International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 04 – Issue 09 || September 2019 || PP. 74-79

Vehicle and Pedestrian Detection Based on Improved Faster R-CNN

***Huailong Yi[1]**

Zhuangzhuang Mao[2]

Mengchao Liu[3]

[1,2,3]School of Mechanical and Power Engineering
Henan Polytechnic University, 454003, Jiaozuo, China
** Corresponding author*

Abstract: The paper proposes a vehicle and pedestrian object detector based on the two-stage Faster R-CNN for complex autonomous driving scenarios, which is an end-to-end deep convolutional neural network. To solve the problem of insufficient real-time performance of traditional two-stage Faster R-CNN, the paper uses ResNet as the backbone network of Faster R-CNN. We compare the effects of different backbone network structures (VGG_CNN_M_1024, VGG-16, ResNet50, ResNet101) on the precision and real-time performance of object detection. At the same time, according to the size of vehicles and pedestrians on the RobotCar and KITTI datasets, different aspect ratios are designed to make the anchor frame generated by RPN more accurate. The experimental results on the RobotCar and KITTI datasets show that the proposed approach obtains significant improvements.

Keywords: Faster R-CNN; Vehicle and pedestrian detection; RPN; ResNet

1. Introduction

With the rapid development of deep learning, object detection method based on deep Convolution Neural Network has been widely used. Especially in the field of autonomous driving, deep learning brings a powerful solution to the environmental awareness stage. Vehicle mounted camera is the most commonly used sensor in autonomous driving system. It is mainly used for detecting static and dynamic objects in road scenes. However, vehicle mounted cameras are susceptible to other factors such as weather. In long-distance driving, due to the influence of light, weather, dynamic objects, seasonal effects and structure, a large number of changes will occur under different conditions and different environments, which also brings great challenges to the autonomous driving algorithm.

In the past, hand-made local invariant feature representations were always dominant, such as Scale Invariant Feature Transforms (SIFT [1]) and Histograms Of Gradient (HOG [2]). The most famous is the multi-scale deformation component model proposed by the Felzenszwalb team (DMP [3]), which has been very successfully for generic object detection. However, these feature representation methods required careful design work and considerable domain expertise. For complex and variable scenarios, these traditional object detection methods are difficult to achieve certain generalization capabilities, and have not achieved satisfactory results in

some public data sets(RobotCar [4] dataset and KITTI [5] dataset).

In recent years, compared with the traditional feature extraction method, the deep neural network has better effect. Common object detection algorithms such as R-CNN [6], Fast R-CNN [7], Faster-RCNN [8], YOLO [9], SSD [10] have achieved good results. The two-stage framework [8] proposed the Faster R-CNN algorithm, which generated object candidate regions through an efficient and accurate Regional Proposal Network (RPN), and further merged RPN and Fast R-CNN into one network by sharing convolution features. The whole algorithm was an end-to-end process, which greatly improved the speed of the algorithm. The one-stage framework [10] proposed a SSD algorithm similar to YOLO. Since the YOLO algorithm simply divided the image equally, the positioning accuracy of the algorithm was not as high as that based on the region proposal algorithm. Therefore, SSD added the anchor mechanism of Faster R-CNN on the basis of the YOLO, which combined the functions of regional proposal, and used the deep network feature maps of different scales to predict the object at each position. Generally speaking, compared to the two-stage framework, the one-stage framework needs to be improved in terms of accuracy. One important reason for the inaccuracy is the serious imbalance between the positive and negative sample numbers of candidate frames. The two-stage framework improves the regression accuracy of the bounding box due to the introduced RPN. Because a large number of negative sample frames are removed to solve the imbalance problem of positive and negative samples, the one-stage framework has greater advantages in real-time.

Previous network Faster R-CNN uses AlexNet [11] or VGG16 [12] as the skeleton structure. Because the parameters of VGG16 are huge, it requires high calculation requirements. The model can't meet the real-time requirement of autonomous driving. Inspired by ResNet[13], the paper uses ResNet as the backbone network of Faster R-CNN. We compare the effects of different backbone network structures (VGG_CNN_M_1024, VGG-16, ResNet50, ResNet101) on the precision and real-time performance of object detection. At the same time, according to the size of vehicles and pedestrians, different aspect ratios are designed to make the anchor frame generated by RPN more accurate. Our improved model is verified on the RobotCar dataset and on the well-known KITTI benchmark dataset.

2. Network Architecture

Faster R-CNN is a two-stage object detector which relies mainly on the Regional Proposal Network (RPN) to generate region proposals. This proposal generator could be learned via supervised learning methods. RPN is a fully convolutional network which takes an image of arbitrary size and generates a set of object proposals on each position of the feature map. The network slid over the feature map using a $n \times n$ sliding window, and generated a feature vector for each position. The feature vector was then fed into two sibling output branches, object classification layer and bounding box regression layer. These feature maps are passed to the second stage detection network for and classification and Regression. In the paper, Resnet101 is used to replace the original VGG16 as the backbone network. The network structure is shown in Figure 1.

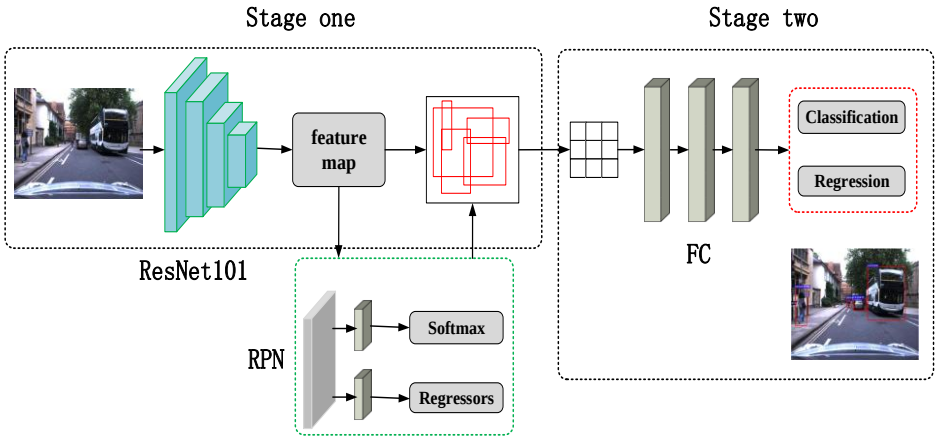

Figure 1. The architecture of improved Faster R-CNN

2.1 ResNet module

With the depth of the network increases, there will be a degradation problem. When the network becomes deeper and deeper, the accuracy of training will tend to be flat and the training error will become larger. This is obviously not caused by over-fitting, because over-fitting means that the training error of the network will continue to decrease, but the test error will become larger. In order to solve this degradation phenomenon, ResNet was proposed. Instead of directly fitting the desired feature mapping with multiple stacked layers, an explicit residual mapping is fitted with them. The difference between the residual network and the ordinary network is the introduction of skip connection, which can make the information of the last residual block flow into the next residual block without hindrance, improve the information flow, and avoid the problem of disappearance gradient and degradation caused by deep network. Residual learning block is shown in Figure 2, which uses the "shortcut connection" method to transfer input x directly to the lower part of the output as the input of the initial result. This output can be expressed as $H(x)=F(x)+x$. When $F(x)$ is 0, $H(x)$ is equal to x. This is the identity mentioned earlier. On this basis, changing ResNet learning goal is no longer through each layer of neural network learning, but the difference of learning objectives $H(x)$ and x. The residual can be expressed as $F(x)=H(x)-x$ The ResNet101 backbone network used in the paper is shown in Figure 3. It is mainly composed of the convolution core with the size of 1×1, 3×3 and the stack of modules with 64,128,256,512 channels.

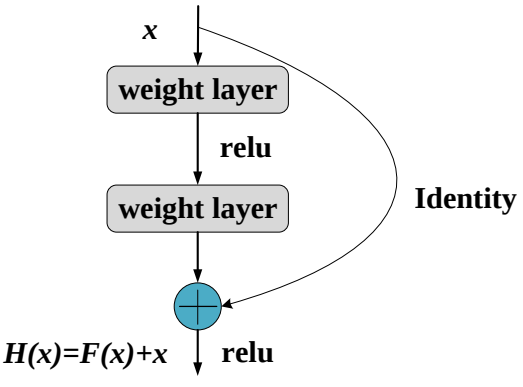

Figure 2 Residual learning block

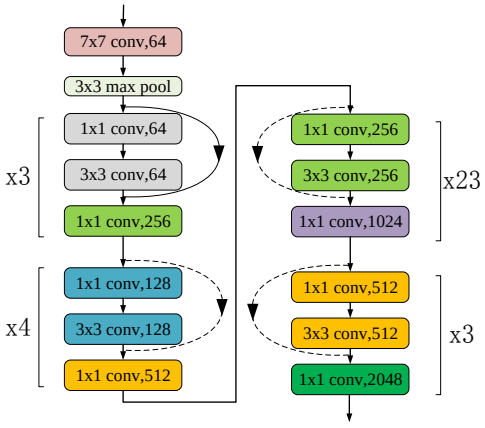

Figure 3 The backbone structure of ResNet101

2.2 Network anchor design

Different levels of feature layers in convolutional neural networks have different receptive field sizes. In the network structure of the paper, different feature layers are used to detect objects of different sizes. Because of the significant differences in size and shape of vehicles and pedestrians, anchor points are designed to have different grid sizes and aspect ratios when detecting pedestrians and vehicles. In these feature maps, each feature layer is associated with a specific scale of anchor points. Four different aspect ratios are designed based on the shape ratio of the vehicle and pedestrian on the RobotCar dataset and the KITTI dataset, each anchor corresponds to sixteen regions of four scales (64^2,128^2, 256^2, 512^2) and four aspect ratios (0.5,0.7,1.0,2.0) of input image prediction.

3. Experiments and Results

3.1 Implementation Details

To evaluate our approach proposed in the paper, we conducted experiments on two representative autonomous driving datasets. The RobotCar dataset and the KITTI dataset are trained and evaluated. The performance of object detection is measured by mean accuracy (mAP). All experiments in the paper are based on the Caffe deep learning framework and are trained on a single NVIDIA GTX1070. The experimental hardware configuration is Intel Core i7 processor. The operating system is Ubuntu 18.04 platform and the programing environment is based on Python. The maximum number of iterations of the Faster R-CNN are set to 80000 times.

3.2 Expeiments on the Robot Car and KITTI Dataset

On the RobotCar dataset, 2,476 images currently labeled are used to train and evaluate the approach. Among them, 1500 images of training set and 976 images of test set are converted into LMDB data format of Pascal VOC, which is convenient for high-speed data reading of the network. On the KITTI dataset. There are 5236 images in the training set and 2245 images in the test set. The recognized the KITTI benchmark dataset is transformed into a combination of cars and pedestrians.

The experimental results are shown in Table 1. Compared with VGG16, ResNet101 used in the paper has considerable comparability in accuracy. At the same time, it takes less time to test a single image than the algorithm in the paper. By comparing different models, this further proves the effectiveness of the method in the paper. We also visualized the Precision-Recall curve for car, pedestrian, as shown in Figure 4.

Table.1 Performance results of different algorithms on the RobotCar/KITTI dataset

Method	Runtimes(s)	Average accuracy (%)		
		mAP	Vehicle (AP)	Pedestrian(AP)
Faster-RCNN+GG_CNN_M_1024	0.3625	59.45/56.58	65.73/62.15	53.16/51.01
Faster-RCNN+VGG16	0.5602	81.54/65.49	87.75/76.65	75.33/54.32
Faster-RCNN+ResNet50	0.4217	59.01/56.92	70.43/68.34	47.59/45.49
Faster-RCNN+ResNet101	0.4902	87.16/68.54	87.77/79.27	76.56/57.82

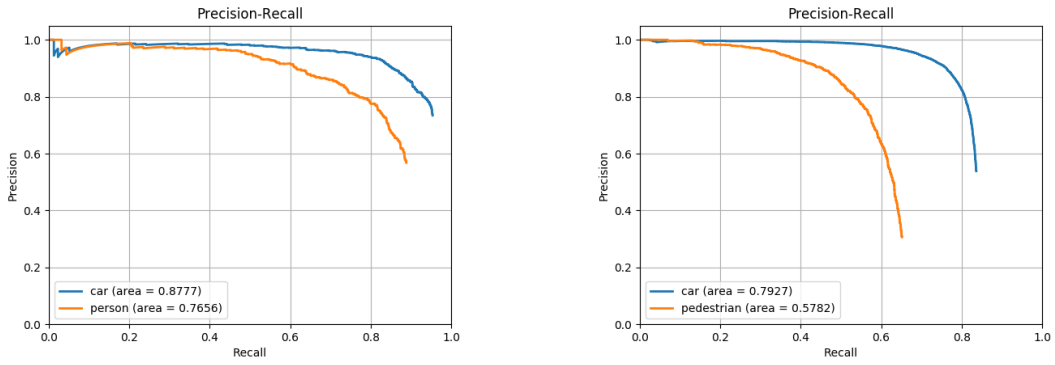

Figure.4 Precision-Recall curve

Figure 5 and Figure 6 show some qualitative results of our approach. Our method can handle different scenes. It is robust to small object in different distance. And when the scene is crowded, our method still performs well in most cases.

Figure.5 Qualitative results of object detection on the RobotCar dataset (ResNet101

Figure.6 Qualitative results of object detection on the KITTI dataset (ResNet101)

4. Conclusion

In this paper, an improved network based on Faster R-CNN is proposed to vehicle and pedestrian detection. Our approach can handle different scenes. It is robust to small object in different distance. The paper uses ResNet as the backbone network of Faster R-CNN. We compare the effects of different backbone network structures on the precision and real-time performance of object detection. At the same time, according to the size of vehicles and pedestrians on the Robot Car and KITTI datasets, different aspect ratios are designed to make the anchor frame generated by RPN more accurate. By comparing different models on the Robot Car and KITTI datasets, this further proves the effectiveness of the proposed approach in the paper.

Reference

[1] D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004

[2] N. Dalal and B. Triggs,"Histograms of Oriented Gradients for Human Detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893, 2005

[3] P. F. Felzenszwalb, R. B. Girshian, D. Mcallester, and D. Ramanan,"Object Detection with Discriminatively Trained Part based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627–1645, 2010

[4] A. Geiger, P Lenz, P. Girshick, and R. Urtasun,"Are we ready for autonomous driving? The KITTI vision benchmark suite," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3354–3361, 2012

[5] W. Maddern, G. Pascoe, C. Linegar, and P. Newman,"1 year, 1000 km: The Oxford Robot Car dataset," International Journal of Robotics Research, vol. 36, no. 1, pp.3–15, 2017

[6] R. Girshick, J Donahue, T. Darrell, and J. Malik,"Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 580–587, 2014

[7] R. Girshick,"Fast R-CNN," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448, 2015

[8] S. Q. Ren, K. M. He, R. Girshick, and J. Sun,"Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi,"You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 779–788, 2016

[10] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, et al.,"SSD: Single Shot MultiBox Detector," Lecture Notes in Computer Science, pp. 21–37, 2016

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton,"ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, vol. 2, pp. 1097–1105, 2012

[12] K. Simonyan and A. Zisserman,"Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference of Learning Representation, 2015

[13] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun,"Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016