Research on vehicle detection algorithm based on improved YOLOV5

Songyu Deng[#,*]
Zhuhai College of Science and Technology
Zhuhai, China
aldrichair@gmail.com

Zixuan He[#]
Nanjing University of Finance & Economics
Nanjing, China
zixuanhe921@outlook.com

Meng Niu[#]
Shihezi University
Shihezi, China
mengmeng0711@outlook.com
[#]These authors contributed equally.

***Abstract*—Given the large cost of the current vehicle detection implementation project, this paper reduces the cost of the project by modifying the Yolov5s backbone model to greatly reduce the number of parameters in the training process under the premise of ensuring accuracy. First, MobileNetV3 is used to replace the Yolov5s backbone network, which reduces the number of parameters from 14.4M to 3.1M and the mAP from 0.984 to 0.977. The accuracy of the model is only reduced by 0.007 with the greatly reduced number of parameters, and the model is greatly improved. The SE attention module of the inverted residual structure is replaced by the CBAM module, and the parameters and mAP are further optimized, thus reducing the parameter quantity to 2.4M and the mAP to 0.978, which makes the model better improved and can better meet the vehicle detection put into use.**

Keywords—vehicle detection, Yolov5, MobileNetV3, attentional mechanism

I. INTRODUCTION

Intelligent transportation, as one of the important components of the current intelligent era, has attracted a lot of attention from contemporary scholars. The detection of vehicle targets has a crucial role and significance in driverless as well as assisted driving. In practical scenarios, vehicle target detection is more complex and requires high accuracy in vehicle detection while the recognition, as well as detection time, must be controlled, and it is particularly important to improve the algorithm efficiency as well as the running time.

Among the traditional target algorithms, which mainly address a small number of detection problems, the mainstream algorithms are mainly Cascade + HOG/DPM + Haar/SVM and many improvements of the above methods. The problems of the mainstream methods are mainly twofold: on the one hand, the sliding window selection strategy is not targeted, has high time complexity, and has redundant windows; on the other hand, the robustness of manually designed features is poor.

With the development of deep learning technology, traditional target detection algorithms have been gradually replaced. The mainstream algorithms based on deep learning today are mainly divided into two categories: one is the two-stage algorithm that divides the detection and into two stages, also known as the target detection algorithm based on the candidate region, mainly the R-CNN series, whose process is divided into the following three steps: extraction of the candidate region, classification of the candidate region, and correction of the extracted and classified candidate region coordinates, compared with the traditional target The DPM performance is improved compared to the traditional target detection algorithm, but its training speed is slow and contains a large number of repetitive operations with large space cache, and the information is easily lost due to the limitation of the size of the input image; the other category is the one-stage algorithm that integrates detection and integration, such as the Yolo (You Only Look Once) series algorithm, the SSD (Single Shot Multibox Detector) algorithm, Over Feat algorithm, etc. These algorithms can maintain a high detection accuracy and simple structure, and significantly improve the speed of detection, in the combination of speed and accuracy is far better than the two-stage algorithm, suitable for edge-based vehicle target detection in the case of vehicle-road cooperation. In the Yolo series of algorithm research, although the detection speed and accuracy are greatly improved, in the training process, the use of a large number of parameters, making the use of this type of algorithm project landing cost is high, landing implementation difficulties and for long-range small targets and more covered targets, the detection effect is poor.

Therefore, in this paper, we adopt the improved Yolov5 algorithm to improve the number of parameters and accuracy layer by layer and do three ablation experiments, namely, the classical Yolov5 algorithm and the improved Yolov5 model with MobileNetV3 as the backbone network, which reduce the number of parameters in the training process to some extent, and replace the SE attention module in the MobileNetV3 inverse residual with the CBAM module to ensure the relative improvement of the number of parameters and accuracy.

II. IMPROVED YOLOV5 ALGORITHM MODEL

A. Yolov5 structure

The Yolo series algorithm divides the input image into S × S grids, and if the center of the target to be detected falls within a grid, this grid is responsible for predicting that target, and each grid predicts B bounding boxes and category information, and the bounding boxes contain target location and confidence information. Based on this detection method, the yolov5 network predicts three bounding boxes for each grid, each containing coordinate information (x, y, w, h), one confidence level, and C conditional category probability information. yolov5 algorithm has four models of target detection network, namely Yolov5s, Yolov5m, Yolov5l, Yolov5x, increasing in network width and depth are increased sequentially.3 The YOLOv5 algorithm makes some improvements on the basis of the YOLOv4 algorithm so that its speed and accuracy are greatly improved in performance. The main improvement ideas are as follows:

Input: Mosaic data enhancement, adaptive anchor frame calculation, and adaptive image scaling in the model training phase; Benchmark network: improved Focus structure and CSP structure; Neck network: target detection network tends to insert some layers between BackBone and the final Head output layer, and FPN+PAN structure is added in Yolov5; Head output layer: the anchor frame mechanism of the output layer is the same as YOLOv4, and the main improvements are the loss function GIOU_Loss during training and DIOU_nms for prediction frame screening.

B. Introduction of algorithm improvement

The scale and difficulty of the vehicle target detection task are very different from the multi-category target detection task, and the YOLOv5 algorithm for the multi-category target detection task is directly applied to the vehicle detection task, which will face the following problems: the overall model volume is large, the computation is large, and the model is easy to cause the waste of resources when it is carried on the embedded equipment for operation; it does not focus on the vehicle detection due to road congestion, the camera angle The problem of occlusion between vehicles caused by road congestion, camera angle and the problem of detection of small and medium-sized targets at a long distance is not considered. In this paper, we design a specific vehicle target detection model based on YOLOv5, so that the improved network can achieve higher performance in the vehicle detection task as follows:

By modifying the attention structure of MobileNetV3, we propose the MobileNetV3_CBAM network and use it as the backbone network of YOLOv5, so as to reduce the number of parameters of the model in the backbone feature extraction network and reduce the training difficulty;

To address the problem of imprecise detection of small and medium targets in the YOLOv5 model, we additionally add a tiny target detection head, observe the distribution characteristics of the dataset, perform K-means clustering according to the real frame size provided by the dataset, and generate a multi-scale anchor frame size with strong adaptability for vehicle detection by this method to improve the detection capability of the model for four different sizes of vehicle targets: large, medium, small and tiny. The specific structure is shown in Figure 1.

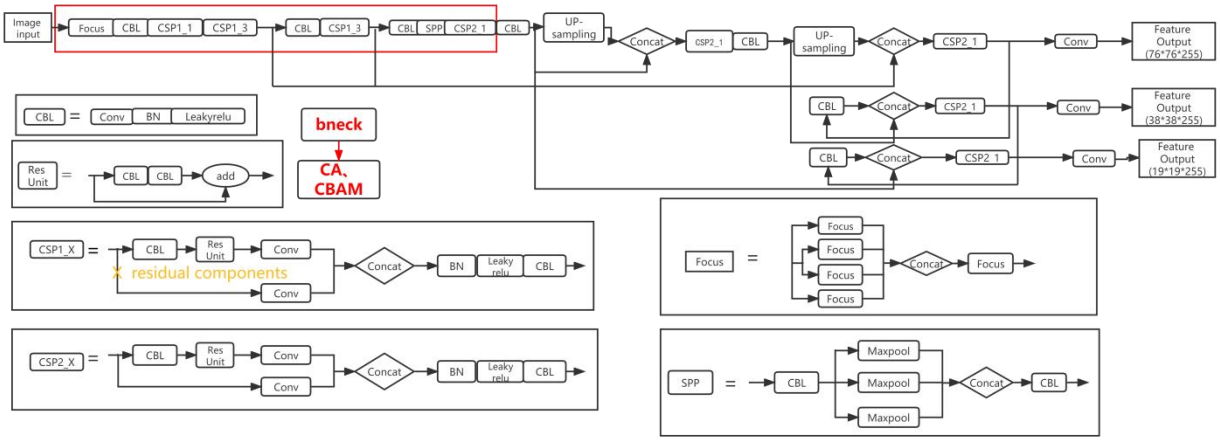

Fig. 1. Improved yolov5 frame diagram

1) MobileNetV3

MobilenetV3 follows the inverse residual structure of the MobilenetV2 network model, and the inverse residual structure has fewer parameters than the same level residual structure due to the DW convolution operation, as shown in Figure 2. which can be divided into the Depthwise Convolution operation and the Pointwise Convolution operation. Pointwise Convolution. The number of convolution kernels is equal to the number of channels in the input feature map, and the output feature map cannot be further expanded after the channel-by-channel convolution is completed. The point-by-point convolution is equivalent to using a normal two-dimensional convolution with a convolution kernel of size 1 × 1 × M, where M is the number of input channels, and using several filters to up-dimension or down-dimension the input features.

The use of DW convolution not only achieves the effect of using two conventional convolutions for feature extraction but also reduces the computational effort by a factor of 8 to 9 using DW convolution compared to the conventional convolution operation.

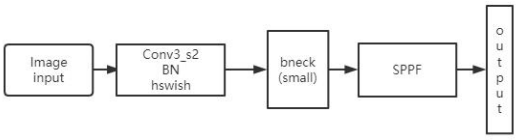

Fig. 2. MobileNetV3 model frame diagram

2) SE module

MobilenetV3 uses the SE attention mechanism (Squeeze-and-excitation) based on the inverse residual structure of MobilenetV2. The SE module mainly obtains the scores of each channel by pooling the input features by global average and then outputs the weights that are equivalent to the original input channels after two fully connected layers and activation functions. The resulting channel weights are multiplied by the corresponding two-dimensional matrix of the original input feature map, and the resulting output features can better focus on the relationship between the channels so that the model can better learn the importance of different channel features. The framework structure is shown in Figure 3.

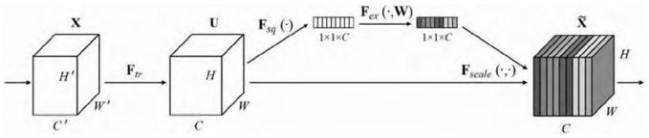

Fig. 3. SE framework structure diagram

C. CBAM attention mechanism

The SE attention mechanism used in MobilenetV3 only considers the relationship between channels unilaterally, but not the information in the spatial dimension. Therefore, we modify the SE attention module of MobilenetV3 into a CBAM module, so that the network can obtain more comprehensive information when extracting features and improve the learning ability.

The CBAM module is mainly composed of CAM and SAM structures, where the CAM (Channel-Attention-Module) structure has an additional global maximum pooling operation compared with the SE module, and the two same dimensional scores obtained from the global pooling are input to the shared MLP to output the corresponding weights, and the weights obtained from different global pooling operations are summed in the same dimensional matrix. The SAM (Spatial-Attention-Module) structure takes the output attention scores of the CAM module and the original input features as the matrix multiplication results as the input and performs the maximum pooling and average pooling of the input features without changing the feature map size (the pooling The two pooling results are deep stacked (concat), and then a convolution operation is performed to reduce the number of channels to 1. The result is then fed into the sigmoid function for activation, and the output is the spatial attention score. The output is the spatial attention score, which is multiplied by the matrix with the input processed by the channel attention score, and the output takes into account the relationship between space and channels.

D. Candidate frame clustering analysis

The YOLOv 5 algorithm combines the Anchor mechanism of Faster R-CNN to set three different candidate frame sizes for different scales of network layers. The COCO dataset contains a total of 80 categories with different sizes and patterns, so the original candidate frame sizes are not representative. In this paper, we use the K-means clustering method to reacquire the candidate frame size only for the vehicle target and take the vehicle dataset as the statistical object. In the clustering process, the intersection ratio between the candidate frame and the rear frame is used as the distance measure of clustering.

Observing the dataset, it is found that the vehicles in the dataset can be divided into three detection targets: large, medium, and small. The original YOLOv5 authors calculate the distance from the cluster center to each real box based on the Keans clustering algorithm and using Euclidean distance, according to (citation), using Cluster IOU instead of Cluster SSE to calculate the distance formula, the average of the resulting anchor box and the predicted box generated based on the offset Cluster IOU is calculated as follows:

$$\text{dis tan}\, ce(box, centroid) = 1 - IoU(box, centroid) \quad (1)$$

III. Experimental results and analysis

A. Data set introduction

Using the publicly available dataset DETRAC-TRAIN-VOC, a total of 4104 images, the dataset contains night and daytime urban road sections, highway sections, and other complex scenes of the image dataset, the vehicle target involves a wide range of scenes, and by the data real frame labeled distribution is more uniform, very suitable for building multi-target vehicle detection model. The dataset is divided into 3692 train_val and 412 test by 9:1, and 70% of train_val is divided into 2586 train and the remaining 30% as val.

Some of the datasets are as shown in Figure 4:

Fig. 4. Part of the dataset

B. Experimental environment and parameter configuration

Experimental environment: OS Ubuntu 18.04, deep learning framework Pytorch1.7, CPU E5-2678v3, memory 62G, GPU single RTX A5000, video memory 24G.

The model parameters are initialized based on the pre-trained weight files obtained from training on large datasets such as the COCO128 dataset as an example migration, based on which the parameters are fine-tuned using datasets for multi-vehicle targets and related scenarios, and optimized by back-propagation algorithm through continuous iterative training to obtain a multi-vehicle target detection model.By considering the influence of Batchsize and optimizer selection strategy on the model progress, this paper conducts quantitative experiments on the relevant hyperparameters using the YOLOv5 model, as shown in Table I:

TABLE I. Hyperparameter Comparison Experiment

Optimizer	Batchsize 8	Batchsize 16	Batchsize 32
SGD	0.973	0.975	0.976
Adam	0.975	0.977	0.980
AdamA	0.979	0.980	0.982

According to Table I, this paper uses the model algorithm to select AdamA in the optimizer, and the Batchsize is set to 32 than using SGD optimizer, and the total accuracy is improved by about 1% when the Batchsize is 8, which is the best training effect.

Based on the above conclusions, this paper uses MobileNetV3_CBAM modified YOLOv5 as the base network for training, the input size is 640×640, the number of channels is 3, the AdamA optimizer is used for training, L2 regularization is used to avoid overfitting, the weight decay parameter is set to 0.0005, the initial learning rate is 0.01, the Batchsize is set to 32, and 200 epochs are iterated.

C. Testing performance evaluation index

The evaluation indexes are mainly considered in terms of the accuracy and speed of the detected targets. Since this task is a single-target detection task, the detection accuracy is used to specifically evaluate the target localization capability,

and the index to measure the accuracy is the average accuracy mean value, and the IoU is taken as 0.45, which is defined as follows:

$$mAP = \frac{1}{n}\sum_{j=1}^{n} AP(j) \tag{2}$$

Where n is the total number of categories, the data set in this paper contains a total of 1 category of targets, so n=1 and AP is the average precision of a category, defined as the mean value of the precision rate under different recall rates, calculated as follows:

$$\begin{cases} AP = \int_0^1 P(R)\mathrm{d} \\ P = \dfrac{TP}{TP + TN} \\ R = \dfrac{TP}{TP + FN} \end{cases} \tag{3}$$

where P denotes the precision rate, which refers to the probability of correct prediction among all positive samples predicted, R denotes the recall rate, which refers to the probability that all positive samples are detected, TP denotes true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative.

The detection speed is measured using FPS, which is defined as the frame rate and indicates the amount of time that can be detected per second.

D. Experimental results and analysis

1) Improved model and its ablation experiment

The improved YOLOv5 vehicle detection algorithm proposed in this paper contains four improvements, which are: using the optimized MobileNetV3_CBAM as the backbone network of YOLOv5, modifying the normal convolutional layer in the partial structure of the neck network to DW convolution, introducing a detection head specifically for detecting tiny targets and using the K-means clustering algorithm to generate different sizes Anchor frame size of feature map.mAP@0.5denotes the average detection accuracy when IoU=0.5; ACC denotes the accuracy, Recall denotes the average recall, and Params denotes the overall parametric number of the model. The comparison test results of the models are shown in Figure 5. From Table II, we can find that MobileNetV3 is used as the backbone network of the YOLOv5 model, because MobileNetV3 uses more DW convolutional modules, which makes the B model 1/5 of the A model in terms of computation, and because MobileNetV3 applies the SE attention mechanism, the accuracy of the model in detecting overlapping targets is improved. The C model further modifies the SE attention mechanism into the CBAM module based on the B model, which further reduces the number of parameters and achieves better detection of overlapping targets than the A model.

TABLE II. COMPARISON OF THE RESULTS OF SEVERAL YOLOV5 MODELS

Models	Params (M)	GFLOPs	ACC	Recall	FPS	mAP@0.5
YOLOv5	14.4	15.8	0.947	0.964	35.7	0.984
YOLOv5 MobileNetV3	3.1	2.3	0.951	0.929	48.5	0.977
YOLOv5 MobileNetV3_CBAM	2.4	2.1	0.953	0.927	33.3	0.978

Fig. 5. Test results

IV. CONCLUSION

In this paper, we address the problems of a large number of YOLOv5 parameters and low accuracy of small and medium-sized target detection and leakage detection encountered in real-time detection, based on the YOLOv5s target detection model, we use MobileNetV3 to modify the backbone network of the original model, based on which the SE attention module of MobileNetV3 inverted residual structure is replaced with CBAM module, and use K The number of parameters of the modified model is about 1/7 of the original model, which greatly reduces the overall volume of the model, while the introduced attention mechanism and anchor frame optimization compensate for the loss of model accuracy due to the reduced number of parameters. Considering the cost and promotion of unmanned driving and related application technologies in the field, the model in this paper is more suitable to be installed in embedded devices, which can better realize the real-time and accuracy of vehicle detection in different environments, and provide the basis and ideas for further research in the field of vehicle detection and unmanned driving, so as to realize faster and more accurate vehicle detection solutions.

REFERENCES

[1] Cao Jingwei. Research on smart car target detection and tracking algorithm under complex scenes [D]. Jilin University, 2022.

[2] Chengjun Zhang, Hu Xiaobing, Niu Hongchao. Research on vehicle target detection based on improved YOLOv5 [J]. Journal of Sichuan University (Natural Science Edition), 2022, 59(05):79-87. doi:10.19907/j.0490-6756.2022.053001.

[3] hang Y. M., Liu J. W., Song X. L., Wang Z. J., Wang M. E., Huang L. H. A deep learning visual vehicle detection method based on YOLOv3 [J]. Automotive Practical Technology, 2022, 47(05):30-33. doi:10.16638/j.cnki.1671-7988.2022.005.007.

[4] Liu Junming, Meng Weihua. Research review of single-stage target detection Algorithm based on Deep Learning [J]. Aeronautical Ordnance,20,27(03):44-53.

[5] Hu Junping, Wang Hongshu, Dai Xiaobiao, Gao Xiaolin. Improved YOLOv5 algorithm for real-time detection of small target traffic signs [J/OL]. Computer Engineering and Applications:1-10 [2022-11-02].

[6] Wang C, Wang H, Yu F, et al. A High-Precision Fast Smoky Vehicle Detection Method Based on Improved Yolov5 Network[C]// 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID). IEEE, 2021.

[7] YANG Hui, QUAN Jichuan, LIANG Xinyu, et al. Weak Supervisory Camouflage Objective Detection Method Incorporating Attention Mechanism [J] . Network security and data governance Rational , 2022 , 41 (3) : 81- 91.

[8] Hu Junping, Wang Hongshu, Dai Xiaobiao, et al. Real-time Detection Algorithm of Small Target Traffic Sign based on Improved YOLOv5 [J]. Computer Engineering and Applications,2023,59(02):185-193.