Review of Target Detection Algorithm Based on Deep Learning

Xiaofang Liao [a], Xianfeng Zeng [b]

School of Information Science and Technology of South China Business College, Guangdong University of Foreign Studies, Guangzhou 510545, Guangdong

[a] fennyliao@163.com, [b] xf_zeng78@163.com

Keywords: Target detection; Deep learning; Computer vision.

Abstract: In recent years, artificial intelligence (AI) technology has developed rapidly, and personal safety, social safety, and national security have attracted more and more attention. Deep learning is widely used in different kinds of fields, among which target detection has made continuous breakthroughs in image detection or video processing. Target detection should be real-time and accurate, which is the requirement of people for the effect of target detection, while traditional target detection has been difficult to meet its requirements. Target detection algorithm based on deep learning has become the mainstream in this field. This paper mainly introduced two-stage models based on region detection classification: R-CNN, SPP-NET, Fast R-CNN, Faster R-CNN, and the advantages and disadvantages of the target detection algorithm YOLO and SSD based on regression single-stage model, and summarized and prospected the development direction of target detection.

1. Introduction

Target detection refers to identifying and locating the target object of interest in a still image (or dynamic video), which is the essential task of image understanding and application. The location of the object is accurately found in the given model, and the category of the purpose is marked. The detection results directly affect the results of subsequent high-level tasks such as target tracking, motion recognition, and behavior understanding. This is one of the core issues in the field of computer vision. The development of the traditional target detection algorithm was slow around 2010. Only in 2013 did researchers introduce Convolutional Neural Networks into target detection to break through the bottleneck of gradual development. Therefore, target detection based on deep learning can be developed rapidly.

Target detection should be real-time and accurate, which is the requirement of people for the effect of target detection. With the continuous development of deep learning technology, the real-time and accuracy of detection are gradually improving. Therefore, the target detection algorithm based on deep learning has attracted the attention of many researchers and has become one of the hot topics in the field of machine learning.

The traditional target detection algorithm needs to extract features manually, while the target detection algorithm based on deep learning uses depth network instead of manual feature extraction. At present, it is mainly divided into two categories: One is the target detection algorithm based on the two-stage model of region detection classification; The other is the target detection algorithm based on the regression single-stage model.

2. Target Detection Algorithm Based on Two-stage Model of Region Detection Classification

This kind of algorithm first generates a series of sparse proposals through selective search or Convolutional Neural Networks, and then classifies and regresses the samples of these proposals through Convolutional Neural Networks. The process is shown in Figure 1.

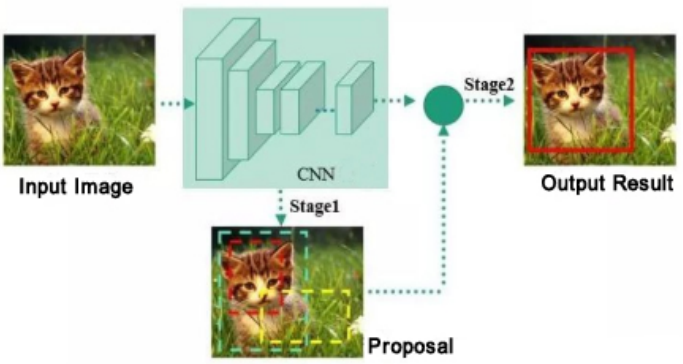

Figure 1 A schematic diagram of a target detection algorithm based on a two-stage model of region detection classification

2.1 R-CNN

R-CNN algorithm [1] is the first work of Region Proposal and CNN, which lays a foundation for the detection algorithm based on CNN in the future. R-CNN first uses Selective Search to predict the target and possible location of the input image, generating about 2,000 top-down proposal regions. These proposal regions are converted into fixed sizes, and then Convolutional Neural Networks are used for feature extraction. Finally, the extracted features are classified by SVM, and the position and size of the boundary frame are fine-tuned through boundary regression.

Although the R-CNN algorithm has achieved 50% performance improvement compared with the traditional target detection algorithm, the real-time performance cannot meet the actual requirements because the training algorithm speed is limited and takes up an ample disk space. There are a lot of repeated operations, which restrict the performance of the algorithm. In the process of normalization, feature truncation or stretching will occur, which will lead to incomplete information input to CNN and loss of some useful feature information. The following SPP-NET algorithm provides solutions to these problems.

2.2 SPP-NET

He Kaiming and others proposed an SPP-Net algorithm in 2014 [2] targeting the question that the Convolutional Neural Networks need to input a fixed size image size. By adding a Spatial Pyramid Pooling between the convolutional layer and the full connection layer, the Convolutional Neural Networks can process proposal regions of various sizes and obtain feature vectors of the same length from feature maps in these regions of different sizes, breaking through the limitation that CNN can only extract input of fixed size. The method solves the problem of CNN's repeated feature extraction of images, dramatically improves the speed of generating candidate regions, and saves the calculation cost. However, SPP-Net also has apparent disadvantages: A large number of results need to be transferred and cannot train parameters as a whole;

2.3 FAST R-CNN

Based on the SPP-NET algorithm structure, Girshich R et al. proposed an improved Fast R-CNN algorithm in 2015 [3], which developed the speed and accuracy. The process is as follows: Firstly, when inputting an image, a plurality of convolutional layers and pooling layers are used to generate a feature map. For each candidate region, the ROI pooling layer is used to extract feature vectors with a fixed lengths. Finally, the feature vectors are fed into a series of full connection layers. There are two output layers in this whole connection layer. One is to use softmax to estimate the probability of classes, and the other is to use a Bframe regressor to detect the coordinates of the target frame and correct the boundary position. The main advantage of Fast R-CNN is to realize real-time end-to-end joint training by establishing a multi-task model and using neural networks to carry out classification operations. Training can update all network layer parameters without additional disk space to cache

features. Fast R-CNN uses synchronization training at the end of the network to improve its accuracy, but it has no significant performance in improving the performance of classification steps.

2.4 FASTER R-CNN

In 2015, Ren Shaoqing and others proposed an improved Faster R-CNN [4] for the R-CNN series. The process of RPN (Region Proposal Network) is as follows: Taking the whole image as input, calculating the total area to get the feature layer, and then inputting convolution features into RPB network to obtain the feature information of the proposal; Then, the classifier is used to judge whether the features extracted from the proposal belong to a specific class. Finally, for the proposal belonging to a certain feature, the regression is used to further adjust its position. In a word, Faster R-CNN abandons the sliding window method and introduces the RPN network, which makes region extraction, classification, and regression share convolution features, thus significantly improving the operation speed. However, this algorithm needs a large number of Anchor Frame to determine the target, and then to recognize the target.

3. Target Detection Algorithm Based on Regression Single-stage Model

This kind of target detection algorithm does not need to generate a proposal, and directly converts the problem of target border location into a regression problem to deal with. Features are directly extracted from the CNN network to predict object classification and location. Uniform and dense sampling are carried out on the feature map of multiple layers of the image. Different scales and aspect ratios can be used in sampling, and CNN is used to extract features and then directly classify and regress. The whole process only needs one step, so its advantage is fast speed ,as shown in Figure 2. However, an essential disadvantage of uniform and dense sampling is that training is challenging, resulting in slightly lower accuracy of the model.

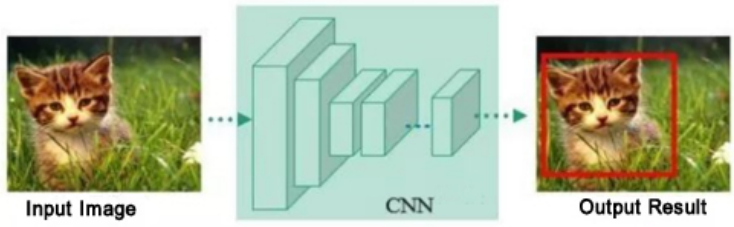

Figure 2 A schematic diagram of target detection based on a regression single-stage model

3.1 YOLO

The YOLO (YOU Only Look Once) algorithm proposed by Redmon et al. [5] solved the target detection as a regression problem. The original image containing multiple targets is input, and the bounding frame and the category of the target are directly regressed on the divided grid. The algorithm is to scale the input image and share SXS-sized networks. If the center of the object to be detected falls into the networks, the bounding frames and the credibility of these bounding frames are predicted. YOLO is based on the Google-Net image classification model, which realizes end-to-end target detection in a real sense. The detection speed is fast, but the accuracy is slightly lower than that of Faster R-CNN.

3.2 SSD

SSD algorithm [6] proposed by Liu W et al. is a single-layer deep neural network, which can directly predict the coordinates and categories of targets without the process of generating candidate frames. The algorithm combines YOLO's regression idea and Faster R-CNN's anchor mechanism to achieve both speed and accuracy. Its core is to use a small convolution filter to predict the class scores and frame offset of a fixed set of default bounding frames in the feature map. In the recognition stage,

the position information and category information of several bounding frames selected from the whole image is input, and several bounding frames with different scales and shapes are used in the feature image to predict the target object. In the testing phase, the network predicts the existence possibility of each category of objects in each boundary frame, and finally obtains the final detection results through NMS (Non-Maximum Suppression) method. The algorithm not only maintains the advantage of YOLO's fast speed, but also is as accurate as Faster R-CNN.

4. Analysis of Experimental Comparison Results

The performance of the target detection algorithm is mainly manifested in real-time and accuracy. Through Table 1, we will understand the performance of the target detection algorithm based on deep learning. Each target detection algorithm makes a comparative analysis of detection accuracy and speed on the public data set PASCAL VOC2007 [7]. As shown in Table 1.

Table 1 Detection Accuracy and Speed Table of Various Algorithms for Target Detection

Algorithm Name	Basic Network	mAP	FPS
R-CNN	AlexNet	66.0	34s/sheet
SPP-NET	ZF-5	59.2	2
FAST R-CNN	VGG-16	70.0	3
FASTER R-CNN	VGG-16	73.2	5
YOLO	custom	63.4	45
SSD300	VGG-16	73.2	59

As can be seen from the table, the detection accuracy of Faster R-CNN, YOLO, and SSD has exceeded 60%, and only YOLO and SSD can meet the real-time requirements in detection speed. When the basic network of the algorithm changes, the detection accuracy, and detection speed will change accordingly. Whether it is Faster R-CNN, YOLO, or SSD, when the basic network is changed, or the size of the input image is changed, the detection accuracy is improved, but the detection speed will also decrease at the same time. It can be seen that the detection accuracy and detection speed achieve good results at the same time. How to balance the detection accuracy and detection speed is the focus of the following research.

5. Conclusion

This paper introduced the advantages and disadvantages of the target detection algorithm based on deep learning, improved the performance of the target detection algorithm based on deep learning, and the traditional target detection algorithm, and then greatly improved the detection speed and accuracy. However, there are still some problems that have not been well handled, such as the problem of small target detection, the detection performance has not yet reached the ideal state, etc., which all require researchers to continue in-depth research and development.

Therefore, target detection based on deep learning is a challenging topic, and its research significance and application value are paramount. With the emergence of more and more comprehensive data sets and various open-source learning frameworks, the topic will develop more rapidly.

References

[1] Girshick R , Donahue J , Darrell T , et al.Rich Feature Hierarchies for AccurateObject Detection and Semantic Segmentation[C]// 2014 IEEE Conference onComputer Vision and Pattern Recognition(CVPR). Columbus, USA: IEEEComputer Society, 2014: 580-587.

[2] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16.

[3] Girshick R. Fast R-CNN[C]// IEEE International Conference on Computer Vision. Washington，USA: IEEEComputer Society, 2015: 1440-1448.

[4] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137-1149.

[5] Redmon J, Divvala S, Girshick R, et al.You Only Look Once: Unified, Real-Time Object Detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA: IEEE Computer Society, 2016: 429-442.

[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiFrame Detector[C]// European Conference on Computer Vision. Springer International Publishing, 2016.

[7] Everingham M, Winn J.The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Development Kit[J].International Journal of Computer Vision,2006,111(1):98-100.