A Target Detection Algorithm Based on Faster R-CNN

XinQing Yan(✉), YuHan Yang, and GuiMing Lu

North China University of Water Resources and Electric Power,
Zhengzhou, People's Republic of China
1161547277@qq.com

Abstract. Target detection is one of the hotspots of image processing research. In the image, due to factors such as distance or light, it will affect the target detection result and increase the error detection rate. Moreover, the existing network training time is too long to meet the actual needs. In order to reduce the lack of light or shadow interference and other factors, based on the Faster R-CNN framework, this paper innovatively proposes a method to improve its feature network ResNet-101 to extract deep features of images.In order to shorten the running time, this paper introduces the region number adjustment layer to adaptively adjust the number of candidate regions selected by RPN during the training process. This paper conducts experiments on the PASCAL VOC data set. The experimental results show that the improved feature network model proposed has an accuracy improvement of 2% compared with the original feature network model. The results show that the target detection algorithm proposed in this paper has higher recognition accuracy than the original algorithm.

Keywords: Target detection · ResNet-101 · Faster R-CNN

1 Introduction

Object detection is an important research direction in the field of computer vision, which is very important for extracting and mining regions of interest. Target detection is a central problem in computer vision [1], and has important research value in the fields of pedestrian tracking, license plate recognition and unmanned driving. The features extracted by traditional algorithms are basically low-level and simple features manually selected. These features are more targeted at specific objects and better represent multiple targets. In addition, some prior knowledge needs to be manually set. In recent years, deep learning has greatly improved the accuracy of image classification, so target detection algorithms based on deep learning have gradually become mainstream. Traditional target detection algorithms are mainly based on the matching of traditional frames or feature points in sliding windows. AlexNet won the championship in the 2012 ImageNet Large-scale Visual Recognition Challenge [2]. Its role goes far beyond traditional algorithms, and brings the general public's vision into the convolutional neural network. Target detection methods based on deep learning have just appeared in the field of computer vision. Target detection based on deep learning

S. Shi et al. (Eds.): AICON 2020, LNICST 356, pp. 502–509, 2021.
https://doi.org/10.1007/978-3-030-69066-3_44

can be roughly divided into two categories: 1) Regression-based detection algorithms; 2) Classification-based detection algorithms. The former is represented by the YOLO system algorithm. Reference [3] proposes the YOLO model. YOLO divides the image into SxS grids, and each grid is responsible for detecting targets centered on the grid. YOLO uses a cell-centered multi-scale area to replace the regional target proposal network, thus giving up some accuracy in exchange for faster detection speed. This method is directly trained on the original image. This method is faster and less accurate. The latter's typical algorithm is Faster-RCNN. Reference [4] proposed a regional convolutional neural network model, which uses a deep convolutional neural network to select some candidate regions in the image to be detected by a selective search method, so as to perform advanced feature extraction, and then use multiple SVM pairs Functions are classified and target discovery tasks are completed. In the article [5], the Faster-RCNN (Faster-RCNN) model is proposed to improve the detection accuracy and speed of the RCNN model. His central idea is to use RPN network to exclude areas of interest and train based on it. The method is characterized by high accuracy and slow speed. Therefore, if the accuracy requirements are high, the Faster-RCNN series algorithm should be adopted based on the classification idea. However, when real-time requirements are high, fast RCNN is slower and not easy to apply [6]. In this article, the Faster-RCNN model is improved. This article innovatively proposes a method to improve its feature network ResNet-101 to extract the deep features of the target, extract the deeper features of the image, and use the PASCAL VOC data. And in order to shorten the running time, this paper introduces the region number adjustment layer to adaptively adjust the number of candidate regions selected by RPN during the training process. The set is compared with the Faster-RCNN model experiment to verify the method proposed in this paper.

2 Related Work

The convolutional neural network obtains the features of the target by learning the data set of manually labeled features [7]. At present, algorithms based on convolutional neural networks can be roughly divided into two modes, namely, two-stage mode and one-stage mode. The detection process of the former is divided into two steps [8]. First, the algorithm generates several candidate boxes, and then classifies the candidate boxes through CNN [9]. The latter is a direct regression to the category probability and location coordinates of the target. The emergence of R-CNN has successfully applied CNN to the field of target detection, but R-CNN also has certain problems.

In order to break through the time bottleneck of the candidate region algorithm, Ren Shaoqing et al. proposed FasterR-CNN in 2016 [10]. FasterR-CNN uses RPN instead of selective search, which greatly reduces the time to extract candidate frames. This algorithm can be roughly understood as a combination of RPN and FastR-CNN. From R-CNN to FasterR-CNN, its detection speed and detection accuracy are constantly improving [11]. Such algorithms are still an important branch of target detection algorithms.

3 Faster R-CNN

The Faster R-CNN target detection network is mainly divided into two steps. The first step is to locate the target. Enter a picture into the feature extraction network, and extract the feature map of the image after a series of convolution and pooling operations. The task of target detection is not only to detect the target, but also to find and locate the candidate target on the feature map through the RPN network. The second step is to classify the specific category of the target. The range box regressor is used to modify the position of the candidate target to generate the final candidate target area. The softmax classifier is used to identify the categories of candidate targets. This paper uses a classification network to determine whether the candidate area belongs to the target of interest, so as to realize the detection of the target of interest. The structure of Faster R-CNN is shown in Fig. 1.

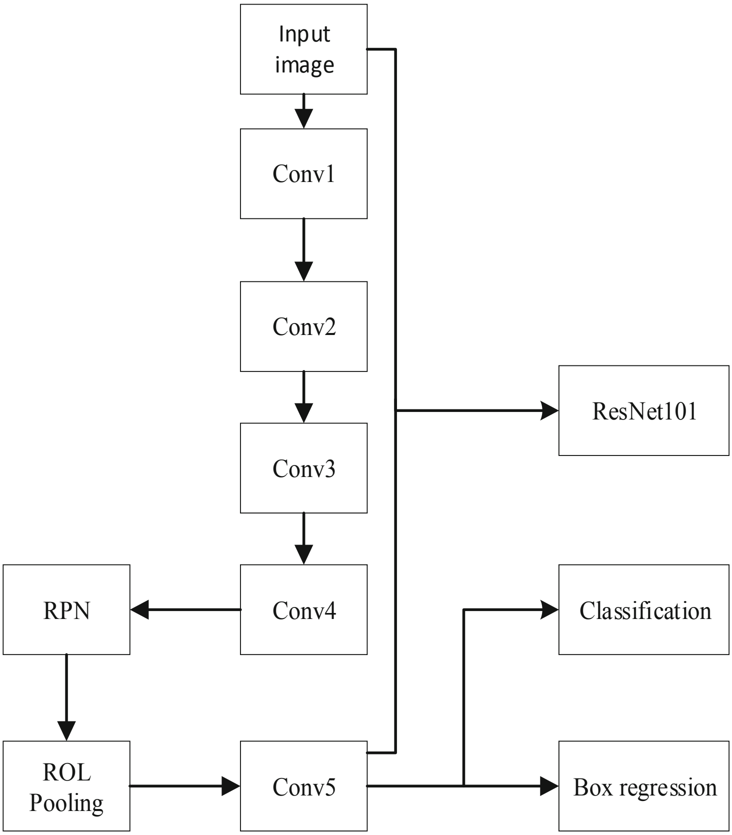

Fig. 1. The structure of Faster R-CNN

3.1 ResNet

In this paper, the feature extraction of the target is completed by improving the ResNet-101 model based on the convolutional neural network. Compared with the feature extraction network such as AlexNet and VGG16, its characteristics are as follows.

ResNet adopts the shortcut connection method of “shortcut” identity mapping network, which neither generates additional parameters nor increases computational complexity. As shown in Fig. 2, the shortcut connection simply performs identity mapping and adds its output to the output of the overlay. Through SGD back propagation, the entire network can be trained in an end-to-end manner.

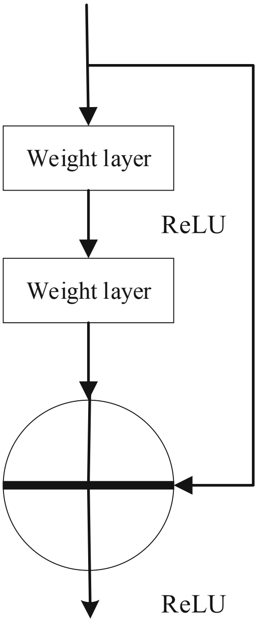

Fig. 2. Residual network model

$$X_{l+1} = X_l + F(X_l, W_l) \tag{1}$$

$$X_{l+2} = X_{l+1} + F(X_{l+1}, W_{l+1}) = X_l + F(X_l, W_l) + F(X_{l+1}, W_{l+1}) \tag{2}$$

$$X_L = X_l + \sum_{i=l}^{L-1} F(X_i, W_i) \tag{3}$$

$$\frac{\partial \varepsilon}{\partial X_l} = \frac{\partial \varepsilon}{\partial X_L}\frac{\partial X_L}{\partial X_l} = \frac{\partial \varepsilon}{\partial X_L} + \left[1 + \frac{\partial}{\partial X_L}\sum_{i=l}^{L-1} F(X_i, W_i)\right] \tag{4}$$

In the formula, the network input value is X_l, and the network output value is $X_{l+1}, X_{l+2}, \ldots, X_{:L}, \varepsilon$. The network residual block is $F(X_l, W_l)$, $F(X_{l+1}, W_{l+1})$.

ResNet adopts the “Bottleneck design” network design structure, uses 1 × 1 convolution to change the dimension, 3 × 3 convolution inherits network performance, and controls the number of input and output feature maps of 3 × 3 convolution. When the number of layers is high, the number of 3 × 3 convolutions is reduced, which greatly reduces the number of convolution parameters and the amount of calculation while increasing the depth and width of the network.

3.2 Improved Feature Extraction Network

This article attempts to improve the residual structure based on the traditional ResNet-101, combined with the characteristics of the picture. This paper increases the width of the network and extracts the deep features of the target so that the network can learn key distinguishable features. Therefore, we propose an improved deep residual network structure to improve the recognition accuracy of the target person. The residual unit used in this paper, using two parallel 3 × 3 convolutional layers. The function of the 1 × 1 convolutional layer is to change the dimension, and the two parallel 3 × 3 convolutional layers inherit the performance of the VGG network. The unit uses a pre-activation method, and all convolutional layers use ReLU as the activation function, That is, perform batch regularization before using the activation function and convolution operation, Compared with the original residual unit, this structure has little difference in training parameters. However, follow-up experiments show that the residual unit structure proposed in this paper has significantly improved the performance of the target person detection and model.

3.3 Improved Region Proposal Network

When the feature extraction network extracts image features, it is a process from low to deep. Each layer has its own RPN, and the candidate regions are generated by extracting feature maps of different scales. The RPNs corresponding to different proportions are different. When the deep neuron’s acceptance range expands, the corresponding anchor box size also increases. The larger the candidate area, the smaller the RPN. After obtaining the candidate area, the features are converted into uniform size through RI pooling. Finally it is sent to the classifier. So as to complete the entire Faster R-CNN process.

This paper introduces the NP (number of proposals) layer to adaptively adjust the number of candidate regions selected by the RPN during the training process.

$$N_{Pi+1} = \begin{cases} N_{Pi}(1+\mu_1) & L_i \geq 2L_{i-1} \\ N_{Pi} & 0.5L_{i-1} < L_i < 2L_{i-1} \\ N_{Pi}(1-\mu_2) & L_i \leq 0.5L_{i-1} \end{cases} \tag{5}$$

In the formula, i represents the sequence number of every N training. N_{Pi} represents the number of candidate regions used from the Nith training to the (N + 1)ith training. L_i represents the average regression loss from the Nith training to the (N + 1) ithtraining. μ_1 represents the penalty factor and μ_2 represents the reward factor.

In the training process, the NP layer is introduced to feedback and adjust the training results. Calculate the average of regression loss every N training intervals. Through the blank control group experiment (fixing $N_{p_{i+1}}$ and taking different values for experiment), every N times L_i reduces by half and self-increases by 1 time as the reasonable change jitter interval. Beyond this interval, it is considered that feedback adjustment is required. When L_i doubles or more, the number of candidate regions is increased by a multiple of $(1 + \mu_1)$. When L_i is reduced by half or smaller, it is considered that the number of selection boxes can be appropriately reduced, and the number of candidate regions becomes a multiple of itself $(1 - \mu_2)$.

Set the upper and lower limits of the number of candidate regions, so that the number of candidates can be adaptively changed from 300 to 2000.

4 Experiment and Analysis

4.1 Data Set and Experimental Environment

The experimental hardware adopts Z440 workstation with 32G memory and NVIDIA P2000 graphics card, and the operating system is Ubuntu 16.04. The programming environment used is as follows, the programming language used is Python language, and the deep learning framework is TensorFlow 2.0. The training samples for the experiment come from the training set in the Pascalvoc dataset.

4.2 Experimental Results and Analysis

Experiment 1. In the training phase, each target in each image in the training set needs to be marked with a rectangular box, and the occluded target should also be marked. During the test, if the overlap between the identified target detection frame and the marked rectangular frame reaches more than 90% of the marked rectangular frame, the test is recorded as successful. The PASCAL VOC 2007 data set is a classic open source data set, including 5000 training set sample images and 5000 test sample images, and a total of 21 different object categories. This article uses the training set and test set of the data set, and the experiment uses the Tensorflow framework to implement the convolutional neural network model. The parameters such as random inactivation, maximum iteration value, batch size in Faster-RCNN generate the average accuracy value (mAP) Greater impact. In order to get a better output, these parameters need to be optimized.

In the experiment, the maximum number of iterations is 70,000, the batch size of the RPN network is 256, and the random deactivation value is selected as 0.6. Experiment one is the performance of the target detection network with the adjustment layer of the number of regions in terms of average detection accuracy and speed. After many experiments, the calculation speed can be increased by 18% on average, with almost no loss of accuracy, saving a lot of overhead. The results of experiment one are shown in Table 1.

Table 1. Detection effect with different number of candidate regions

Network model	Number of candidate regions	Mean average precision	Total time (s)
VGG16	2000	81.86%	54135
VGG16	1000	81.86%	50967
VGG16	500	80.13%	47843
VGG16	100	79.22%	56901
VGG16	10	56.81%	44697
VGG16 + NP	Trained	83.14%	45631
Improved ResNet-101 + NP	Trained	85.62%	45531

Experiment 2. Experiment setup is the same as experiment one. Experiment 2 mainly compares the detection accuracy under different feature extraction models. The results are shown in Table 2.

Table 2. Detection Accuracy under different characteristic networks

Feature network	Mean average precision	Total time (s)
AlexNet	85.26%	54135
ResNet-101	86.23%	50967
VGG16	83.14%	45631
Improved ResNet-101	85.62%	45531

It can be concluded from Table 2 that the improved model has the highest accuracy rate, which is 3% higher than that of the original model. That is, for AlexNet and the original feature network, the improved feature network can better extract image features, which can be more accurate. At the same time, the improved ResNet101 network model has the highest value of mAP, that is, the target network model has relatively good performance, thereby increasing the recognition accuracy of the model. Therefore, the improved model detection effect in this paper is relatively good.

5 Conclusion

Aiming at the problem of target recognition accuracy and target positioning accuracy affecting detection, this paper proposes a method to improve its feature network ResNet-101 based on the Faster R-CNN framework, and improves the Region Proposal Network. The following conclusions can be obtained.

Improve the design of the ResNet-101 feature extraction network based on convolutional neural network in order to make the model more sensitive to the target.

This paper increases the width of the network on the basis of the original network. This improvement enables the network to learn key distinguishable features, thereby improving the accuracy of target recognition. Improvements to the Region Proposal Network greatly increase the running time. The introduction of the region number adjustment layer also improves the accuracy rate to a certain extent.

In short, compared with the original model, the improved Faster RCNN model has improved the accuracy of target recognition.

References

1. Liu, Z., Lyu, Y., Wang, L., et al.: Detection approach based on an improved Faster RCNN for brace sleeve screws in high-speed railways. IEEE Trans. Instrum. Meas. **100**(1), 39–46 (2019)
2. Girshick, R., Donahue, J., Darrell, T,. et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
3. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
4. Duan, L., Zhang, D., Xu, F., et al.: A novel video encryption method based on faster R-CNN. In: 2018 International Conference on Computer Science, Electronics and Communication Engineering, pp. 112–116 (2018)
5. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: and resnet features are equivalent with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 10, no. 3, pp. 91–99 (2015)
6. McNeely-White, D., Beveridge, J.R., Draper, B.A.: Inception and resnet features are equivalent. Cogn. Syst. Res. **59**(1), 312–318 (2020)
7. Khan, S.H., Hayat, M., Bennamoun, M., et al.: Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans. Neural Netw. Learn.g Syst. **29**(8), 3573–3587 (2017)
8. Ouyang, W., Wang, X., Zhang, C., et al.: Factors in finetuning deep model for object detection with long-tail distribution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 12, no. 1, pp. 864–873 (2016)
9. Zhang, D., Li, J., Xiong, L., et al.: Cycle-consistent domain adaptive faster RCNN. IEEE Access **12**(7), 123903–123911 (2019)
10. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 10, no. 4, pp. 91–99 (2015)
11. Chen, X., Zhang, Q., Han, J., et al.: Object detection of optical remote sensing image based on improved faster RCNN. In: 2019 IEEE 5th International Conference on Computer and Communications, vol. 12, no. 3, pp. 1787–1791 (2019)