Target Detection Based on Deep Learning

Yuchen Zhu*

1College of Artificial Intelligence and Data Science, Hebei University of Technology, TianJin, China.

*Corresponding author’s e-mail: gaoming@cas-harbour.org

Abstract. As an important research direction of computer vision, target detection has been widely used in face recognition, intelligent driving, robot navigation and other fields. In recent years, with the deepening research on deep learning, great progress has been made in the field of computer vision, such as image acquisition, image processing and target detection. Compared with the traditional target detection algorithm based on candidate regions, it has the problems of poor timeliness and slow detection speed. Recently, the popular target detection algorithm based on regression realizes the real sense of end-to-end detection and greatly improves the detection efficiency. However, the accuracy of small target detection and dense target detection has not been solved. In the future, we still need to improve the efficiency and accuracy of recognition on the existing basis, and solve the problem of small target and dense target detection to make it more widely used in practical application scenarios. In this paper, the principle, advantages and disadvantages, accuracy and other aspects of the above algorithms are introduced in detail, the problems existing in the target detection algorithm are summarized, and the future development direction has been prospected. In short, both algorithms have advantages and disadvantages, but the regression-based target detection algorithm has better practicality and development prospects.

1. Introduction

Target detection is mainly to judge the position and feature of input images, and the essence of it is to combine image segmentation and recognition into one part. Target detection is an important direction of the field of computer vision. In recent years, target detection is widely used in face recognition, intelligent driving, robot navigation and so on[1].

Before the rise of deep learning, traditional target detection algorithms are mostly detected by artificial features. Three steps of traditional target detection are shown in Fig 1: (1) Generate a large number of candidate regions by using exhaustive image segmentation techniques. (2) Feature extraction for each candidate region (such as HOG[2], SIFT[3], Haar[4] etc.). (3) Transfer these features to a classifier (such as SVM[5], Adaboost[6], Random Forest[7]) and it determines the category of candidate region. Due to the untargeted sliding window and the need for manual feature extraction, target detection algorithms have high time complexity, low robustness, poor accuracy and real-time performance, which makes it into a bottleneck.

Fig 1: Three Steps of Traditional Target Recognition
With the development of deep learning, replacing the traditional manual selection and extraction process by introducing Convolutional Neural Network (CNN) [8] self-learning target features. Furthermore, many algorithms are innovated on this basis, which improved the efficiency and accuracy of target detection. At present, the target detection algorithm based on deep learning is divided into two categories: Based on candidate regions and based on regression. In this paper, the mainstream algorithms of two categories are reviewed and compared. Finally, the existing problems in the field of target detection are summarized and prospected.

2. Overview of target detection algorithms based on region proposal
Professor Yan LeCun [9] firstly proposed that CNN is the core component of the target detection algorithm bans on candidate regions. Because of the advantages of local connection and weight sharing in convolutional neural networks, it has good robustness in the application of object classification. In addition, this also makes convolutional neural networks have translation, scaling and distortion invariance to some extent. Target detection algorithms based on region proposal generate candidate regions that may contain objects to be tested. This process mainly uses Selective Search [10], EdgeBoxes [11] and other algorithms. Then these candidate regions will be tested and verified, so that final test results can be obtained.

2.1. R-CNN
The R-CNN [12] model proposed by Girshick et al., which made a great breakthrough in target detection. By experiment, mean Average Precision (mAP) for all categories of R-CNN on VOC 2007 data set [13] increased to 58.5%.

![Fig 2: R-CNN Model Process](image)

The R-CNN steps are shown in Fig 2: (1) Input the image to be tested. Generate around 2000 candidate boxes by using the Selective Search algorithm. (2) Scale each candidate box to the same size (Original text adopts the 227×227 scaled size). (3) Extract features vector from CNN. (4) Send the feature vectors to SVM classifiers and fully connected networks, then classify vectors and fine-tune borders.

Although the R-CNN algorithm improves the accuracy, it still exists many disadvantages. For example, CNN requires forward propagation for feature extraction when the Selective Search algorithm extracts every candidate box. This situation leads to insufficient real-time performance. In addition, force scaling to a fixed size for the input image retrieval box makes images be distorted. This also leads to lower accuracy. However, SPP-NET provides solution to these problems.

2.2. SPP-NET
SPP-NET was proposed by Kaiming He [14] on the basis of R-CNN. The first full connection layer in R-CNN requires fixed input image size, but the convolution layer does not require image size. So He et al. added SPP layer after the last convolution layer. This setting avoids the complex process of normalizing candidate regions by R-CNN, which reduces the operation time. The detection speed is increased by about 100 times. The results show that the mAP of all categories on the VOC2007 data set [13] is increased to 60.9%.

Specific steps are shown in Figure 3: (1) Input the image to be tested. Generate around 2000 candidate boxes by using the Selective Search algorithm. (2) Input the image to CNN for one-time feature extraction of the whole image. To get a feature vector of fixed size, pyramid space pooling for
each candidate box for feature mapping after getting the feature map. (3) Send the feature vectors to SVM classifiers and fully connected networks, then classify vectors and fine-tune borders.

![A network structure with a spatial pyramid pooling layer.](image)

**Fig 3:** A network structure with a spatial pyramid pooling layer.

Here 256 is the filter number of the conv5 layer, which is the last convolution layer. [14]

The disadvantage is that inheriting R-CNN needs to store a large number of features and multi-stage training. And because of the multi-scale of pyramid pooling layer, added a new problem that all convolution layers before the pyramid pooling layer cannot be fine-tuned.

### 2.3. Fast R-CNN

Ren et al.'s Fast R-CNN [15] model introduces a single pyramid pooling layer based on R-CNN model. This model solves the problem of repeated calculation of the candidate box and allows fine-tuning of all layer parameters. Finally, SVD is used to decompose the parameter matrix of the full connection layer and compress them into two small full connection layers. In this model, SVM classifier is changed to SoftMax classifier.

The specific steps are shown in Figure 4: (1) Input the image to be detected. Use CNN to extract features and use a sliding window to find objects in the image. Use Generate around 2000 candidate boxes by using Selective Search algorithm. (2) The single pyramid pooling layer is used to pool the candidate regions to the feature representation of a fixed size. (3) Softmax is used for multi-classification target recognition, and the border position is fine-tuned by border regression.

![The model of Fast R-CNN](image)

**Fig 4:** The model of Fast R-CNN

Later using VOC 2007 combined with VOC 2012 data set, the test result is 70% (The expansion of data sets greatly increases the performance of target detection). Although the detection speed is greatly improved, the end-to-end training test is still not realized, which cannot meet the real-time application.
2.4. Faster R-CNN

The time-consuming nature of the algorithm described above is mainly due to the use of an independent candidate box generation algorithm. In order to solve this problem, Ren et al.[16] proposed the FasterR-CNN target detection framework. The algorithm uses Region Proposal Networks (RPN)[17] instead of Selective Search to generate candidate regions. Fully convolutional neural network - RPN can extract proposals by sharing the basic features of volume. The core idea is to generate candidate regions directly using CNN. The method used is essentially a sliding window. Because each task cooperates with each other and shares parameters in the Faster R-CNN training process, the detection time is greatly shortened. The mAP reached 69.9% on the VOC 2007 dataset.

The specific steps are shown in Figure 5: (1) Input the image to be tested, extract the feature of any size of the input image by CNN, and output a set of rectangular Object Proposal, which greatly improves the real-time performance of target detection. (2) The sliding network of the last layer of shared conv layer formed with RPN slides once. Using K anchors of different sizes to locate sliding windows (usually K = 9), candidate regions are predicted and feature region candidates are generated. (3) ROI Pooling uses the candidate regions generated by RPN and the shared feature map extracted by the convolutional neural network to obtain the proposal feature map of a fixed size. (4) Softmax is used to classify the specific categories of the proposal, and the accurate position of the object is obtained by using the border regression.

Fig 5: The model of Faster R-CNN

The use of RPN enables Faster R-CNN to integrate multiple steps of region recommendation, feature extraction, classification and location in the network, and truly become end-to-end training. However, because an anchor box on the Faster R-CNN feature map corresponds to a large area in the original map, the Faster R-CNN is not good for small target detection.

3. Overview of target detection algorithms based on regression

Target detection algorithms based on candidate regions are all input images to extract features firstly, and then identify objects in feature space by classifier and locator. This process causes poor timeliness and slow detection speed. The subsequent target detection algorithm (as YOLO, SSD) based on regression makes full use of the idea of regression. They extract features directly from the original image and regress at multiple locations and it will appear target borders and target categories.

3.1. YOLO

In 2016, Redmon et al.[18] proposed the YOLO algorithm. Its network design strategy continues the core idea of GoogleNet[19]. YOLO integrates classification, location and detection functions into a network, which can predict multiple Box locations and class probabilities at one time, and achieve end-to-end target detection in the true sense.

The YOLO algorithm zooms the input image to a uniform size and divides it into s x s grids. Each grid is responsible for only the target object whose center point is located on the grid and predicts the value of the B(self-defined) scale boundary candidate box. The final one-time prediction contains the target object center grid. Due to the rough grid design, it meets the real-time requirements in the time dimension. However, the accuracy is far below the accuracy requirement of target detection. And each grid can only identify one object, so in small objects and high density (such as flying birds, crowded streets) problems.
3.2. SSD
In 2016, LIU[20] proposed an SSD algorithm. The algorithm draws on a Faster R-CNN anchor frame design mechanism. Firstly, it generates a series of default boxes with different sizes and aspect ratios on feature maps. In the meantime, generate different types of confidence scores on the default box, then non-maximum suppression is performed to complete the detection process. Moreover, SDD based on YOLO's return thought and thought of multi-scale detection, extract feature maps of different scales for detection. SSD follows the principle that larger feature maps detect smaller targets and smaller feature maps detect larger targets. This principle improved detection accuracy significantly.

SSD is faster and more accurate than Faster R-CNN. But since SSD uses information from shallow networks with fewer feature convolution layers to detect small targets, caused insufficient representation and a lack of deep semantic features. The recall rates for small targets are also low. Therefore, it is suggested that the performance of SSD can be improved by enhancing the precision data of small targets in data sets. Since this algorithm inputs the features of different convolution layers into their respective detection branches independently, their connection is not strong enough. Therefore, it is easy to detect the same object with different sizes of boundary frames at the same time, that is, the problem of repeated detection has not been solved.

4. Comparison of target detection based on candidate regions and target detection based on regression

Table 1: Comparison of two types of target detection

| Target detection algorithm | target detection based on candidate regions | target detection based on regression |
|----------------------------|-------------------------------------------|-------------------------------------|
| Principle                  | Input the image, generate candidate regions by selective search and region generation network, and then obtain the feature map by convolution and pooling. Finally, classify the image by classifier and refine the boundary box by linear regression. | The input image is processed directly, and the position coordinates and category information are generated directly on the original image. Finally, the position is regressed. |
| Advantage                  | Applicable to multi-scene High accuracy of target detection | Fast testing velocity Less data stored |
| Disadvantage               | Complex calculation process More data stored Slow detection speed | Low accuracy of small target detection High density things have low accuracy |
| Real-time                  | ×                            | ✓                                |

For the two popular target detection algorithms, I compare their principles, advantages, disadvantages and real-time performance, and list them in the table, making it easier to understand their similarities and differences.

5. Conclusion
The target detection algorithm is a very important research field. This paper reviews two main target detection algorithms based on deep learning. Target Detection Architecture Based on Candidate Region has high accuracy and low missing rate. But the detection speed is too slow to meet real-time application scenarios and accomplish end-to-end target detection. Target detection architecture based on regression directly classifies and returns the target images, which improves detection speed greatly. Missing rate and accuracy are also improved as data sets are updated and improved in new
versions. Although the target detection algorithm based on deep learning has a great performance improvement in speed and accuracy compared with the traditional target detection algorithm, there are still many directions worthy of our study: (1) How to improve the utilization rate of context information, thereby enhancing the accuracy of small target detection in complex scenes. (2) The above algorithms are trained in a large number of data sets. But large-scale instance data annotation is an expensive project. So how to train network models with weakly supervised joint large scale image-level classification and small scale Instance data annotation is a breakthrough problem. (3) Future research can explore better or special design for detection task feature extraction network, and find better detection box selection method.

References

[1] Shunfei Wang, Junhua Yan, Zhigang Wang. (2015). Improved moving target detection method based on local joint features. Chinese Journal of Scientific Instrument, 36(10), 8.
[2] Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In: Computer Society Conference on Computer Vision & Pattern Recognition. IEEE.
[3] Lowe, D. (2004). Distinctive image features from scaleinvariant keypoints. International Journal of Computer Vision, pp: 60.
[4] Lienhart R, Maydt J. (2002) An extended set of Haar-like features for rapid object detection. Proceedings. In:International Conference on Image Processing. IEEE
[5] Cristianini N, Shawe-Taylor J. (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods: Generalisation Theory[J]. In: Conference on Artificial Intelligence. 16(2), 337-337.
[6] Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm DRAFT. Morgan Kaufmann Publishers Inc, San Francisco.
[7] Liaw, A., & Wiener, M. (2001). Classification and regression by randomForest. R News 2:18-22.
[8] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Tan, R. S. (2017). A deep convolutional neural network model to classify heartbeats. Computers in Biology and Medicine, 89.
[9] Lecun, Y., & Bottou, L. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324.
[10] Van, d. S. K. E. A., Uijlings, J. R. R., Gevers, T., & Smeulders, A. W. M. (2011). Segmentation as selective search for object recognition. IEEE.
[11] Zitnick, C. L., & Dollar, P. (2014). Edge Boxes: Locating Object Proposals from Edges. ECCV. Springer International Publishing, West-Berlin.
[12] Tongtong Liu, Rubing Wang, Jikang Chen, Shengliang Han, Jimin Yang. (2018) Fine-Grained Classification of Product Images Based on Convolutional Neural Networks. Advances in Molecular Imaging., 4
[13] Everingham, M., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International Journal of Computer Vision. Netherlands. 88(2), 303-338.
[14] Ren SH, Girshick R et al. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. LOS ALAMITOS, USA.39(6) 1137-1149
[15] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
[16] REDMON J, DIVVALA S, GIRSHICK R, et al. (2016) You only look once:Unified, real-time object detection. In:IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA.
[17] REDMON J, DIVVALA S, GIRSHICK R, et al. (2016) You only look once: Unified, real-time object detection. In:IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA.
[18] Zhuoyao Zhong, Lianwen Jin, Zecheng Xie. (2015) High performance offline handwritten Chinese character recognition using GoogLeNet and directional feature maps. In: ICDAR.France.pp. 846-850

[19] Liu, W., Anguelov, D., Erhan, D., et al. (2016) SSD: Single Shot MultiBox Detector. European Conference on Computer Vision, Springer, Cham, 21-37.