Multi-Scale Pedestrian Detection through Feature Fusion in Faster RCNN with VGG16

Lucas Thomas[1],Charlotte Moore[2],
Springfield College[1],Springfield College[2]
Lucaskkk@gmail.com[1],charloott0909@gmail[2]

Abstract:

Addressing the issue of multi-scale challenges in pedestrian detection, this study introduces a novel pedestrian detection model that integrates Faster R-CNN with multi-scale feature fusion. The model employs VGG16 as the foundational feature extraction network, which is utilized for feature fusion. Subsequently, the model generates features characterized by high resolution and rich semantics, which are then used for prediction, followed by classification and regression processes. Experimental results on a custom pedestrian dataset indicate that the proposed method enhances the capability of multi-scale pedestrian detection to a significant degree.

Keywords:

Feature fusion, Pedestrian detection, Deep learning, Multi-scale detection.

1. Introduction

With the rapid development of computer vision based on deep learning, object detection, as oneof the directions of computer vision, has been widely applied in many fields. Pedestrian detection is a classic research field in target detection. Although the general target detection algorithms have achieved relatively good detection results in recent years, there is still a big gap between the results obtained by pedestrian detection and those obtained by human eyes. Since Hinton [1] proposed to use neural network to extract features of objects in images, target detection under deep learning has gradually become a hot research direction [2]. Similarly, the combination of computer vision technology and deep learning for pedestrian target detection has also become the mainstream algorithm in this application field.

Pedestrians vary in scale due to the distance from the shooting lens, so it is easy for convolutional neural network to lose small-sized pedestrians in the image when extracting pedestrian features. The feature map at the lower level of convolutional layer has high resolution and contains more positioning information such as texture boundary, while the feature map at the higher level has low resolution and contains more semantic information. At present, the feature information extracted by the detection algorithm is single, and generally it can only carry out target detection on the feature map obtained by a certain layer, which often leads to missed detection and false detection. Faster RCNN [3] is adopted as the basic detection algorithm to solve the problem of multi-scale pedestrian detection, and the algorithm is improved on this basis.

2. Related Work

In recent years, target detection algorithms are mainly divided into two categories. One is a two-stage detection algorithm, and the representative algorithms mainly include RCNN [4], Fast RCNN [5] and Faster RCNN [3]. Among them, RCNN USES Selective Search algorithm to get a series of candidate boxes by iteratively merging superpixels, and then inputs them as samples to the convolutional neural network to form corresponding feature vectors, and finally classifies them by SVM. Fast RCNN algorithm mainly designs the pooling layer structure of region of interest (ROI) on the basis of RCNN, and proposes the idea of multi-task loss function, which effectively combines the classified loss with the border regression loss. Faster RCNN creatively puts forward candidate region network (RPN), effectively improves the problem of positive and negative sample candidate box generated by Selective Search algorithm, and the proposed FPN [10] solves the problem of multi-scale prediction. The other is a single-stage detection algorithm, which mainly includes SSD [6], YOLO [7], YOLOv2 [8], and YOLOv3 [9]. SSDalgorithm firstly generates image feature graph, then generates different candidate boxes based on feature graph, and finally classifies and regresses the candidate boxes in the same network. YOLO and others divide the grid on the image, and each grid predicts an object, which is easy to cause the missing detection of the target. Two-stage detection algorithm has relatively high detection accuracy, while single-stage detection algorithm has relatively fast detection speed.

3. Multiscale Feature Fusion Model

3.1. Faster RCNN Detection Model

Faster RCNN mainly creatively puts forward candidate region network (RPN) and effectively improves the problem of positive and negative sample candidate box generated by Selective Search algorithm. Therefore, the Faster RCNN algorithm is divided into TWO parts: RPN and Fast RCNN, and the main process is shown in Figure 1

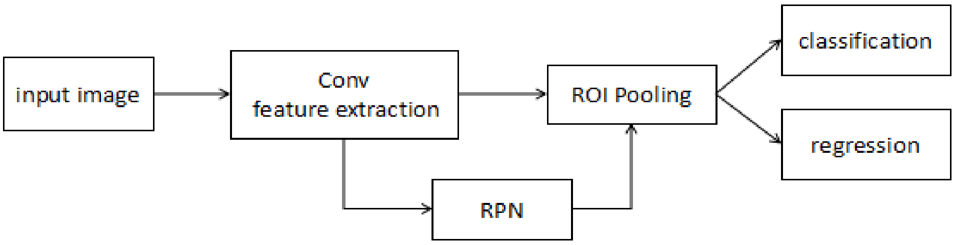

Figure 1. Detecting process of Faster RCNN algorithm

(1) First, the original image is preprocessed, and then the image is input into the detection model.

(2) The convolutional neural network is used to extract image features. There are many networks for feature extraction, including AlexNet, VGG, ResNet, etc. In this paper, the VGG16 network is mainly adopted, and the convolutional layer network is improved to solve the multi-scale problem in pedestrian detection.

(3) The candidate box is extracted by candidate area network (RPN), and the candidate box is classified and regressed.

(4) The candidate box output is corresponding to the feature map, and the feature map of the candidate box with different sizes is processed into the same size through the ROI pooling layer. Finally, the feature map of the candidate box is input to the full connection layer for classification regression.

3.2. MS - Faster RCNN Model

Faster RCNN creatively proposed candidate region network (RPN), effectively improving the deficiencies of previous candidate region extraction, and RPN network directly extracted the candidate box of the feature map obtained through the convolutive layer. It can be seen from the Faster RCNN algorithm process shown in Figure 1 that the extraction of candidate box is usually to input the feature map obtained from the last layer of the convolutive layer into the RPN network, while the feature map at this time usually has high semantics and lacks the positioning information such as low-level texture.

In order to improve the multi-scale feature fusion algorithm proposed in this paper, the main idea is to combine high-level semantic information with low-level detailed information. In this paper, VGG16 feature extraction network is selected as the backbone network. Feature extraction network has five stages, and the convolutive layer generated at the end of each stage is defined as {C1, C2, C3, C4, C5} respectively. Fusion of the top-down process of each layer to 1x 1 channel convolution dimension reduction, and before the feature fusion on sampling by bilinear interpolation operation to expand the resolution of the C5, and then an element of blend and C4, iterative feature fusion process above, in turn finally through 3 x3 convolution layer to reduce aliasing effect brought by the characteristics of the fusion process to get the final figure.The multi-scale feature fusion diagram is shown in Figure 2

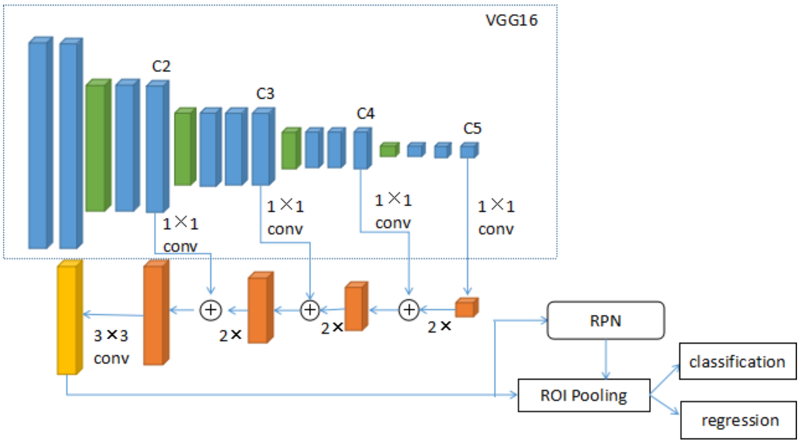

Figure 2. Multi-scale feature fusion model

4. Experiment and Analysis

4.1. Data Set

In this paper, the self-made pedestrian data set is adopted. The data set mainly comes from Internet and mobile phone photos. When collecting samples, it involves as many scenes as possible to increase its generalization ability. First, the data set was preprocessed and the useless photos in the data set were cleaned to obtain a total of 2402 images. Then the images in the data set are marked manually according to certain rules, and the training set and test set are divided. There are 1754 pictures of training set samples, including 648 pictures of test set samples.

4.2. Experimental Details and Evaluation Criteria.

The experiment in this paper is carried out on the basis of Tensorflow deep learning framework. Firstly, the network weight parameters are initialized with ImageNet pre-training. In the experiment, the training methods of RPN and Fast RCNN were alternately trained. In this paper, the optimization method of stochastic gradient descent is adopted to conduct iterative training for the network, and Dropout method is adopted to prevent overfitting. The value is set as 0.5 and the impulse factor is set as 0.9. The initial learning rate was set as 0.01, and the learning rate was attenuated to 0.0001 after 20,000 iterations, and the total number of iterations was 50,000.

The performance of pedestrian detection algorithm is usually evaluated by the generalization ability on the test set, and there are certain evaluation criteria to measure the generalization ability of the algorithm. For basic dichotomies, the categories can be divided into True Positive(TP), False

Positive (FP), True Negative (TN), and False Negative (FN) according to the combination predicted by the classifier. Two common standards for evaluating algorithm performance are Precision and Recall. Precision and recall are defined as

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

The two measures of precision and precision are relatively contradictory. In general, when the precision is high, the precision is often low. On the contrary, when the recall rate is high, the precision rate will be low. If the precision is taken as the vertical axis and the recall as the horizontal axis, "P-R curve" can be obtained. When more than two algorithms are compared, a reasonable criterion is to compare the area (average accuracy, mAP) under the PR curve. The greater the area, the better the performance.

4.3. Analysis of Experimental Results

In order to better compare the overall performance of the multi-scale feature fusion model, the experimental results are obtained by comparing with other pedestrian detection algorithms as shown in the Table1 below

Table 1. Comparison of experimental results

Method	Backbone	MR	Recall	mAP
Fast RCNN	VGG-16	17.4	80.13	78.4
Faster RCNN	VGG-16	14.5	83.22	80.8
MS-Faster RCNN	VGG-16	12.9	85.38	83.1

It can be seen from the experimental results that the MS-faster RCNN proposed in this paper is superior to the other two algorithms in performance. Compared with Fast RCNN and MS-faster RCNN, the false detection rate (MR) is 2.9 and 1.6 lower, the Recall rate (Recall) is 3.09 and 2.16 higher, and the average accuracy rate (mAP) is 4.7 and 2.3 higher.

5. Conclusion

This paper constructs a pedestrian detection model based on Faster RCNN, and introduces a multi-scale processing module to process the pedestrian detection of different scales in the image. Compared with the original Faster RCNN, the MS-faster RCNN model reduces the missing and false detection of small-scale pedestrians to some extent, and effectively improves the performance of multi-scale pedestrian detection.

References

[1] Krizhevsky A, Sutskever I, Hinton G E, et al. ImageNet Classification with Deep Convolutional Neural Networks[C]. Neural Information Processing Systems, 2012: 1097-1105.

[2] Y.Q. Zhao,Y. Rao , S.P. Dong , J.Y. Zhang . A review of deep learning target detection methods [J]. Chinese journal of image and graphics,2020,25(04):629-654.

[3] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]. Neural Information Processing Systems, 2015: 91-99.

[4] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]. Computer Vision and Pattern Recognition, 2014: 580-587.

[5] Girshick R. Fast R-CNN[C]. international conference on computer vision, 2015: 1440-1448.

[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. European Conference on Computer Vision, 2016: 21-37

[7] Redmon J, Divvala S K, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]. Computer Vision and Pattern Recognition, 2016: 779-788

[8] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]. Computer Vision and Pattern Recognition, 2017: 6517-6525.

[9] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. arXiv: Computer Vision and Pattern Recognition, 2018.

[10] Lin T, Dollar P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]. Computer Vision and Pattern Recognition, 2017: 936-944.