Survey of Target Detection Based on Neural Network

Bao Deng¹,²,*, Hao Lv¹,²

¹School of Computer Science and Engineering, Northwestern Polytechnical University, Xi’an, China
²Xi’an Aeronautics Computing Technique Research Institute, Aviation Industry Corporation of China, Xi’an, China

*Corresponding author: dengb@avic.com

Abstract. Target detection is a hot topic in the field of artificial intelligence, which is widely used in robot, UAV, aerospace and other fields. In this paper, the research background and significance of target detection are summarized, and two categories of target detection algorithms based on deep learning, i.e., candidate region based and regression based, are described. For the first category, a series of region with convolutional neural network (r-cnn) algorithms are introduced, this paper introduces the researchers' research on the basis of r-cnn algorithm: the improvement of feature extraction network, pooling layer of region of interest and non-maximum suppression algorithm. The second algorithm is divided into Yolo (you only look once) series, SSD (single shot multibox detector) algorithm and its improvement. According to the current target detection algorithm in the development of more efficient and reasonable development trend, the research hotspot of target detection algorithm in the future is prospected, including unsupervised and unknown class object detection.

Keywords: Target detection, Deep learning, CNN, feature extraction.

1. Introduction

The main task of target detection is to locate the target of interest from the input image, and then accurately determine the category of each target of interest. At present, target detection technology has been widely used in daily life safety, robot navigation, intelligent video surveillance, traffic scene detection, aerospace and other fields. At the same time, object detection is the basis of other advanced visual problems such as behavior understanding, scene classification and video content retrieval. However, there may be great differences between different instances of the same kind of objects, while different objects may be very similar, and different imaging conditions and environmental factors will have a huge impact on the appearance of objects, which makes target detection very challenging.

In recent years, target detection algorithms based on deep learning form two categories: candidate region based and regression based. The target detection algorithm based on candidate region is also called two-stage method, which divides the target detection problem into two stages: one is to generate candidate region, the other is to generate candidate region, the second is to put the candidate regions into the classifier for classification and position correction. The target detection algorithm based on regression has only one stage, which directly regresses the predicted target objects.
2. Two stage target detection algorithms

2.1. R-CNN Model
In 2014, girshick et al. [1] successfully applied the convolutional neural networks (CNN) in the field of target detection, and proposed r-cnn algorithm, which combines alexnet with selective search algorithm [2] to decompose the target detection tasks into several independent steps (as shown in Figure 1). Firstly, 2000 candidate regions are extracted by selective search algorithm, then each candidate region is normalized, and then the features are extracted in CNN one by one. Finally, SVM classification and region regression are carried out for the features.

The detection accuracy of r-cnn algorithm on Pascal voc2007[3] data set has reached 58.5%. Compared with the traditional target detection algorithm, it has made great progress. But there is still much room for improvement. For example, 2000 candidate regions extracted from a single image need to be input into CNN one by one, which leads to huge computational cost and seriously affects the detection speed; Moreover, before the candidate region is input into CNN, it must be clipped or scaled to a fixed size, which will deform the candidate region and lose more information, resulting in the decline of network detection accuracy.

2.2. Fast R-CNN Model
In 2015, girshick et al. [4] proposed the fast r-cnn algorithm (as shown in Figure 2). Inspired by the SPP net algorithm, they simplified the SPP layer into a single scale ROI pooling layer to unify the size of candidate region features, and further proposed the idea of multi task loss function to unify the training and learning of classification loss and boundary box regression loss, the classification and location tasks can not only share convolution features, but also promote each other to improve the detection effect.
Although fast r-cnn effectively speeds up the detection rate, it still relies on the selective search algorithm to generate candidate regions. Studies have shown that the convolution layer of convolutional neural network has a good ability to locate the target, but this ability is weakened in the full connection layer. Therefore, Ren et al. Proposed the framework of fast r-cnn algorithm in 2015. RPN is a kind of full convolutional network (FCN) structure. It takes any size of feature map as input and produces a series of candidate regions that may contain the target after convolution operation, which makes the algorithm realize end-to-end training and greatly improves the detection speed.

2.3. Research on Improvement

Fast r-cnn algorithm has achieved good results in the accuracy and speed of detection. It is mainly composed of four modules: feature extraction network is used to extract image features; ROI pooling layer normalized the candidate region features of different sizes; RPN generates candidate regions of the target according to the image characteristics; NMS [5] algorithm is used to remove redundant detection box.

Lin et al. [6] proposed featurepyramid network (FPN) in 2017. FPN constructed a hierarchical structure with horizontal connection from top to bottom, extracted multiple different scale features for detection, each scale feature is obtained from the fusion of high-level features and shallow features, which not only has strong semantic information, but also has rich geometric information.

ROI pooling, that is, the pool of interested regions is to divide the corresponding feature graphs of candidate regions into fixed number of space blocks, and then maximize or average pool each space block, so that different candidate regions can output the same size feature map.

RPN is the main innovation of fast r-cnn algorithm, which is mainly based on the anchor mechanism to generate a large number of target candidate regions. In recent years, the improvement research aims to produce more accurate candidate areas to improve the detection effect.

In 2017, Zhao et al. Proposed cascade r-cnn algorithm. Through cascading r-cnn detection model with three regions intersection over Union (IOU) threshold increasing, the candidate regions generated by RPN were screened, leaving candidate areas with high IOU value, which effectively improved the detection accuracy of the model. In 2018, Chen et al. Introduced context information into RPN stage to fine tune candidate areas, making network positioning more accurate.

NMS algorithm firstly sets an IOU threshold manually, sorts all detection boxes of the same class according to classification confidence, selects the detection results with the highest classification confidence score, removes the adjacent results with IOU value exceeding the threshold, so that the network model can get a better balance between recall rate and accuracy.

3. Single stage target detection algorithm

The regression-based target detection algorithm does not need candidate region to generate branches. For given input images, the candidate boxes and categories of the target are returned directly in multiple positions of the image. This paper will be divided into two series to summarize the regression-based target detection algorithms: Yolo series and SSD series.

3.1. YOLO Model

In 2015, Redmon et al. [7] proposed the Yolo algorithm, which integrates the functions of classification, location and detection in a network. The input image only needs one network calculation to directly obtain the boundary box and category probability of the target in the image. As shown in Figure 3, the whole input image is divided into S-type × Each grid is only responsible for the object whose center falls on the grid and only predicts the information of B bounding boxes, and then selects the appropriate confidence threshold to remove those bounding boxes with low probability of target existence. Although Yolo algorithm completely abandons the step of candidate region generation, it greatly improves the detection speed and can meet the speed requirements of real-time target detection. However, due to its rough network design, it cannot meet the accuracy requirements
of real-time target detection, and there are some problems, such as the target cannot be accurately located, easy to miss detection, small target and multi-target detection effect is not good.

**Figure 3. YOLO Algorithm structure**

In 2017, Redmon et al. Proposed yolov2 algorithm, which made a series of improvements to Yolo algorithm, focusing on solving the problem of low recall rate and poor positioning accuracy. It used anchor mechanism of fast r-cnn algorithm to remove the full connection layer in the network, and used convolution layer to predict the position offset and category information of detection frame. Moreover, it is different from the manual design of original anchor mechanism, it uses K-means clustering to learn the best initial anchor template in the training set. In addition, yolov2 adds a pass-through layer to connect the shallow feature map to the deep feature map, which makes the network have fine-grained characteristics. In addition, yolov2 can use multiple data sets to optimize training mode, and use wordtree method to train simultaneously on Imagenet classification data set and MS coco detection data set, the real-time detection task of more than 9000 target categories is realized.

In 2018, Redmon et al. Proposed yolov3 algorithm, which uses the idea of jump connection in residual network for reference, and constructs a 53-layer benchmark network named dartenet-53, which only uses $3 \times 3$ and $1 \times 1$ convolution layer of 1 has the same classification accuracy as resnet-152 [8], but greatly reduces the amount of calculation; In order to deal with multi-scale targets, three different scale feature maps are used for target detection. Each feature map is the fusion of high-level and shallow level feature maps; In the prediction of categories, logistic regression method is used instead of softmax method, so that each candidate box can predict multiple categories and support the detection of objects with multiple labels. Yolov3 algorithm can meet the requirements of accuracy and speed of real-time detection tasks, and has become one of the most preferred target detection algorithms in the current engineering field.

### 3.2. SSD Model

Liu et al. [9] proposed SSD algorithm in 2016. Based on the regression thought, combined with the idea of multi-scale detection, multiple feature maps of different scales were extracted for detection, and the larger feature map was used to detect relatively small targets. The strategy of smaller feature map to detect larger targets significantly improved the detection effect of large targets. At the same time, by using the anchor mechanism of fast r-cnn algorithm, the fixed number of prior frames with different scales and aspect ratio are preset at each position of the extracted feature graph. The network can directly extract candidate boxes on feature maps for intensive sampling and extraction. At the same time, the real-time detection speed is maintained, the positioning accuracy of the model is improved. As shown in Figure 4, SSD network is based on the structure of full convolution network. It replaces the full connection layer of vgg16 with convolution layer, and adds several auxiliary convolution layers at the end of vgg16 network to reduce the size of feature map, which is used to
extract feature maps of different scales, and the convolution operation is used to detect the feature map of different scales.

SSD algorithm surpasses fast r-cnn algorithm in speed and accuracy of detection. However, different volume layer features extracted by SSD algorithm are input into their detection branches independently, so it is easy to see that the same object is different in size.

The small boundary frame can detect the same time, that is, the problem of repeated detection. Moreover, the detection branch of each layer only focuses on the target of specific scale on its own branch, and does not consider the relevance between different layers and different scale targets, so the detection effect of small targets is general.

![SSD Algorithm structure](image)

**Figure 4. SSD Algorithm structure**

3.3. Research on Improvement

In 2017, Jisoo et al. [10] proposed RSSD algorithm. On the basis of SSD algorithm, it adopted a special feature fusion method for the extracted features of different scales: for each specific scale feature, the larger scale feature was pooled, and the smaller scale feature was deconvoluted. In the same year, Cheng et al. Proposed dssd algorithm to replace vgg16 with resnet101, two special modules are designed: prediction module and deconvolution module. The structure of prediction module is similar to residual module. The fusion of different layers of features is realized by jumping connection, so as to improve the representation ability of features.

In 2018, Liu et al. [11] proposed the RFB net algorithm, designed the receptive field block (RFB) to increase the feature extraction ability of the network by simulating the human visual receptive field. RFB structure uses the idea of concept for reference and introduces three different expansion rates of \( 3 \times \times \) In addition, Zhang et al. Proposed refinedet algorithm, which combines the advantages of one-stage and two-stage detection algorithm, and designed two modules: object detection module and anchor fine-tuning module. The former filters dense anchors to remove some negative samples without objects, At the same time, the position and size of the screened anchors are roughly adjusted, and the latter further regresses the anchors output from the object detection module, which makes the network perform two regression tasks, effectively improving the network positioning ability, and the sample screening effectively alleviates the imbalance between positive and negative samples.

The latest improvement research on SSD [12] focuses more on the reasonable and efficient use of FPN structure, extracting multi-scale features with rich context information and spatial information, and solving the problem of target scale change. In addition, Zhao et al. Proposed a multi-level feature pyramid network (mlfpn) by cascading multiple small-scale FPN subnetworks, the formation of different levels of different scale features, and the full use of features and fusion, so that the network performance and small target detection are greatly improved.

4. Conclusion

Currently, the popular data sets in common target detection tasks are Pascal VOC 2007, Pascal VOC 2012, Ms coco, Imagenet, open images, livs, etc [13].
Target detection is a very important research field and has a wide application prospect. In this paper, the emerging object detection algorithms based on deep learning in recent years are divided into candidate region and regression based. The development and improvement of these two algorithms are reviewed in detail. The popular data sets in the field of target detection are introduced. Although the current target detection algorithm has been widely used in real life, there are still many challenges.

How to effectively combine the context information to solve the detection of small targets and occluded targets in complex real scenes; The second is to explore a better feature extraction network or a feature extraction network specially designed for the detection task, as well as a better algorithm. The selection method of detection frame; Third, the current target detection algorithms are based on supervised learning, there are a large number of unlabeled data in reality, so it is very valuable to study how to use the weak supervised learning target detection algorithm. The fourth is to explore how to move from the known category of target detection, combined with effective semantic information, to the unknown category of target detection.

References

[1] SHAWE-TAYLORJ, CRISTIANINI N. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. England: Cambridge University Press, 2000.
[2] FREUND Y, SCHAPIRE RE. Experiments with a new boosting algorithm. International Conference on Machine Learning. USA: IMLS, 1996.148-156.
[3] LIAW A, WIENERM. Classification and regression by random forest. Rnews, 002, 2(3): 18-22.
[4] HEK, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. Proceedings of the European Conference on Computer Vision. Switzerland: IEEE, 014.346-361.
[5] GIRSHICK R. Fast R-CNN. Proceedings of IEEE International Conference on Computer Vision. USA: IEEE, 2015.1440-1448.
[6] LONG, SHELHAMER, DARRELL T. Fully convolutional networks for semantic segmentation. Computer Vision and Pattern Recognition. USA: IEEE, 2015.3431-440.
[7] SIMONYANK, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations. USA: IEEE, 2015.714-723.
[8] JEONG J, PARK H, KWAK N. Enhancement of SSD by concatenating feature maps for object detection. British Machine Vision Conference. UK: BMVA, 2017.
[9] FU Cheng-yang, LIU Wei, RANGA A, et al. DSSD: Deconvolutional Single Shot Detector. http://arxiv.org/abs/1701.06659, 2017.
[10] LAW H, DENG J. CornerNet: detecting objects as paired keypoints. European Conference on Computer Vision. Germany: IEEE, 2018.765-781.
[11] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation. European Conference on Computer Vision. Netherlands: IEEE, 2016.483-499.
[12] ZHOU Xing-yi, ZHUO Jia-cheng, KRAHENBUHL P. Bottom-up object detection by grouping extreme and centerpoints. Computer Vision and Pattern Recognition. USA: IEEE, 2019. 850-859.
[13] LUO Hui-ji an, CHEN Hong-kun. Survey of Object Detection Based on Deep Learning. ACTA ELECTRONICA SINICA. China. Vol. 48, 2020.