International Core Journal of Engineering Volume 7 Issue 3, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0042

Static Vehicle Detection Based on Improved Faster RCNN

Junjie Gong, Hongqiong Huang

Shanghai Maritime University, College of information engineering, Shanghai 200135. China.

Abstract

Faster RCNN convolutional neural network to study the problems existing in vehicle detection: 1. Different scales of vehicle, especially when the remote vehicle detection rates are low. 2. The shape and size of the vehicle would change due to motion or camera, which makes it impossible to detect the vehicle. 3. When the vehicle overlaps with other things or the distance is too close, the frame regression algorithm leads to missing detection. In this paper, the convolution neural network model is improved by introducing the multi-connection feature pyramid model, variable convolution and softening non-maximum linear suppression algorithm, and then the online difficult case mining strategy and multi-scale training strategy is introduced to improve the training strategy. The experimental results show that the average accuracy (mAP) of the improved faster RCNN on COCO2019 data set reaches 54.8%, which is nearly 11.4% higher than that of the unimproved model, and it is superior to other mainstream detection networks in detection accuracy.

Keywords

Static Vehicle Detection; Variable Convolution; Feature Pyramid; Non-Maximum Suppression; Online Difficult Case Mining; Multi-Scale Training.

1. Introduction

Today, with the rapid development of intelligence, motor vehicles have become the main means of human transportation, and various intelligent systems with motor vehicles as research objects have emerged at the historic moment[1]. Vision-based vehicle detection as the basis of subsequent operations, such as tracking, ranging, path planning and so on, which has great significance of research. Only by accurately detecting the target can it lay a solid foundation for subsequent operations, such as vehicle tracking and path planning. The most important feature of the two-step detection method represented by faster regional convolution neural network algorithm is the suggestion of regional networks. Compared with SSD and YOLO single-step detection algorithms, Faster RCNN uses the suggestions of regional network by using the calibrated frame, so that RPN can distinguish the foreground from the background, and provides several frames about suggestions for the next specific category detection network. The accuracy of Faster RCNN is higher than that of other detection networks.In this paper, Faster Rcnn is used as the detection network for vehicle detection. Zhou Feiyan and others mentioned that because geometric structure of the sampling in convolution network is fixed, its geometric transformation modeling ability is limited, and the network may not adapt to geometric changes of the target[3]. Wang Huabin and others put forward that an offset variable can be added to the position of each sampling point in the convolution kernel, the sampling geometry is not fixed, and the convolution kernel, that is, the block of sampling, can sample near the current position instead of a sampling block with a fixed geometry. Although it would increase a small amount of computation, it can make the network adapt to geometric changes of images. On traffic monitoring image dataset, the average accuracy of the deformable convolution neural network can be improved from 70% to 75%. Xu Zhijing and others mentioned that Faster R-

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0042

CNN used convolution layer with four times of down sampling to carry out subsequent object classification and bounding box regression. But there is an obvious defect in doing so, small objects have less pixel information and are easily lost in the process of down sampling. In order to deal with the detection problem of such objects with obvious size differences[5]. Guo Qifan and others put forward that using the image pyramid to enhance multi-scale changes can obviously improve the detection accuracy of small objects, but it would bring more computation. Kuang Xianyan and others mentioned that Faster RCNN would use non-maximum suppression (NMS) for post-processing to screen candidate frames. The method is to sort the detection boxes according to the scores, and keep the boxes with the highest scores. According to the traditional NMS, when the target distance is too close, it would be deleted. Resulting in missed detection[7]. Hu Hui and others proposed that softening non-maximum suppression can improve the situation of false deletion caused by too close target distance[8].

2. Improved Faster RCNN based on training strategy and model

2.1 Model improvement

2.1.1. Introducing variable convolution into convolution layer

The geometric structure of sampling in convolution network is fixed, and its ability of geometric transformation modeling is limited, so the network may not adapt to geometric changes of the target. In view of this, this paper introduces variable convolution, and further adjusts the displacement of the position information sampled in space. The displacement can be learned in training, and no extra supervision signal is needed[11]. An offset variable is added to the position of each sampling point in the convolution kernel, that is, the displacement of the original sampling point and the offset point of sampling.

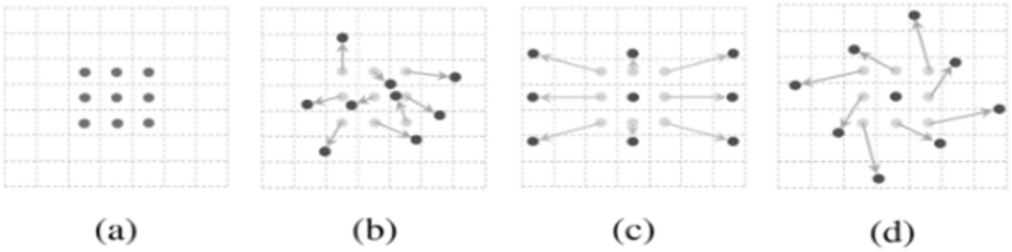


Figure 1. (a) Ordinary convolution (b), (c) and (d) Variable convolution

The sampling mode of variable convolution is shown in Figure 1 above. (a) sampling according to the common convolution rule, (b), (c) and (d) are deformable convolutions, and a displacement is added to the normal sampling coordinates. The schematic diagram of the variable convolution layer is shown in Figure 2.

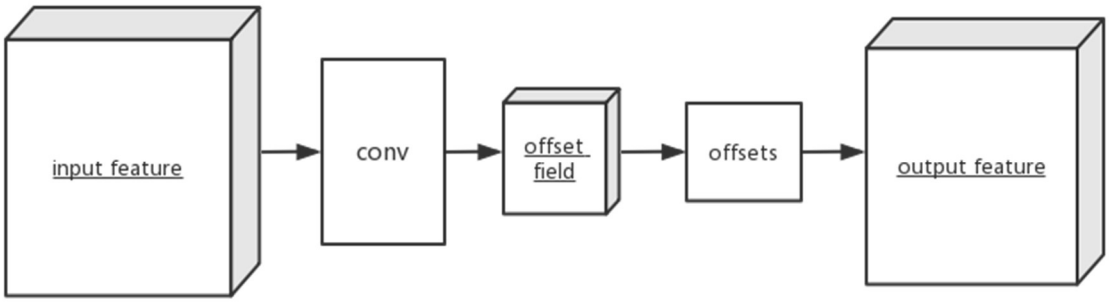


Figure 2. Schematic diagram of variable convolution layer

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0042

2.1.2. Introduce feature pyramid after convolution layer

In convolutional neural network, the spatial resolution of low-level features is high, and the learning is basically contour features or target textures, and the location information is rich. High-level features are more abstract semantic features, used for target classification. The steps of feature extraction using FPN are shown in Figure 3.

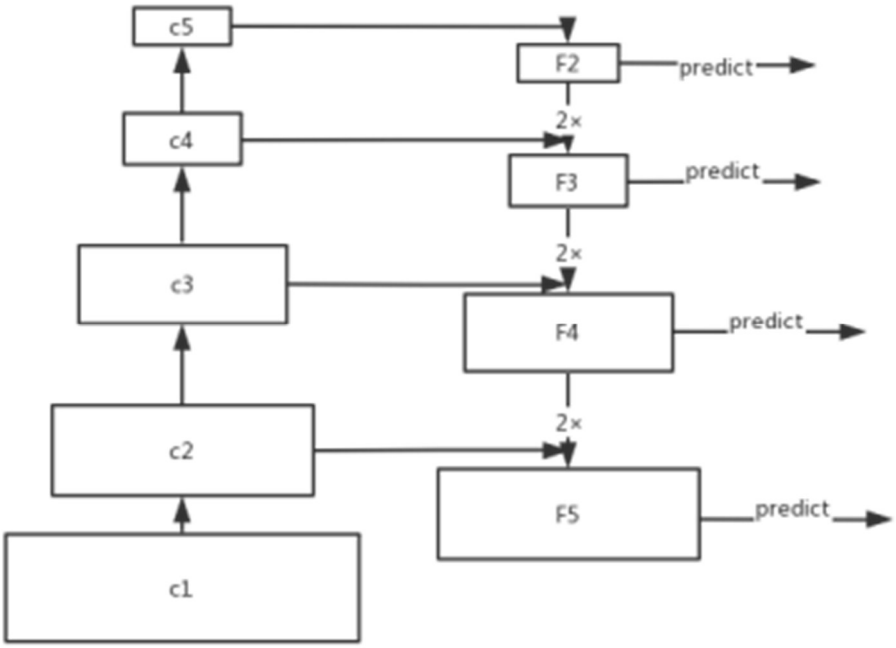


Figure 3. FPN feature extraction process

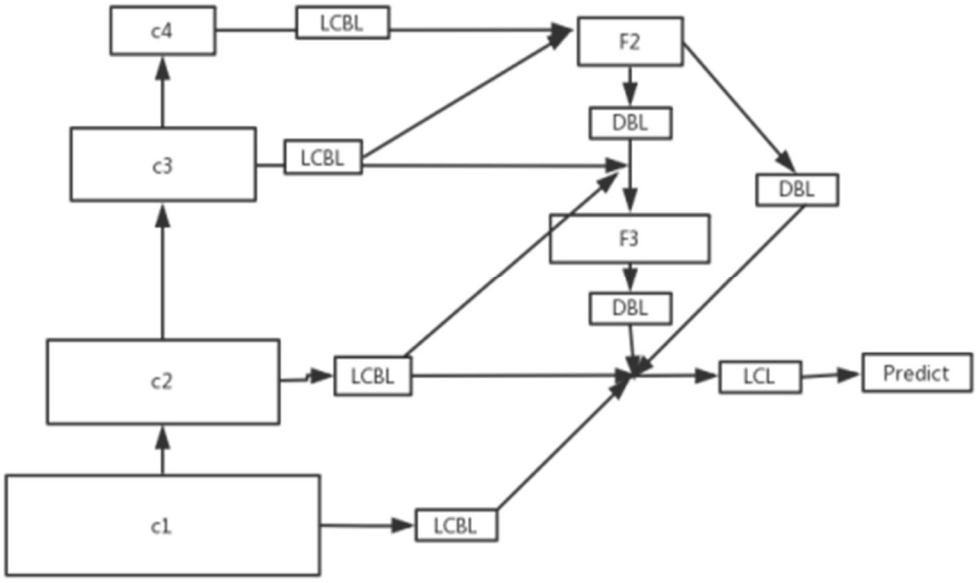


Figure 4. Multi-connected pyramid structure

However, the feature fusion in FPN has the following two problems. Firstly, the generation of side feature layer is only related to this layer and higher layer, ignoring the lower feature layer, while the feature map of lower layer is more important for vehicle detection. Secondly, for Faster RCNN network, the path between the upper layer and the lower layer is relatively long, which easily leads to the loss of the characteristic information of the upper layer in the propagation process.To solve the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0042

above problems, this paper proposes a multi-connection feature pyramid model. Compared with the original feature pyramid structure, this paper adds two connection paths. As shown in Figure 4. In order to reduce the loss of high-level feature information and improve the accuracy of classification, the idea of residual network is used to fuse the higher level feature map with the lower level feature map. The feature layer Conv2 is taken as an example. Firstly, Conv1, Conv2, Conv3 and Conv4 layers required for fusion are convolved, and the number of channels of feature graph is unified to 256 by using a convolution kernel with the size of 3×3, so as to compress the dimensions and reduce the amount of parameters. BN layer is added to increase the generalization ability of the network and prevent over-fitting. Then, by deconvolution, the feature maps of Conv3 and Conv4 are sampled, and the low-level Conv2 is down sampled, so that the four feature layers have the same size after sampling. Features are fused by pixel-by-pixel summation, and fused by 3×3 convolution kernel. Finally, the generated feature map is output for detection and classification tasks.

2.1.3. Softening non-maximum suppression algorithm

NMS algorithm is shown in formula (1). In the formula, Si is the score of each border, M is the current box with the highest score, BI is one of the remaining boxes, and Nt is the threshold value set. As $iou(A,B) = \frac{A \cap B}{A \cup B}$, when IoU is greater than Nt, the score of the frame is directly set to 0, which is equivalent to being discarded, which may cause the missing of the frame[14].

$$\mathrm{si} = \begin{cases} si, iou(M,bi) < Nt \\ 0, iou(M,bi) \geq Nt \end{cases} \tag{1}$$

Because NMS algorithm is a non-zero algorithm, it is easy to miss vehicle detection. In this paper, soft-NMS (Soft-NMS) is used to suppress the missing inspection of vehicles when they overlap. Soft-NMS uses an attenuation function to gradually reduce the candidate box score, as shown in formula (6).

When the two frames are closer, the function value would gradually decay to 0, which effectively improves the detection accuracy when the vehicle is blocked.

$$si \begin{cases} si, & iou(M,bi) < Nt \\ si(1 - iou(M,bi)), & iou(M,bi) \geq Nt \end{cases} \tag{2}$$

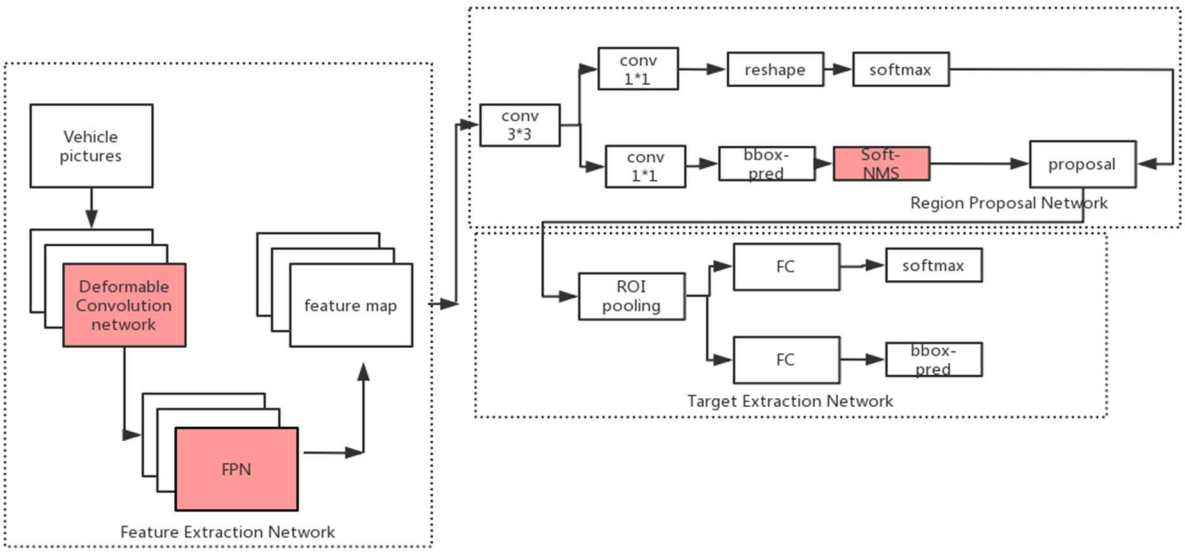


Figure 5. Improved network model

2.2 Overall network design

The improved network of the model is shown in Figure 5. Firstly, in the feature extraction stage, the backbone trunk convolution network Resnet-50 adds offset instead of the standard convolution

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0042

network, which is used to calculate the offset value of pixels and achieve the function of variable convolution. Secondly, after convolution, FPN module is added to fuse feature maps of different scale outputs of Resnet-50, and NMS algorithm is used for screening. Finally, in the operation of the proposal, the candidate frames would be screened by linear regression, and the original NMS algorithm would be replaced by Soft-NMS algorithm to achieve the purpose of reducing missed detection.

3. Experimental design and simulation results

3.1 Experimental platform and data set

The experimental environment is ubuntu18.04 operating system, inter (r) core (TM) i7-7800x CPU @ 3.50ghz processor 16GB memory, NVIDIA GeForce GTX1080Ti 11GB graphics card. All experiments are implemented in pytorch deep learning framework, and the programming language is Python3.7.The experimental data set comes from the public data set in 2019, and 5524 pictures are selected. 2762 pictures were used as training samples, and the remaining 2762 pictures were used as test sets.

3.2 Experimental results and analysis

3.2.1. Experimental evaluation index

Average precision is adopted as the evaluation standard of vehicle target inspection. As shown in equation 3, AP is the average accuracy.

$$mAP = (Ap_1 + AP_2 + \cdots + APk) / K \tag{3}$$

3.2.2. Comparison under different algorithms and strategies

In this experiment, Faster RCNN is used as the basic detection network and resnet-50 as the backbone network. On this basis, multi-scale mining (mstrain) and online hard case mining strategy (ohem) is introduced, and the algorithms are shown in Table 1.

Table 1. Time-consuming comparison of mAP and single picture of different algorithms

Method	mAP	Frames/sec
Original Faster RCNN	42.2%	16
Faster RCNN+DCN	48.1%	16
Faster RCNN+FPN	45.6%	15
Faster RCNN+MCFPN	47.7%	15
Faster RCNN+Soft-NMS	46.7%	18
Faster RCNN+ohem+mstrain+DCN+MCFPN+Soft-NMS	54.8%	15

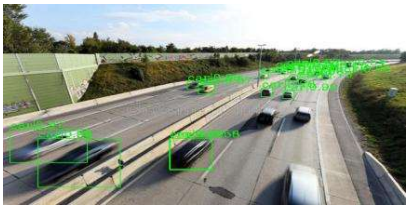
The average accuracy of Faster RCNN after multi-fusion improvement is nearly 11 percentage points higher than that of Faster RCNN before improvement. Although the accuracy is higher than that of mainstream networks such as YOLO and SSD, the speed is still slower than that of other mainstream networks.

3.2.3. Detection Effect in Different Scenes

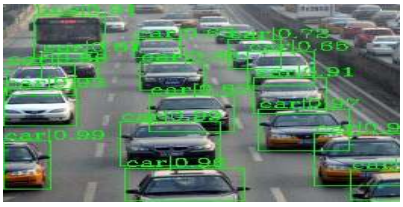
In order to verify the robustness and generalization ability of this experiment, several pictures are selected from the test data set for vehicle detection. As shown in Figure 7. It is mainly the detection results of different light intensity, different vehicle speed and different degrees of occlusion in traffic. Typical detection of results are shown in Figure 6 above, with the frame as the vehicle target and the upper left corner as the target label and confidence. As shown in Figure 6(a), when observing distant vehicles and vehicles near the camera, on the one hand, the scale is small, on the other hand, the speed is fast, which leads to deformation, but this method can still detect them. Observe that in Figure 6(b), vehicles, bicycles and the flow of people cross each other and overlap, resulting in partial occlusion

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0042

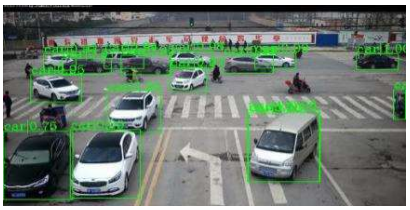
or only partial image of edge, but this method can also detect all vehicles. Looking at Figure 6(c), it can be seen that most of the vehicles can be detected because the vehicles in the distance are seriously distorted and blocked. As shown in Figure 6 (d), (e) and (f), the vehicle without a block can still be detected at night. Based on the above experimental results, this model can identify vehicles with different illumination, different shooting distance and different contour features more accurately than the Faster RCNN detection model, which shows that this vehicle detection model has higher accuracy and robustness when applied to vehicle detection tasks than the original Faster RCNN.



(a) The deformation caused by the excessive speed of the vehicle



(b)Overlapping of vehicles with each other



(c)Miscellaneous traffic between vehicles and pedestrians

(d)Night conditions

(e)Night conditions

(f)Night conditions

Figure 6. Detection results in different scenes

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0042

References

[1] Krizhevsky A,Sutskever I,Hinton G E.Imagenetclassification with deep convolutional neural networks[C]. Advances in Neural Information Processing

[2] Systems: Proceedings of the Franke U,GavrilaD,GorzigS,etal.Autonomousdriving goes downtown [J]. IEEE Intelligent Systems, 1998,13(6):40-48.

[3] Feiyan Zhou, Linpeng Jin, Jun Dong. Review of Convolutional Neural Networks [J]. chinese journal of computers, 2017,40(06):1229-1251.

[4] Huabin Wang, Min Han, Guanghui Wang, Yu Li. Variable structure convolution neural network method for remote sensing image element extraction [J]. Journal of Surveying and Mapping, 2019,48(05):583-596.

[5] Zhijing Xu, Hai Huang. ship target detection in SAR images based on multi-connected feature pyramids [J/OL]. advances in laser and optoelectronics: 1-14 [2020-11-12]. http://kns.cnki.net/kcms/detail/31.1690.tn.20200.200000000006

[6] Qifan Guo, Lei Liu, Jun Zhang, Wenjuan Xu, Wenfeng Jing. Multi-scale feature fusion network based on feature pyramid [J]. Journal of Engineering Mathematics, 2020,37(05):521-530.

[7] Hui Hu, Chen Zeng. Vehicle target detection method based on improved R-FCN [J]. Computer Engineering and Design, 2020,41(04):1164-1168.

[8] Maotao Zhu, Hongxiang Zhang, Ruihua Fang. Research on vehicle detection method based on RCNN [J]. Mechanical and Electrical Engineering, 2018,35 (8): 880-885.

[9] Kaijing Shi, Hong Bao, Bingxin Xu, et al. Vehicle detection method in front of intelligent vehicle based on Faster RCNN [J]. Computer Engineering, 2018,44 (7): 36-41.

[10] Kaijing Shi, Hong Bao. Research on vehicle ahead detection based on improved FAST R-CNN [J]. Computer Science, 2018,45 (S1): 179-182.

[11] Fukai Zhang, Feng Yang, Ce Li. fast vehicle detection method based on improved YOLOv3 [J]. computer engineering and application, 2019,55 (2): 12-20.