Discussion on Target Detection Algorithm in Computer Vision

Fujian Qiu
21821194
School of Computer Science and Technology,
Zhejiang University
21821194@zju.edu.cn

Abstract

Computer vision is a wide range of disciplines. In general, computer vision includes five major technologies: image classification, target detection, target tracking, semantic segmentation, and object segmentation. This article focuses on the detection of targets, and discusses the main target detection algorithms in recent years and some comparisons between these algorithms.

1. Introduction

Target detection usually outputs the border and label of a single target from the image. We can use a graph to illustrate the effect of target detection as shown in Figure 1. Of course, Figure 1 not only illustrates the process of target detection, but also illustrates the essence of the three major tasks of computer vision: image classification is to answer the question that this image is a cat, and target detection not only needs to answer what is in the image, but also It is also necessary to give the positional problem of these objects in the image. In the figure, for example, not only the A-cat and the dog in the figure are identified, but also the specific positioning of the A-cat and the dog. Image segmentation requires pixel-level image segmentation. In the figure, for example, each object is segmented by pixel-level criteria, which requires higher algorithms.

Beginning with this section, we will conduct a comprehensive and detailed study and explanation of the second major task of computer vision, target detection. It is not only an extension of the content of the previous CNN image classification, but also a must-see for further research on image algorithms. Before the formal learning of various target detection algorithms, this section will first review the main algorithms that affect the development history of the target detection algorithm, in order to play an important role in the later content.

Before the deep learning was officially involved in the computer vision target detection task in 2012, the traditional target detection algorithm has been the steps of region selection, feature extraction and classification regression in a more traditional way such as sliding window convolution,

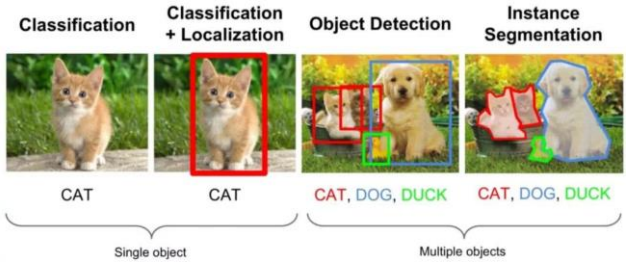

Figure1.

for example, before the rise of deep learning. The Deformable Component Model (DPM) method is excellent in the detection field. After deep learning has emerged and gradually become the core method of computer vision, a series of target detection algorithms based on deep learning algorithms can be roughly divided into two major schools:

- Two-stage algorithm: first generate candidate regions and then perform CNN classification (RCNN series)
- One-stage algorithm: Apply algorithms directly to the input image and output categories and corresponding positioning (yolo series)

Whether it is a two-stage or one-stage algorithm, they are looking for a balance point or extreme point on the fast and quasi-points of recognition. Either faster or faster, but as deep learning and computer vision move forward, algorithms that are both fast and accurate are gradually being implemented. This section evaluates the target detection algorithms of the two major schools of target detection.

2. Two-stage algorithm

2.1. R-CNN

As a pioneering approach to introducing deep learning into target detection algorithms, R-CNN[1] is of great significance in the history of target detection algorithms. The R-CNN algorithm is representative of the two-step method, that is, the region is proposed as a region proposal, and then the CNN is used for recognition and classification. Since the candidate box plays a key role in the success or failure of the algorithm, the method is named after the initial

letter R of the Region plus CNN. Compared with the traditional sliding convolution window to judge the possible region of the target, R-CNN uses the selective search method to pre-extract some candidate regions that are more likely to be target objects, and the speed is greatly improved, and the calculation cost is also significantly reduced. Overall, the R-CNN approach is divided into four steps:

- Generating candidate regions
- Feature extraction using candidate CNN for candidate regions
- Send the extracted features to the SVM classifier
- Finally, use the regression to correct the target position.

Although R-CNN was born very early in 2013, there are still many shortcomings: the positive and negative sample candidate regions of the training network generated by the selective search method are very slow in speed, which affects the overall speed of the algorithm; CNN needs to separately A generated candidate region performs a feature extraction, and there are a large number of repeated operations, which restricts the performance of the algorithm.

2.2. SPP-Net

In response to the R-CNN issue, He Yuming, who once proposed ResNet, proposed SPP-Net[2]. The algorithm clips and scales the candidate region before convolution feature extraction with CNN by adding space between the convolutional layer and the fully connected layer of the network to make the input image size of the CNN. Consistent. Spatial pyramid pooling solves the problem of inconsistent input candidate region size, but more importantly, it reduces the repetitive calculation in R-CNN, greatly improving the speed and performance of the algorithm. The disadvantage of SPP-Net is that after the processing of the spatial pyramid layer, although the input sizes of CNN are the same, the perceptual field of view of the candidate frame is also very large, so that the convolutional neural network cannot update the model weight effectively during training.

2.3. Fast R-CNN

In response to the problem of SPP-Net, in 2015, Microsoft Research conducted an effective improvement on the R-CNN algorithm based on the spatial pyramid layer of SPP-Net.

The structure of the Fast R-CNN[3] is shown in the figure 2. The improvement of Fast R-CNN is that a pooling layer structure of ROI Pooling is designed, which effectively solves the problem that the R-CNN algorithm must crop and scale the image area to the same size. A multitasking loss function is proposed. Each ROI has two output vectors: a

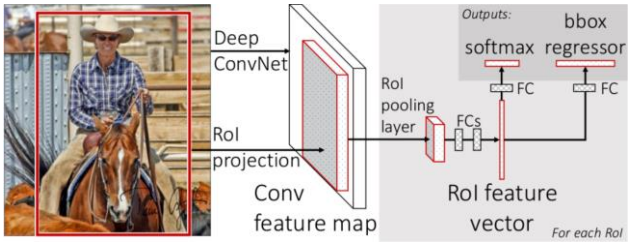

Figure2

softmax probability output vector and a bounding box regression position vector for each class.

Although Fast R-CNN draws on the idea of SPP-Net, there is still no improvement in the candidate box generation method for R-CNN's selective search, which makes the Fast R-CNN still have a lot of room for improvement.

2.4. Faster R-CNN

In order to solve the problem of candidate box generation left over from R-CNN, several authors of the R-CNN series proposed the Faster R-CNN[4] method. The key of Faster R-CNN is to design a region candidate network of RPN (Region Proposal Network), and the selection and judgment of candidate frames are handed over to RPN for processing, and the candidate regions after RPN processing are classified and classified based on multi-task loss. The advantage of Faster R-CNN is that the feature information extracted by CNN can share the weight of the whole network, which solves the problem of slow speed caused by a large number of candidate frames. However, because the RPN network can generate multi-size candidate frames in a fixed-size convolution feature map, the variable target size and the fixed receptive field are inconsistent.

2.5. Mask R-CNN

In 2017, He Yuming continued to improve the R-CNN algorithm based on the previous one and proposed the Mask R-CNN algorithm. The overall architecture of Mask R-CNN[5] is shown in the figure 3.

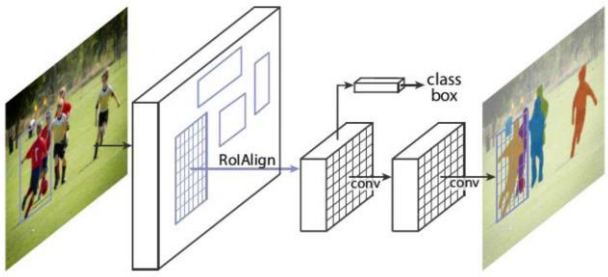

Figure 3

Mask R-CNN upgrades the ROI Pooling layer of Fast R-CNN to ROI Align layer, and adds branch FCN layer, i.e.

mask layer, for semantic mask recognition, and generates target candidates through RPN network. The box then classifies and predicts the frame for each target candidate frame, while predicting the segmentation for each target candidate frame using a full convolutional network. Mask R-CNN is essentially an instance segmentation algorithm. Compared to Semantic Segmentation, instance segmentation has a more subtle segmentation for similar objects.

Even though the two-stage target detection algorithm is constantly evolving and the detection accuracy is getting higher and higher, there are always speed bottlenecks in the two stages. In some real-time scenarios of target detection requirements, the R-CNN series of algorithms is ultimately lacking. Therefore, a one-stage algorithm has emerged, which is represented by the yolo algorithm series, which demonstrates the real-time target detection effect of an end-to-end deep learning system. The main idea of the yolo algorithm series is to get the category and specific location of the target object directly from the input image, and no longer generate candidate regions like the R-CNN series. The direct effect of doing this is fast.

3. One-stage Algorithm

3.1. yolo v1

The core idea of the yolo v1 [6] algorithm is to use the entire image as the input of the network, and output the target object's category and the specific position coordinates of the bounding box directly on the output layer of the network. Yolo v1 divides the input image into S*S grids. Each grid predicts two bounding boxes. If the target object falls into the corresponding mesh, the mesh is responsible for detecting the target object. The paper can understand the three steps of the yolo v1 algorithm: scaling the image, running the convolutional network, and non-maximum suppression. Although yolo v1 is fast, the disadvantages are obvious: since a grid can only predict two bounding boxes, this makes yolo v1 not very good for detecting very small objects, often with large deviations in positioning, in addition to yolo V1 also has problems such as weak generalization performance.

3.2. SSD

In view of the inaccurate positioning of yolo v1, the SSD algorithm (Single shot Multibox Detector)[7] proposed at the end of 2016 is to combine yolo's bounding box regression method with Faster R-CNN's anchor box mechanism, through different convolutions. The target object area is predicted on the feature map of the layer, and the output is discretized multi-scale and multi-scale bounding box coordinates with different aspect ratios. These improvements enable the SSD to guarantee the accuracy of the inspection when inputting images with lower resolution. This also makes the SSD detection accuracy exceed the previous yolo v1.

3.3. yolo v2/yolo 9000

In view of the inaccurate positioning problem of yolo v1, yolo v2[8] focuses on the solution to this problem: using Darknet-19 as a pre-training network, adding the BN (Batch Normalization) layer, and proposing a new training method - joint A training algorithm that blends the two data sets together. Objects are categorized using a hierarchical view, and a large number of categorical dataset data is used to augment the test dataset to mix two different datasets. On the other hand, compared to yolo v1, the bounding box coordinates are directly predicted by the fully connected layer. yolo v2 draws on the anchor boxes in the R-CNN. Using the Anchor Box will slightly reduce the accuracy, but it can be used to make YOLO Can predict more than one thousand boxes, while recall reaches 88%, mAP reaches 69.2%.

3.4. yolo v3

In order to achieve higher positioning accuracy while maintaining speed, yolo v3[9] uses a more complex network structure. Compared to previous networks, yolo v3's improvements include multi-scale prediction (FPN), more complex network structure Darknet53, and cancellation of softmax as candidate frame classification, which makes yolo v3 faster and more accurate. improve. Unlike R-CNNs that require thousands of single-target images, yolo v3 predicts through a single network assessment. This makes YOLOv3 very fast, under the same conditions it is 1000 times faster than R-CNN and 100 times faster than Fast R-CNN. The speed comparison between yolo v3 and each algorithm is shown in Figure 4.

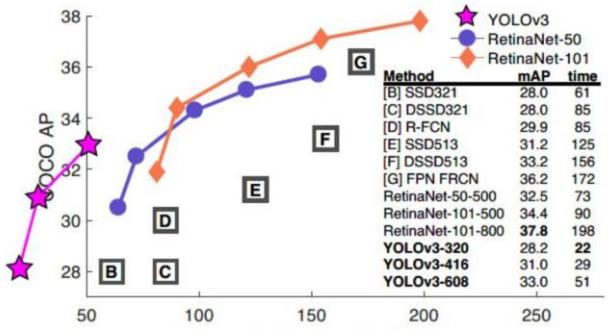

Figure 4

4. Conclusion

In the previous article, we discussed the development of the target detection algorithm in computer vision. The

algorithms of the two genres are all moving forward in an orderly manner. At the same time, the two schools are also learning and promoting each other. With the deepening of research in related fields such as deep learning, the target detection algorithm will become more and more perfect and efficient. But no matter which target detection algorithm is used, we need to find a balance between efficiency and precision before the algorithm does not fully meet people's needs.

References

[1] Girshick, R., et al. "Rich feature hierarchies for object detection and semantic segmentation." 2014 IEEE Conference on Computer Vision and Pattern Recognition IEEE, 2014.A. Alpher. Frobnication. Journal of Foo, 12(1):234–778, 2002.

[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. (2015). Fast r-cnn. Computer Science.

[4] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. 2015.

[5] Kaiming H, Georgia G, Piotr D, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018:1-1.

[6] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015.

[7] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 2015.

[8] Redmon J, Farhadi A. [IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Honolulu, HI (2017.7.21-2017.7.26)] 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - YOLO9000: Better, Faster, Stronger[J]. 2017:6517-6525.

[9] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. 2018.