Irregular Target Object Detection Based on Faster R-CNN

Bin Zhang, Yubo Zhang and Qinghui Pan
Zhengzhou University, Zhengzhou, 450006, China

Abstract. For the shortcomings of traditional target detection algorithms can only extract specific target features for detection, propose the Faster R-CNN target detection model of deep learning, combined with VGG16 and ResNet101 convolutional neural network methods, to detection of irregular target objects. Experiments established two types of irregular target data sets, walnut and jujube, use the network training and testing, verified the feasibility of deep learning network for detecting irregular target objects. The experimental results show that the Faster R-CNN target detection network, if training on the self-built data set, the final detection result reaches 95%, which proves the effectiveness of the network for detecting irregular target objects.

1. Introduction

With the continuous improvement of the level of technology, image target detection technology has been widely applied to various fields of life such as industry, agriculture, medical care, and transportation. However, in various fields of application, feature extractors are designed for specific detection targets, cannot extract deeper target features, resulting in lack of generalization ability of target detection algorithms. This paper mainly uses the convolutional neural network to extract the features of the deep features of the target object, applying the deep learning algorithm Faster R-CNN to learning feature and detection of irregular target objects.

The target detection algorithm generally consists of three parts: Region proposal, feature extraction and target classification and regression. The regional feature extraction is the core part of the algorithm, and the feature extraction ability directly affects the accuracy of the target classification and regression, influencing the performance of the entire detection network. Traditional feature extraction methods, such as HOG [1] (histogram of oriented gradient), SIFT [2-3] (scale invariant feature transform), Harr-like [4-5] etc., the essence belongs to the hand-designed feature extractor. Different feature extractors need to be designed for different recognition problems. The recognition performance depends on the feature selection, which has high requirements for designers. Adoption DPM [6-8] (deformable part-based model) algorithm, combine multiple target features to improve the disadvantages of using only a single target feature for detection. However, the artificially designed feature extractor can only extract the texture gradient information and the original pixel information in the image, can not learn deeper semantic information and learn the essential information of the target object, so that the artificially designed feature extractor can not meet the target detection task in complex scenes.

In recent years, the target detection method based on deep learning has developed rapidly. R. Girshick proposed R-CNN [9] (region with convolutional neural network) in 2014. The algorithm first generates about 2k regional suggestion boxes by the Selective Search [10] method. The generated suggestion boxes were classified by pre-trained convolutional neural networks, and by using a deeper CNN model (VGG16 [11]), get a result of 30% higher than the traditional method on the PASCAL VOC.
2012 data set. Compared with the traditional method, the convolutional neural network model is used to obtain more expressive target features to replace the manually designed feature extractors. In addition, convolutional neural networks use thousands of class-independent regional proposals to reduce image search space. However, due to the use of the Selective Search method, thousands of regional suggestion boxes are generated. Each box takes about 2 seconds to generate, and each suggestion box will be extracted by CNN, resulting in a large amount of time consumption that cannot meet real-time requirements. In response to the existing problems, the subsequent SPP-Net [12], Fast R-CNN [13], Faster R-CNN [14-15], etc. are continuously improved, the detection speed and accuracy are gradually improved to meet the target detection in practice.

2. FasterR-CNN Introduction

The algorithm consists of three parts: feature extraction, region suggestion box generation, and Fast R-CNN detection. Firstly, the feature extraction network is used to extract the feature of the original image, and the feature input RPN [16] generates a series of region suggestion boxes, the generated region suggestion boxes will be input into the Fast R-CNN network to finally generate the target object class and position.

2.1. Regional proposal network (RPN)

The proposal network (RPN) input is an arbitrary size image, and outputs a series of rectangular area candidate frames, each candidate box corresponding to a target probability and a rectangular frame position information. A 3x3 convolution kernel is used to convolve with the feature map on the generated feature map, and each pixel point gets a low-dimensional feature. Each pixel in the feature map generated by the convolutional network is called an anchor, each anchor will predict three scales (128, 256, 512) and three aspect ratio (1:1, 1:2, 2:1) candidate boxes in the original image. These candidate boxes will be input into two parallel fully connected layers, which are the candidate frame regression layer and the classification layer, these two networks are used to regress the target position in the candidate box and to predict the score of the target belonging to the foreground object. The loss function of the RPN consists of two parts, position loss and class loss. The training uses an end-to-end gradient descent method. Finally, sort according to the level of the region proposal score, select the first 300 candidate boxes, as the Fast R-CNN input, the structure of RPN diagram is shown in Figure 1.

![Figure 1. RPN structure diagram](image)

2.2. RPN and Fast R-CNN

Both RPN and Fast R-CNN, trained independently, will modify their convolutional layers in different ways, adopt a pragmatic four-step training algorithm to learn shared features via alternating optimization. In the first step, train the RPNs initialized with an ImageNet-pre-trained model and fine-tuned end-to-end for the region proposal task. In the second step, train a separate detection network by Fast R-CNN using the proposals generated by the step-1 RPN. This detection network is also initialized...
by the ImageNet-pre-trained model. At this point the two networks do not share convolutional layers. In the third step, use the detector network to initialize RPN training, but fix the shared convolutional layers and only fine-tune the layers unique to RPN. Now the two networks share convolutional layers. Finally, keeping the shared convolutional layers fixed, fine-tune the unique layers of Fast R-CNN.

3. Experimental design and results analysis

3.1. Test data set
Collecting two types of objects, walnut and jujube, as dataset targets, in order to collect the image of the target object, the data is collected within a range of 10 to 20 cm from the target object using an autofocus camera. To comprehensively collect data, a single jujube, a single walnut, a plurality of jujube, a plurality of walnuts, a walnut and a jujube are separately arranged in each image, and different backgrounds and illuminations are set for data collection. Collection method is shown in Table 1.

| Simple background       | Complex background                                           |
|------------------------|-------------------------------------------------------------|
| Wood grain desktop     | Add round and other debris                                  |
| Black paper            | Advertising paper with objects printed                      |
| Gray floor             | Black and white paper                                       |
| Blue paper             | Cover the middle part with white paper                      |
| Wood color paper       | Cover the edges with white paper                            |

Under 10 different backgrounds, a total of 1026 images were collected, including about 1,700 target objects. A total of 857 images were selected as the original training images, and the rest were used as test images. Due to the small amount of data collected, enhance the data by the left-right flip method before the data is trained, and the number of data sets is doubled as the final data set. Part of the data set is shown in Figure 2.

3.2. Experimental results and analysis
The VGG16 and ResNet101 feature extraction networks pre-trained by the Imagenet data set are used for the initialization of the convolutional layer in Faster R-CNN, and the target detection network is fine-tuned for the established data set. The experiment was trained on the Nvidia GeForce GTX1060 GPU and the entire experiment was based on the deep learning framework Tensorflow.

In the Pascal VOC dataset, using VGG16 and ResNet101 feature extraction network experiments, the results are shown in Table 2.
Table 2. Test results in the Pascal VOC data set

| Method                     | mAP  |
|----------------------------|------|
| Faster R-CNN + VGG16       | 70.8 |
| Faster R-CNN + ResNet101   | 75.7 |

In the self-built dataset, using VGG16 and ResNet101 feature extraction network experiments, the results are shown in Table 3.

Table 3. Test results in self-built data sets

| Method                | jujubeAP | walnutAP | mAP  |
|-----------------------|----------|----------|------|
| Faster R-CNN + VGG16  | 0.996    | 0.911    | 0.954|
| Faster R-CNN + ResNet101 | 0.997    | 0.911    | 0.954|

Test results with actual images, as shown in Figure 3.

Figure 3. Actual picture test results

Analysis of experimental results: comparing the different feature extraction networks of table 2, on the Pascal VOC datasets, use deeper network, have better ability to extract features. However, compared with the test results of self-built data sets, it can be seen that different feature extraction networks do not greatly improve the accuracy of target detection. So on small sample data sets, a deeper network layer does not effectively improve the recognition rate. On self-built datasets, the recognition accuracy rate can reach more than 95%, which indicates that the detection performance of Faster R-CNN for irregular targets meets the actual needs.

4. Conclusion

This paper for the shortcomings of traditional target detection algorithms. Use the deep learning target detection network Faster R-CNN to detect irregular target objects in the natural environment, the final test result is approximately 95%. Prove that Faster R-CNN have stronger ability to detect irregular target objects. But there is still have problems of slow detection speed and lost small target object. In the subsequent research, we will study how to improve the detection speed and accurate detection of small target objects.

References

[1] Taigman Y, Yang Ming, Ranzato M A, et al. Deepface: closing the gap to human-level performance in face verification [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. i.]: IEEE Press, 2014: 1701-1708.

[2] Ma Xiaoyu, Grimson W E L. Edge-based rich representation for vehicle classification [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. i.]: IEEE Press, 2005: 1185-1192.

[3] LOWED G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60 (2): 91-110.
[4] SCHWEGMANN C P, KLEYNHANS W, SALMON B P. Synthetic aperture radar ship detection using Haar-like feature [J]. IEEE Geosciences & Remote Sensing Letters, 2017, 14 (2): 154-158.

[5] NGUYEN V T, LE T L, TRAN T H, et al. A method for hand detection based on internal Haar-like features and cascaded adaboost classifier [C] //Proc of the International Conference on Communications and Electronics, 2016: 608-613.

[6] FELZENZSWALB P F, GIRSHICK R B, MCALLESTER D. Cascade object detection with deformable part models [C] //Proc of the Computer Vision and Pattern Recognition, 2010: 2241-2248.

[7] RUAN Z, WANG G, LIN X, et al. Deformable part-based model transfer for object detection [J]. IEICE Trans. On Information & Syst-ems, 2014, 97 (5): 1394-1397.

[8] XIE Y, PEI M, LIU Z, et al. Tracking pedestrian with multicomponent online deformable part-based mode [J]. Journal of Information Science &Engineering, 2016, 32 (3): 664-676.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[10] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In ICCV, 2011.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[12] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [C] //Proc of the 13th European Conference on Computer Vision, 2014.

[13] GIRSHICK R. Fast R-CNN [C] //Proc of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[14] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2015, PP (99): 1-9.

[15] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C] // Proc. Of the International Conference on Neural Information Processing System, 2015: 91-99.

[16] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatorally trained part-based models [J]. IEEE T-rans on Pattern Analysis and Machine Intelligence, 2010, 32 (9): 1627-1645.