Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090316
ISSN 2371-9524 Vol. 9 Num. 3

Road sign detection algorithm based on improved YOLOV4

Xude Zhang*

Micro-nano and Intelligent Manufacturing Engineering Research Centre of Ministry of Education, Kaili University, Kaili, China
**Corresponding author: zhangxude@kluniv.edu.cn*

Keywords: Target detection, Depth separable convolution, Attention mechanism

Abstract: The detection of traffic signs is an important part of the research on automatic driving. Road traffic signs occupy the edge of the image, the image is small, and the detection accuracy is low. The improved YOLOv4 target detection algorithm is used to detect road traffic signs. The original activation function is modified to the h-swish activation function. The input image is convolved by 1x1 to obtain the image feature concentration. The main feature extraction network adds depth separable convolution and residual edge parts, and introduces attention mechanism to enhance the feature extraction performance. The road sign prior frame is regenerated using K-means clustering algorithm, The clustering algorithm can achieve network convergence. After the test, it is shown that by training and evaluating the CCTSDB dataset, MAP@0.5 83.47%, 2.78% higher than the original YOLOv4; The parameter quantity of the network model is 45.60M, which is 18.5% of the size of the original YOLOv4 model. The network becomes lightweight, and the target detection of road signs can be well achieved through testing.

1. Introduction

With the continuous development of intelligent transportation technology, road traffic signs have gradually entered the public eye as the most important component of autonomous driving. The traditional machine learning used for road signs is based on color discrimination and image segmentation. However, traditional recognition methods have problems such as low recognition accuracy and efficiency [1-3]. In recent years, with the formation of computer hardware computing power and the maturity of deep learning CNN networks, there are higher requirements for real-time and accuracy of object detection algorithms. Deep learning CNN networks have become a research hotspot in the field of computer vision [4]. Object detection algorithms are divided into two-stage object detection algorithms [5] and one-stage object detection algorithms. Among them, the two stage algorithm has high accuracy, but the training speed is slow and the network is relatively complex. The One stage algorithm has fast speed and relatively simple network training. Although the training accuracy is reduced, it can meet the requirements well. The One stage algorithm includes YOLO [6] series algorithms, SSD [7] algorithms.

2. YOLOv4 Object Detection Algorithm

The YOLOv4 algorithm is an improved version of YOLOv3. The YOLOv4 backbone feature extraction network CSPDarknet53 uses CSPNet [8] based on Darknet53 to enhance network learning capabilities, ensure accuracy, and reduce memory consumption and computational costs. The Mish activation function is used as the activation function, which is smoother in the negative value region compared to the LeakyReLU activation function [9], which is beneficial for gradient calculation and updating, and can achieve better accuracy and generalization ability. The backbone network utilizes different pooling kernels through spatial pyramid pooling to effectively fuse shallow detail information and deep semantic information, increasing the receptive field. The path aggregation network consists of two parts: the feature pyramid network FPN and the path enhancement network PAN. FPN transmits deep semantic information, while PAN transmits shallow localization information, enhancing the network's representational ability. The detection head adopts the Head [10] from YOLOv3, consisting of 3 × 3 and 1 × 1 convolutional layers. The PAN processed output is used for result prediction. The YOLOv4 network structure is shown in Figure 1. The CSP-Residual Block in the backbone network uses CSP structure and Darknet53, which enhances the information extraction ability compared to the original YOLOv3. The other part is the residual edge [11], which is directly connected to the end after a small amount of processing, which can enhance the learning ability of CNN and effectively improve accuracy to a certain extent while meeting the requirements of lightweight, low computational complexity, and low memory access cost.

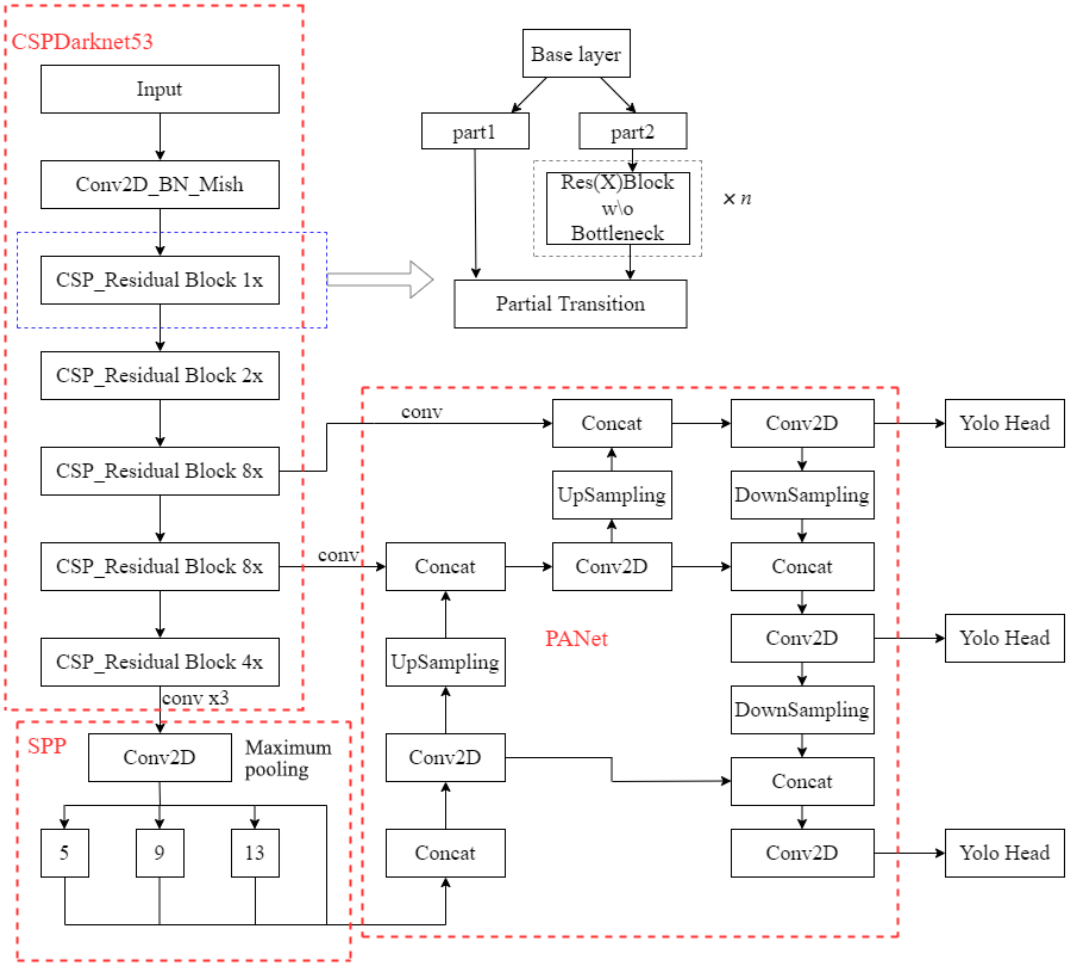

Figure 1: YOLOv4 Network Architecture Diagram

3. Improved YOLOv4 algorithm

3.1. Improvement of Backbone Feature Extraction Network

The YOLOv4 backbone feature extraction network CSPDarknet53 uses stacking of residual blocks and residual edges to enhance the learning ability of CNN and meet the requirements of feature extraction. However, there are high computational requirements that are not conducive to lightweight and low computational deployment at the edge. The feature extraction network generates redundant feature maps during the convolution process. The improvement of the backbone feature extraction network should reduce the number of model parameters and improve the execution speed of the model while meeting the detection performance requirements [12].

The improved form of CSP_Sesidual Block in YOLOv4 is that the feature map is convolved by 1x1, and the output channel is half of the input channel. The output layer is then passed through the BN layer and ReLU activation function, and feature concentration can be achieved through this convolution; Perform depthwise separable layer by layer convolution. For a step size of 1, add attention mechanism output directly. For a stride of 2, the output is obtained through depthwise separable convolution, and then concatenated through 1x1 convolution before being output through attention mechanism. The output result is subjected to 1x1 convolution and depth wise separable layer by layer convolution, and finally the output and input are connected by residual connection; The improved Block structure is shown in Figure 2.

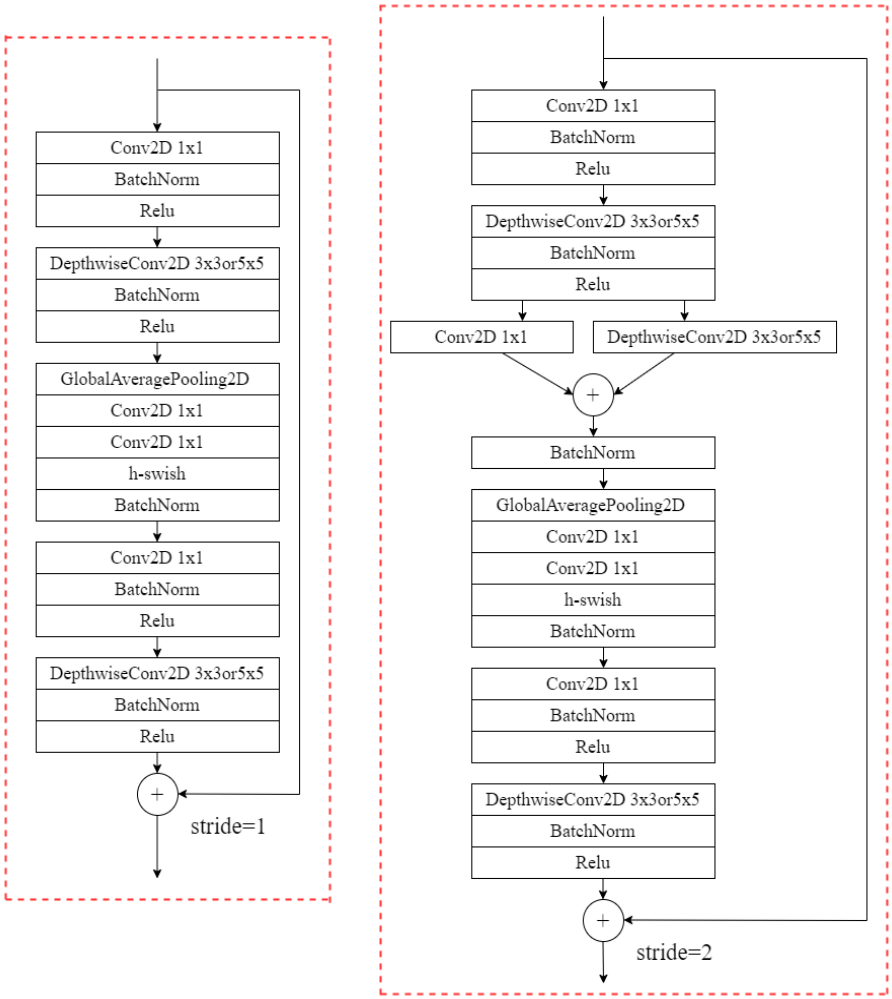

Figure 2: Improved New Block Structure

The improved YOLOv4 network includes 16 layers of NewBlocks [13], and the backbone feature extraction network reduces 7 layers compared to the original CSP_SesidualBlock. The original backbone feature extraction network of YOLOv4 is CSPDarknet-53. The three feature layers output by the backbone feature extraction network are used as inputs for SPP and PANet prediction networks. The three feature layers are the results of downsampling the length and width of the input image by 8 times, 16 times, and 32 times, respectively. Based on the improved YOLOv4 network, it is also necessary to find a new backbone feature extraction network to extract effective feature layers for 8x, 16x, and 32x downsampling. The result of 8x downsampling of the input image is the output of Block3, which is used as the first layer output out1 of the improved feature extraction network. The result of 16x downsampling of the input image is the output of Block4, which is used as the second layer output out2 of the improved feature extraction network. The result of 32x downsampling of the input image is the output of Block5, which is used as the third layer output out3 of the improved feature extraction network.

3.2. Improvement of activation function

The activation function in deep neural convolutional networks is used to incorporate nonlinear factors to enhance the expressive power of the model. The improved YOLOv4 used in this study adopts the h-swift activation function, which is calculated as shown in formula 1

$$h-swish[x]=x\frac{RELU6(x+3)}{6} \tag{1}$$

The activation function h-flash image is shown in Figure 3. When the value is negative, the gradient changes smoothly. Using the h-flash activation function can reduce the number of memory accesses, thereby significantly reducing latency costs and improving performance.

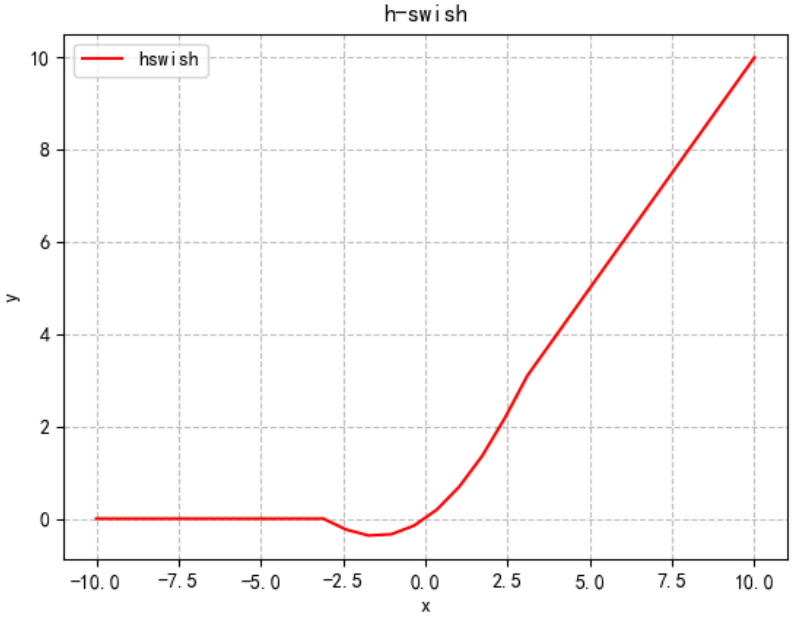

Figure 3: h-swift activation function image

3.3. K-means clustering generates prior boxes

The YOLO series object detection is based on anchors, which require pre-set boxes of different sizes and aspect ratios on the image in advance to make the model easier to learn. By using different

sizes and aspect ratios, a larger intersection to union ratio can be obtained, which can detect a higher probability of occurrence. A prior box with good matching for the target object can help determine the width and height of common targets. When making predictions, using this already determined width and height processing can help us make predictions. Clustering research on the width and height of the annotated dataset can be achieved through the k-means clustering algorithm, which can calculate prior boxes with good matching degree for the target object. This study uses the CCTSDB dataset for training and updates the prior boxes using the K-means clustering algorithm to improve localization accuracy [14].

This study used K-means clustering algorithm to regenerate 9 new anchor box parameters, namely (7, 20), (9, 15), (13, 20), (10, 28), (18, 28), (14, 39), (26, 39), (36, 55), and (62, 87), for traffic sign detection. Figure 4 shows the clusters of road signs obtained using K-means clustering, which are composed of different colors.

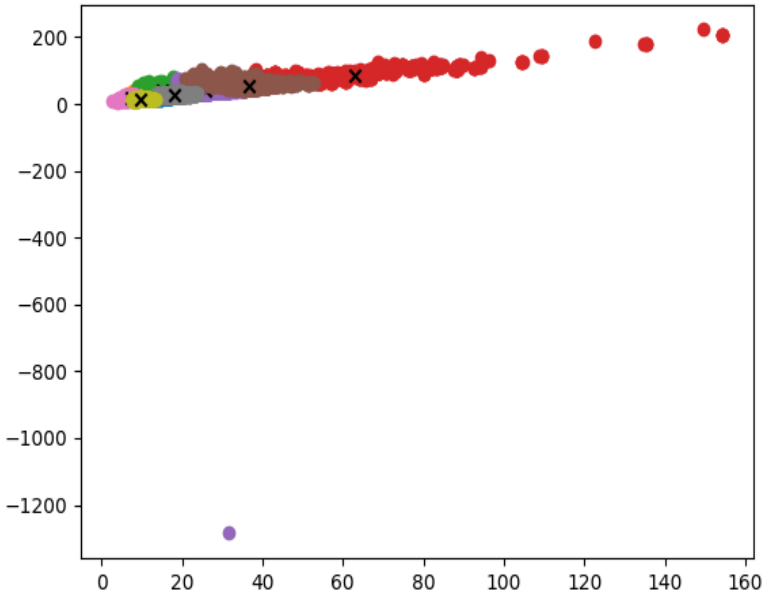

Figure 4: Distribution of road sign annotation boxes

4. Road Sign Detection Experiment and Analysis

4.1. Experimental environment configuration

The experimental environment for object detection based on road signs is Windows 10. During training, the GPU used is NVIDIA GeForce RTX 2080Ti, with a graphics card memory of 11GB. The software environment is PyCharm script editor, and the learning framework is PyTorch.

4.2. Road Sign Dataset

The selection of the dataset should be based on characteristics such as balanced categories, widespread usage scenarios, and large data volume. The Chinese traffic sign detection dataset CCTSDB contains 15724 images of various resolutions collected from urban roads and highways [15]. Some of the datasets in the experiment are shown in Figure 5.

Figure 5: Partial dataset images

The proportion of labeled images in the dataset is relatively small, and the images contain a large amount of irrelevant background information, making image detection a significant challenge. Due to the large number of photos in the dataset, in order to save training time, 5000 images were randomly selected and divided into three categories for training.

4.3. Model training and evaluation metrics

The training process of the improved YOLOv4 is divided into 100 epochs, with Batchsize set to 16. Mosaic is used for data augmentation during the training process, and Focal Loss is used to control the weights of positive and negative samples. When detecting road signs, the learning rate of hyperparameters is set to 0.01, and cosine annealing is used to attenuate the learning rate during training.

The improved YOLOv4 training model uses detection target accuracy P (Precision), detection target standard recall R (Recall), average accuracy AP (Average Precision), mean average precision mAP (Mean Average Precision), and model parameter size to determine the performance of the model before and after improvement.

4.4. Result and Analysis

Using an improved YOLOv4 network model for training and validation on the CCTSDB dataset, the detection targets are divided into three categories: prohibitory, warning, and mandatory.

When using the improved YOLOv4 for detection on the Chinese traffic sign detection dataset, a comparison of the detection performance of different networks using the CCTSDB dataset is shown in Table 1. The accuracy P and recall R are evaluated using mAP with an IOU of 0.5 as the evaluation criteria.

Table 1 shows the parameter performance of our algorithm, YOLOv4 algorithm, Resnet50 YOLOv4 network algorithm, and SSD algorithm. By comparing and analyzing the target detection accuracy P, recall R, average accuracy AP, average accuracy mean mAP, and model size, the improved YOLOv4 algorithm achieved an accuracy of 83.47% in detecting road signs, with a model parameter size of 45.60M, making it more lightweight. Compared with the original YOLOv4 network model, our algorithm improved the detection accuracy by 2.78%, and reduced the model size by 199.93MB. Compared with the SSD model and Resnet50 YOLOv4 model, the detection accuracy of our algorithm also increased by 6.64% and 3.91%, respectively. According to Table 1 analysis, the improved YOLOv4 network shows a significant improvement in object detection performance, with a reduction in the number of network model parameters and an improvement in

object detection accuracy. It can better meet the requirements of lightweight and real-time traffic sign detection.

Table 1: The effectiveness of different networks in CCTSDB detection

algorithm	category	Precision (%)	Recall (%)	AP (%)	MAP (%)	Parameter quantity (M)
	Prohibitory	94.78	29.62	71.98		
SSD	warning	95.16	35.98	80.42	**76.83**	**100.27**
	Mandatory	97.30	16.82	78.10		
	Prohibitory	89.67	73.10	85.94		
YOLOv4	warning	92.31	65.85	87.41	**80.69**	**245.53**
	Mandatory	84.31	40.19	68.73		
	Prohibitory	87.73	64.13	80.80		
Resnet50-YOLOv4	warning	86.61	67.07	78.51	**79.56**	**128.48**
	Mandatory	85.62	58.41	79.37		
	Prohibitory	91.72	72.28	82.76		
This article's algorithm	warning	94.31	70.73	86.78	**83.47**	**45.60**
	Mandatory	89.44	59.35	80.86		

5. Conclusion

Based on the YOLOv4 algorithm, we aim to address the requirements of road traffic signs occupying adjacent positions, being small in size, and having high detection difficulty in images. Propose an improved YOLOv4 object detection algorithm, which reduces the generation of redundant feature maps through 1x1 convolution and depthwise separable convolution in the backbone network. The detection accuracy of the model is improved by introducing attention mechanism and residual module. After training CCTSDB, the YOLOv4 detection accuracy is improved by 2.78%, and the model size accounts for 18.5% of the original size, making it easier to meet the requirements of lightweight use.

Acknowledgements

Foundation of Micro-nano and Intelligent Manufacturing Engineering Research Centre of Ministry of Educa-tion (No.2024WZG03); Supported by the Foundation Research Project of Kaili University (No.2024YB010).

References

[1] Pan Huiping, Wang Minqin, Zhang Fuquan. Traffic sign detection and recognition method based on optimized YOLO-V4 [J]. Computer Science 2022, 49(11): 179-184.

[2] Wang Jingyi, Liu Shuhui. Traffic Sign Recognition Method Based on Improved YOLOv4 [J]. Electronic Design Engineering 2022, 30(18): 184-188.

[3] Shen Zhi, Xu Li, Fu Xiangyuan. Traffic sign detection based on improved YOLO v4 light blur scene [J]. Computers and Modernization 2022, (07): 27-32

[4] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of

the IEEE, 1998, 86(11): 2278-2324.
[5] GIRSHICK R, DONHUE, 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.
[6] JOSEPH REDMON, SANTOSH DIVVALA, ROSS GIRSHICK, ALI FARHADI. You Only Look Once: Unified, Real-Time Object Detection. [C]Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779-788.
[7] WEI LIU, DRAGOMIR ANGUELOV, DUMITRU ERHAN, CHRISTIAN SZEGEDY, SCOTT REED, CHENGYANG FU, ALEXANDER SSD: Single Shot MultiBox Detector [C], ECCV 2016: Computer Vision-ECCV 2016 pp 21-37.
[8] CHIEN-YAO WANG, HONG-YUAN MARK LIAO, I-HAU YEH, YUEN-HUA WU, PING-YANG CHEN, JUN-WEI HSIEH. 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, 2019.
[9] Gezgin H , Alkan R M .Traffic sign detection and recognition based on MMS data using YOLOv4-Tiny algorithm[J].Neural Computing and Applications, 2024, 36(33):20633-20651.DOI:10.1007/s00521-024-10279-y.
[10] JOSEPH REDMON, ALI FARHADI.YOLOv3: An Incremental Improvement. [C].Computer Vision and Pattern Recognition, 2018.
[11] KAIMING HE, XIANGYU ZHANG, SHAOQING REN.Deep Residual Learning for Image Recognition. [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp.770-778.
[12] KAI HAN, YUNHE WANG, QI TIAN, JIANYUAN GUO, CHUNJING XU, CHANG XU. GhostNet: More Features from Cheap Operations, [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp.1580-1589.
[13] Yin Songlin, Tan Fei, Zhou Qing, Xianyang. Traffic Sign Detection Based on Improved YOLOv4 Model [J]. Radio Engineering, 2022, 2022, 52 (11): 2087-2093.
[14] Wang Zehua, Song Weihu, Wu Jianhua A Lightweight Traffic Sign Detection Model Based on Improved YOLOv4 Network [J] Computer Knowledge and Technology, 2022,18 (05): 98-101+104.
[15] JIANMING ZHANG, ZHIPENG XIE, JUAN SUN, XINZOU, JIN WANG. A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection[C]. IEEE Access, 2020, vol.8, pp.29742-29754.