A Real-time Traffic Sign Detection Model Based on Improved YOLOv3

XiongFei Liu¹, Fan Xiong²

¹School of Physics and Electronics, Central South University, Changsha, Hunan, 416000, China
²Corresponding author’s e-mail: 809649938@qq.com

Abstract. Advanced driver assistance system, which can quickly and accurately detect driver assistance personnel, ensure traffic safety, and play an important role in guiding road traffic signs. However, traditional detection models often have false detection, missed detection, and predicted low accuracy of the target box, especially for the Identification of small targets, its accuracy is difficult to meet the requirements. Therefore, based on the target detection algorithm model YOLOv3, this paper replaces the common two-dimensional convolution with dilated convolution and expands the receptive field, At the end of the model, a three-layer spatial pyramid cascade structure is added to make the multi-scale feature map deep fusion. finally, the improved model is tested for performance on the GTSDB dataset. The results show that the improved model implemented in this paper can effectively solve some problems in the actual road traffic sign detection, and meet the requirements of real-time detection.

1. Introduction

Whether it is an automated driving assistance system or the current hot driverless field, traffic sign detection is undoubtedly very important. However, among a large number of traffic sign detection methods, algorithms based on deep learning have drawn great attention from scholars[1-2], the improvement of computer computing performance has made convolutional neural network(CNN) models widely used in the field of object detection. CNN have become an application trend in the field of computer vision with their high efficiency. Such as R-CNN[3-4], YOLO[5-6]. Many researchers have also applied it to the field of traffic sign detection, and have improved them.

YOLOv3 algorithm obvious deficiencies, such as the target prediction box accuracy is not high, small target detection scenario occurs erroneous detection, missed detection[7-8]. As for these problems, this paper improves YOLOv3, and proposes a fast traffic sign detection model YOLOv3-pro, which is a deep fusion of feature maps. Its structure shown in Figure 3: Dilated convolutional in deep residual networks; 3 layers convolution of cascade structure. Together, they form a feature pyramid containing 6 convolutional layers of different scales; Meantime, perform an upsampling operation on the feature pyramid in 2 times steps and fuse with the previous deep residual network to form a deep fusion detection model. This paper uses the GTSDB[9] dataset to compare and test different algorithms. Experimental results show that our method achieves good results in small target detection.
2. YOLOv3 model
YOLOv3, the latest version of the YOLO series, is an end-to-end one-stage universal object detection algorithm. It uses the K-means clustering to automatically select the best initial regression box for the dataset, and uses a multi-scale anchor structure to improve the accuracy of target detection.

\[
\text{confidence} = p_r(\text{object}) \times \text{IOU}_\text{truth}^{\text{pred}}, \quad p_r(\text{object}) \in \{0,1\} \quad (1)
\]

Figure 1 shows the processing flow. It divides each input image into \( s \times s \) grids. If the center of the real target box falls on a grid, the grid is responsible for detecting the goal. Each grid needs to predict the probability value of the class \( C \) target, \( B \) boxes, and corresponding confidence score. Formula (1) shows the confidence score calculation, where \( p_r(\text{object}) \) is 1 for the grid containing the target, \( \text{IOU}_\text{truth}^{\text{pred}} \) represents the error between the prediction box and ground truth, it reflects the grid about the accuracy of the prediction. If a target has multiple prediction boxes, NMS is used to select the best.

| Type   | Filter | Size       | Output               |
|--------|--------|------------|----------------------|
| conv   | 32     | 3×3        | 256×256×16           |
| conv   | 64     | 3×3/2      | 128×128×32           |
| conv   | 32     | 1×1        | 128×128×32           |
| conv   | 64     | 3×3        | 64×64×64             |
| res    |        |            | 64×64×64             |
| conv   | 128    | 3×3/2      | 64×64×64             |
| conv   | 64     | 1×1        | 64×64×64             |
| conv   | 128    | 3×3        | 64×64×64             |
| res    |        |            | 64×64×64             |
| conv   | 256    | 3×3/2      | 32×32×128            |
| conv   | 128    | 1×1        | 32×32×128            |
| conv   | 256    | 3×3        | 32×32×128            |
| res    |        |            | 32×32×128            |
| conv   | 512    | 3×3/2      | 16×16×256            |
| conv   | 512    | 3×3        | 16×16×256            |
| res    |        |            | 16×16×256            |
| conv   | 1024   | 3×3/2      | 8×8×512              |
| conv   | 512    | 1×1        | 8×8×512              |
| conv   | 1024   | 3×3        | 8×8×512              |
| avgpool|        |            | 8×8×512              |
| connected |      |            |                      |
| softmax |        |            |                      |
| global | 1000   |            |                      |

Figure 2. YOLOv3 model structure
The structure YOLOv3 model is shown in Figure 2. Backbone is Darknet-53. It uses upsampling and residual networks. In the feature maps of the last 3 scales, 3 channels are used to perform target detection independently.

3. Improved YOLOv3 model

3.1. Candidate box generation strategy based on K-means

Anchor box mechanisms have been proposed in target detection models such as Fast R-CNN\cite{4}, but the size of the prediction box is manually set based on experience, which will cause the network to converge slowly during training and be prone to local optimization. The traditional K-means clustering method uses the Euclidean distance function, which makes larger boxes have more error clustering than smaller boxes, and the clustering results may be biased. For this reason, we use IOU\cite{10} scores to evaluate the clustering results, thereby avoiding errors caused by the scale of the box. The formula for the distance is as follows:

\[
d(box, centrod) = 1 - IOU(box, centrod)
\]

The central particle is the center point of the cluster; \(IOU(box, centrod)\) is the overlap ratio of the cluster box and the central box. The K-means clustering method is used to compare the IOU scores of different k values to find the best number and size of anchor boxes.

3.2. Backbone Network Based on Dilated Convolution and FPN Structure

For the detection of small targets, high-resolution feature maps are required. In order to obtain more detailed semantic features and accurately determine the location of the target, a wider receptive field and more global information are required. This paper combines dilated convolution and FPN structure, and proposes a high-resolution, large receptive field deep residual backbone network for feature extraction.
Figure 5. Two types of residual unit with dialated convolution

3.2.1. Dilated convolution structure
Dilated convolution is to expand the convolution kernel by changing the internal spacing receptive field. Figure 4 shows three differently spaced convolution kernels, where rate represents the interval of holes in the convolution kernel. (a) is 3*3, rate = 1 dilated convolution, the convolution kernel receptive field range is 3*3. (b) is 3*3, rate = 2 dilated convolution, convolution receptive field increases to 7*7. (c) is a range of 3*3 and rate = 3. The receptive field of the convolution kernel increases to 15*15, which ensures that the convolutional network can extract feature information in a larger field of view [11-13]. Figure 5 shows two application types.
### 3.2.2. Multi-scale spatial pyramid structure

For error detection, undetected presence, analyze the reasons for the lack of features for the target model of learning. Therefore, here we try to add 3 scales of target detection channels after the main network of YOLOv3 (as shown by the dotted line in Figure 6), together with the original 3 convolutional layers of YOLOv3 to form a convolution feature pyramid with 6 different scales. That is: 64*64, 32*32, 16*16, 8*8, 4*4, and 2*2 resolutions; at the same time, perform upsampling operations on the feature pyramid in 2 times steps with the previous residual network layer to form a deep feature fusion detection model[14-15], and the detection is performed independently on the final 6 independent fusion feature maps.

Based on the above improvements, the final YOLOv3-pro model feature extraction network structure shown in Figure 6.

### 4. Experiments and results

Follow the idea YOLOv3, the target detection task as an end to end regression model using stochastic gradient in training drop the loss function to achieve the optimal solution. From actual training dataset, generated a priori box, training model and the final results of the analysis elaborated on.
4.1. GTSDB dataset
The model in this paper uses German Traffic Sign Detection Benchmark (GTSDB) dataset for training and testing. The data set were 900 road traffic signs panoramas, contains a total of 1206 traffic signs, each picture contains 0, 1 or more traffic signs, of which the first 600 as a training set (including 846 traffic signs), 300 as a test set (including 360 traffic signs). The data set labels are divided into 4 categories: mandatory, prohibition, danger, and other. Figure 7 shows some.

![Figure 7. Some examples in the GTSDB dataset](image)

4.2. Generation of prior box
For the selected anchor frame, is generated by K-means clustering. When doing clustering, the width and height of all ground truth boxes in the dataset are used as feature inputs for K-means clustering, we use the aforementioned IOU to measure the distance between two bounding boxes, and when calculating the iou of two bounding boxes, we only need to use their 4 position parameters (xmin, ymin, width, height).

According to the magnitude of the GTSDB and the experimental results of K-means, it is finally determined that the position coordinates are predicted using convolution feature maps of 9 scales. The prior boxes are: 10*13, 16*30, 33*23, 30*61, 62*45, 59*119, 116*90, 156*198, 373*326.

4.3. Model training
YOLOv3-pro training decreased by stochastic gradient algorithms end. In order to make the training converge quickly, the pre-trained model Darknet-53 is used to initialize the shared convolutional layer in this model. In training, we need to match the bounding box anchor predict the real box in order to establish their correspondence. If the predicted bounding box a priori has the highest overlap with the real box, mark the traffic sign as a positive sample while ignoring the non-optimal bounding box; it is worth noting that for each real box, only one bounding box is assigned to Reduce duplication in your target audience. Training process parameters set as shown in Table 1.

| Size of input | Batch | Momentum | Initial learning rate | Decay | Training steps |
|---------------|-------|----------|-----------------------|-------|----------------|
| 512*512       | 64    | 0.9      | 0.001                 | 0.0005| 20000          |

Figure 8 shows the corresponding Loss curve of the method in the training process. the number of training iterations, and the ordinate represents the Loss value during training. It can be seen that the model will achieve a convergence effect. That is, at the beginning of training, the method in this paper decreases rapidly; after 10,000 iterations of training, the Loss curve declines steadily; when iterating to 20000 times, Loss converges to 0.2 and ends the training.
4.4. Results and analysis

4.4.1. Test on small targets
From the GTSDB dataset, a sample of small target traffic signs at 6 scales is selected, and its ground truth is around 10*20, 20*40, 30*60, 40*80, 50*100, and 60*120. Select 50 pictures each scale together to form a small target sample test set, and use Faster-RCNN, YOLOv3 and this method for comparison experiments.

Figure 9 shows YOLOv3-pro detects an actual effect on small targets, from an image point of view, all of the small target are properly marked and classified, no missed or false detection cases, to achieve the desired. The results of different model algorithms for small target detection are shown in Figure 10.

It is obvious that the average accuracy of this model is higher than the other two models in the small target data sample. As the target size increases, the AP value also increases. When the traffic sign box...
is about 60*120, the average accuracy of YOLOv3 is 77.3%, while the method in this paper can reach 81.6%, an increase of 4.3%, and performance has improved.

4.4.2. Performance comparison of different detection algorithms

| method of detection | mAP (%) | FPS (f/s) |
|---------------------|---------|-----------|
| YOLOv3              | 80.44   | 46.55     |
| Faster R-CNN        | 78.42   | 2.51      |
| Article method      | 85.26   | 38.52     |

Table 2 shows the results of the three models on a test dataset GTSDB after training. As can be seen from the table, the model proposed in this paper obtained 85.26% mAP, which is superior to Faster R-CNN in terms of accuracy. Compared with the original YOLOv3, the detection accuracy is improved by 4.82%. This indicates that cascade through multi-scale features can get more detailed target feature, the detection accuracy is significantly improved, to some extent, solve undetected, error detection of the problem. Of course, can also be found from the table, YOLOv3 detection rate improved compared to the decline before the improvement, mainly due to the deeper layers of the iterative model, bulky structure of the cascade model of training time has increased. In addition, Dialted convolution is more time-consuming and computationally intensive than 2D convolution. But from the perspective of detection speed, 38.52 f/s has reached the requirements of real-time detection.

5. Conclusion
This paper proposes an improved real-time traffic sign detection model based on YOLOv3. The extended convolution method is applied to the backbone network to effectively expand the receiving domain and improve the recognition effect of small target traffic signs on the road. Multi-level spatial pyramid The feature fusion backbone network can obtain more detailed semantic features during feature extraction, which effectively improves the problem of missed detection and false detection in the road traffic sign recognition scene. The experimental results show that combined with dilated convolution and multi-level spatial pyramid feature fusion, the performance of the final model target detection has been improved. mAP is 4.82% higher than YOLOv3 and 36.01 fps faster than Faster R-CNN.

References
[1] A. De La Escalera, L.E. Moreno, M.A. Salichs, J.M. Armingol, Road traffic sign detection and classification, IEEE Trans. Ind. Electron.44 (6) (1997) 848 – 859,doi:10.1109/41.649946.
[2] Á. Arcos-García, J.A. Álvarez-García, L.M. Soria-Morillo, Deep neural network for traffic sign recognition systems: an analysis of spatial transfor.
[3] Girshick R.Fast R-CNN[C]//ICCV, 2015: 1440-1448.
[4] Ren S Q, He K M, 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.
[5] Redmon J. Divvala S, Girshick R, et al.You only look once: unified, real-time object detection[C]//CVPR, 2016.
[6] Redmon J, Farhadi A.YOLOv3: an incremental improvement[J].arXiv: 1804.02767v1, 2018.
[7] Gong An, Li Chenggian, Niu Bo. Real-time pedestrian detection method based on convolutional neural network [J]. Journal of Computer System Applications, 2017, 26 (9): 215-218.
[8] Guo Zhi, Song Ping, Zhang Yi, et al. Aircraft target detection method for remote sensing image based on deep convolutional neural network [J]. Journal of Electronics & Information Technology, 2018, 40 (11): 149-155.

[9] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, C. Igel, Detection of traffic signs in real-world images: the German traffic sign detection benchmark, in: Proceedings of the 2013 International Joint Conference on Neural Networks, IJCNN, IEEE, 2013, pp. 1 – 8.

[10] F. Ahmed, D. Tarlow, D. Batra, Optimizing expected intersection-over-union with candidate-constrained CRFs, IEEE International Conference on Computer Vision (2016).

[11] Zhang Fukai, Yang Feng, Li Ce. Fast vehicle detection method based on improved YOLOv3 [J]. Computer Engineering and Applications, 2019, 55 (02): 18-26.

[12] 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-.

[13] Guo Jinxiang, Liu Libo, Xu Feng, Zheng Bin. Aeroplane scene detection method based on YOLO v3 [J / OL]. Laser and Optoelectronics Progress.

[14] Yi, Z., Yongliang, S., & Jun, Z. (2019). An improved tiny-YOLOv3 pedestrian detection algorithm. Optik. doi:10.1016/j.ijleo.2019.02.038

[15] Yunong Tian, Guodong Yang, Zhe Wang, Hao Wang, En Li, Zize Liang. Apple detection during different growth stages in orchards using the improved YOLOv3 model[J]. Computers and Electronics in Agriculture, 2019, 157.