International Journal of Computer Applications Technology and Research
Volume 10–Issue 06, 161-165, 2021, ISSN:-2319–8656

Traffic Sign Detection and Recognition Based on Improved YOLOv4 Algorithm

Gaoli Hu
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Chengyu Wen
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Abstract: Traffic sign detection and recognition play an important role in intelligent transportation. In this paper, a traffic sign detection framework based on YOLOv4 is proposed. The original CSPDarkNet53 backbone network model is replaced by RepVGG, and the SPP module is added in the feature pyramid part to improve the expression ability of information. The CCTSDB traffic sign data set is used to detect three categories of indication signs, prohibition signs and warning signs. In order to further improve the performance of YOLOv4 network, K-means++ algorithm was used to perform cluster analysis on the experimental data to determine the size of the priori box suitable for CCTSDB dataset. The experimental results show that the map value of the improved framework is increased by 4.1%, which indicates that the improved YOLOv4 network has a high practical value in traffic sign detection and recognition.

Keywords: Deep learning; Traffic sign detection; YOLO4; RepVGG

1. INTRODUCTION

In recent years, the intelligent transportation system has developed rapidly, and the detection and recognition of traffic signs in natural scenes are important components of it. Traffic signs provide valuable traffic information such as road names, instructions and warnings, which can help drivers to comply with traffic signs according to law, greatly prevent traffic accidents and reduce traffic congestion. Therefore, the detection and identification of traffic signs is of great significance.

Generally speaking, the traditional target detection methods usually use manual features such as color [1] and shape [2] to extract regions of interest in the image. It is difficult to achieve ideal detection results in the field of traffic sign detection and recognition. With the vigorous development of artificial intelligence and computer vision, deep learning is widely used in the image field, and significant progress has been made in target detection, and it is one of the most effective solutions for traffic sign detection. Many well-known networks based on region generation, such as the R-CNN series [3-5], have good performance in target detection. There are also single regression-based networks including SSD[6], YOLO[7-9] series, etc, which simultaneously predict the bounding box and target probability from the input image. Although many achievements have been obtained in traffic sign detection, it is still a challenge to accurately and quickly locate and classify traffic signs in the face of relatively small traffic signs, unfixed shapes, and unstable characteristics in different situations.

Based on the above related work, this paper designs an improved traffic sign detection and recognition algorithm based on the YOLOv4 model [10]. Finally, the difference between the improved algorithm and the original algorithm is compared through experiments, and the results are verified on the CCTSDB dataset. The experiment shows that the method used in this paper effectively improves the detection accuracy and obtains good detection results.

2. YOLOv4 NETWORK STRUCTURE

YOLOv4 is improved on the basis of YOLOv3, which combines the best algorithm model and training skills in the current neural network. It can better distinguish the target information and the background area through the whole image training. It belongs to a single-stage target detection algorithm with strong real-time performance. YOLOv4 is mainly composed of backbone feature extraction network (CSPDarknet53), feature pyramid structure (SPP[11], PANet[12]) and a prediction result layer (Yolo Head). The structure is shown in Figure 1.

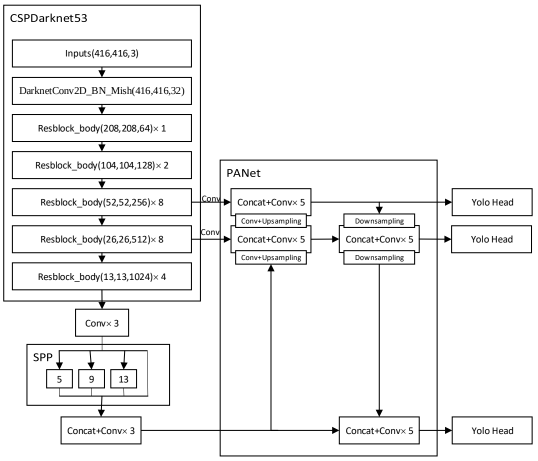

Figure 1. YOLOv4 Network Structure

CSPDarkNet53 integrates CSPNet [13] network on the basis of DarkNet53, reduces the disappearance gradient problem in deep network, and makes the structure lightweight and reduces the amount of data calculation. The feature pyramid structure adopts the path aggregation network with spatial pyramid pooling layer, which improves the problem of shallow information loss in the transmission process, increases the receptive field and improves the detection ability of the model. Yolo Head continues to use the method of YOLOv3 to extract three feature layers to complete the target detection task.

The input image is extracted from the CSPDarknet53 network, and the last three layers of semantic information are passed to the feature pyramid structure. After repeated feature extraction and feature fusion, the input image is fed to the Yolo Head network to generate the object category and the prediction boundary box to complete the target detection task.

In the task of traffic sign detection, the problem is that the image resolution is low, the background is complex, the noise is large and the information is small. Therefore, the feature information extraction in the target detection task is particularly important. This paper considers the fusion of shallow and deep semantic information to improve the effect of target detection. RepVGG [14] network model is used to replace the original CSPDarknet53 backbone network, and SPP module is added to the feature pyramid to improve the expression ability of information. The K-means++ clustering algorithm is used to process the experimental data to solve the problem that the original anchor frame scale is not suitable for the data set used in this paper and realize the demand of traffic sign detection.

3. IMPROVED YOLOV4 TARGET DETECTION ALGORITHM

3.1 RepVGG Feature Extraction Network

Since RepVGG has no complex branch structure and model design, and its performance is improved by re-parameterization, which is equivalent to the effect of multi-branch structure in accuracy and speed, we use a new network RepVGG to perform feature extraction task.

The reasoning time subject of RepVGG is affected by VGG [15], which is only composed of 3x3 convolution and ReLU activation function. The training time model is similar to ResNet using multi-branch structure. The structure reparameterization method is used to realize the conversion of training time and reasoning time model. The core structure of the model is shown in Figure 2.

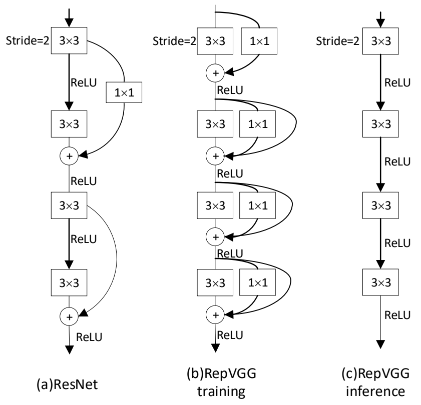

Figure 2. RepVGG Core Structure Diagram

The training time model uses identity structure and 1 × 1 branch structure block, and the information flow is y = x + g (x) + f (x). After the training is completed, it is equivalently converted to y = h (x). Finally, the training block is converted into a 3 × 3 convolution reasoning time block by using the structural reparameterization technology, as shown in Figure 3.

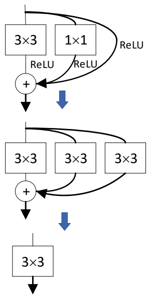

Figure 3. Reparameterization of RepVGG structure

$W^{(3)} \in R^{C_2 \times C_1 \times 3 \times 3}$ is used to represent the kernel of the 3×3convolution layer of the C_1 input channel and the C_2 output channel, and $W^{(1)} \in R^{C_2 \times C_1}$ is used to represent the kernel of the 1 × 1 branch. $\mu^{(3)}$, $\delta^{(3)}$, $\gamma^{(3)}$, and $\beta^{(3)}$ are used to represent the mean, standard deviation, learning factor and deviation of BN layer after 3 × 3 convolution, respectively. $\mu^{(1)}$, $\delta^{(1)}$, $\gamma^{(1)}$, and $\beta^{(1)}$ are 1 × 1 convolution branches, and $\mu^{(0)}$, $\delta^{(0)}$, $\gamma^{(0)}$, and $\beta^{(0)}$ are identity branches, respectively. Let $M^{(1)} \in R^{N \times C_1 \times H_1 \times W_1}$ and $M^{(2)} \in R^{N \times C_2 \times H_2 \times W_2}$ be input and output, respectively, and * be a convolution operator. If $C_1 = C_2, H_1 = H_2, W_1 = W_2$, there is :

$$\begin{aligned} M^{(2)} = {} & \mathrm{bn}\left(M^{(1)} * W^{(3)}, \mu^{(3)}, \delta^{(3)}, \gamma^{(3)}, \beta^{(3)}\right) \\ & + \mathrm{bn}\left(M^{(1)} * W^{(1)}, \mu^{(1)}, \delta^{(1)}, \gamma^{(1)}, \beta^{(1)}\right) \\ & + \mathrm{bn}\left(M^{(1)}, \mu^{(0)}, \delta^{(0)}, \gamma^{(0)}, \beta^{(0)}\right) \end{aligned} \qquad (1)$$

If you do not use an identity map, just use the first two of the equation. where bn is the inference time bn function :

$$W^{i}_{i,:,:,:} = \frac{\gamma_i}{\sigma_i} W_{i,:,:,:} \qquad (\ 2\ )$$

$$b'_i = -\frac{\mu_i \gamma_i}{\sigma_i} + \beta_i \qquad (\ 3\ )$$

And each BN and its front convolution layer are converted into a convolution layer with bias vector. Let$\{W', b'\}$ be the kernel and bias transformed from {W, μ, σ, γ, β}:

$$\mathrm{bn}(\mathrm{M}, \mu, \sigma, \gamma, \beta)_{:,i,:,:} = \left(M_{:,i,:,:} - \mu_i\right) \frac{\gamma_i}{\sigma_i} + \beta_i \quad (\ 4\ )$$

Using the above structure re-parameterization method, a 3 × 3 convolution is obtained for reasoning operation. By stacking the reasoning structure blocks, the main network of RepVGG reasoning can be constructed to complete the reasoning process.

3.2 SPP Embedded Characteristic Pyramid Structure

The semantic information contained in different feature layers is different, and the contribution to the output features after fusion is different. The feature extraction only in the output layer of the network will lead to the decline in the detection performance of small objects. Therefore, it is necessary to make full use of the semantic information at different levels to achieve the task of multi-scale detection.

YOLOv4 uses SPP and PANet structure in the feature pyramid, SPP module is shown in Figure 4.

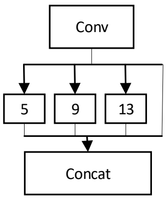

Figure 4. SPP structure

After the SPP module convolutions the output feature layer of the backbone network, four maximum pooling cores are respectively used for maximum pooling, and then the obtained different feature maps are channel spliced. The size of the output feature map, and the number of channels becomes four times that of the original. SPP structure can increase the receptive field of feature layer, capture effective context features, and improve the detection performance of the model.

The PANet structure is shown in Fig. 5. A bottom-up path aggregation network is added to the original structure of FPN, which further improves the detection effect. The three feature layers are extracted repeatedly through PANet network, and the more abstract top-level features are fully integrated with the underlying information. The feature maps of the same size generated in the forward propagation process are fused through horizontal connection, which makes full use of the feature layer information of different scales and effectively improves the target detection ability.

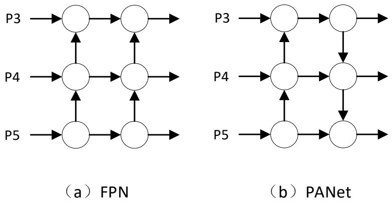

Figure 5. Characteristic pyramid structure

The YOLOv4 network sends the last layer of output to the SPP structure for up-sampling operation. In order to increase the multi-scale receptive field and improve the performance of the model, this paper adopts the SPP structure for the three-layer output on the basis of the above, and integrates it into the PANet network model to strengthen the expression ability of the output characteristic information. The structure is shown in Fig. 6.

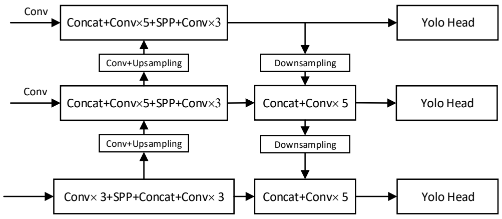

Figure 6. Improved structure diagram

3.3 Optimal anchor frame size acquisition

The Anchor box obtained by clustering can reduce the difficulty of target detection in prediction. The preset prior box of YOLOv4 network is obtained on PASCAL VOC dataset. However, for the self-set dataset, the use of the original preset Anchor box may make Yolo Head fail to select the appropriate target boundary box, which seriously affects the detection effect of the target. Therefore, this paper first clusters the real annotation boxes in the dataset.

In this paper, the K-means++ [16] clustering algorithm is used to cluster the boundary frame of the target in CCTSDB data set. Compared with the traditional K-means [17] clustering algorithm, the K-means++ clustering algorithm optimizes the selection of the initial point, and can select the optimal clustering center in the acquisition of the clustering center, effectively reducing the clustering deviation, so as to obtain a prior frame more suitable for the target data set and improve the accuracy of the target detection.

The K-means++ algorithm first randomly selects a sample point from the data set X as the initial clustering center, and then calculates the nearest distance between each sample point and the selected clustering center, and identifies it with D (x). Then the probability P (x) of each sample point selected as the next clustering center is calculated. Finally, the sample point corresponding to the maximum probability value is selected as the next clustering center.

$$\mathrm{P}(x)=\frac{D(x)^2}{\sum_{x\in X}D(x)^2} \qquad （5）$$

Repeat the above steps until K cluster centers are selected and K-Means algorithm is used to calculate the final clustering results of k cluster centers until the size of the Anchor box is no longer changed.

4. EXPERIMENT AND RESULT ANALYSIS

4.1 Experimental environment

The experimental platform is Windows10 (64bit) operating system, intel i7-8700 CPU, 16G memory, NVIDIA GeForce GTX 1060 6G memory, CUDA version 10.2, CUDNN7.5. Build a network model using the PyTorch framework.

4.2 Dataset processing

The dataset used in this paper is CCTSDB China Traffic Sign Detection Benchmark (CSUST Chinese Traffic Sign Detection Benchmark) [18]. The annotation data of CCTSDB dataset has three categories: indication mark, prohibition mark and warning mark, a total of 15723 pictures.

In the experiment, the original image in CCTSDB dataset was converted into jpg format, and the original label was converted into an xml file in VOC format suitable for YOLOv4 network, which was convenient to read the image annotation information. The proportion of training set and test set was 9 : 1.

4.3 Experimental parameters

In this experiment, YOLOv4 is used as the algorithm detection framework, and the pre-training weight is used as the basic feature extraction model by using the migration learning method. The momentum and weight attenuation are set to 0.9 and 0.0005, the batch size is set to 16, and the learning rate is 0.001. In order to train convergence, SGDM gradient optimization method is used, and the loss function is CIOU Loss.

4.4 Experimental results

Firstly, the K-means++ clustering algorithm is verified to optimize the detection accuracy of CCTSDB dataset targets. On the dataset of this paper, three classes and nine priori boxes are set up. The sizes obtained by K-means++ clustering algorithm are : (7,18), (9,24), (11,29), (13,36), (16,43), (18,27), (21,55), (30,44), (52,81). Under the same parameter settings, the average accuracy of the anchor box obtained by clustering is 1.1 % higher than that of YOLOv4 algorithm. Therefore, the target box obtained by K-means++ clustering algorithm is easier to fit the real target and obtain better detection results. As shown in table 1:

Table 1. Comparison of optimization results

Model	Map(%)	FPS
Improved ago	94.5	19.5
The improved	95.6	20

YOLOv4-R using RepVGG as the backbone network and the YOLOv4-S based on the improved feature pyramid structure are trained and tested on the CCTSDB dataset, and compared with the original target detection network. The experimental results are as follows.

Table 2. Performance comparison of different improved algorithms

Model	Map(%)	FPS
YOLOv4	95.6	20
YOLOv4-R	96.7	19
YOLOv4-S	97.9	20

It can be seen from table 2 that the performance of the new target detection network is improved compared with the original network.

The YOLOv4 detection algorithm using RepVGG as the backbone network YOLOv4-R improves the detection accuracy by 1.1 % compared with the original network, indicating that RepVGG is simple in structure but powerful in performance. As a backbone network, RepVGG can effectively complete the feature extraction task, improve the accuracy of target detection, and the frame rate is not significantly reduced.

Compared with the original network, the average accuracy of the YOLOv4-S detection algorithm based on the improved feature pyramid structure is increased by 2.3 %. It shows that adding SPP network in the feature pyramid structure can increase the receptive field of the feature layer, capture effective context features, and enrich feature information through feature fusion at different scales to improve the detection performance of the model.

In order to further verify the effectiveness of the improved algorithm, the network model integrating all the above improved methods is trained and verified on the CCTSDB dataset. The test results are shown in Table 3.

Table 3. Improved model detection comparison

Model	Map(%)	FPS
Improved ago	94.5	19.5
The improved	98.6	20

Compared with the experimental results of table 3, this paper proposes an improved YOLOv4 target detection algorithm, and its average accuracy is increased by 4.1 %, which verifies the effectiveness of the improved algorithm and shows the feasibility of the improved algorithm, which can meet the performance requirements of current traffic sign detection.

The accuracy and recall rate of the improved network model are improved compared with the original network. The P-R curve comparison diagram is shown in Figure 7:

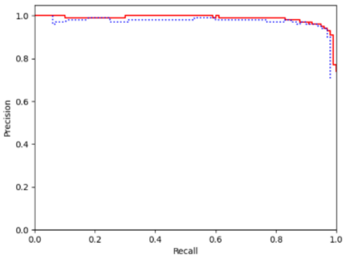

Figure 7. P-R curve comparison diagram

Test results are shown in Figure 8：

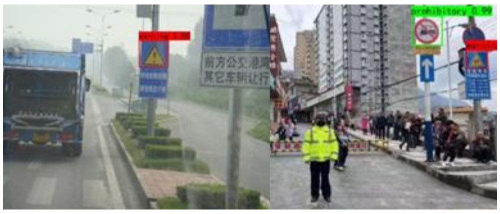

Figure 8. Detection effect diagram

The improved traffic sign recognition effect is more accurate than the original network positioning, the probability of missed detection and false detection is lower, and the recognition accuracy is higher, which effectively meets the detection requirements of traffic scenes and verifies the effectiveness of the improved algorithm.

5. CONCLUSIONS

This paper proposes an improved target detection algorithm based on YOLOv4, and verifies the effectiveness of the improved algorithm in CCTSDB dataset. The YOLOv4 detection algorithm using RepVGG as the backbone network has a reasonable trade-off between depth, accuracy and speed. Based on the improved feature pyramid structure, the expression ability of information is effectively improved. Finally, the boundary frame of CCTSDB dataset obtained by K-means++ clustering algorithm is more likely to fit the real target and obtain better detection results. The results show that the proposed network model effectively improves the detection performance and meets the detection requirements of traffic scenes.

6. REFERENCES

[1] Li H, Sun F. Liu L, et al.A novel traffic sign detection method via color segmentation and robust shape matching[J].Neurocomputing, 2015, 169:77-88.. .

[2] Berkaya S K, Gunduz H, Ozsen O, et al.On circular traffic sign detection and recognition[J].Expert Systems with Applications, 2015,48:67-75.

[3] Girshick R, Donahue I, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C].Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.

[4] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2015: 1440-1448.

[5] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards realtime object detection with region proposal networks[C].Advances in neural information processing systems, 2015: 9199.

[6] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. European conference on computer vision. Springer,Cham, 2016: 21-37.

[7] Redmon I, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.

[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. Computer Vision and Pattern Recognition, 2018, 22(4): 17-23.

[10] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[C]. Computer Vision and Pattern Recognition, 2020, 17(9): 198-215.

[11] Ghiasi G, Lin T Y, Le Q V. Dropblock: A regularization . method for convolutional networks[C]. Advances in Neural Information Processing Systems, 2018: 10727-10737.

[12] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[13] Wang C Y, Mark Liao H Y, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of cnn[C].Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 390-391.

[14] Xiaohan Ding, Xiangyu Zhang,Ningning Ma.RepVGG: Making VGG-style ConvNets Great Again[C]. Computer Vision and Pattern Recognition,2021.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 1, 2, 6.

[16] ARTHUR D. VASSILVITSKII S. K-Means ++: The Advantages of Carefull Seeding[J]. Proceedings of Theghteenth Annual Acm Siam Symposiumon Discrete Algorithms Society for Industrial & Applied Mathematics,2007, 11(6): 1027-1035.

[17] Luo Xiaoquan, Pan Shanliang, improved YOLOV3 fire detection method[J]. Computer engineering and application, 2020, 56 (17) : 187-196.

[18] Zhang J, Huang M, Jin X, et al. A Real-Time Chinese Traffic Sign Detection Algorithm Based on Modified YOLOv2. Algorithms, 2017, 10(4):127.