network for road object detection using a feature fusion approach [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 51(8): 4716-4728.

2. Nguyen V D, Trinh T D, Tran H N. A Robust Triangular Sigmoid Pattern-Based Obstacle Detection Algorithm in Resource-Limited Devices [J]. IEEE Transactions on Intelligent Transportation Systems, 2023.
3. Lu X, Zhong Y, Zhang L. Open-Source Data-Driven Cross-Domain Road Detection From Very High Resolution Remote Sensing Imagery[J]. IEEE Transactions on Image Processing, 2022, 31: 6847-6862.
4. Xiuli Xin, et al. "SwinT-YOLOv5s: Improved YOLOv5s for Vehicle-mounted Infrared Target Detection". Proceedings of the 41st Chinese Control Conference (13).Ed. , 2022, 210-215.
5. Guo C., Fan B., Zhang Q., Xiang S., Pan C.. AUGFPN: Improving multi-scale feature learning for object detection [J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020.
6. Sunkara R, Luo T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part III. Cham: Springer Nature Switzerland, 2023: 443-459.
7. Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [J]. arXiv preprint arXiv:1704.04861, 2017.
8. Hou, Qibin, Daquan Zhou, and Jiashi Feng. "Coordinate attention for efficient mobile network design." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.