International Core Journal of Engineering Volume 7 Issue 3, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0015

Traffic Sign Detection in Complex Environment based on Improved YOLOv3

Qiyuan Xiao, Weiwei Yu

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China.

Abstract

Aiming at the problems of traffic signs in complex environments in real life or detection failures caused by too small targets, the target detection technology for traffic signs has slow detection speed, low detection accuracy and poor robustness. And other issues. The experiment is based on the YOLOv3 network architecture. In order to reduce the loss of information layer by layer in the network transmission process, using the idea of residual dense network, an improved YOLOv3 algorithm is proposed to realize the multiplexing and fusion of network multi-layer features. The resnet101 network replaces the original darknet-53 network to improve the detection performance of YOLOv3.

Keywords

YOLOv3 Network; Feature Extraction; Residual Block; Residual Dense Block; Target Detection.

1. Introduction

Target detection is one of the hot research topics in the field of computer vision, which has received widespread attention. Fast and effective target detection methods play a very critical role in many practical applications. Traffic sign detection is a real-time detection task based on specific object categories in images. In recent years, it has developed rapidly with the emergence of deep convolutional neural networks [1] (deepCNN). GirshickR et al. [2] proposed the RCNN model, which first applied the deep convolutional neural network to the target detection [3] scene, and greatly improved the performance of the traditional DPM model. FasterR-CNN combines FasterR-CNN [4,5] with RPN (region proposal network) to successfully implement a new end-to-end learning optimization unified network architecture in the application of target detection field [6], which improves the candidate region Extraction speed and accuracy. However, there is a problem with FasterR-CNN and various improved FasterR-CNN methods, that is, only a single convolutional layer feature map is used, resulting in the lack of information at different resolutions, making it impossible to detect some small objects. At present, most traffic sign detection methods are carried out in videos or images with relatively simple backgrounds. In real life, the performance of the detector is affected by interferences such as the complex background of traffic signs, different postures and complex target environments. ZhangL et al. [7] proposed an improved FasterR-CNN (Faster region-based convolutional neural networks) to detect traffic signs, using RPN as a detector to generate initial candidate regions, and using cascaded random forest classifiers for classification. This method can reduce the false detection rate of traffic signs, but the detection speed is slow. Chen Guangxi et al. [8] proposed the use of aggregate channel feature algorithm to generate more initial candidate frames, and the use of CNN for deep feature extraction. The traffic sign detection method has good performance in the existing traffic sign detection data set, but the detection is complex in the environment. The accuracy of traffic signs is low. Compared with the YOLO [10] algorithm, although the YOLOv2[9] algorithm has higher detection accuracy and detection speed and other indicators, it

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0015

still has problems such as insensitive to small target detection. Therefore, RedMomJ et al. [11] proposed a new target detection method, the YOLOv3 algorithm, which combines the advantages of YOLOv2 and the residual network [12], and effectively solves the problem of difficult detection of small targets. However, there are still problems such as missing information due to the multi-layer feature extraction process. In order to solve the problem of loss of information extracted layer by layer due to the increase of network depth, this paper draws on the idea of residual dense network (RDN) proposed by ZhangY et al. [13], based on the YOLOv3 network structure, and proposes an improved YOLOv3 algorithm. Traffic sign detection in complex environments. That is, the residual dense block (RDB) is introduced in the deep layer of the network, so that each convolutional layer makes full use of the hierarchical features of all convolutional layers connected to it. The experimental results show that the entire network can effectively reduce information loss, realize the multiplexing and fusion of network multi-layer features, and can quickly and effectively detect traffic signs in complex environments.

2. Related theoretical knowledge

2.1 Residual dense block

Residual dense block extracts rich local features through densely connected convolutional layers, making full use of the hierarchical features of all convolutional layers. Residual dense network stitches each RDB output feature map and introduces global feature fussion (GFF). As shown in Figure 1, assuming that there are D residual dense blocks, the input is the F_0 feature map, that is, the output F_d of d residual dense blocks is expressed as

$$Fd=Hd(Fd-1)=Hd(Hd-1(\ldots(H1(F0))\ldots)) \quad (1)$$

In the formula, Hd(·) is the feature extraction of d residual dense blocks after a series of operations such as convolution. The output Fd is generated by feature extraction of the d-th RDB internal convolution block, that is, Fd is regarded as a local feature. GFF extracts the global feature FGF by fusing all the RDB output feature maps, which can

$$FGF=HFGF([F1,\ldots,Fd]) \quad (2)$$

Where $[F_1,...,F_d]$ is the series of feature maps generated from the residual dense blocks 1,...,D, and HFGF is the 1×1 and 3×3 convolution composite function. The 1×1 convolution layer is used to adaptively fuse a series of features at different levels, and the 3×3 convolution layer further extracts features.

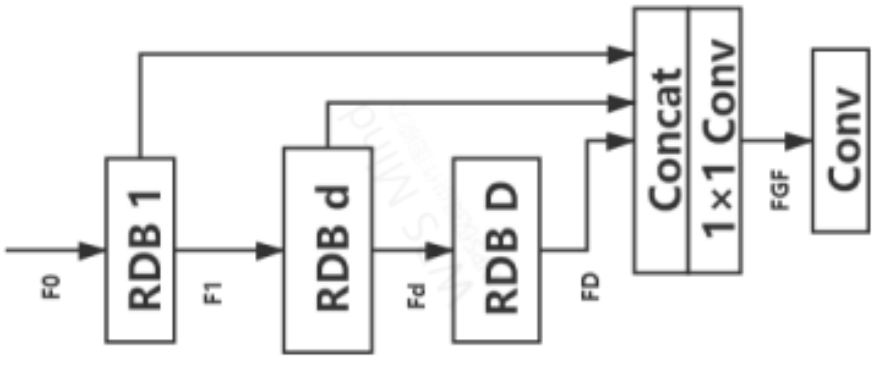

Figure 1. Schematic diagram of residual dense block

2.2 YOLOv3 algorithm

2.2.1. Darknet-53 feature extraction model

The YOLOv3 feature extraction model uses multiple 3×3 and 1×1 convolutions, and the entire network has 53 convolutional layers, so it is referred to as Darknet-53. as shown in figure 2.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0015

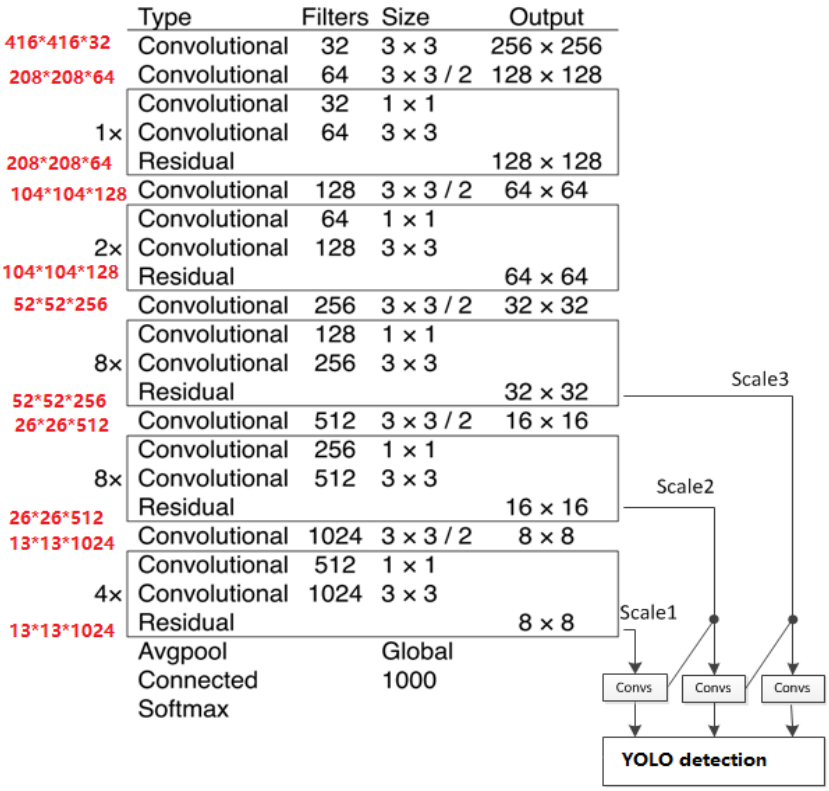

Figure 2. Darknet-53 feature extraction network structure diagram

2.2.2. Multi-scale prediction

Three scale prediction methods are used, namely scale 1: using 13×13 feature maps to output box information after 7 layers of convolution; scale 2: using scale 1 convolutional layer to output feature maps for the fifth layer, and perform one time Convolution and one × 2 up-sampling, connect the up-sampling feature map with the 26×26 size feature map, and output the box information through 7 layers of convolution again. The feature map size is twice as large as scale 1; scale 3: use 52 ×52 feature map, using a similar method of extracting feature maps in scale 2. The feature size obtained for each prediction task is N×N×[B×(4+1+C)], where N is the size of the grid, B is the number of bounding boxes obtained for each grid, 4 is the number of bounding box coordinates, and 1 is the target predicted value, C is the number of categories.

3. Improved YOLOv3 network

3.1 Improvement based on residual dense block

There are two main ways to achieve multi-layer feature multiplexing and reduce layer-by-layer information loss. One is to connect the network structure feature map in series, and connect different convolution outputs into one to obtain a rough representation of each target feature by cascading (14], the second is based on the pyramid feature hierarchy [15] (pyramidal feature hierarchy) method, which is a combination of different convolutional layer outputs in a pyramid manner. Each combination will give a prediction result, and then merge all the detection results by non-maximum suppression. YOLOv3 uses a multi-scale prediction method to predict the resolution of 13×13, 26×26, and 52×52 in the network respectively, where the smaller resolution corresponds to the larger scale, but according to this article, it is aimed at the detection of traffic signs in complex environments. The experimental results of the data set show that scale 1 uses a 13×13 size resolution for prediction, which is easy to cause certain false detections and missed detections. In order to enable the network to more effectively use multi-layer feature information, reduce false detections and missed detections. And avoid the complicated calculation caused by the new structure at the same time. This paper draws

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0015

on the idea of the residual dense network proposed by ZhangY et al. [13], and only introduces the residual dense module in the lower layer of the feature map resolution in the YOLOv3 network to replace the original residual block. That is, the method of introducing dense residual blocks into the last 4 consecutive residual blocks. At the same time, local feature fusion and global feature fusion are introduced for feature extraction. After replacement, the input of the residual dense block is generated by 1024 3×3 convolution kernels to generate 1024 feature maps, and secondly, after 512 1×1 convolution kernels in the residual dense block are convolved to generate 512 feature maps. In this regard, this paper uses a shortcut connection to perform a linear mapping WS to match the dimensions of the two (through 1×1 convolution). And so on in subsequent connections. In addition, each RDB output feature map is spliced, and 1×1 and 3×3 convolutions are used for feature extraction. In this paper, the improved YOLOv3 network is named YOLOv3-ResDB network as shown in Figure 4, where some parameters of the convolutional layer are omitted, please refer to the Darknet-53 feature extraction model in Figure 2.

3.2 Improvement based on resnet network

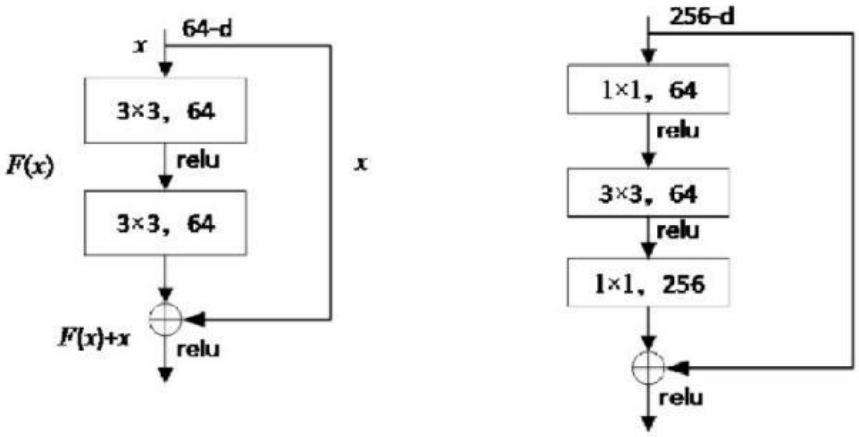

Figure 3. Residual block network structure diagram

Deepening the convolutional neural network can strengthen the feature extraction ability, but at the same time the phenomenon of gradient disappearance will be more obvious, and the training effect of the network will decrease instead, and the residual network allows the network to be deepened as much as possible. Its structure is shown in Figure 3. Therefore, this article uses resnet101 to replace the original darknet53 to improve the detection performance of yolov3. In addition to the normal output of the convolutional layer, the residual module has a branch to connect the input directly to the output. The output and the output of the convolution are arithmetic added to get the final output, as shown in equation (1), x is For the input of the structure shown, F(x) is the output of the convolution branch, and H(x) is the output of the entire structure. The residual structure artificially creates an identity map, which makes the entire structure converge in the direction of the identity map, ensuring that the final error rate will not get worse due to the increase in depth.

$$H(x) = F(x, w_i) + x \tag{1}$$

$$H(x) = F(x, w_i) + W_i x \tag{2}$$

$$F(x, w_i) = W_2 \sigma(W_1 x) \tag{3}$$

Each residual unit usually consists of several superimposed convolutional layers Conv, batch regularization layer BN, and ReLU activation. In equation (3), F(x, wi) represents the residual learning mapping, where σ represents the ReLU activation function, and the bias term is omitted here. In equation (2), when the F and x dimensions are different, the bias is added Wx, its advantage is that it will not increase the training parameters.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0015

The improved network structure is shown in Figure 4. Since YOLOv3 has a weak ability to detect small targets, corresponding adjustments have been made in the process of improvement. After deep convolution, the representation is stronger, but detailed information is easy to lose. If deep features and shallow features can be combined, the feature description of the feature map for small targets will be improved. Starting from this idea, this paper uses the feature pyramid method to fuse shallow features and deep features, uses the deep output in the residual network as the guide of global semantic information, and uses bilinear interpolation to obtain feature maps of the same scale as the low-level feature maps. The cascading method integrates multi-scale context information. After the dimensionality reduction of 1×1 convolution, the detection and recognition operation of the fusion feature map is performed.

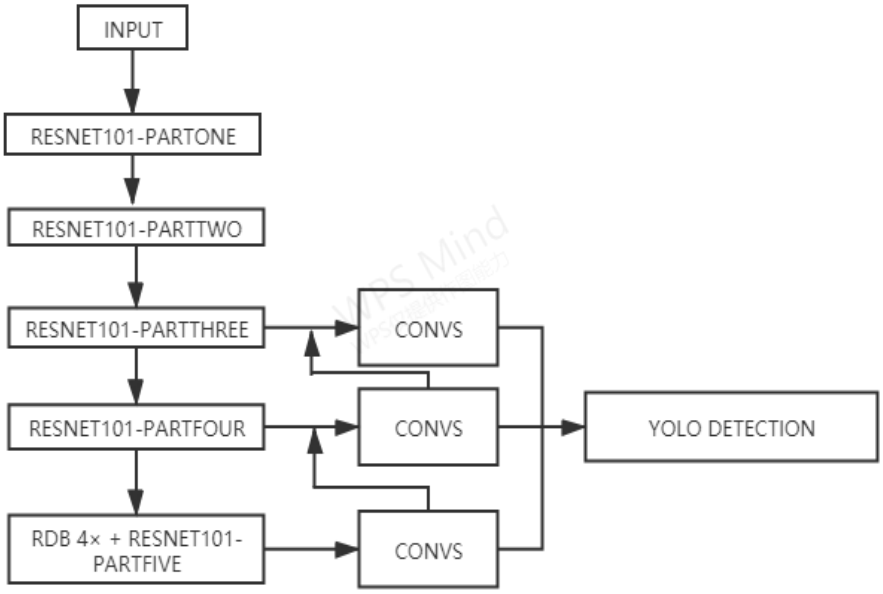

Figure 4. YOLOv3 improved structure diagram

4. Experimental simulation and analysis

4.1 Experimental platform

The experimental environment of this article is: Intel(R)Core (TM)i7-9750H CPU @ 2.60 GHz, 16G memory, graphics card RTX2060, Window10, 64-bit operating system.

4.2 Experimental data set

The experimental data in this paper are all kinds of traffic sign pictures collected in real scenes in complex environments, which are used as samples of traffic sign detection data in complex environments in this experiment. There are various scales of traffic signs in complex environments in the experimental data set, which can enhance the robustness of the model to various size changes during the training process. In this experiment, a total of 2000 traffic sign pictures in a complex environment were collected. Each image contains a number of different debris in a complex environment or a traffic sign with a very small target. 1000 pictures were randomly selected as the training set and 400 pictures as the test set. The test set contains a total of 735 traffic signs in different environments.

4.3 Detection results and analysis of traffic signs in complex environments

The detection effect of YOLOv3-RDB obtained through experiments is shown in Figure 5. It can be seen from Figure 5 that the network model designed in this paper can detect traffic signs with extremely small targets under a certain degree of interference from a complex environment.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0015

Figure 5. Small target detection in various complex environments

In order to verify the effectiveness of the method in the traffic sign detection in complex environments, this paper compares the Faster-RCNN algorithm in the literature [6], the YOLOv2 algorithm in the literature [9] and the YOLOv3 algorithm in the literature [11]. In the complex environment, the detection accuracy and recall rate of traffic sign data set, the experimental results of each detection algorithm are shown in Table 1.

Table 1 Comparison of several different algorithms

Network model	total people	Number of correct detections	Total detection number	Accuracy %	Recall rate%
Faster R-CNN	735	586	698	83.9	79.7
YOLOv2	735	566	659	85.9	77.0
YOLOv3	735	652	715	91.2	88.7
YOLOv3-ResDB	735	655	704	93.1	89.1

As can be seen from Table 1, this article introduces the residual dense block in the deep layer of the YOLOv3 network feature extraction model. Compared with the previous network model, the detection accuracy and recall rate are improved. The Faster-RCNN algorithm in the traditional machine learning mainstream algorithm is in the detection process. The region generation network is used to generate more than 2000 target candidate regions, and then CNN is used to classify the candidate regions. The YOLO algorithm directly uses CNN to process the entire image, reducing the computational complexity, and its detection speed is faster than Faster-RCNN. Both YOLOv2 and YOLOv3 use multi-scale and cross-scale feature fusion, but simple multi-scale is not the key to improving target detection performance. In YOLOv2, 32× downsampling is used for regression prediction box with anchors. The larger downsampling factor has a larger receptive field. This is beneficial to the classification task, but it damages the target detection and localization, that is, the small target will disappear during the downsampling process, and the boundary of the large target cannot be located accurately. In response to the above problems, the YOLOv3 network uses three different scales for box prediction, and draws on the idea of FPN, and performs target prediction in the early stage of downsampling to improve the detection and location of small targets. In addition, the input size of YOLOv2 is an image, and finally only the feature map is used, which leads to poor traffic sign detection performance in complex environments. However, the relatively simple network structure has a faster detection speed. This paper mainly introduces residual dense blocks based on the YOLOv3 network structure. Compared with residual blocks, there are more cross-layer connections, and additional 1×1 and 3×3 convolutions are added for local feature fusion and global feature fusion, which can effectively Using multi-layer convolution features to reduce the loss of information layer by layer, etc., to achieve multi-layer feature reuse and fusion. At the same time, because the resnet101 network is used instead of the original darknet-53 network, the feature extraction becomes better, and the final detection result is also improved.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0015

5. Conclusion

This paper proposes to use the target detection network YOLOv3 to detect traffic signs in complex environments. In this network, the residual dense block unit in the residual dense network is deeply combined, and the resnet101 network is replaced by the darknet-53 network. This not only effectively reduces The multi-layer feature information of the network is lost, and the computational complexity caused by the new structure is avoided, and the multiplexing and fusion of multi-layer information is realized. The research in this paper not only improves the performance of YOLOv3 network in traffic sign detection and small target detection in complex environments, but also provides a new method for other target detection, such as traffic vehicles, ships at sea, and animal identification in forests. However, this article only manually collects images of traffic signs in complex environments for training and testing. However, in the complex environment with more debris and dense traffic signs overlap, its detection performance still needs to be improved and improved.

References

[1] SILVER D, HUANG A, MADDISON C J, et al.Mastering the game of GO With deepneural networks and tree search[J]. Nature, 2016, 529(7587):484-489.

[2] GIRSHICK R,DONAHUE J,DARRELL T,et al.Feature hierarchies for accurate objecte detection and semantic segmentation [C]// IEEE Conference on Computer Vision and Pattern Recognition ,Vloumbus, 20014:580-587.

[3] ENGELCKE M, RAO D, WANG D Z, et al.Vote3Deep:Fast Object detection in 3D point clouds using efficientconvolutional neural networks [C]// IEEE Iinternationalconference onRobotics andAutomation, IEEE, 2017:1355-1361.

[4] SUN C,WANG X,LU P,et al.Object ranking on deformable part models with bagged Lambda MART [C]// Asian Conf on computer Vision,ACCV 2014,Springer,2014:59-71.

[5] GIRSHICK R. Fast R-CNN [C]// IEEE Iinternational conference on computer Vision, Santiago, 2015: 1440-1448.

[6] REN S, HE K, GIRSHICK R, et al. Faster R-CNN:Towards realtime Object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis &Machine intelligence,2015, 39(6):1137-1149 .

[7] ZHANG L, LIN L, LIANG X, et al.Is Faster R-CNN doing well for pedestrian detection [C]// European conference on computer Vision, Las Vegas, 2016:443-457.

[8] Chen Guangxi, Cai Tianren, Huang Yong, et al. Pedestrian detection based on aggregated features and convolutional neural networks [J]. Computer Engineering and Design, 2018, 39(7): 2059-2063, 2068.

[9] REDOM J, FARHADI A.YOLO9000:Better,Faster,stronger[J].arXiv preprint arXiv:1612.0824,2016.

[10] REDOM J, DIVVALA S, GIRSHICK R, et al. You only look once:Unified,realtime Object detection [C]// IEEEconference on computer Visionand pattern Recognition,Las Vegas,2016: 779-778.

[11] REDOM J, FARHADI A. YOLOv3: An incremental improvement [C]// IEEE conference on computer Vision and Pattern Recognition, 2018.

[12] Cao Chuan, Zhang Hongying. Face recognition algorithm based on improved residual network [J]. Sensors and Microsystems, 2018, 37(8): 127-129, 133.

[13] ZHANG Y,TIAN Y,KONG Y,et al.Residual dense network for image super-resolution [C]// IEEE conference on computerVision and Pattern Recoginition,2018.

[14] KE W, CHEN J,JIAO J,et al.SRN:Side-output redisual network for Object symetry detection in the wild [C]// IEEE conference on computer Vision and Pattern Recoginition,2017.

[15] LINT Y,DOLLAR P,FIRSHICK R,et al.Feature pyramid networks for Object detection [C]// IEEE conference on computer Visionand Pattern Recoginition,IEEE computer Society,2017: 936-944.