Improved YOLO v5 with balanced feature pyramid and attention module for traffic sign detection

Linfeng Jiang[*], *Hui* Liu, *Hong* Zhu, and *Guangjian* Zhang

School of Artificial Intelligence, Liangjiang, Chongqing University of Technology, Chongqing, China

Abstract. With the development of automatic driving technology, traffic sign detection has become a very important task. However, it is a challenging task because of the complex traffic sign scene and the small size of the target. In recent years, a number of convolutional neural network (CNN) based object detection methods have brought great progress to traffic sign detection. Considering the still high false detection rate, as well as the high time overhead and computational overhead, the effect is not satisfactory. Therefore, we employ lightweight network model YOLO v5 (You Only Look Once) as our work foundation. In this paper, we propose an improved YOLO v5 method by using balances feature pyramid structure and global context block to enhance the ability of feature fusion and feature extraction. To verify our proposed method, we have conducted a lot of comparative experiments on the challenging dataset Tsinghua-Tencent-100K (TT100K). The experimental results demonstrate that the mAP@.5 and mAP@.5:0.95 are improved by 1.9% and 2.1%, respectively.

Keywords: Traffic sign detection, Convolutional neural network, Feature fusion.

Introduction

Since the rapid development of automatic driving technology, great changes have taken place in people's daily life. At the same time, a large number of new technological developments are in an urgent demand. Traffic sign detection is one of them. The mission of traffic sign detection is to locate traffic signs from given pictures or videos, and then predict the category information of traffic signs correctly. However, it is still a challenging task due to the complex background, various kinds of shapes, together with the shelter of the trees.

With the evolution of deep learning, many people attempt to use convolutional neural network (CNN) based object detection methods to detect traffic sign, such as YOLO [1] and SSD [2] (Single Shot MultiBox Detector). However, the results they achieved are not

[*] Corresponding author: linfengjiang@cqut.edu.cn

very satisfactory. Besides the accuracy of detection, the models they proposed are always complex, together with a large number of parameters, which is computationally expensive, and is hard to be embedded into the automatic driving terminal.

To get higher prediction speed with lightweight network model, YOLO v5 is employed as the baseline method, which is one of the best object detection methods with excellent detection accuracy and very low time complexity, especially suitable for intelligent driving. Additionally, for a higher detection precision, we have made some improvement on YOLO v5. The proposed method mainly makes the following contributions: (i) Balance feature pyramid structure is used to improve our model, which can enhance the ability of feature fusion, so that both semantic information and position information of traffic sign in small size are taken into account. (ii) Attention module is also added in our model to help feature extraction.

The experimental results show that the accuracy and recall of our model are obviously improved over the baseline method YOLO v5, mAP@.5 and mAP@.5:0.95 are improved by 1.9% and 2.1%, respectively.

The rest of the paper is organized as follows. Section 2 introduces some related work in traffic sign detection. Section 3 gives the proposed method. Section 4 gives the results of our methods and the comparison with other methods. The last section gives the conclusion.

2 Related work

In this section, we briefly review the methods that can be used for traffic sign detection. The study of traffic sign detection can be divided into: traditional approach and deep learning based method.

2.1 Traditional approach

The detection methods based on traditional methods are mainly based on the physical characteristics of targets, the most common methods are color based detection algorithm and shape based detection algorithm. The shape based methods are always based on Hough Transform. Besides, the color based detection method obtains the color information points from the image, then connects them into regions, and finally obtains the interested region.

2.2 Deep learning based method

Object detection methods based on deep learning can be divided into two categories: one-stage methods and two-stage methods. Classic algorithms like SSD, YOLO, and RetinaNet belong to one-stage methods. Some other methods like Fast R-CNN, Faster R-CNN are the symbols of two-stage methods. Usually, two-stage methods are excellent in accuracy, while one-stage methods are better in speed. In this section, we will briefly introduce some classic object detection algorithm.

YOLO [1] was proposed by R. Joseph et al. in 2015. It is the first one-stage object detection algorithm. One-stage object detection methods do not have the process of classification on the region proposal, but directly regresses the output category. YOLO is well known for its accuracy, together with the extremely speed, which is one of the most commonly used algorithms in industrial circle. The core idea of YOLO is to transform the object detection into a regression problem. It feeds pictures into a neural network, and then outputs the bounding boxes and categories of objects directly. Later, Yolo is continuously optimized and improved. Thus, YOLO v2, v3, v4, v5 were proposed. In particular, YOLO v5 greatly improves the accuracy and reduces the size of the model, which will be introduced in the next section. SSD [2] is another famous one-stage object detection methods. The main contribution of SSD is using small convolutional filters to predict category information and box offset. SSD is superior to the first version YOLO in both accuracy and speed.

Two-stage method is another technology road map for object detection. Different from one-stage methods, two-stage methods extract the depth features of images through backbone network, and then generate region proposal through RPN network. Finally, it determines the class information through two branches of classification and regression. In 2013, Ross et al. proposed RCNN [3] network. It is one of the earliest object detection methods based on deep learning. It made a breakthrough in object detection, and achieved 58. 5% mAP in Pascal VOC 2007 dataset, while DPM [4] only get a mAP of 34. 3% in the same dataset. To improve RNN, Ross g et al. proposed Fast RCNN [5] in 2015. They improved RCNN by inserting SPP-Net module, and use VGG 16 as its backbone. Thus, it gets a better detection accuracy. Later, Faster RCNN [6] occurs in June, 2015. It is the first end-to-end deep learning based object detection method. It introduces Region Proposal Network (RPN) and breakthrough the speed limit of two-stage methods and make a great improvement in detection results. In the next few years, a lot of object detection methods appear, promoting the development of object detection technology.

3 Proposed method

3.1 Brief introduction to YOLOv5

As introduced in section 2.2, YOLO v5 is the 5th generation of YOLO. It is famous for its detection accuracy and prediction speed. As shown in Figure 1, YOLO v5 has a simple network structure, consisting of input, backbone, neck and prediction.

A) Input: As YOLO v4 did, YOLO v5 adds Mosaic data augmentation method to the training pictures. Through random scaling, random cutting, and random layout, four different pictures are mixed into one picture. By these means, the background information of training image is enriched, which is very beneficial to small target detection. Additionally, when calculating batch normalization, the data of four pictures are calculated at one time, therefore the mini batch size does not need to be very large, especially suitable for single GPU training.

B) Backbone: The backbone of YOLO v5 is the combination of focus module and CSP darknet53 structure. Focus module slices one data information into four, and then enate them on channel dimension. This module is designed for reducing FLOPS and increasing speed, rather than mAP increase. CSPDarknet53 contains 29 convolutional layers, and a 725 × 725 receptive field. Its ability of feature fusion is much better than original Darknet53.

C) Neck: Its neck combines Feature Pyramid Networks (FPN) and Path Aggregation Network (PAN). It includes four connection layers, four convolution layers, and five CSP layers. It can speed up the transmission of feature information and feature fusion.

D) Prediction: Given the size of input is 640*640*3, by feature partitioning, three outputs with sizes of 20 * 20 * 255, 40 *40* 255 and 80*80*255 are produced. They are used for detection of different sizes.

3.2 Improved detection model

YOLO has achieved great results in object detection. However, in order to get more feature information, YOLO v5 uses 8 times, 16 times, and 32 times down samplings to detect objects of different sizes respectively. Thus, a large numbers of position information has been lost. This makes it difficult to detect small objects.

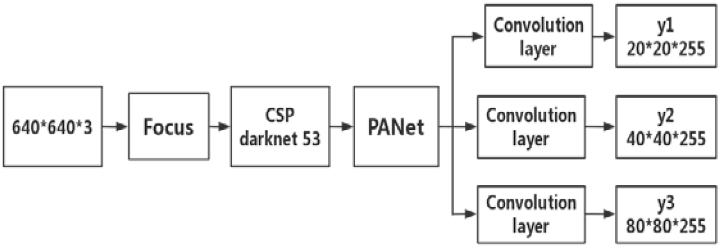

Fig. 1. YOLO v5 network structure.

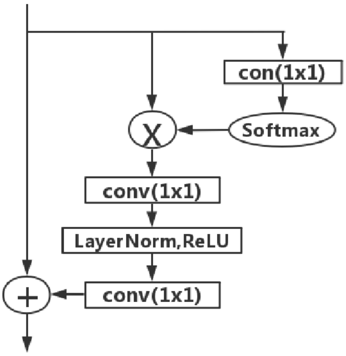

Fig. 2. Global context (GC) block.

In order to obtain multi-scale information, giving consideration to both semantic information and position information of small targets, we fuse the feature information of three different scales. The three outputs of YOLO v5 (y1, y2 and y3) derive from down sampling of different depths. They possess different semantic information and position information. Considering that majority target in our dataset are in a small size, motivated by Libra R-CNN [11], we use balanced feature pyramid to improve our model. As Figure 3 shows, we first operate up sampling and down sampling on y1 and y3 respectively, afterwards cat them in the channel dimension. For a better feature extraction, we embed the attention module. We use GC block to further refine the network. The architecture of GC block shows below in Figure 2. It can capture inter channel dependencies, so that beneficial to feature fusion. We then output the new y1', y2' and y3' with the operation of up sampling, down sampling and 1*1 convolution, respectively.

In order to verify our method, we conducted experiments on datasets. However, a new problem occurred. The convergence rate of the new model became very slow, and the optimal value was hard to obtained. Based on the idea of Resnet [7] , we further optimized our model. We enate the originate outputs y1, y2 and y3 with the new outputs y1', y2' and y3'. By these means, a) we protect the originate feature information, and fuse it with the new feature information. b) our model can promote optimization, speed up convergence and prevent the situation of no convergence.

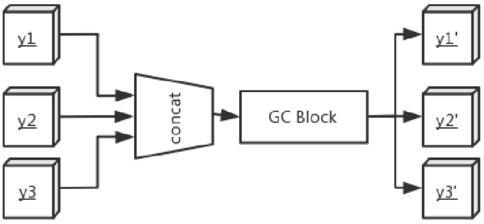

Fig. 3. Improved module.

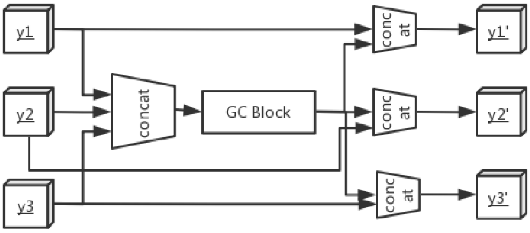

Fig. 4. Further improved module.

4 Experiments& results

4.1 Datasets & equipment

In order to verify our proposed method, we have conducted a lot of comparative experiments on the challenging dataset Tsinghua-Tencent-100K [7] (TT100K). The TT100k dataset contains almost 10, 000 pictures, and all of them are in the size of 2048*2048 pixels, while majority traffic sign are in a small size. In order to achieve better training and prediction results, we selected those 45 classes with more than 100 samples.

The experiments were run on a computer with Intel(R) Xeon (R) CPU, 32GB main memory and one Nvidia Quadro RTX5000 GPU with 16GB memory. The implementation environment is under the Pytorch1. 8. 1.

4.2 Results

For evaluating the effect of proposed model, we use recall, precision and mAP to quantitatively analyze our model. First, we compare improved YOLO v5 with the original YOLO v5. We list the comparative data in table 1.

Table 1. Comparison with the original method.

	P	**R**	**mAP@. 5**	**mAP@. 5:0. 95**
YOLOv5	0.85	0.828	87.8%	67.2%
Ours	0.874	0.861	89.7%	69.3%

From table 1, we can conclude that compared with original YOLO v5 network, the precision of our model increases by 2.4%, and the recall increases by 3.3%. Also, mAP@.5and mAP@. 5:0.95 increase by 1.9%, 2.1%, respectively.

In addition, we compare our methods with the state-of-the-art object detection methods. The results are shown in table 2.

Table 2. Comparison with other methods.

	SSD300 [10]	**Faster RCNN [9]**	**YOLOv3**	**YOLOv4**	**YOLOv5**	**Ours**
mAP@.5	63.71%	79.1%	82.4%	86.8%	87.8%	89.7%

In table 2, our method gets the best results on the TT100K dataset. Therefore, we can conclude that our method is effective on the dataset, and has made progress compared with the original method.

5 Conclusion

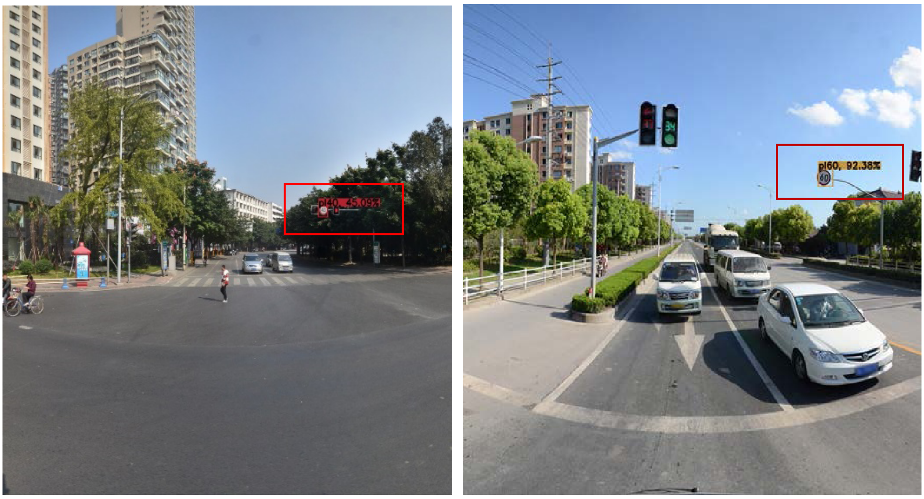

Fig. 5. Detection results in TT100K dataset. The detection results of small targets are marked with red boxes.

In this paper, we proposed an improved YOLO v5 to solve the problems existing in traffic sign detection. Aiming at traffic sign in small size, we chose the TT100K dataset. By comparing it with state-of-the-art methods mentioned above, our methods are better in accuracy. In the future, we will keep attempting to modify our model, so as to more suitable for automatic driving.

References

1. Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. Computer Vision & Pattern Recognition, 2016.
2. Liu W, Anguelov D, Erhan D , et al. SSD: Single Shot MultiBox Detector[J]. European Conference on Computer Vision, 2016.
3. Girshick R, Donahue J, 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. Felzenszwalb P F, Mcallester D A, Ramanan D . A discriminatively trained, multiscale, deformable part model[C]// 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008.
5. Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
6. 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 & Machine Intelligence, 2017, 39(6):1137-1149.
7. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
8. Zhe Z, Liang D, Zhang S, et al. Traffic-Sign Detection and Classification in the Wild[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.
9. Yao Z, Song X, Zhao L, et al. Realtime method for traffic sign detection and recognition based on YOLOv3tiny with multiscale feature extraction[J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2021, 235(7): 1978-1991.
10. Pan W, Liu B, Chen Y, et al. Traffic sign detection and recognition based on YOLO v3[J]. Transducer and Microsystem Technologies, 2019.
11. Pang J, Chen K, Shi J, et al. Libra R-CNN: Towards Balanced Learning for Object Detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.