Research on Small Target Detection Technology Based on Improved Faster R-CNN

Zhongyuan Wang

Changchun University of Science and Technology, ChangChun, JiLin, China.

17808050776@163.com

Abstract

Aiming at the problems of low detection accuracy and poor recognition ability of small target, a small target detection technology based on improved Faster R-CNN is proposed. Based on Faster R-CNN algorithm, the structure of feature extraction network VGG16 is improved. By adjusting the network model and fusing the output features of the convolution layer, different levels of semantic information are obtained. The feature extraction ability of small target is enhanced. At the same time, the RPN network anchor size ratio is redesigned. K-means algorithm is used to cluster the data and select more suitable anchors. We can obtain the size of candidate frame accurately matched to small targets, which is conducive to improve the detection effect of small targets. The experimental results on dota dataset show that the proposed algorithm has good performance for small target detection, and the average accuracy is about 5% higher than the original Faster R-CNN, which verifies the superiority of the algorithm.

Keywords

Faster R-CNN, Small Target Detection, VGG-16, Anchors, K-means Clustering.

1. Introduction

With the rapid development of deep learning technology, more and more target detection and recognition technologies have been developed, and become a hot research object. Girshick et al. Proposed a regional convolution neural network (R-CNN). On the VOC 2012 dataset, the average accuracy of target detection reaches 53.3% [1]. On this basis, fast r-cnn and fast r-cnn have been proposed [2-3]. Small target detection technology is a difficult and hot topic in the field of target detection. Small target objects account for a small proportion in the image, with low resolution, rough contour and noise interference. Small targets in the image contain less information, which is easy to be confused with background information, and the training data is difficult to mark. The features extracted by the detection technology based on deep learning are weak, which can easily lead to missing or false detection, and the detection accuracy is very low [4-7]. The existing target detection algorithms can not effectively solve the existing problems. This paper proposes a small target detection algorithm based on improved fast r-cnn. Aiming at improving the accuracy and speed of small target detection, the following improvement strategies are proposedOrganization of the Text:

(1) The feature expression ability of small target is enhanced. The output features of the convolution layer in the feature extraction network VGG16 are fused, and the combined features are reduced in dimension, and then classified and regressed. The essential reason for the low accuracy of small target detection is that the feature extracted by neural network is weak. Feature fusion is used to enhance the feature extracted by feature network.

(2) Redesign the size and number of Anchors. Based on K-means algorithm, the data is clustered and analyzed, and anchors are redesigned. Choosing more suitable anchors will have better detection effect.

2. Basic Structure and Principle of Faster R-CNN

Faster R-CNN is developed from the original r-cnn framework, and is currently the most widely used in the field of target detection [8]. Traditional algorithms such as opencv AdaBoost use sliding window and image pyramid network to generate candidate frames, and r-cnn uses selective search method to generate candidate frames. These methods are time-consuming and time-consuming [9]. Fast r-cnn abandons the traditional method, directly uses the region proposal network (RPN) to generate candidate frames, and integrates feature extraction and RPN into a network model. The network shares convolution layer for feature extraction to reduce redundant candidate frames, significantly shorten the generation time of candidate frames, and greatly improve the speed. This is fast The huge advantage of r-cnn. The basic structure of fast r-cnn network model is shown in Fig. 1, which is mainly divided into four parts :

(1) Shared convolution layer: it is composed of a set of basic convolution, relu function, pooling, etc. the shared convolution layer extracts features from the image, and the feature map is shared by the subsequent RPN network and full connection layer. Generally, VGG16 network is used, and more target features can be obtained by deep network layers, which has better detection effect for small targets.

(2) RPN: RPN network is used to batch generate candidate frames and set candidate anchors on the feature map. According to softmax classification, which anchors are positive examples with targets and which are negative anchors without targets are judged. Then, the candidate boxes corresponding to anchors are modified by bounding box regression to obtain more accurate coordinates.

(3) ROI Rooling: the ROI Rooling layer integrates the feature map and candidate frame information, maps the input candidate frame data of different sizes to a fixed scale candidate frame feature vector, transforms it into data of uniform size, and sends it to the full connection layer for classification and regression judgment.

(4) Classification: calculate the candidate feature map, determine the target category, and obtain the final accurate position of the candidate box through bounding box expression.

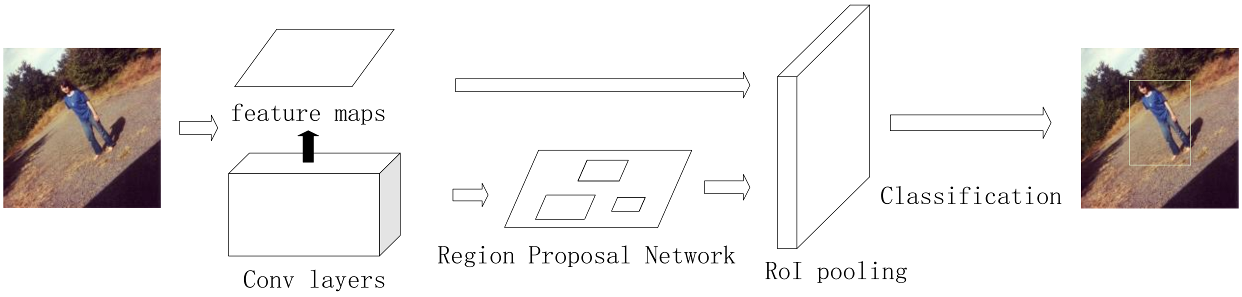

Fig.1 Basic structure of Faster R-CNN Network model

3. Research on improved small target detection technology

3.1 Feature Extraction Network VGG

VGG16 network has 13 convolution layers, 4 maximum pooling layers, 3 full connection layers and softmax layers. The convolution layer uses 3 × 3 convolution kernel, and extracts the input image features by sharing convolution. The network structure is shown in Fig. 2.

The 13 convolution layers and 4 pooling layers can be divided into five parts. The first part includes two layers of convolution layers and one maximum pooling layer. The two convolution layers contain 64 convolution kernels with the size of 3 × 3, the step size of 1, the maximum pooling window of 2 × 2 and the step size of 2. The second part, the third part, the fourth part and the fifth part are basically the same. The difference is that the number of convolution kernels in the convolution layer increases exponentially, and other parameters remain unchanged. After the fifth part pooling, connect the three full connection layers. The first two fully connected layers have 4096 eigenvalues, and the third layer has 1000 eigenvalues. The last layer of the network is the softmax layer for classification, which is mainly used to calculate the probability of a certain category of objects in the input image.

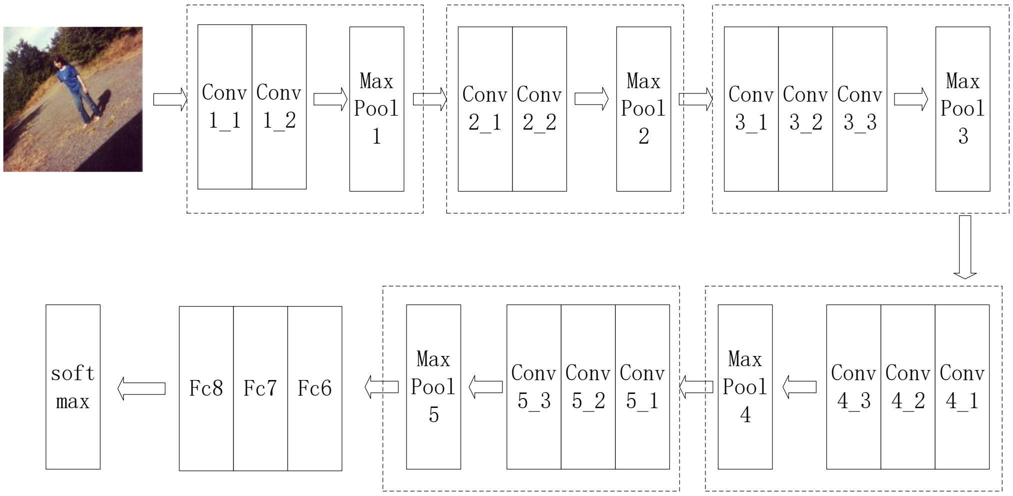

Fig. 2 VGG16 network structure

Faster R-CNN network has a serious problem of missing detection for small objects, especially for the small targets among a large number of objects. The real size of the small target is too small, usually only contains dozens or fewer pixels, or the proportion of the target in the whole image is small due to the shooting distance. Because the scale of small targets in the image is relatively small, the features of these small targets are extracted through multiple convolution pooling layers of VGG16. VGG16 contains four pooling layers, so the generated feature map is only 1 / 16 of the original image scale. Therefore, when these feature maps are sent to the subsequent RPN network, due to the small target pixels are too few, the target space information is relatively small, so through the subsequent classification When regression layer is used, it will cause missing detection or decrease recognition effect.

As shown in Fig. 3, the output features of the convolution layer in VGG16 are fused. Firstly, the RPN network generates candidate frames, and the ROI pooling layer synthesizes the feature map and candidate frame information, and maps the input candidate frame data of different sizes to a fixed scale candidate frame eigenvector, and transforms it into data of uniform size. Secondly, this result is mapped to the third, fourth and fifth parts of VGG. The feature map is shared for the subsequent RPN layer. At this time, the candidate frame contains the feature information of the third, fourth and fifth parts, and then the ROI pooling process is performed to obtain the candidate frame feature vector with fixed scale. Then, the feature information of the third, fourth and fifth layers is normalized to improve the convergence speed of the network, and these information are combined in the channel to improve the calculation efficiency of the fitting process. Finally, the dimension of the merged feature information is reduced, and the 1 × 1 convolution kernel is used to send the final feature vector to the full connection layer for classification and regression judgment. The parameters generated by network training are generated by the full connection layer. There are three layers of full connection layer in VGG network, so too many network parameters will increase the amount of calculation, and it is easy to produce over fitting phenomenon. In order to reduce the size of the network, save the cost of calculation, and do not affect the extracted high-level semantic information, we modify the three-layer full connection layer. Finally, the two full connection layers are replaced by a full connection layer with 1024 features, and other parameters remain unchanged. This improvement significantly reduces the network size, reduces the number of network parameters, and reduces the detection time under the premise of ensuring the detection accuracy.

The essential reason for the low accuracy of small target detection is that the feature extracted by neural network is weak. By using the method of feature fusion, the output features of the third, fourth and fifth parts of the VGG16 network are fused. The features of different levels are fused, and the semantic information of different levels is obtained at the same time of prediction. The features extracted from the feature network are enhanced The feature expression ability of small target. Using the fusion features to detect small targets, and adjust the network model, reduce the number of parameters, significantly improve the detection accuracy and detection time.

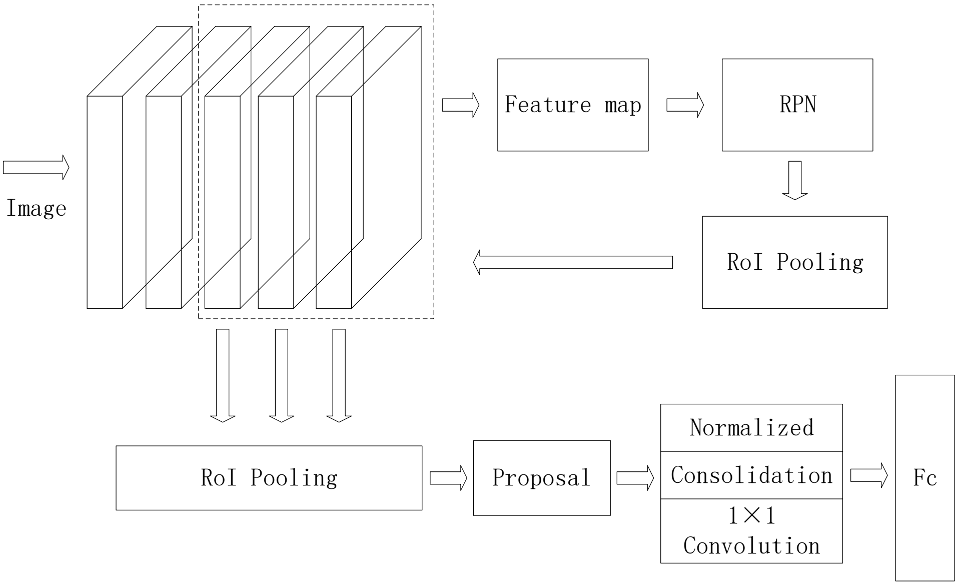

Fig. 3 improved VGG network work flow chart

3.2 Size and Quantity Design of Anchors

In RPN network, anchors correspond to (128,256,512) three scales and (1:1, 1:2, 2:1) three aspect ratios. These scales are reasonable for general object detection. However, for small targets, the size of the target marked on the original image is too large, which will cause the size of candidate frame unable to adaptively and accurately match with the small target in the image As a result, the effect of small target detection is not obvious, and the performance of the network can not meet the requirements. In addition to the scale and aspect ratio of anchors, the number of anchors also has an important impact on the detection accuracy. In order to meet the accuracy requirements of small target detection task, this paper redesigns the size and number of anchors, so that the network has a better detection effect for small targets within the scope of anchors. Through the cluster analysis of the sample data set of the training model, the anchor size is obtained, and the candidate frame size which can accurately match the small target is obtained. In this paper, the K-means clustering algorithm is selected for the experiment. The process of K-means clustering algorithm is as follows:

(1) K samples were randomly selected as cluster centers in the data set.

(2) For each other sample in the data set, the distance to the cluster center is calculated, and it is divided into the cluster centers with the nearest distance.

(3) For all samples in each category in the dataset, the center position of each category is recalculated and the new cluster center is relocated.

(4) Root out the new cluster center and repeat steps (2) and (3) until the cluster center is no longer changed.

Table 1 Cluster Center Size and Aspect Ratio of Candidate Frames

Numeber	1	2	3	4
Length	40	20	23	18
Width	40	40	69	72
Ratio	1:1	1:2	1:3	1:4

The above algorithm is used to cluster the DOTA data set, and 12 clustering centers are obtained. The target size of the data set is mainly concentrated between 30-80 pixels. 12 anchor points are obtained from the statistics of the cluster center size and aspect ratio of candidate frames. The information statistics are shown in Table 1.

In order to meet the size requirements of small target samples in the dataset, the size of candidate frames for prediction is determined to be 32 and 64 pixels, and the aspect ratio is 1:1, 1:2, 1:3, 2:1, 3:1, 4:1, 1:4. Using the new anchors to train the network is conducive to fully learning the characteristics of small objects and improving the effect of small target inspection.

4. Experiment and Analysis

4.1 Data Set

The DOTA dataset is an aerial image data set produced by simultaneous interpreting and Remote Sensing Laboratory of Wuhan University and Huazhong University of Science and Technology Telecom school. Aerial images from different sensors and platforms include 15 categories and 2806 aerial images. Different from other traditional data sets, the image scale of DOTA data set varies greatly, and a small area may contain many dense small targets. Because of the particularity of the shooting angle, most of the images in the dataset are small targets, with only dozens or even a few pixels. Therefore, it is very suitable for the small target detection task data set, which can well measure the effect of the improved fast r-cnn algorithm for small target detection.

4.2 Training and Parameter Assessment

The training environment is based on ubuntun16.04, tensorflow-gpu1.10.0, python 3.5, cuda9.0, nvidla geforce GTX 1080 Ti, Intel i5-4210 CPU, 8G memory

In order to optimize the parameters of the network, a rotation iterative training method is used to train on the basis of the trained model (improved VGG). The training process was completed on GTX 1080ti, and the optimization method was SGD.

Similar to other target detection tasks, precision, recall, accuracy, loss value, AP and map are used as parameters for evaluation.

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

$$recall = \frac{TP}{TP+FN} \tag{2}$$

$$accurcacy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

4.3 Experimental Results and Analysis

4.3.1 Loss value analysis

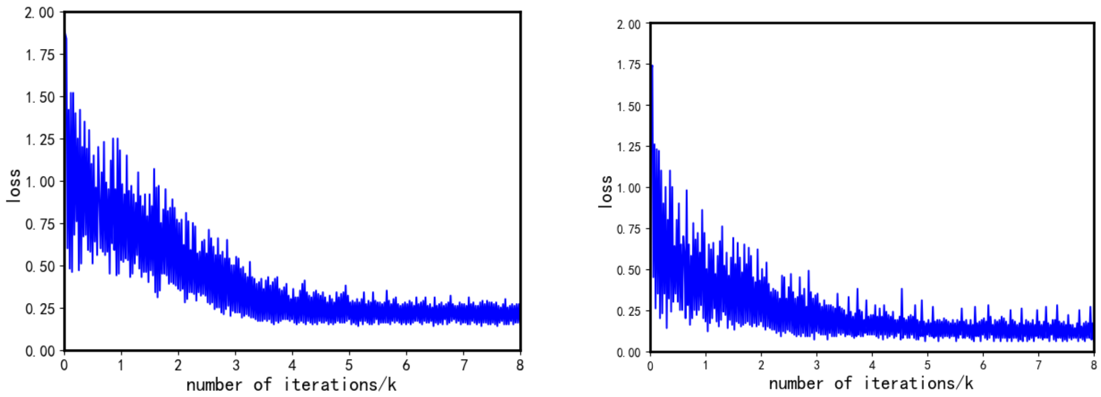

Fig. 4 Change of loss value before and after change

The loss function and simulation results in the improved model training are shown in Fig. 4. The left figure shows the convergence chart of the loss value before the improvement, and the right figure shows the change chart of the loss value after the improvement. The data includes 15 categories. Take 1500 pictures as the training set and 500 as the test set. Normalize the samples to avoid over fitting. From the change trend of the loss value of the former 80K iterations, the convergence speed of the loss value after the improved algorithm becomes faster, the curve tends to be stable gradually, and there is no over fitting phenomenon. When the training iteration reaches about 20K, the curve reaches stability, and the average loss function value reaches 0.05. The improved algorithm model shortens the time for curve to reach stability, and the accuracy is also significantly enhanced.

4.3.2 Analysis of accuracy and recall

In this paper, PR (precision recall) curve is used as the evaluation performance index, and the recall rate is taken as the x-axis and the accuracy coordinate is as the y-axis. Fig. 5 visually shows the recall rate and accuracy rate of the model on the training samples, and the area under the PR curve and the coordinate axis represents the average accuracy.

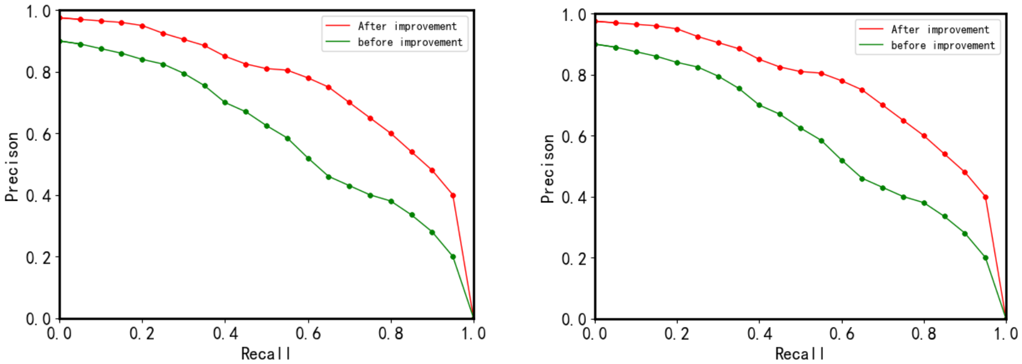

Fig. 5 PR curve of harbor and aircraft before and after improvement

Figure 5 selects the PR curves of harbor and aircraft in the data category. The recall value is not strictly increasing, and the overall trend of the curve is downward. Compared with the PR curve before and after the improvement, the precision decreases with the increase of recall. The recall of the front harbor reaches the equilibrium point at 0.5, the recall of the improved harbor reaches the equilibrium point at 0.7, the recall of the aircraft before the improvement reaches the equilibrium point at 0.6, and the recall of the improved aircraft reaches the equilibrium point at 0.8. The AP value after the improvement is larger than that before the improvement, and the model detection performance is good.

4.3.3 Model experiment results and comparative analysis

Table 2 Cluster Center Size and Aspect Ratio of Candidate Frames

Network Model	mAP	FPS
Faster R-CNN	68.77	5
YOLO	56.9	45
Improved Faster R-CNN	73.67	7

From Table 2, the improved Faster R-CNN model has the highest map value, the model detection performance is relatively good, and the number of frames per second is relatively high, which can extract more features of the target Compared with YOLO, although the FPS of YOLO is high and the running time is short, the performance of YOLO detection is poor, and the recognition rate of small targets is low. Generally speaking, the improved Faster R-CNN model has the best comprehensive performance in small target detection, which is better than most mainstream detection algorithms.

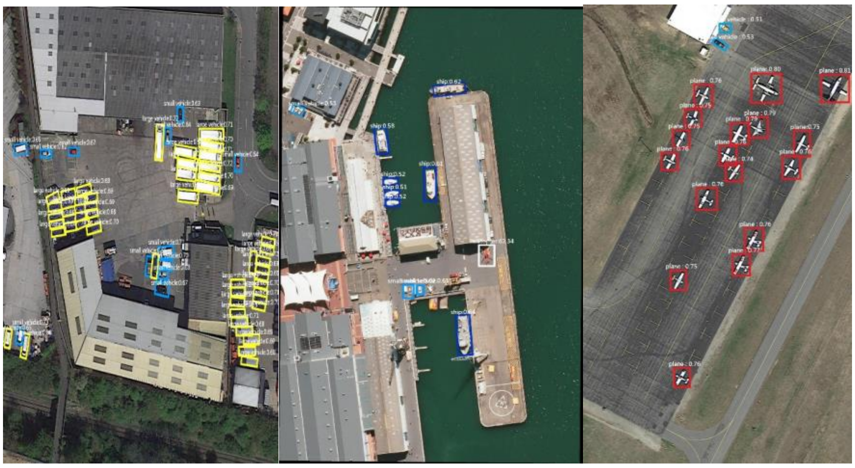

Fig. 6 Example of improved algorithm model detection effect

5. Conclusion

This paper proposes a small target detection algorithm based on improved Faster R-CNN. Based on the original Faster R-CNN network, a series of improvement and optimization are carried out. Firstly, the output features of the convolution layer in the feature extraction network VGG16 are fused. It can achieve the purpose of enhancing the feature extraction ability of small target. At the same time, the VGG16 network structure is updated to reduce the network size without affecting the ability of feature extraction. Secondly, K-means clustering analysis is used. The size and number of anchors are redesigned to make the candidate frames corresponding to anchors better match the small targets in the image. The detection accuracy of small target is further improved. Through the experimental analysis on dota data set, the detection accuracy of small objects in the image has been significantly improved, which verifies the effectiveness and feasibility of the improved Faster R-CNN.

References

[1] Donahue J, Jia Y, et al. Decaf: A deep convolutional activation feature for generic visual recognition, Proceedings of the 31st International Conference on International Conference on Machine Learning(2014)p. I-647.

[2] Girshick R. Fast R-CNN, IEEE International Conference on Computer Vision IEEE Computer Society (Xi'an,China,2015),p. 1440-1448.

[3] Ren S,He K, Girshick R,et al.Faster R-CNN:towards real-time object detection with region proposal networks International Conference on Neural Information processing Systems. (MIT Press, 2015) p.91-99.

[4] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S.E, Fu C, Berg A.C:Ssd:Single shot multibox detector. European conference on computer vision(2016)p.21-37.

[5] Uijlings J.R.R, De Sande, KEAV, Gevers T, Smeulders AWM: Selective search for object recognition. International Journal of Computer Vision 104(2013) No.3 p.154-171.

[6] Fang Pengcheng. Research on small target detection in images. Beijing University of Posts and telecommunications, 2019

[7] Kampffmeyer M,Salberg A B,Jenssen R.Semantics egmentation of small objects and modeling of uncertain-ty in urban remote sensing images using deep convolu-tional neural networks. Computer Vision and PatternRecognition Workshops,(2016) p. 680-688.

[8] Ren S,He K,Girshick R,et al. Faster R-CNN: towardsreal-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, (2015),Vol39(6) p. 1137-1149.

[9] Girshick R,Donahue J,Darrell T,et al. Rich feature hierarchies for accurate object detection and semantic seg-mentation. IEEE Conference on Computer Vision andPattern Recognition,(2014) p.580-587.