Research on Traffic Sign Detection Algorithm Based on Improved YOLOv3

Dan Zhao, Li Zhao, Rong Ma and Xiaohan Zhao
32269 Troops, Lanzhou, Gansu, China
Email: 575915179@qq.com

Abstract. As the traffic sign target detection has the defects of small size, low resolution and unobvious features, an improved YOLOv3 network model is proposed. Firstly, using a K-means algorithm TT100K traffic sign data set clustering analysis and redefine the initial candidate frame of the network. Secondly, this paper improves the FPN structure in the original network, retain the large-scale prediction of 52×52 in the original network, and then use the feature map output from the second down-sampling in the YOLOv3 network to establish a larger-scale prediction of 108×108. For the purpose of solving the size of image and distortion problems, the pyramid pooling operation with fixed block sizes of 5, 9, 13 is used before the detection layer, and then the output features are merged with the original feature map, so as to achieve the same size for inputs of different sizes output. We use the improved YOLOv3 network model with its original model and other small target detection algorithms to conduct comparative experiments on the TT100K data set. The results show that the improved YOLOv3 network model can detect traffic signs more effectively, with better detection accuracy and real-time performance.

Keywords: Target detection; traffic signs; YOLOv3; multi-scale prediction.

1. Introduction
Traffic sign detection is one of the main research contents in the intelligent driving system. As the deep convolutional neural networks developing faster and faster, its detection methods are emerging one after another. SSD [1], DSSD [2], FCOS [3], etc. have received close attention from academia and industry. Although the detection accuracy of small targets has been improved to a certain extent, the operating speed still needs to be studied. Target detection object is to locate and identify the target image of interest, and determine their category and location. YOLOv3 [4] proposed by Redmon J et al. uses the entire image as the input of the network, and directly returns the position and category of the target space in the output layer, and improves YOLO9000 [5], adding multi-scale predictor categories to correct Classification [6]; The detection effect of small targets including the location method is better than that of YOLO9000 [7]. On the basis of the ResNeXt model, it uses a multi-dimensional feature fusion method to improve the detection precision of small target traffic signs [8]. Taking the ESPNetv2 structure as the basic network, the deconvolution structure is used to fuse the low-level and high-level feature maps to generate more characteristic feature maps, which can quickly and accurately detect traffic signs [9]. The idea of a dense residual network is introduced into the YOLOv3 network to realize the multiplexing and fusion of the network’s multi-layer features. The Resnet-101 network is used to replace the original Darknet-53 network, which improves the detection property of the YOLOv3 network [10]. In the real scene, the main network in this paper is the YOLOv3 target detection network. In the traffic sign detection task, it is necessary to quickly detect the traffic signs, as
far as the original YOLOv3 network is concerned. First of all, traffic signs occupy only small pixels in the image, and the feature fusion map of the third scale in the original network can be used for small target detection [11]. Bell [12] defined a small target as COCO number sampling, and performed feature fusion with the 2 times down-sampling feature map in Darknet-53 to obtain a target whose feature map size is less than or equal to 32×32 pixels, and input the detection layer to scale for the prediction of 108×108, the size of the same standard is small, the resolution is low, and the features are not obvious. When the target is performed, the prediction of the original network with a scale of 52×52 is retained. The detection of two rulers requires special data and greater accuracy high algorithm [13]. Then, Reference Space Pyramid Pooling (SPP) [14]. Nowadays, as the deep learning technology continues to evolve, it is a method that uses a fixed block size of 5, 9, 13 before the detection layer. It has a good effect in realizing the same size detection, obstacle detection and other fields for different sizes of input. Finally, compare the detection effects of the target with different improved networks, and compare the improved YOLOv3 network with the original network.

2. YOLOv3 Algorithm

YOLOv3 is an object detector proposed by Joseph et al. Its backbone network uses Darknet-53 instead of Darknet-19, and has a total of 53 convolutional layers. Figure 1 is the frame of network. As shown in figure 2, the YOLOv3 network’s hierarchical structure is shown clearly.

![YOLOv3 network model](image)

The smallest component DBL module of Darknet-53 includes convolution, batch normalization and Leaky ReLU activation function. YOLOv3 divides the prediction into three scales of 13×13, 26×26, and 52×52. These three scales output feature maps of three different scales to the detection layer. The low-level feature maps have a smaller field of view and are responsible for detection. For small targets, the deep feature map has a large field of view, making it easy to detect large targets. Therefore, YOLOv3 has a good performance in detecting large and small targets. Because the YOLOv3 network has the advantages of high training efficiency, strong adaptability to different scale targets, and suitable for complex traffic scenes, this paper improves the YOLOv3 network and uses the traffic sign data set TT100K [14] for training and detection.
Figure 2. YOLOv3 hierarchical structure diagram.

3. Improve YOLOv3 Detection Algorithm

Traffic sign detection in a real scene is small target detection, and the initial candidate frame and its network structure preset for the COCO data set by the YOLOv3 model are not suitable for small target detection. Therefore, this paper uses K-means clustering to perform cluster analysis on the traffic sign data set, redefine the initial candidate frame size, and then improve the YOLOv3 model to achieve traffic sign detection.

3.1. k-means Cluster Analysis

YOLOv3 network initial candidate block width and height set to a fixed value, it will affect the precision and speed of target detection. Therefore, the TT100K traffic sign data set’s cluster analysis is achieved by the k-means clustering algorithm in this paper, and the average degree of overlap (AvgIOU) is used as the target cluster analysis measure. The clustering AvgIOU objective function \( f \) can be expressed as:

\[
 f = \arg \max_{n,k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} I_{IOU} (b, c)}{n}
\]

where: \( b \) represents the sample, that is, the target in the ground truth; \( c \) represents the center of the cluster; \( nk \) represents the number of samples in the \( k \) cluster center; \( n \) represents the total number of samples, and \( k \) represents the number of clusters; \( I_{IOU} (b, c) \), represents the intersection ratio between the center box of the cluster and the cluster box; \( i \) represents the sample number; \( j \) represents the sample number in the cluster center.

Make \( k=1 \sim 9 \), then the cluster analysis is performed on the samples in the TT100K traffic sign data set, and the relationship between the \( k \) value and AvgIOU is shown in figure 3.

As the value of \( k \) increases, the objective function changes more and more steadily, and the changing inflection point can be considered as the optimal number of initial candidate frames. When the \( k \) value is greater than 6, the curve begins to become stable, so we choose 6 as the number of initial candidate frames. Not only the convergence of the loss function can be accelerated, but also the error caused by the candidate frame can be eliminated. These 6 initial candidate frames correspond to the input picture size in the data set of 416×416, 608×608, 1024×1024, set the width and height to \([4×4, 5×6, 8×8, 11×11, 15×16, 22×24], [5×6, 7×8, 10×11, 13×14, 19×20, 30×32], [9×11, 14×15, 19×21, 16×19, 36×40, 56×59]\).
3.2. Data Enhancement of Traffic Signs

In order to achieve data enhancement of traffic signs, this paper uses a color enhancement scheme based on red, yellow, and blue traffic signs to enhance the features of traffic signs in the image, and vice versa, weaken other features in the image, so that the network can learn more during training. About the characteristics of traffic signs. By filtering the image’s R, G, and B values of each pixel, weaken the R, G, and B pixel values in areas other than the traffic signs, thereby enhancing the traffic signs in the image, as shown in figure 4. As a preprocessing process before the image is input to the network, the use of color enhancement can enhance the characteristics of the traffic signs in the image. At the same time, since only the color enhancement scheme is used, it will not increase too much extra calculation and will not affect the detection speed.

![Figure 3. K-means cluster analysis results.](image)

**Figure 3.** K-means cluster analysis results.

![Figure 4. Comparison of results before and after image enhancement.](image)

**Figure 4.** Comparison of results before and after image enhancement.
Scheme based on red, yellow, and blue traffic signs to enhance the features of traffic signs in the image, and vice versa, weaken other features in the image, so that the network can learn more during training. About the characteristics of traffic signs. By filtering the image’s R, G, and B values of each pixel, weaken the R, G, and B pixel values in areas other than the traffic signs, thereby enhancing the traffic signs in the image, as shown in figure 4. As a preprocessing process before the image is input to the network, the use of color enhancement can enhance the characteristics of the traffic signs in the image. At the same time, since only the color enhancement scheme is used, it will not increase too much extra calculation and will not affect the detection speed.

3.3. Improve YOLOv3 Network Model

In order to take advantage of the more information of the small target contained in the shallow features in the network, the second down-sampling feature map in Darknet-53 is used to realize up-sampling the feature fusion map of the third scale in the YOLOv3 network, and then compare it with Darknet 2 times of down-sampling feature map fusion in -53 is input to the detection layer to achieve a prediction with a scale of 108×108.

Make an example of the input size of 416 pictures, the initial candidate frame corresponding to the scale of 52×52 in the TT100K dataset is [13×14, 19×20, 30×32], and the initial candidate frame corresponding to the scale of 108×108 is [5×6, 7×8, 10×11], figure 5 shows the improved network, and the hierarchical structure is shown in figure 6. Compared with the original YOLOv3 algorithm that predicts on three scales, the improved YOLOv3 algorithm only needs to predict two scales, which can detect targets in images faster.

| type          | filters | size          | output         |
|---------------|---------|---------------|----------------|
| convolutional | 32      | 3x3           | 256x256        |
| convolutional | 64      | 3x3/2         | 128x128        |
| convolutional | 32      | 1x1           | 128x128        |
| convolutional | 64      | 1x1           | 64x64          |
| convolutional | 128     | 3x3           | 32x32          |
| convolutional | 256     | 3x3/2         | 16x16          |
| convolutional | 512     | 3x3/2         | 8x8            |
| convolutional | 1024    | 3x3/2         | 4x4            |
| convolutional | 512     | 3x3           | 16x16          |
| convolutional | 256     | 3x3           | 64x64          |

**Figure 5. Improved YOLOv3 network model.**

High-resolution traffic sign images may cause loss of information or inconsistencies in scale during preprocessing and multi-scale prediction, which will affect the detection effect. Spatial pyramid pooling uses different block pooling for a picture, and each block extracts a feature as a dimension to ensure that the dimensions of the final features are consistent, thereby solving the problems of information loss and scale inconsistency. Therefore, referring to the spatial pyramid pooling method, a fixed block size pooling operation is used before the detection layer. In order to achieve the fusion of the feature map level of local features and global features, the largest pooling core of the spatial pyramid pooling structure must be exhausted. It may be close to the size of the feature map that needs to be pooled (13×13), so the maximum pooling core is set to 13, and the remaining two cores are
sequentially reduced by 4, and set to 9 and 5. In this way, various features of each picture are extracted to improve the detection accuracy of traffic signs. Although the spatial pyramid pooling with three different blocks increases the complexity of the model and affects the model speed, experiments show that the model speed decreases less and the accuracy improves more. Therefore, it is worthwhile to add the spatial pyramid pooling.

![Figure 6. Improved YOLOv3 hierarchical structure.](image)

### 4. Experiment and Result Analysis

To verify the correctness and effectiveness of the improved YOLOv3 traffic sign detection algorithm proposed in this paper, this paper conducts two experiments to verify the two experiments in terms of average detection accuracy (mAP) and detection frames per second (FPS).

#### 4.1. Standard Data Set and Experimental Platform

This article uses the traffic sign data set TT100K, deletes the pictures without annotated files in the data set, uses 7150 pictures for training, 1050 images for verification, and 3070 pictures for testing. The experimental platform is as follows: the operating system Ubuntu16.04, the deep learning framework pytorch1.4, the CPU is AMD-R2700, the memory is 32GB, the GPU is NVIDIA GeForce 2080ti×2, and the video memory is 22G.

#### 4.2. Model Training and Evaluation Indicators

Train the YOLOv3 network and improve the YOLOv3 network separately, use the built-in parameter evolution method of YOLOv3 to adjust the parameters, set the initial learning rate to 0.001, the maximum number of iterations to 300 epoch, and set the learning rate to 75 epoch, 150 epoch and The attenuation is 10 times at 250 epochs.

**FPS** (Frames Per Second): The frame image rate detected per second.

**Precision**: precision rate, can be abbreviated as $p$, as in equation (2).

$$\text{Precision} = \frac{TP}{TP + FP}$$  \hspace{1cm} (2)

**Recall**: Recall rate, which can be abbreviated as $r$, as in equation (3).
The specific meanings of FP, TP, FN, and TN in equations (2) and (3) are that TP means predicting the positive class as a positive class number, FP means predicting a negative class as a positive class number, and FN means a positive class number. The forecast is a negative class number. By setting a fixed threshold, the detector's prediction results are sorted in descending order according to the confidence score, and the samples are output as positive examples for prediction one by one.

Average Precision, AP: Average recognition efficiency, as in equation (4).

$$AP = \int_{0}^{1} p(r)dr$$

Mean Average Precision, mAP: The average recognition efficiency of multiple categories, as shown in equation (5):

$$mAP = \frac{\sum_{e=1}^{Q} AP(q)}{Q}$$

4.3. Experimental Results and Analysis

The three models are trained and tested on the TT100K data set under the input image size of 416×416, and compared with the original YOLOv3 network. Table 1 are the experimental results.

|         | mAP  | FPS  |
|---------|------|------|
| YOLOv3  | 0.691| 54.2 |
| YOLOv3-A| 0.707| 52.9 |
| YOLOv3-B| 0.736| 56.4 |
| YOLOv3-C| 0.752| 57.3 |

Table 1 can show us that the detection effect of the improved YOLOv3 (ie YOLOv3-C) on the traffic sign TT100K data set is higher than that of the original YOLOv3 network model. Among them, YOLOv3-A represents the method of data enhancement, which increases the detection accuracy by 1.6% mAP, indicating that the data enhancement based on traffic signs can make the characteristics of traffic signs more obvious, so that the network can better detect traffic signs in the image; The detection accuracy of YOLOv3-B is increased by 4.5% mAP, and the detection speed is also improved to a certain extent, which shows that large-scale prediction can make better use of the information of small targets in the image, and can better detect small targets of traffic signs. At the same time, the improved YOLOv3 detection layer has changed from the original three scales to two scales, which reduces the amount of network calculations, so the FPS of the network can be improved. The YOLOv3-C model adds a spatial pyramid pooling layer, which increases mAP by 0.8% and FPS to 57.3FPS. Therefore, adding a spatial pyramid module has necessary practical significance and basically achieves the purpose of real-time detection. Figure 7 show the test results.
5. Concluding Remarks
In terms of traffic sign detection, the main improvements made in the process of improving the YOLOv3 road target detection network are as follows: Aiming at the problem that the detection rate of traffic signs based on YOLOv3 is not high. In the image preprocessing part, data enhancement is used; optimization of the YOLOv3 algorithm is proposed. Firstly, the original YOLOv3 algorithm predicts on three scales, and the improved YOLOv3 algorithm only needs to predict two scales, which can detect targets in images faster. Secondly, use spatial pyramid pooling to perform different Block pooling, each block extracts a feature as a dimension to ensure that the dimensions of the final feature are consistent, thereby solving the problem of information loss and inconsistent scale. In the case of only improving the network structure, the improved network also has good detection performance, Finally the two optimization methods are embodied in a network structure, the detection performance of the algorithm is greatly improved, and it can also meet the application requirements of road target detection.

References
[1] Liu W, Anguelov D, Erhan D, et al. 2016 SSD: Single shot multibox detector European Conference on Computer Vision pp 21-37.
[2] Fu C, Liu W, Ranga A, et al. 2017 DSSD: Deconvolutional single shot detector https://arxiv.org/abs/1701.06659.
[3] Tian Z, Shen C, Chen H, et al. 2019 FCOS: Fully convolutional one-stage object detection International Conference on Computer Vision pp 9627-9636.
[4] Redmon J and Farhadi A 2018 YOLOv3: An incremental improvement https://arxiv.org/abs/1804.02767.
[5] Redmon J and Farhadi A. 2016 YOLO9000: Better, faster, stronger IEEE Conference on Computer Vision and Pattern Recognition pp 6517-6525.
[6] Jiao L, Zhang F, Liu F, et al. 2019 A survey of deep learning-based object detection IEEE Access 128837-128868.
[7] Zhao Q, Sheng T, Wang Y, et al. 2019 M2Det: A single-shot object detector based on multi-level feature pyramid network National Conference on Artificial Intelligence pp 9259-9266.
[8] Zhang Y, Gong Z Y and Wei W W 2020 Traffic sign detection based on improved Faster R-CNN model Progress in Laser and Optoelectronics 57 (18) 173-181.

[9] Nguyen H 2020 Fast traffic sign detection approach based on lightweight network and multilayer proposal network Journal of Sensors 2020 (2020) 25-33.

[10] Qi Y X and Yu W W 2021 Traffic sign detection in complex environment based on improved YOLOv3 International Core Journal of Engineering 7 (3)145-153.

[11] Bell S, Zitnick C L, Bala K, et al. 2016 Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks IEEE Conference on Computer Vision and Pattern Recognition pp 2874-2883.

[12] Zhu Z, Liang D, Zhang S, et al. 2016 Traffic-sign detection and classification in the wild IEEE Conference on Computer Vision and Pattern Recognition pp 2110-2118.

[13] Purkait P, Zhao C, Zach C, et al. 2017 SPP-Net: Deep absolute pose regression with synthetic views https://arxiv.org/abs/1712.03452.