Research on Object Detection Algorithm based on Deep Learning

Xin Zhang

School of North China Electric Power University, Baoding 017000, China.

Abstract

With the development of artificial intelligence, object detection as one of the important research directions, object detection technology based on deep learning has also made rapid development. Now object detection technology is applied in various fields, such as intelligent medical, face recognition, intelligent monitoring, automatic driving and so on. This paper is mainly based on the two-stage object dection algorithm represented by RCNN series and the one-stage object dection algorithm represented by Yolo series, and summarizes their development, improvement and shortcomings. Finally, this paper summarizes and prospects the two series of object detection algorithms.

Keywords

Object Dection; Deep Learning; RCNN; YOLO.

1. Introduction

Before the introduction of deep learning, the traditional object detection algorithm is mainly divided into three steps, which are candidate region extraction, feature extraction and classifier classification. In traditional object dection, the selection of candidate regions is generally based on sliding window. In the image, because the size of the target is different, it needs different size windows to slide on the image, resulting in too many candidate regions and high computational complexity. Feature extraction is the most important part of object dection, which is especially important The quality of feature extraction will ultimately affect the results of object dection. The traditional feature extraction methods usually extract sift [1], hog[2], Haar[3] and other features. These features will change in diversity because of the shape of the target, the influence of lighting and background and other factors, and also need to manually select the suitable features of the target area, so this method is very important On the one hand, the extracted features are not robust, on the other hand, they are not universal. After feature extraction, it is sent to SVM [4] and AdaBoost [5] for classification.

Due to a series of problems existing in the traditional object dection technology, the development of object dection encountered a bottleneck. Until alexnet [6] achieved good results in Imagenet[7] classification challenge, researchers saw the advantages of convolutional neural network in feature extraction, which made object dection enter a new era, so the object dection method based on deep learning began to develop With the rapid development, many outstanding object dection algorithms have emerged in the later stage. Object detection algorithms based on deep learning are mainly divided into two categories. The first category is two-stage object detection algorithm, and the second category is one-stage object detection algorithm.

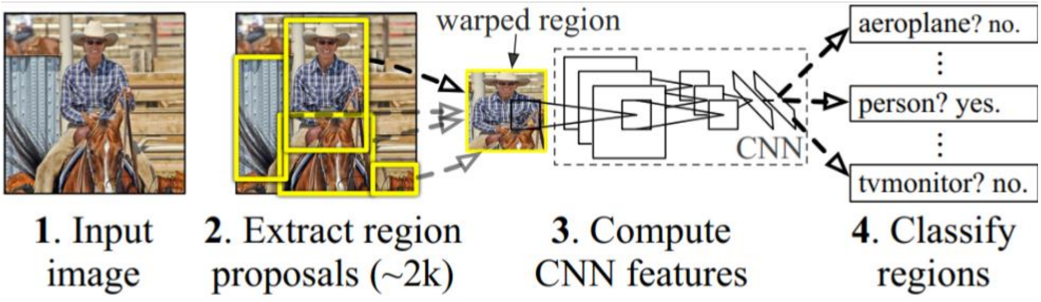

Figure 1. RCNN

2. Two stage object detection algorithm

2.1 RCNN

In 2014, girshick [8] and others proposed RCNN object detection algorithm. RCNN applied deep learning to object detection task for the first time, which is a pioneering work of object detection based on deep learning. The network structure of RCNN is shown in Figure 1

The whole operation process of RCNN is mainly divided into four steps. The first step is to select the candidate regions of the input image. The second step is to extract the features of these candidate regions by using convolution neural network. The third step is to classify the features extracted in the second step into the SVM classifier. Finally, the regression operation is carried out for the detected candidate boxes The specific operation process is as follows.

Rcnn uses the selective search algorithm to find the most likely sub region of the target: first, it uses a simple region division algorithm to divide the image into many small regions to generate a region set, and then aggregates the adjacent small regions according to the similarity and size of the adjacent regions until the region set is empty. The idea of selective search algorithm is similar to clustering, which embodies a relatively exhaustive idea. Finally, about 2000 candidate regions are obtained, which improves the efficiency compared with the candidate regions obtained by sliding window.

The feature extraction network in RCNN uses alexnet, and the size of the input image must be 227 × 227, so the candidate region needs to be cropped or warp before it is input into the network, and then it is sent into the network for feature extraction after it meets the input conditions.

The RCNN uses SVM two classifiers, and uses as many two classifiers as it needs to judge the number of categories. For the final output of RCNN, the 2000 * 4096 dimension feature is multiplied by the weight matrix 4096 * 20 composed of 20 SVM classifiers to get the 2000 * 20 dimension matrix, which indicates that each suggestion box is the score of an object category. For each category of 20 columns, non maximum suppression is performed to eliminate the overlapped suggestion box, and the suggestion box with the highest score in each category is obtained.

After many detection frames pass NMS, although the candidate frame with the highest score in each category has been obtained, there are still some errors with the real frame. Therefore, the frame regression technology is used to further correct the obtained detection frame to make it closer to the real frame and improve its detection accuracy.

2.2 Fast RCNN

Although RCNN has a great improvement compared with the traditional object detection algorithm, there are still some problems. Firstly, for more than 2000 candidate frames extracted by selectsearch method, all of them need to be sent to CNN for feature extraction. For these 2000 candidate regions, there must be overlapping parts, so it will involve the process of repeated calculation. Secondly, for the proposed algorithm The candidate regions need to be processed with uniform size, and then input into neural network for feature extraction. In RCNN, the image is processed by the method of crop / wrap, which will result in the loss of some regions and the distortion of the image.

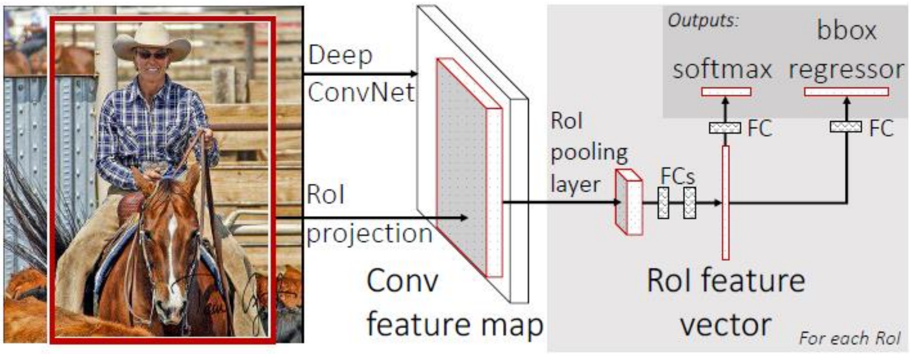

Figure 2. Fast RCNN

Aiming at a series of problems existing in RCNN, Ross B. girshick and others adopted spp net [9] method to improve RCNN on the basis of RCNN network structure model in 2015, and proposed fast RCNN [10] network structure model. The overall block diagram of fast RCNN is shown in Figure 2.

Compared with RCNN, sppnet has two improvements. Firstly, it reduces the computational complexity of convolution. Secondly, it solves the problem of image distortion caused by image deformation in RCNN.

RCNN is to make the candidate region image get a fixed size after the candidate region is cropped / wraped, and then input each processed candidate region image into the convolutional neural network for feature extraction. Sppnet is to input the whole image into the convolutional neural network to get the feature map of the whole image, Then more than 2000 candidate regions are directly mapped to the feature map, and the feature vectors of more than 2000 candidate regions are obtained, which greatly reduces the operation of CNN in feature extraction of candidate regions and reduces the computational complexity. After the candidate regions are cropped / wrapped by RCNN, the size of the feature vectors of each candidate region obtained by CNN is the same, Then the following classification and regression operations are carried out. However, the size of the feature vectors obtained by feature mapping in sppnet is not fixed. Sppnet inputs these different size feature vectors into spp (spatial pyramid transformation layer) to get the fixed size feature vectors, and then carries out the subsequent classification and regression.

Based on the idea of sppnet, fast RCNN proposes ROI pooling to output fixed size eigenvectors according to different size eigenvectors. ROI pooling obtains fixed size eigenvectors from input eigenvectors by maximizing pooling operation.

Fast RCNN has three main improvements: 1. Convolution is no longer for each region proposal, but directly for the whole image, which reduces a lot of repeated calculation. The original RCNN is to convolute each region proposal separately, because there are about 2000 region proposals in an image, and the overlapping rate between regions is very high, so repeated calculation is generated. 2. ROI pooling is used for feature size transformation, because the input of full connection layer requires the same size, so region proposal cannot be used as the input directly. 3. Each class corresponds to a regulator, and softmax is used to replace the original SVM classifier

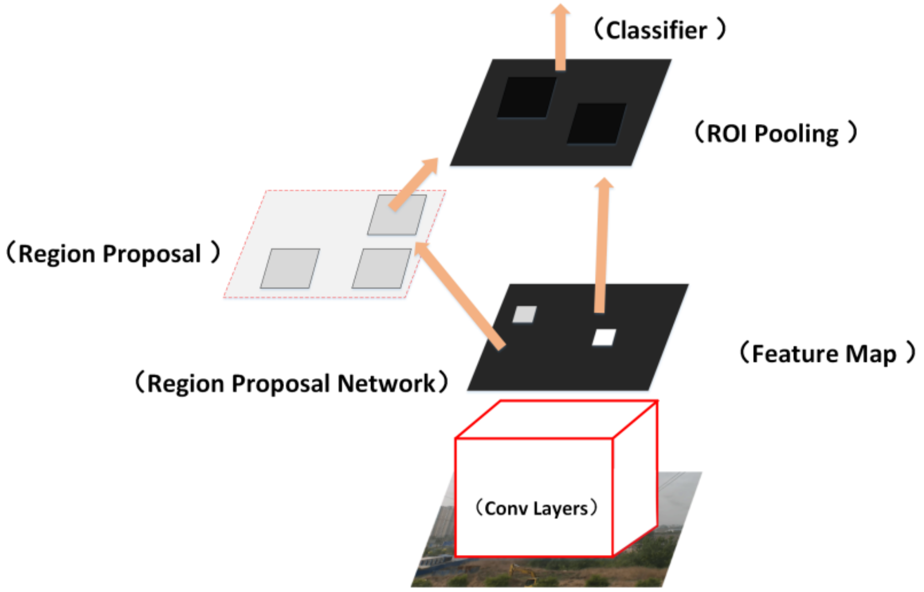

Figure 3. Faster RCNN

2.3 Faster RCNN

Fast RCNN [11] proposes a ROI pooling, then integrates the whole model, trains several modules including CNN, spp transform layer, classifier and bbox regression. It can be said that fast r-cnn algorithm realizes end-to-end training of neural network in a certain sense, but it is not the end-to-end training in the real sense, because select seal is used in candidate region selection, This is handled separately. Fast RCNN introduces an RPN network to replace select search to generate candidate regions. In the real sense, it combines four steps, namely, region recommendation generation, feature extraction, classification and border regression to train together. The network structure of fast RCNN is shown in Figure 3.

The core of the whole innovation of Fast RCNN is the introduction of RPN region generation network. RPN and object detection network share convolution network, so that the selection of candidate regions and subsequent object detection are carried out in the same network, and the original select is no longer used Search method is used to select candidate regions, which reduces computational complexity and improves execution efficiency. Fast RCNN proposes anchor mechanism, which uses nine anchors composed of three areas (128,256,512) and (1:1,1:2,2:1) to classify and regress borders. The composition of RPN is shown in Figure 4.

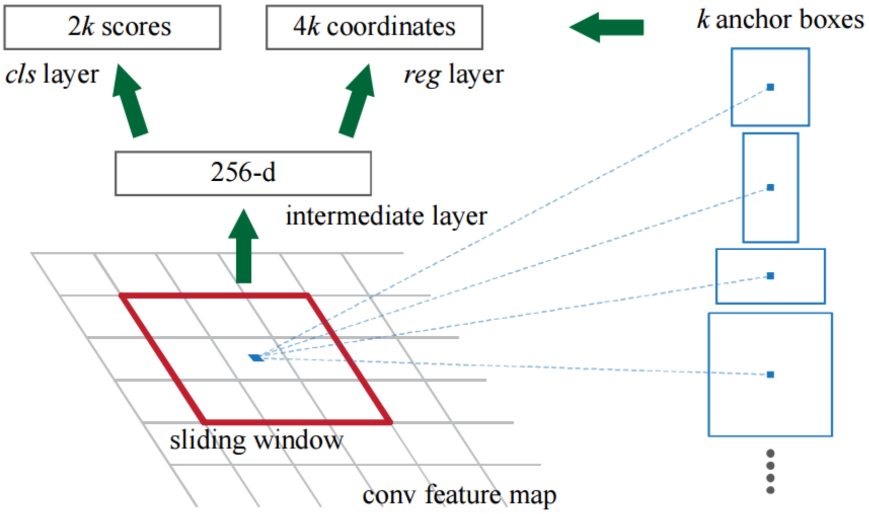

Figure 4. RPN

The input image is obtained by convolution vgg-16 network, and then input into RPN network. RPN network processes the extracted convolution feature graph. RPN is a fully convolutional neural network. Firstly, 256channel is used, The convolution layer of 3 × 3kernel is followed by two parallel 1 × 1 convolution layers. In the two parallel 1 × 1 convolutions, the left side classifies the convolution. The classification here is whether there is a target, and does not care what the target is, but for distinguishing the background and the foreground. Since each anchor corresponds to K candidate boxes and each candidate box takes two values, which means whether there is or not probability of passing, the output corresponding to each anchor should be a 2K dimension vector. Therefore, 2K channels are used for classification convolution on the left side and convolution network for obtaining border position information on the right side, Since each anchor corresponds to K candidate boxes, each candidate box has 4 position values (x, y, W, H), so the output corresponding to each anchor should be a 4K dimension vector, so the convolution on the right uses 4K channels. When the anchor box with target is translated and scaled by four position regression values, a large number of candidate boxes can be generated. At this time, some candidate boxes with higher prediction score can be screened by using non maximum suppression as the final region positions.

3. One stage object detection algorithm

3.1 Yolov1

Yolov1 [12] directly treats the problem of object detection as regression, and directly predicts the physical category and location information. Compared with the candidate region selection of RCNN series, the final classification and regression are completed based on the candidate region, which greatly reduces the complexity and improves the detection speed.

Yolov1 divides the image into 7 × 7 grid shape first. For the input image, after convolution neural network, a 7 × 7 × 30 feature map is obtained. Each pixel point in the feature map corresponds to each area in the 7 × 7 grid of the original image. Each region is responsible for predicting two boundary frames, and finally 7 × 7 × 2 Detection boxes will be obtained. If the center of a target falls into the region, the region will be responsible for detecting the target. Each bounding box is responsible for predicting five values, x, y, W, h, and confidence s. Where (x, y) indicates the predicted deviation of the bounding box from the upper left corner of the cell where the target center is located, (W, H) represents the width and height of the bounding box relative to the whole image, and whether the boundary box represented by s-confidence contains the target and a judgment on its prediction accuracy. The formula for confidence s is shown in 3-1, where Pr (object) represents the possibility of the target in the box, Pr (object) is 1 if there is an object in the box, and 0 if there is no target in the box. Each cell also produces C conditional probabilities Pr (class| object) and only one group of class probabilities are predicted on a grid cell. Regardless of the number of boundary boxes B, the non maximum suppression stage index is confidence score.

$$s_i = Pr(\,Object) * IOU_{pred}^{truth} \tag{1}$$

$$P_r(Class_i|Object) * P_r(Object) * Iou_{pred}^{truth} = P_r(Class_i) * Iou_{pred}^{truth} \tag{2}$$

The end-to-end training can be carried out through the loss function. Loss function design of Yolov1 algorithm is mainly composed of three parts.

$$Loss = E_{coord} + E_{IOU} + E_{class} \tag{3}$$

3.2 Yolov2

Although Yolov1 has greatly improved the detection speed compared with RCNN series, it also has many problems. Firstly, yolov1 preprocesses the images of input network through lossy way, which will lead to some picture distortion and affect the detection accuracy. Second, each box of yolov1 is only responsible for predicting two boundary boxes. For dense small target centers, there are same problems This situation in a box can not be handled well, and many goals will be missed, and some problems such as the like will be missed.

In order to solve some problems existing in yolov1 and to solve the error problems in recall rate and positioning accuracy of yolov1, Joseph Redmon et al. Proposed the yolov2 [13] detection algorithm in 2016, which can make it better balance between speed and accuracy.

Batch normalization is helpful to solve the problem of gradient disappearance and gradient explosion in the process of back propagation of neural network, and reduces the sensitivity to some super parameters. When each batch is normalized separately, it has a certain regularization effect, so it can obtain better convergence speed and convergence effect.

Therefore,yolov2 adopts this strategy which adds batch normalization after each build up layer instead of using drop.By using this method,the map of yolov2 improve by 2.4%.

Because there are many training samples for image classification, and the supervised training method is used in image detection, it is necessary to label the image in advance. However, the labor cost of labeling the bounding box is relatively high, so fewer samples are labeled for training object detection. Therefore, at that time, yolov1 used the samples of image classification to train the convolution layer, At that time, the classification samples on imaget used by yolov1 input 224 * 224 images, and then train the convolution layer. In the training object detection, a higher resolution image 448 * 448 was used as the input. In this way, the model could not adapt well, thus affecting the detection effect.

Yolov2 is aimed at this problem. Firstly, it uses low-resolution image to train the convolution layer, then uses high-resolution image training convolution layer to fine tune the network model, and then uses high-resolution image as input for detection and training. Through this crossover mode, the model can adapt to the change of image resolution. In the end, yolov2map increased by 3.7 percent in this way.

Yolov2 uses the idea of Fast RCNN for reference and uses anchor boxes. Yolov2 abandons the method of using full connection layer to predict the boundary box in yolov1, but uses anchor boxes and convolution layer to predict the boundary box. Compared with the need to manually set the size and aspect ratio of the anchor frame in fast RCNN, yolov2 uses the K-means clustering method to automatically generate the size of the anchor frame, and uses the IOU distance function to replace the Euclidean distance function, as shown in the figure. Considering the complexity and recall of the model, the number of anchor frames is 5.

3.3 Yolov3

Yolov3 [14] further optimizes the detection accuracy of Yolo series and the shortcomings of small object detection. First of all, it uses the change of basic network to replace darknet-19 with darknet-53. In addition, it uses the idea of FPN to use feature fusion for object dection, and uses logic to replace softmax to complete the target classification.

In darknet-53, pooling layer and full connection layer are canceled. In the process of forward propagation, the change of tensor size is realized by changing the step size of convolution kernel. In yolov2, the image is reduced by 1 / 32 of the final image after five times. In yolov3, the image is also scaled five times to 1 / 32 of the original image.The whole network model is composed of five Darknet modules, each module has the same structure. The combination of 1 * 1 and 3 * 3 convolution layers is used as the residual unit of the network. The convolution network part in darknet-53 is composed of kernel + BN + leakyrelu as the standard component. The standard component and residual unit are shown in the figure below.

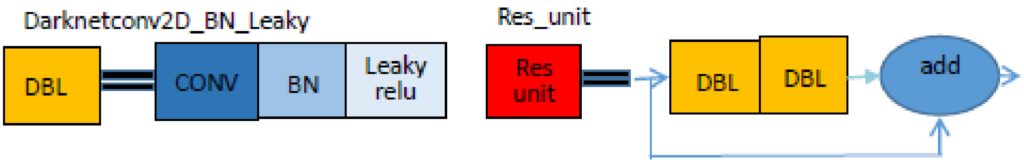

Figure 5. Standard components and residual units

Using the idea of FPN for reference, yolov3 fuses the three smallest size feature maps in the network, which are 13, 26 and 52 respectively, to complete the object dection, so as to improve the detection accuracy of small targets. Because we know that the deeper the network level is, the richer the semantic information is, which plays a great role in the detection of large targets. However, with the deepening of the network level, the geometric features of the image will be less and less, which is not conducive to the detection of small targets. On the contrary, the shallow network does not have higher semantic information, but it contains more The resolution is relatively high. Therefore, if the two are combined to do the work of object dection, the detection of small targets will be greatly improved. Therefore, three different scale feature maps are used in yolov3 After 5 times of down sampling, the 13 * 13 size feature map is obtained, while the 26 * 26 and 52 * 52 size feature maps are retained. The semantic information of 13 * 13 feature map is relatively rich, which is used to detect large targets. By up sampling the 13 * 13 feature map and fusing the 26 * 26 size feature map, the medium target is detected, and then the small target is detected by fusing the 52 * 52 size feature map. The way of feature fusion is shown in the figure 6.

Yolov3 continues to use kmean method to generate nine anchors, which are evenly distributed to three scales for detection, that is, each scale has three anchor frames for detection. Therefore, three bounding boxes are predicted for each mesh in each feature map. Each bounding box predicts (x, y, W, h, confidence) five basic parameters, and then there are 80 category probabilities.The output of logistic is used instead of softmax. In this way, multi label objects can be supported.

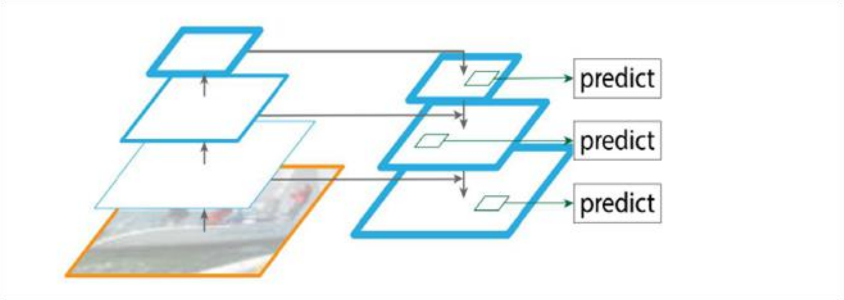

Figure 6. Principle of FPN network detection

4. Summary

This paper systematically describes the popular one-stage object detection algorithm Yolo series and two-stage object dection algorithm RCNN Series in recent years. The speed of one-stage object dection algorithm is better than that of two-stage object dection algorithm, and the accuracy of two-stage object dection algorithm is higher than that of one-stage object dection algorithm. Compared with the traditional detection algorithm, the object dection algorithm based on deep learning has been greatly improved in accuracy and real-time performance. However, due to the complexity and variability of the real scene, there are still many problems. How to reduce the impact of complex background on object dection and how to reduce the accuracy degradation caused by the change of target size and shape has become a research hotspot in the field of object detection.

References

[1] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[2] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection [C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE, 2005.

[3] Lienhart R, Maydt J. An extended set of Haar-like features for rapid object detection [C]// Proceedings. International Conference on Image Processing. IEEE, 2002.

[4] Joachims T. Making Large-Scale SVM Learning Practical[J]. Technical Reports, 1998, 8(3):499-526.

[5] R tsch G, Onoda T, Müller K R. Soft margins for AdaBoost[J]. Machine learning, 2001, 42(3): 287-320.

[6] RussakovskyO, Deng J, Su H, et al. Imagenet large scale visual recognitionchallenge[J]. International Journal of Computer Vision, 2014:1-42.

[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. doi:10.1145/ 3065386.

[8] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[J]. arXiv preprint arXiv:1311.2524,2013.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[C]. arXiv:1406.4729,2014

[10] Ross Girshick. Fast R-CNN[C]. arXiv:1504.08083, 2015.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal.

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, et al. You Only Look Once:Unified, Real-Time Object Detection[C]. arXiv:1506.02640, 2016.

[13] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[J]. 2016.

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