Traffic Sign Detection Algorithm based on improved YOLOv4

Huibai Wang, Hao Yu
College of Information Science and Technology North China University of Technology Beijing, China
wanghb@ncut.edu.cn, 18613316415@163.com
Corresponding Author: Hao Yu   Email:18613316415@163.com

2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) | 978-1-7281-5244-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ITAIC49862.2020.9339181

***Abstract*—In the deep learning algorithm, YOLOv4 and Faster R-CNN have achieved excellent target detection performance. To improve the real-time detection of small targets in traffic signs, this paper presents an improved YOLOv4 target detection algorithm, we use K-means clustering to design anchor box, which is used to adapt to different small and medium scale. According to the size of small and medium scale signs, one more feature layer is extracted, and four different feature layers are fused to detect. Experimental results show that the improved algorithm can improve the detection accuracy and real-time performance.**

Keywords—YOLOv4; Traffic Sign Recognition; K-means clustering; Anchor box; Faster R-CNN;

I. Introduction

Traffic sign detection and recognition technology are one of the key technologies in the field of automatic driving, and it is widely used in the assistant driving system, so the study of traffic sign detection and recognition method is based on YOLOv4 in this paper is of great significance to regulate driving.

The traditional method of machine learning is mainly based on the combination of color and shape, but the result is not good. In recent years, with the development of deep learning in target detection and image classification, there are two kinds of Algorithms: detection-based Algorithm (such as R-CNN [1] , Fast R-CNN [2]), Faster R-CNN [3], and regression-based methods (such as Yolo [4], Yolov2[5], Yolov3[6]).

In summary, this paper improves the Algorithm of YOLOv4 to solve the problem that the character of the traffic sign is not easy to be influenced by illumination, real road, and other complex factors. The structure of the network is optimized, and a layer of feature pyramid is added to detect the multi-scale feature, and the anchor box corresponding to the multi-scale is obtained by the K-means clustering method according to the characteristics of the marker scale, to improve the positioning accuracy. The experimental results show that the improved model based on YOLOv4 has a good effect on traffic sign recognition.

II. YOLOv4

A. Target detection principle of YOLOv4

Based on the original object detection architecture of Yolov3, the algorithm introduces some optimization methods from data processing, backbone network, network training, activation function, loss function, and so on, the model achieves the best detection speed and precision up to now. YOLOv4 first extracts the features of the input image through the feature extraction network backbone network, then divides the input image into s * s Grid, then each cell is responsible for detecting the target that the center falls into the grid. To complete the detection of n objects, each cell first predicts the confidence of three anchor boxes and the anchor box. The size and position of the anchor box are represented by four values (x, y, w, h), where (x, y) is the central coordinate of the anchor box, and W and h are the width and height of the anchor box. Thus, the predicted value of each anchor box contains five parameters (x, y, w, h, c), where (x, y) represents the offset from the center of the anchor box to the center of the real box, and (w, h) represents the width and height of the anchor box. The final anchor box confidence consists of two parts: the probability that the box contains the target and the accuracy of the bounding box.

Definition 1. There exists the following formula about confidence

$$c = p_r(object) \times IOU_{pred}^{truth}(p_r(object) \in \{0,1\}) \quad (1)$$

The prediction still needs to be decoded, setting the top left corner of the bounding box to t_x, t_y, the width and height of the anchor box to t_w, t_h .the adjusted center and width and height to and p_x, p_y, p_w, p_h.

Definition 2. The final prediction box to be positioned as follows

$$p_x = \sigma(x) + t_x \quad (2)$$

$$p_y = \sigma(y) + t_y \quad (3)$$

$$p_w = e^{w} t_w \quad (4)$$

$$p_h = e^{w} t_h \quad (5)$$

The final output of this model should be (1,3, S*S, 4 +1+N). To get the final prediction results, we need to sort the scores and restrain the non-maximum value to get the final prediction results.

B. The core structure of YOLOv4

The main feature network structure of YOLOv4 is CSPDarkNet53(as shown in figure 1), and the structure of residuals is modified, and the CSPnet structure is used to separate the original residuals into two parts, the trunk continues to stack the remaining residuals, while the other is connected directly to the end like a different edge. The CSPnet structure is shown in figure 2.

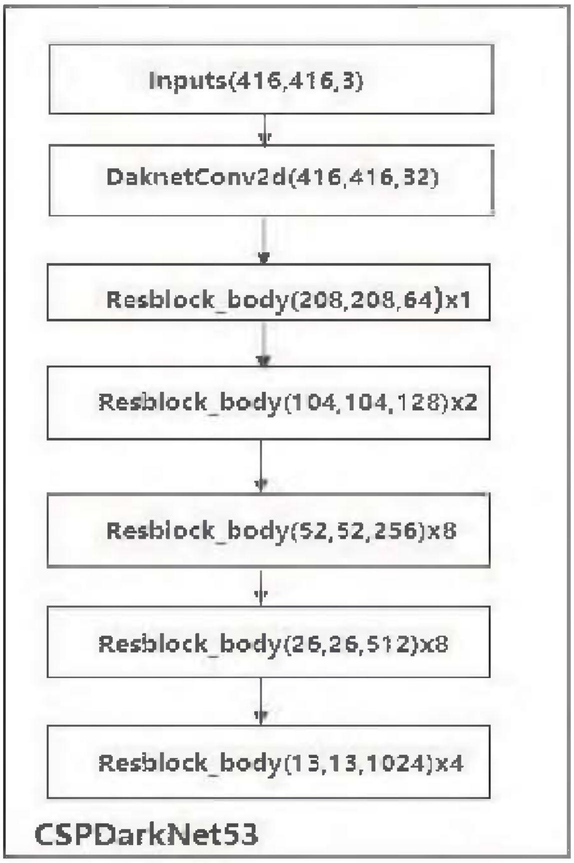
Inputs(416,416,3)
DaknetConv2d(416,416,32)
Resblock_body(208,208,64)x1
Resblock_body(104,104,128)x2
Resblock_body(52,52,256)x8
Resblock_body(26,26,512)x8
Resblock_body(13,13,1024)x4
CSPDarkNet53

Fig.1 backbone feature network

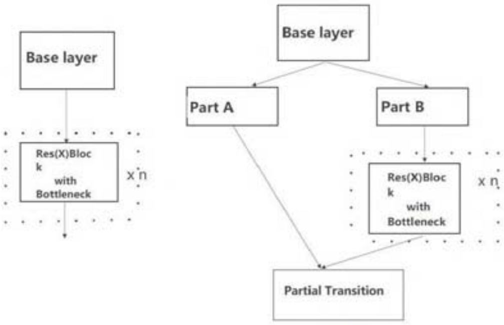

Fig.2 CSPnet structure

In the part of Feature Pyramid, YOLOv4 combines two kinds of improvements: using the PANet and SPP structure, the SPP structure is involved in the convolution of the last feature layer of CSPDarkNet53, and after three convolution of the last feature layer of CSPnet53 by DarknetConv2D, the maximum pool size is 13x13,9x9 and 5x5, and the maximum pool size is 13x13,9x9 and 5x5. It can greatly increase the receptive field and isolate the most salient contextual features. One of the most important characteristics of the PANet structure is the feature extraction repeatedly, in (a) it is the traditional feature pyramid structure after the feature pyramid is extracted from the bottom to the top, we still need to implement the feature extraction from the top to the bottom in (b). Figure 3 below is the PANet architecture.

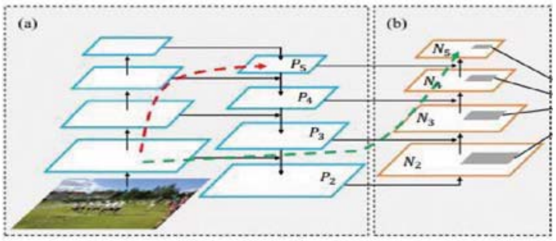

Fig.3 PANet architecture

III. IMPROVEMENTS BASED ON YOLOV4

A. Design of multi-scale Feature Fusion Network

The traffic sign image in the natural scene is affected by color, shape, and various complex environmental factors. The traditional algorithm detector can not realize real-time detection and recognition, and the rate of missing detection is high. In the vehicle-mounted camera, the object of a traffic sign is too small in the whole image, so we add one more layer of feature layer output, and finally, four layers of feature fusion are used to forecast to solve the problem of small object missing detection. Input Picture 416×416 as input, figure 4 describes the specific implementation process.

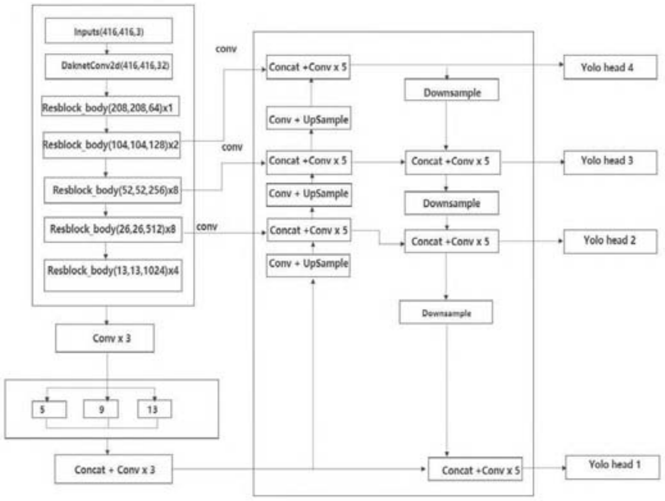

Fig.4 Improved Network Structure Diagram

The improved YOLOv4 uses four different scales to detect targets of different sizes. The feature map is sampled up and then down, and connected with the output of the second residual block of YOLOv4, in this way, a four-layer feature fusion target detection layer is successfully established. For example, the size of the input image is 416 ×416. The shape of the four feature layers is (13,13,1024) , (26,26,512) , (52,52,256) , (104,104,128) .

B. K-means cluster computing anchor box

The introduction of anchor box transforms the problem of target detection into the problem of whether there is a target inside a fixed lattice and the deviation between the prediction box and the real box, the ratio of length to width is not the same, but in the process of traffic, sign data set recognition, the ratio of sign size is the same because the Kmeans algorithm is good for the data set annotation box clustering effect, so we use this method to get the anchor box that matches the ratio of the signage. Comparing the intersection of Cluster Center and label $IOU_{(1,\ C)}$ as the similarity parameter of K-means clustering.

Definition 3. There exists the following calculation formula

$$d=1-IOU_{(1,C)} \quad (6)$$

The similarity parameter of K-means clustering is calculated by comparing the intersection of the cluster center and Label, and the formula is as follows: based on the TT100K traffic sign data set, K-means++ clustering algorithm is used. A total of four feature layers were output, so 12 anchor boxes were selected. Table I shows the corresponding cluster number range values for the width and height of the TT100K traffic sign Dataset.

Table I characteristic receptive field and anchor box

Feature map	13×13	26×26	52×52	104×104
Scale	big	middle	small	Very small
Anchors	(46,75) (69,90) (100,124)	(19,35) (27,43) (37,55)	(15,32) (23,64) (20,25)	(8,17) (12,25) (13,20)

IV. EXPERIMENTAL RESULTS AND ANALYSIS

To verify the detection effect of the improved YOLOv4 on the traffic sign data set, a network framework based on CSPDarkNet53 and a detection framework based on YOLOv4, which has Intel (R) Xeon (R) Silver 4110 CPU@2.10 GHz, 64 GB Ram, Nvidia GeForce RTX 2080Ti GPU, are tested in this paper. The results show that the improved YOLOv4 can be used to detect the traffic sign data set effectively. The operating system is windows 10.

A. Data pre-processing

The data set of this paper is the TT100K data set. There are 9170 images in total. To get a better training effect, four images are selected randomly by using the method of Mosaic data enhancement, the four images were flipped, zoomed, gamut changes, and other operations, and placed following the four directions. This not only greatly expands the original data set, but also enriches the picture's background. Another advantage is that you can directly calculate the number of 4 images while training, thereby increasing the Batch-Size so that the size of the Minibatch set does not have to be so large that the training is less difficult.

B. Model Training

In this paper, the ratio of the training set, verification set, and test set is divided into 8:1:1. Using the transfer learning idea, the training is carried out based on pre-training weight trained by the COCO data set. During the training process, the initial learning rate is set to 0.001, the half-attenuation Coefficient is set to 0.0005, and the Batch-Size to convolutions up the training and prevent the weight from being destroyed at the beginning of the training, the backbone feature network is first frozen and trained for 30 generations, then it is thawed and trained for 30 generations, the number of iterations is 220k, and the relationship between training loss and Training Times is as follows:

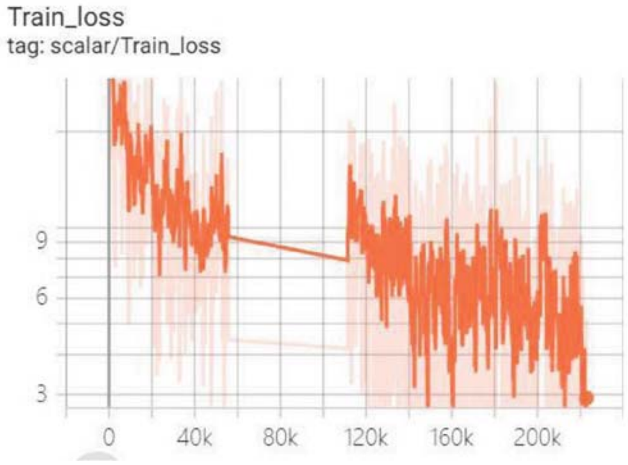

Fig. 5 relationship between training loss and training times

C. Evaluation of Experimental Results

Using the improved YOLOv4, five kinds of signs, such as no parking (pn), no honking (p11), deceleration (pne), the speed limit of 5km/h (pl5), and no motorcycle(p12), were tested under various complicated conditions, such as strong light, rainy day, afternoon and night, and with the improvement of the detection results were compared, as shown in the following figure:

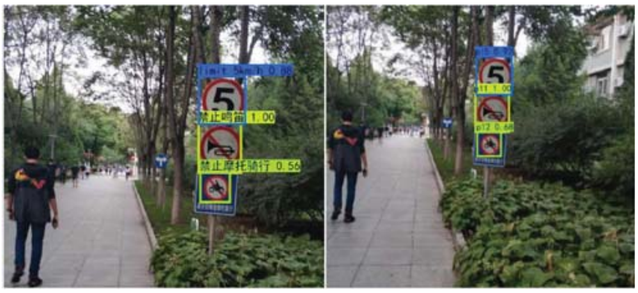

Fig. 6 before and after improvement (Afternoon)

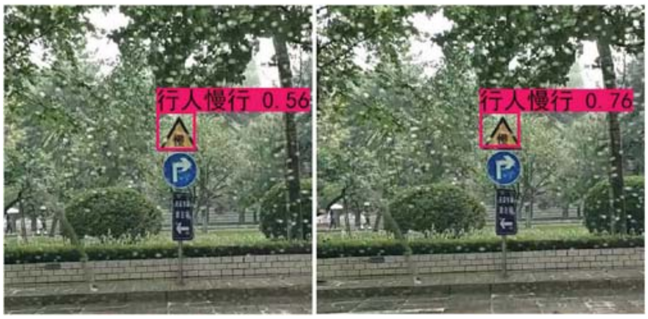

Fig. 7 before and after improvement (Rainy Day)

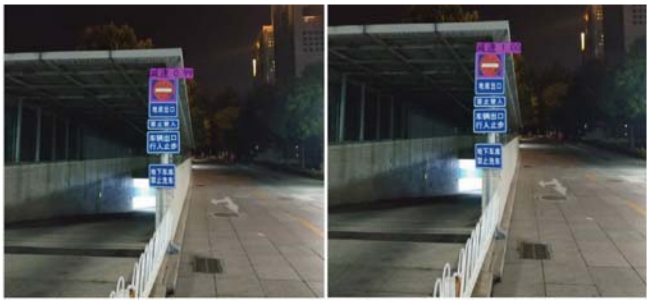

Fig. 8 before and after improvement (Night)

By the comparison of the graph above, it is obvious that the confidence of the improved algorithm is improved under various conditions, and the algorithm shows high robustness.

The AP (average precision) is calculated for each of the five categories, and the mAP of the five categories is calculated, and the average time t of one frame is detected by the improved algorithm. The statistics are shown in Table II below.

Table II Comparison of effect between YOLOv4 and improved YOLOv4

Method	pn	p 11	pne	pl5	p12	mAP	T/ms
YOLOv4	0.65	0.57	0.86	0.81	0.69	71.6	67.7
Improved YOLOv4	0.86	0.74	0.97	0.95	0.83	81.7	68.1

As can be seen from the above table, the AP of each category has been significantly improved, the detection speed improvement is not very large, but also slightly improved. The results show that the improved algorithm is effective.

V. CONCLUSION

In this paper, an improved algorithm based on YOLOv4 is proposed to recognize traffic signs in the driving process. Aiming at the small collection data set, the K-means method is used to cluster the tags, which improves the location precision, improves the network structure for small size marks, adds an extra layer of an output feature layer, and finally gets four scale feature layer, then, four feature layers are fused, and the prediction on these four scale feature layers can effectively solve the problem of missing detection due to the small target, compared with the original YOLOv4, the improved algorithm in this paper has better performance in accuracy and speed. In the follow-up work, it will be necessary to refine more classification, can identify more types of signage, enhance the algorithm robustness, can make it recognize more adverse conditions of signage.

REFERENCES

[1] Girshick R, Donahue J, Darrell T, et al.Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus,2014:580-587.

[2] GIRSHICK R.FastR-CNN[C]//Proceedings of 2015 IEEE International Conference on Computer Vision,2015:1440–1448.

[3] Ren S, He K, Girshick R, et al.Faster r-CNN: Towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.

[4] Redmon J, Divvala S, Girshick R, et al.You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition,2016:779–788.

[5] Redmon J, Farhadi A.YOLO9000: Better, Faster, Stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017:6517652

[6] Redmon J,Farhadi A.YOLOv3: An Incremental Improvement[J].arXiv:1804.02767,2018.