A Fast Traffic Sign Detection Algorithm Based on Modified YOLOv3

Huazhou Dou\textsuperscript{1,a,*}, Haiyun Zhang\textsuperscript{1,b} and Bo Li\textsuperscript{1,c}

\textsuperscript{1} Xi’an Jiaotong University, Xi’an 710000, China

\textsuperscript{a,*}huazhoudou@stu.xjtu.edu.cn;\textsuperscript{b}haiyunzhang@stu.xjtu.edu.cn;\textsuperscript{c}bolee@xjtu.edu.cn

Abstract. Traffic sign detection is an important component of both automatic driving systems and driver assistance systems. However, it is still a big challenge to detect traffic signs quickly and accurately. In this paper, in order to solve this problem, a lightweight algorithm based on YOLOv3 is proposed, which can achieve real-time and accurate traffic sign detection. Firstly, a new lightweight backbone network is proposed by modifying the darknet53 network, which is the original backbone network of YOLOv3. Secondly, in order to further speed up the detection time, depthwise separable convolution is introduced into the proposed backbone network to further reduce the amount of parameters and computations of the model. Thirdly, dense connection is adopted to maintain the accuracy. Experiments on CCTSDB dataset illustrate that the algorithm has great real-time performance and accuracy, which is better than other state-of-the-art algorithms. The algorithm can detect a 416*416 image in only 3.4 milliseconds, which is 44.3% faster than YOLOv3. In addition, the mAP and F1 value are 97.2% and 91.6% respectively, which are better than YOLOv3.

1. Introduction

With the rapid increase in the number of vehicles, people pay more and more attention to traffic safety. As an important part of both automatic driving systems and driver assistance systems, the traffic sign detection system can provide the driver and the automatic driving system with warnings and instructions regarding traffic signs to reduce dangerous driving behaviours and make correct decisions.

Many traffic sign detection approaches have been proposed, which can be divided into two categories, one is the traditional approach and the other is the deep learning based approach. Traditional methods generally contain three steps: obtaining region proposals containing traffic signs, extracting manual features from these regions (such as HOG, SIFT, etc.), then sending these features to the classifier (such as SVM, Adaboost, etc.) [1]-[3]. Although traditional methods can achieve high accuracy, it is complicated to extract hand-craft features and these features cannot represent object well. In addition, traditional methods are time-consuming and not robust in complex environments.

In contrast, deep learning based methods can learn high-level features from large amounts of data automatically, which means manual features are not required. Researchers usually improve existing general object detection algorithms for traffic sign detection, such as Fast R-CNN [4], Faster R-CNN [5], YOLOv3 [6], etc.

There are two types of approaches based on deep learning, one is the one-stage detection approach and the other is the two-stage detection approach. For the two-stage detection approach, region proposals are generated in the first stage, and then the category and coordinates of the object are obtained in the second stage. Li et al. [7] adopted basic framework of Faster R-CNN as the detector...
and adopted a convolutional neural network with asymmetric kernels as the classifier for traffic signs. Yang et al. [8] combined convolution networks with conventional computer vision algorithms to detect traffic signs, a detection network based on Fast R-CNN was proposed. Lu et al. [9] provided a detection network which introduces visual attention mechanism. Although two-stage detection algorithms can achieve high accuracy, the detection speed is insufficient to meet real-time requirement.

For the one-stage detection algorithm, the category and coordinates of the object can be obtained in one stage directly. Zhang et al. [10] provided a Chinese traffic sign dataset called CCTSDB, and proposed a traffic sign detection network which bases on the YOLOv2 [11]. Zhu et al. [12] provided the TT100K dataset which is another Chinese traffic sign dataset with more detailed categories, and provided a network which can detect and classify traffic sign simultaneously. Li et al. [14] proposed a fast traffic sign detection algorithm based on YOLOv3-tiny. Ren et al. [15] proposed an algorithm for small traffic sign detection which bases on MobileNetv2-SSD and adopts the channel attention mechanism. Although one-stage detection algorithms can achieve great balance between accuracy and speed, it is difficult for them to detect small objects in complex environments.

In this paper, we propose a fast traffic sign detection algorithm based on YOLOv3 to improve the detection speed without losing accuracy. The contribution of this paper can be summarized as follows:

1. A lightweight backbone network is designed by modifying the darknet53 network which is the original backbone network of YOLOv3. The new backbone network is called darknet29, for it has 29 convolutional layers.

2. In order to further accelerate the detection speed, depthwise separable convolution [13] is introduced to construct convolutional layers of the darknet29 network, which can further decrease the amount of computations and parameters of the network.

3. Dense connection is adopted to maintain the accuracy of the model.

2. Approach

2.1. Dense Connection
Dense connection is the main idea of densely connected convolutional network (DenseNet). Dense connection can reuse features from preceding layers, and then combine the low-level feature with the high-level feature for object detection, so it can enhance feature propagation and weaken the problem of vanishing gradient. Since dense connection can make better use of low-level features, it is useful for detecting small objects. Figure 1 shows specific structure of dense connection.

2.2. Depthwise Separable Convolution
Depthwise separable convolution means that a standard convolution is factorized into a depthwise convolution followed by a pointwise convolution. For the depthwise convolution, only one filter is adopted for every channel of the inputs. For the pointwise convolution, one 1×1 convolution is used to associate the output of the depthwise convolution. In this way, the amount of parameters can be decreased greatly. The specific structure of depthwise separable convolution is given in figure 2.

2.3. Traffic sign detection algorithm based on modified YOLOv3
In order to achieve real-time detection of traffic signs without loss of accuracy, an algorithm based on YOLOv3 is proposed. Darknet29 network is proposed to replace the darknet53 network as the new backbone network of YOLOv3. Darknet29 network has a total of 29 convolutional layers, and adopts depthwise separable convolution to decrease the amount of parameters and computations, thereby reducing the detection time. Besides that, dense connection is adopted to ensure the network accuracy by reusing the low-level feature. The specific structure of darknet29 network is shown in table 1.

Explanation of definitions in table 1: “Con” denotes the standard convolution, “Con dw” denotes the depthwise convolution, “S1” denotes the stride is 1, “S2” denotes the stride is 2, “Residual” denotes the shortcut operation, which means two feature maps with same size and channels can be added together. “Route” denotes the implement of dense connection, which can concatenate feature-
maps of same size from different layer through channels. For this network, outputs of the upper first layer and the upper third layer will be concatenated together.

**Figure 1.** The structure of dense connection. The circle denotes convolutional layer, the color of the circle means the dimension of the feature map.

**Figure 2.** The left denotes standard convolution. The right denotes the depthwise separable convolution which consists of the depthwise convolution and the pointwise convolution with batchnorm and ReLU.

| Table 1. The structure of the darknet29 network. |
|---|---|---|---|
| Layer Type/Stride | Filter Size | Filters | Output Size |
| Con/S1 | 3x3 | 32 | 416x416x32 |
| Con dw/S2 | 3x3 | 64 | 208x208x64 |
| Con/S1 | 1x1 | 32 | 208x208x32 |
| 1× | Con/S1 | 3x3 | 64 | 208x208x64 |
| Residual | | | 208x208x64 |
| Con dw/S2 | 3x3 | 64 | 104x104x64 |
| Con/S1 | 1x1 | 32 | 104x104x32 |
| 2× | Con/S1 | 3x3 | 64 | 104x104x64 |
| Residual | | | 104x104x64 |
| Con dw/S2 | 3x3 | 128 | 52x52x128 |
| Con/S1 | 1x1 | 64 | 52x52x64 |
| 1× | Con/S1 | 3x3 | 128 | 52x52x128 |
| Residual | | | 52x52x128 |
| Con/S1 | 1x1 | 64 | 52x52x64 |
| 1× | Con/S1 | 3x3 | 128 | 52x52x128 |
| Route | | | 52x52x256 |
| Con dw/S1 | 3x3 | 256 | 52x52x256 |
| Con/S1 | 1x1 | 128 | 52x52x128 |
| 1× | Con/S1 | 3x3 | 256 | 52x52x256 |
| Layer Type/Stride | Filter Size | Filters | Output Size |
|-------------------|-------------|---------|-------------|
| Residual          |             |         | 52×52×256   |
| Con/S1            | 1×1         | 128     | 52×52×128   |
| 1× Con/S1         | 3×3         | 256     | 52×52×256   |
| Route             |             |         | 52×52×512   |
| Con dw/S2         | 3×3         | 512     | 26×26×512   |
| Con/S1            | 1×1         | 256     | 26×26×256   |
| 1× Con/S1         | 3×3         | 512     | 26×26×512   |
| Residual          |             |         | 26×26×512   |
| Con/S1            | 1×1         | 256     | 26×26×256   |
| 1× Con/S1         | 3×3         | 512     | 26×26×512   |
| Route             |             |         | 26×26×1024  |
| Con dw/S2         | 3×3         | 1024    | 13×13×1024  |
| Con/S1            | 1×1         | 512     | 13×13×512   |
| 2× Con/S1         | 3×3         | 1024    | 13×13×1024  |
| Residual          |             |         | 13×13×1024  |

3. Experiment

3.1. Dataset and Experiment Setting
The detection performance of our proposed algorithm is evaluated on CCTSDB dataset, which has 13087 training images and 400 testing images. All of these images are collected from real Chinese traffic scenes with complex background. All traffic signs are classified into three categories: danger, mandatory and prohibitory.

Instead of using default anchors which are not well suitable for CCTSDB dataset, k-means algorithm is adopted to get anchors on the CCTSDB. Table 2 shows new anchors for three output layers of the proposed algorithm.

Experiments are performed with the Nvidia RTX 2080ti GPU. The input size of training and testing images is resized to 416×416. The model is trained for 200 epoches.

Table 2. New anchors for three output layers.

| Output Layer 1    | Output Layer 2     | Output Layer 3     |
|-------------------|--------------------|--------------------|
| (7,7) (10,10) (12,12) | (15,15) (10,24) (22,20) | (20,50) (34,35) (65,74) |

3.2. Evaluation Metrics
Precision, recall, mean average precision (mAP) and speed of detection per image are adopted to evaluate the detector. Because precision and recall cannot get the best value simultaneously, so F1 is also adopted to evaluate the model.

3.3. Results
The performance of the proposed algorithm is compared with other state-of-the-art algorithms. As shown in table 3, the F1 and mAP value of the proposed method can reach 97.2% and 91.6% respectively, the mAP value is higher than all other algorithms. Although the F1 value is not the best, it is comparable to the best one. More importantly, since the proposed algorithm adopts a more lightweight backbone network which adopts the depthwise separable convolution, the detection time per image is 3.4 milliseconds, which outperforms most algorithms and is 44.3% faster than that of
YOLOv3. Although the detection speed of the proposed model is slower than YOLOv3-tiny, the mAP and F1 value is much higher than that of YOLOv3-tiny. The result proves that the proposed method has great real-time performance and accuracy.

Figure 3 further shows several samples of the detection result. It can be seen that the method can detect small objects well.

Table 3. Performance comparison of different algorithms on CCTSDB.

| Method    | Precision/% | Recall/% | F1/%  | mAP/% | Detection time/ms |
|-----------|-------------|----------|-------|-------|-------------------|
| YOLOv3    | 85.6        | 98.4     | 91.5  | 97.0  | 6.1               |
| YOLOv3-tiny | 82.8        | 86.1     | 88.0  | 84.4  | 2.3               |
| Method A [10] | 99.2        | 77.2     | 87.67 | 91.95 | 33                |
| Method B [10] | 98.9        | 82.3     | 89.56 | 94.19 | 20                |
| Method C [10] | 96.44       | 80.95    | 87.69 | 93.11 | 25                |
| Ref [14] | 87.75        | 97.6     | 92.4  | 94.84 | 5                 |
| Ref [15] | -            | -        | -     | 93.2  | 22                |
| Ours      | 86.0        | 98.1     | 91.6  | 97.2  | 3.4               |

4. Conclusion
In this paper, a fast traffic sign detection algorithm based on modified YOLOv3 is proposed for detecting traffic signs efficiently and accurately. Firstly, a new backbone network is designed by modifying the darknet53 network. Then depthwise separable convolution is introduced to decrease the number of parameters, thereby further speeding up the detection speed. In addition, dense connection is adopted to maintain the accuracy, which is useful for detecting small objects. Experimental results prove that our method can achieve great real-time performance and accuracy.

Although the algorithm we proposed has achieved great performance, the F1 value needs to be further improved. In addition, CCTSDB dataset has only three types of traffic signs, the proposed method can be evaluated on other datasets with more categories to further improve the robustness of the method.

References
[1] Y. Yang, H. Luo, H. Xu, and F. Wu, “Towards real-time traffic sign detection and classification,” IEEE Trans. Intell. Transp. Syst., vol. 17, no. 7, pp. 2022 – 2031, Jul. 2016.
[2] M. Liang, M. Yuan, X. Hu, J. Li, and H. Liu, "Traffic sign detection by ROI extraction and histogram features-based recognition," in Proc. IEEE IJCNN, Dallas, Aug. 2013, pp. 1–8.

[3] S. K. Berkaya, H. Gunduz, O. Ozsen, C. Akinlar, and S. Gunal, “On circular traffic sign detection and recognition,” Expert Syst. Appl., vol. 48, pp. 67–75, Apr. 2016.

[4] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis.(ICCV), Dec. 2015, pp. 1440–1448.

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[6] Joseph Redmon, Ali Farhadi, “YOLOv3: An Incremental Improvement”, [Online] Available: https://arxiv.org/abs/1804.02767, 2018

[7] Li J , Wang Z . Real-Time Traffic Sign Recognition Based on Efficient CNNs in the Wild[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(3):975-984.

[8] T. Yang, X. Long, A. K. Sangaiah, Z. Zheng, and C. Tong, "Deep detection network for real-life traffic sign in vehicular networks," Comput. Netw., vol. 136, no. 8, pp. 95–104, May 2018, doi: 10.1016/j.comnet.2018.02.026.

[9] Y. Lu, J. Lu, S. Zhang, and P. Hall, "Traffic signal detection and classification in street views using an attention model," Comput. Vis. Media, vol. 4, no. 3, pp. 253–266, Sep. 2018, doi: 10.1007/s41095-018-0116-x.

[10] Jianming Z , Manting H , Xiaokang J , et al. A Real-Time Chinese Traffic Sign Detection Algorithm Based on Modified YOLOv2[J]. Algorithms, 2017, 10(4):127.

[11] Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. arXiv 2016, arXiv:1612.08242.

[12] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR), Jan. 2016, pp. 2110–2118.

[13] A. G. Howard et al. (2017). “MobileNets: Efficient convolutional neural networks for mobile vision applications.” [Online]. Available: https://arxiv.org/abs/1704.04861

[14] X. Li, J. Zhang, Z. Xie, et al. A Fast Traffic Sign Detection Algorithm Based on Tthree-Scale Nested Residual Structure[J]. Journal of Computer Research and Development, 2020, 057(005):1022-1036.

[15] K. Ren, L. Huang, Ch. Fan, X. Gao. Real-time Small Traffic Sign Detection Algorithm Based on Multi-Scale Pixel Feature Fusion[J]. Journal of Signal Processing, 2020, 36(09):1457-1463.