Journal of Physics:
Conference Series

Overview of two-stage object detection algorithms

Overview of two-stage object detection algorithms

Lixuan Du, Rongyu Zhang, Xiaotian Wang

International School, Beijing University of Post and Telecommunication, BEIJING, 100876 CHINA.

Abstract. Nowadays, object detection has gradually become a quite popular field. From the traditional methods to the methods used at this stage, object detection technology has made great progress, and is still continuously developing and innovating. This paper reviews two-stage object detection algorithms used at this stage, explaining in detail the working principles of Faster R-CNN, R-FCN, FPN, and Casecade R-CNN and analyzing the similarities and differences between these four two-stage object detection algorithms. Then we used HSRC2016 ship dataset to perform experiments with Faster R-CNN, R-FCN, FPN, and Casecade R-CNN and compared the effectiveness of them with experimental results.

1. Introduction

AI has gradually been practiced in many life applications, such as: self-driving cars, predictive analysis application software, and face recognition. As one of the important directions of artificial intelligence, object detection attract many people to the field. The current mainstream object detection algorithms mainly consist of two major classes, one-stage object detection algorithms and two-stage object detection algorithms, both of which are based on deep learning methods [1]. The main distinction between these two methods is whether to generate a region proposal. One-stage object detection algorithms don't need to generate a region proposal. It can directly obtain the classification accuracy of the object and its coordinate position. Two-stage object detection algorithms need to generate a region proposal before classifying and positioning them. Generally speaking, one-stage algorithms have speed advantage, and two-stage algorithms have advantages in accuracy. In this paper, we select two-stage object detection algorithms for overview, which have higher accuracy. We used the images in the horizontal box of part2 of the HSRC2016 ship dataset to test the performance of the Faster R-CNN, R-FCN, FPN, and Casecade R-CNN two-stage object detection algorithms, then analyzed and compared the effectiveness of each algorithm. Finally, we came to the conclusion: Faster R-CNN algorithm has the lowest effectiveness and Casecade R-CNN algorithm has the highest effectiveness. Although R-FCN and FPN algorithms are less effective than Casecade R-CNN, they still have an improvement over Faster R-CNN.

2. Background

2.1. Traditional object detection methods

Traditional object detection methods usually consist of three stages. The first stage is to frame candidate regions on a specified image in the data set and locate the object. Initially, the method of traversing the entire image using a sliding window is used to frame. The second stage is to extract features of these candidate regions. The main features are SIFT [2], HOG [3], etc. The advantages and disadvantages of feature extraction will directly affect the results of the next stage; the third stage is to use the trained classifier for classification. The main classifiers are SVM, AdaBoost [4], etc.

The traditional object detection algorithms include HOG + Cascade detection algorithm, HOG + SVM detection algorithm [3], and DPM algorithm [5].

The characteristic of HOG + SVM detection algorithm is that for objects of the same class, with the increase of training samples, classification accuracy and recall rate of the algorithm

both increase to a certain extent

Compared with HOG + Cascade, HOG + SVM has the following advantages and disadvantages:

Methods	Detection effect	Detection time
HOG+SVM	Good (high detection rate, low false detection rate)	Long (average 500ms, not suitable for real-time applications)
HOG+Cascade	Poor (poor generalization ability, resulting in a low detection rate; related to the training XML, training for specific scenarios can achieve better results)	Short (average 50ms, for real-time applications)

The DPM algorithm is more robust to pedestrian pose changes, so it has higher detection accuracy in complex scenes. However, due to the large amount of calculation, the DPM algorithm has a slow matching speed and poor real-time performance.

2.2. Object detection methods used at this stage

The object detection methods used at this stage mainly include one-stage object detection algorithms and two-stage object detection algorithms. Because the one-stage algorithm does not need to generate a region proposal, its algorithm flow is more concise and clearer. Generally, it is divided into two main lines, one for testing and the other for training. Testing directly converts the input image into output and forms a corresponding detection frame through decoding; Training encodes the labeled data (ground truth) to make it consistent with the corresponding output format of CNN, so as to calculate the corresponding loss. Common one-stage algorithms are YOLO [6], SSD [7], CornerNet [8], and so on.

The YOLO algorithm reduces the rate of missed detection and false detection, improves positioning accuracy, and the speed. The detection performance of SSD is relatively better, and it has two advantages of real-time and high accuracy. However, SSD is more suitable for detecting large objects, and its detection performance for small targets is poor. The CornerNet detection algorithm cleverly converts the detection frame into key points, that is, an object frame can be represented by two points, then when predicting, you can directly predict the key points of the two classes, and then combine the key points.

The two-stage algorithm as a more accurate algorithm is also relatively more complicated. It mainly includes two stages. In the first stage, preliminary tests are performed first, all positive samples are screened out, and regions of interest (RoIs) are generated; In the second stage, the algorithm performs regional classification and location refinement on the RoIs generated in the previous stage. Common two-stage algorithms include Faster R-CNN, R-FCN, FPN, etc.

Faster R-CNN algorithm proposes Region Proposal Network (RPN) [9] based on Fast R-CNN, and integrates the two into a complete network that can learn end-to-end, which not only guarantees accuracy but also Increased speed. However, because the RoI pooling layer in faster R-CNN is located in the middle of the entire network, calculations of each RoI inside every convolution layer behind the RoI pooling layer cannot share parameters. So, the detection speed of Faster R-CNN is still not very fast and cannot meet the needs of real-time tasks. R-FCN algorithm uses a full convolutional network structure. Calculations of each RoI inside every convolution layer can share parameters. So, it has high detection accuracy and very fast speed. Taking into account, the detection effect of R-FCN method is better than Faster R-CNN and far better than Fast R-CNN method. FPN algorithm is a method that uses conventional CNN models to efficiently extract features of various dimensions in a picture, and is an enhancement of the traditional CNN network to express and output picture information.

3. Overview of two-stage object detection algorithms

3.1. Faster R-CNN algorithm

Faster R-CNN can be considered as a system combining RPN [9] and Fast R-CNN. Compared with Fast R-CNN, Faster R-CNN replaces the Selective Search method [10] in the original algorithm with RPN. Its network structure diagram is as follows:

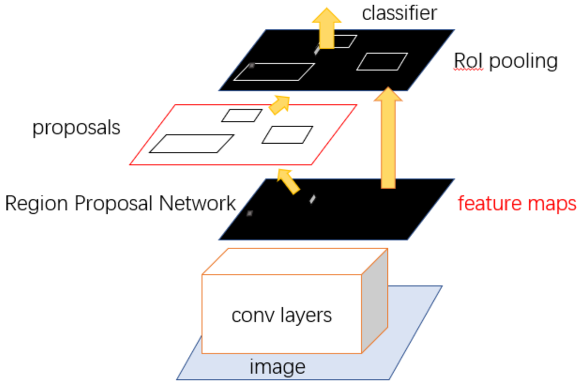

Figure 1 Network structure diagram of Faster R-CNN

Faster R-CNN is mainly divided into the following parts in structure:

(1) Convolution layer: It can extract the feature map of the entire picture.

(2) Region Proposal Network (RPN): As the core part of Faster R-CNN, RPN supersedes Selective Search method in Fast R-CNN algorithm to generate a region proposal. Using RPN to obtain a region proposal can more quickly and efficiently use the CNN network. RPN generates anchors while generating a region proposal. The judgement function determines whether the anchors are foreground or background, and then adjusts the anchors through border regression to obtain an accurate region proposal.

(3) RoI pooling: It can deal with the problem that different sizes of feature maps input to the network with fully connected layer. The fixed size is obtained by up-sampling.

(4) Classification layer and regression layer: The classification layer is responsible for judging which class an object belongs to; The regression layer fine adjusts the positions of regions of interest (RoIs) to obtain the final object detection result.

3.2. R-FCN algorithm

R-FCN algorithm is based on full convolutional networks. It proposes the concept of Position-Sensitive RoI polling base on the Faster R-CNN, then uses position-sensitive score maps to solve the position sensitivity problem of object detection. The core of R-FCN is to change the position of RoI pooling so that more convolution layers can implement parameters sharing when calculating, thereby reducing calculation time.

The impact of RoI pooling position is as follows:

(1) The closer RoI pooling position is to the Input, the deeper RoI-Wise detection subnet of the network, and the higher detection accuracy of the algorithm. However, because calculations of each RoI inside every convolution layer behind the RoI pooling layer cannot share parameters, the algorithm require more calculation and it takes longer to detect.

(2) The closer RoI pooling position is to the output, the shallower detection subnet of the network, the smaller the calculation amount for each RoI, and the higher the efficiency. But it may cause lower detection accuracy of the algorithm.

R-FCN changes the RoI pooling position to the end of entire network, so that calculations of each RoI inside every convolution layer can share parameters. It significantly reduces the amount of calculation and improves detection efficiency. However, since inserting RoI pooling layer breaks the translation invariance of the original convolutional network, changing the RoI pooling position to the end of entire network may also cause the detection network to be insensitive to location. So, R-FCN applies position-sensitive score maps to solve this problem, which contain the location information. To achieve this, a special convolution layer is inserted into the detection network after all other convolution layers, which is responsible for outputting position-sensitive score maps. Then R-FCN completes RoI pooling, which is different from RoI pooling in Faster R-CNN, it is a kind of Position-Sensitive RoI pooling.

3.3. FPN algorithm

FPN is a method that uses conventional CNN models to efficiently extract features of various dimensions in pictures, and is an enhancement of traditional CNN networks to express and output picture information. The purpose of FPN is to make information of each dimension of the input picture better represented by the output features through improving the feature extraction method of CNN network. Therefore, in essence, combining the FPN structure with Faster R-CNN algorithm improves the feature extraction part of the original Faster R-CNN algorithm, so that more feature layers are obtained, and RoI pooling is also increased. FPN can effectively empower conventional CNN models to generate feature maps which are more expressive for the next stage (object detection or classification analysis). In essence, it is a method to enhance CNN feature expression in backbone network. Its three basic processes are bottom-up path, that is, different dimension feature generation from bottom to top; top-down path, that is, feature supplement and enhancement from top to bottom; Correlation expression between CNN network layer features and the features of each dimension of the final output. For each level of CNN network, FPN generates features that reflect the information of this CNN level, then it combines features of each level together to generate the final expression feature combination.

3.4. Casecade R-CNN algorithm

Casecade R-CNN applies a method of training detection models based on different IoU thresholds. Caecade R-CNN achieves the goal of continuously optimizing prediction results by cascading several detection networks together, but detection networks of Cascade R-CNN are trained on positive and negative samples determined based on different IoU thresholds, which is different from ordinary cascading. The difference is a highlight of the algorithm.

Most of the experiments of Cascade R-CNN are completed on COCO dataset and it works well. Cascade R-CNN is composed of a series of detection models. The output of the former detection model is the input of the latter. It can be seen that Cascade R-CNN adopts the training mode of stage by stage, and the IoU threshold value is rising. That is to say, the later the detection model, the higher the IoU threshold it uses.

Casecade RCNN is designed into this cascade structure for two reasons.

(1) The method of training detection models based on different IoU thresholds is used in Casecade R-CNN. Every detection model trained using this method has a large difference in detection effect of input proposals with different IoU values. So, we hope that the IoU threshold selected for training a detection model should be as close as possible to the IoU value of proposal inputted into this model, so as to improve detection effect of the model.

(2) For different thresholds, the output IoU is generally greater than the input IoU. Therefore, the output of the previous IoUs stage can be used as the input of the next stage, so that the IoU obtained is getting higher and higher. All in all, it is difficult to make a detection model trained on a train set defined by a specified IoU threshold achieve the best effect on the proposal input with a large IoU span. The Cascade R-CNN structure can make every detector of the stage focus on the detection of the IoU within a certain range, which makes the detection effect better.

4. evaluation

4.1. Dataset Introduction

In this experiment, the detected ships are extracted from the horizontal box of part 2 of the HSRC2016 ship dataset and includes 256 train images and 98 validation images.

Five types of ships were tested in this experiment: Warcraft, Arleigh Burke, Tarawa, Nimitz and Ticonderoga.

4.2. Model test results

In this experiment, we train the model based on Faster R-CNN, R-FCN, FPN, and Casecade R-CNN, and then apply the results to the validation set. The detection model is evaluated according to the mean average precision (mAP) obtained from different model tests. The higher mAP value, the better performance of the network model in detecting targets in the satellite images.

	mAP
Faster R-CNN	93.63%
R-FCN	94.20%
FPN	94.72%
Casecade R-CNN	95.32%

The experimental result shows that: among the four object detection algorithms, Faster R-CNN's network model has the smallest mAP value and the worst performance; R-FCN's network model is improved by 0.57% compared with Faster R-CNN because it uses a full convolutional network structure; FPN improves the feature extraction method of traditional CNN and thus improves the performance of models. Its mAP value is also increased by 1.09% compared to Faster R-CNN. Casecade R-CNN uses the method of training model based on different IoU thresholds, which greatly improves the performance of its network model. The map value is 1.69% higher than that of fast R-CNN, which is also the highest of the four algorithms.

5. conclusion

In this paper, we first introduce traditional object detection methods, such as HOG + Cascade detection algorithm, HOG + SVM detection algorithm [3]. The detection effect of HOG + SVM is better while its detection time is longer. Although HOG + Casecade requires shorter detection time, its detection effect is unsatisfactory. Later, this paper introduces one-stage object detection algorithms and two-stage object detection algorithms used at this stage, and focuses on explaining and analyzing various two-stage algorithms, including Faster R-CNN, R-FCN, FPN, and Casecade R-CNN. Finally, the train images and the validation images in the horizontal box of part 2 of the HSRC2016 ship dataset are used to train and test the network models of above four algorithms, and the results of experiment are as follows: The effectiveness of Faster R-CNN algorithm is the lowest, and R-FCN algorithm uses a full convolutional network structure, which improves the effectiveness compared with Faster R-CNN; FPN algorithm improves the effectiveness on the basis of Faster RCNN by improving the feature extraction method of the conventional CNN network; Casecade R-CNN algorithm applies the method of training network models based on different IoU thresholds, which is most effective.

This paper has completed the overview of two-stage object detection algorithms, and we will try to review one-stage object detection algorithms in the future. Moreover, we will further study object detection algorithm based on the Anchor-free method [11].

Reference

[1] Lecun Y, Bengio Y, Hinton G. Deep learning.[J]. 2015, 521(7553):436.

[2] Lowe, D.G. Object recognition from local scale-invariant features [A]. In: International Conference on Computer Vision[C], Corfu, Greece, 1999: 1150-1157.

[3] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893.

[4] Shiguang Shan, Peng Yang, Xilin Chen, 等. AdaBoost Gabor Fisher Classifier for Face Recognition[C]// Analysis and Modelling of Faces and Gestures, Second International

[5] Workshop, AMFG 2005, Beijing, China, October 16, 2005, Proceedings. Springer-Verlag, 2005.

[6] FELZENSZWALB P. MCALLESTER D. RAMANAN D. A discriminatively trained. multiscale. deformable part model [CJ // Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Re ogn ition. Anchorage.

[7] REDMON J, DIVVALA S, GIRSHICK R, et a1. You only look once: unified, real—time object detection[c] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA; IEEEComputer Society: 2016: 779—788.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C]//European Conference on Computer Vision, 2016: 21-37.

[9] Law H, Deng J. CornerNet: Detecting Objects as Paired Keypoints[J]. 2018.

[10] Ren, Shaoqing, He, Kaiming, Girshick, Ross, etc. 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.

[11] Kulkarni A, Callan J. Selective Search: Efficient and Effective Search of Large Textual Collections[J]. 2015, 33(4):1-33.

[12] Li Y, Wu Y, Yang G, et al. Anchor-Free Distributed Location Method in Wireless Sensor Network with Range Errors[J]. Sensor Letters, 2016, 14(8):794-799(6).