Research on algorithms based on object detection

Zeshun Zhang

School of Electronic Information Engineering, Langfang Normal University
Langfang, 065000, China

631401120208@mails.cqjtu.edu.cn

Abstract. Object detection technology is a hot research direction in computer vision field technology, which is broadly used in face recognition, vehicle navigation, aviation and other important fields, with broad development prospects. In recent years, object detection algorithms based on deep learning have also appeared as computer science technology has advanced quickly. Comparing modern object detecting algorithms to conventional ones, this algorithm has gradually highlighted the advantages of high precision and good real-time performance. This article reviews traditional object detection algorithms, and focuses on the HOG algorithm of traditional object detection, reviews deep learning-based two-stage and one-stage object identification systems and weighs the benefits and drawbacks of each. A summary and outlook for the future object detection algorithm development is also provided.

Keywords: object detection, traditional methods, two-stage, one-stage, YOLO.

1. Introduction

In computer vision and digital image processing, this is a well-liked direction, object detection is the foundation of many visual tasks and has been applied in various aspects of our lives, such as autonomous driving, facial recognition, industrial inspection, aerospace, etc. Therefore, the object detection algorithm is particularly important. Finding things in an image and determining which category they fall under is the basic goal of object detection. It has significant practical value since computer vision reduces the consumption of human capital. As a result, object detection has recently been a hot topic for theoretical and application study. It is a crucial area of study in the fields of image processing and computer vision as well as an essential component of sophisticated monitoring systems.

Currently, deep learning-based detection algorithms and conventional detection methods can be used for object detection, the latter of which can be further categorized into single-stage and two-stage detection algorithms. Although we have taught computers to locate and recognize objects in images using various methods over the past twenty years and have achieved pretty good results on large image datasets, there is still much work to be done to realize object detection at the level of human detection and recognition. This article first analyzes the traditional object detection method and its pros and cons, and then summarizes the object detection based on deep learning algorithm.

2. Traditional object detection

Before deep learning emerged, because of the scarcity of computational resources, data was usually subjected to complex feature design. Various machine learning algorithms were used for functional design of image processing and then trained. Traditional object detection methods mainly consist of three parts: region selection (sliding windows), feature extraction (SIFT, HOG, etc.), and classifiers (SVM, Adaboost, etc.). Its schematic diagram is shown in Figure 1.

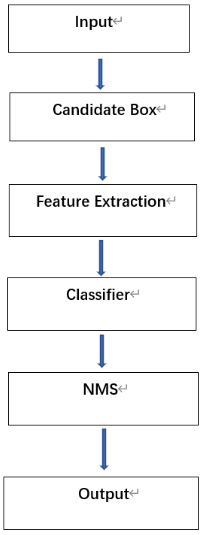

Figure 1. Traditional object detection process

Firstly, the input image is subjected to region selection, where the entire image is searched using methods such as sliding windows to select potential locations containing the object. Secondly, feature extraction is performed on the selected candidate regions for classification. This is done through methods such as SIFT and HOG. The extracted features are then inputted into classifiers such as LDA and SVM for training. Lastly, the findings are filtered using the NMS algorithm according to a predetermined threshold, producing the final detection results. However, traditional object detection has two major issues: The sliding window selection approach lacks focus and requires a lot of time, and the manually designed features are not robust. With the emergence of deep learning, significant advancements have been made in object detection.

The HOG detector is used for feature extraction in object detection. Compared to other detectors, the HOG is an improved version and is a significant improvement over the Sift and Shape Contexts. The density distribution of gradients or edges in an image can effectively explain the representation and geometry of a local object. In 1998, A trainable framework for object detection in static photos in cluttered settings was proposed by Papageorgiou et al., which learned Harr features from images. Such a classifier was used to detect faces in an image by computing the pixel sum of each detection window, taking their difference, and using SVM to learn the extracted features [1]. The advantage of this method is its high speed. In 2004, Lowe designed the SIFT for identifying and describing local features in images [2]. It searched for extremal points in space-scale and extracted their positions, scales, and rotation invariants for reliable matching of objects or scenes between different viewpoints. The SIFT algorithm can also address the issue of image registration and object identification and monitoring that is affected by various factors such as the object's state, the scene's surroundings, as well as the imaging capabilities of the apparatus. In 2013, Kosuke Mizuno and others implemented an object detection processor based on the gradient direction histogram (HOG) on an FPGA prototype board [3]. It made use of a parallelized module in a cell-based pipeline design, a reduced version of the HOG method, and simultaneous SVM calculation and cell scanning. The results showed that this architecture could generate HOG features and detect objects at a speed of 72 fps for SVGA resolution videos (800 × 600 pixels) at 40 MHz.

Compared to other feature description methods, HOG has many advantages. It can be used to convey local shape information since it represents the structural characteristics of edges (gradients). The influence of translation and rotation can be somewhat suppressed through the quantization of position and direction space. Furthermore, the technique for splitting an image into blocks and cells allows for a well-represented characterization of the relationships between local pixels in the image. Therefore, HOG features are particularly suitable for human body detection in images. However, it also has many drawbacks. First, a large amount of useless bounding boxes will be generated due to the massive search through the entire area using the sliding window approach, resulting in high computational complexity, lengthy description generation process, poor real-time performance and slow speed. Second, it is hard to handle occlusion problem, which is not favorable for practical applications.

3. Deep network

3.1. Two-stage detection methods

Deep learning-based object detection algorithms can be broadly categorized into two types: two-stage detection algorithms and one-stage detection algorithms. The two-stage detection algorithm implements detection in two main processes. The input image is used to create a number of regions in the first stage. The second stage is to create an object classifier by using the CNN to extract features from the created areas, and finally, classify and regress on the candidate regions. Therefore, the "two-stage" method is also called region proposal-based object detection. It is the pioneer of deep learning-based detection algorithms and representative algorithms include the RCNN series (RCNN, Fast RCNN, Faster RCNN) and SPPNet.

3.1.1. RCNN. RCNN was the pioneering object detection system that successfully employed deep learning. Its object detection technology is based on a combination of CNN, linear regression, and SVM. RCNN follows the traditional idea of object detection and also uses four steps: selecting a box, picture categorization, extracting features for each box, and object recognition using non-maximum suppression. The only difference is that in the step of feature extraction, deep convolutional networks are used instead of traditional features such as SIFT and HOG.

Classic object detection algorithms use sliding windows to iteratively check all possible regions. Here, only features on pre-selected candidate regions—those that are more likely to include objects—are extracted and evaluated, thereby greatly reducing computational complexity. Firstly, the selective search method is used to generate approximately 2000 candidate areas in the image. Secondly, each region is adjusted to a uniform size and generates a feature vector through a CNN model. Finally, a multi-class SVM classifier analyzes the feature vector to determine the probability that it belongs to a specific object category within that region. To achieve this, an SVM classifier is trained for each category. In addition, RCNN includes bounding box regression training to improve localization accuracy by correcting the precise position of each bounding box. Two databases are used during training: a large recognition database (ImageNet ILSVC 2012) to label the object category in each image. Ten million images, 1000 classes. A smaller detection database (PASCAL VOC 2007) to label the object category and position in each image. Ten thousand images, 20 classes. Pre-training is done using the recognition database, followed by fine-tuning the parameters using the detection database. Finally, evaluation is done on the detection database. In 2014, Girshick et al. proposed the R-CNN model based on AlexNet's research on image extraction [4]. It uses region proposal modules that are not related to classes together with CNN to transform detection tasks into classification and localization problems.

The algorithm mainly consists of three modules: (1) generating 1000~2000 new candidate regions; (2) the extracted candidate regions are cropped and fed into CNN to produce fixed-length feature vectors; and (3) using AlexNet as the backbone architecture of the detector, then transferring various features to the SVM classifier, which can judge the category and finely adjust the candidate region location through the regressor. Although the recognition framework of R-CNN is not very different from traditional methods, thanks to the excellent feature extraction ability of CNN, R-CNN's performance is much better than traditional methods. It also has significant improvements in mAP on VOC 2007, with R-CNN achieving a mAP of 58.5% compared to around 40% for traditional methods. However, R-CNN's drawback is that it has a large computation cost. The R-CNN process is more complicated, including the selection of region proposals. Since one image might provide more than 2000 area suggestions, the majority of which overlap, convolution is performed for each region proposal during the convolutional neural network training process. This results in numerous needless calculations. Therefore, the calculation cost based on region proposal convolution is too large, and this is also the main problem that Fast R-CNN solves later.

3.1.2. Fast R-CNN. In 2015, Ross Girshick proposed Fast R-CNN, which has a clever design, a more compact approach that significantly accelerates object detection [5]. Using the largest-scale network, Fast R-CNN trains 9.5 hours (84-hour reduction) and tests in 0.32 seconds (47-second reduction) compared to R-CNN. Similar to PASCAL VOC 2007, accuracy, about 66%-67%. This model proposes a single-stage training algorithm that combines SVM classification and BBox regression in the CNN stage, solving the multi-stage pipeline training problem existed in R-CNN and SPP-Net. Fast R-CNN solves the three problems of the R-CNN method: first, slow testing speed, with a large overlap between the candidate boxes in one image and redundant feature extraction operations. Second: slow training speed. During training, the candidate regions that were extracted from one image are fed into the network first, then the original image. It is not necessary to repeatedly calculate the top few layers of characteristics in these potential regions. Third, training needs a lot of storage space. R-CNN's independent classifier and regressor need a lot of different features to train on. Fast R-CNN eliminates the requirement for additional storage by combining category judgment and position fine-tuning using a deep network.

3.1.3. Faster R-CNN. Faster R-CNN was introduced shortly after Fast R-CNN to address the issue of previous detection algorithms relying on region proposal algorithms to generate candidate boxes [6]. The prerequisite for Fast R-CNN to achieve real-time detection speed is to ignore the time spent on region proposal, so the calculation of generating candidate box sets has always been the reason why the speed of two-stage detection algorithms cannot be significantly improved. Faster R-CNN uses the ZF model (Zeiler and Fergus model) and the VGG16 model (Simonyan and Zisserman model) as the backbone networks. Faster R-CNN achieves a speed of 5fps (including all steps) using the deep neural network model VGG16. The two-stage network is more accurate and is better able to handle the issue of multi-scale and small objects when compared to conventional one-stage detection networks. As altering the object class in the dataset can effectively change the test model, faster R-CNN performs well on a variety of datasets and is simple to transfer. There are still some problems, though; NMS is employed as a post-processing method based on classification scores to prevent overlapping candidate boxes when RPN generates suggestions. In fact, because the proposals of two objects may be filtered out, this strategy is not very accommodating to occluded objects, causing missed detections. Therefore, improving this NMS mode can improve detection performance. The original RoI pooling of Faster R-CNN causes loss of accuracy due to the two rounds of rounding. Thus, improving this localization pooling or feature scale output problem also requires improvement.

3.2. One-Stage

Detection Methods One-stage object detection algorithm is an image processing technique that recognizes, locates, and extracts object features, and classifies and locates objects with the focus on object areas in the image, ultimately achieving computer vision applications. Considering how the economy and technology have progressed, the newly developed one-stage object detection algorithm directly utilizes deep neural networks to detect objects in images. It can combine the object detection task into an end-to-end structure, eliminating the need for complex classification and positioning steps, greatly simplifying the calculation process. In the image detection process, the one-stage object detection algorithm can more completely capture the object areas because it can effectively detect complex boundaries and overlapping objects. In addition, the one-stage object detection algorithm can adaptively adjust the model to achieve fast convergence during training and effective object detection. Compared to traditional two-stage object detection algorithms, one-stage object detection algorithms can greatly simplify the calculation process, reduce parameter adjustment and thus reduce the consumption of computational resources, achieving higher detection efficiency. One-stage object detection algorithms can better capture object areas and achieve accurate object detection results by adaptively adjusting the model and achieving fast convergence during training.

3.2.1. YOLO series. In 2015, Joseph et al. proposed the YOLO algorithm, taking into account it as a regression-based detection issue. Without specifically extracting candidate boxes, it completes object location and classification from image input in a single network, and can predict the probability of object categories on a complete image. The implementation process is shown in Figure 2 [7].

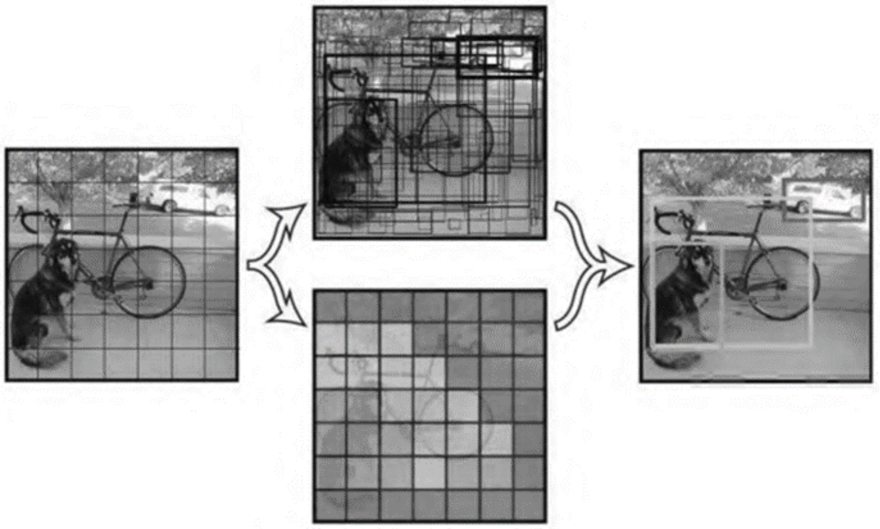

Figure 2. Implementation process of YOLOv1.

Despite the fact that the YOLOv1 object detection algorithm has the benefits of quick detection and simplicity, it also has many disadvantages. For example, during training, only one object can be predicted for each bounding box, which makes it difficult to detect multiple objects in one grid and limits the detection accuracy of the object. YOLOv2, proposed by Redmon et al. in 2017, mainly improved some of the shortcomings of YOLOv1, achieving significant improvements in accuracy and the number of object detections [8]. The algorithm uses the Darknet-19 network and a series of design decisions to improve speed and accuracy. YOLOv2 uses the DarkNet-19 network structure and abandons the GoogLeNet network structure used in YOLOv1. It provides a simple balance between speed and accuracy by improving YOLO, and is also known as YOLO9000. YOLOv3, an improved version of previous YOLO versions, uses a larger Darknet-53 network to replace the feature extraction network. ResNet residual network connections are added to the network to reduce negative gradient effects. It achieves simultaneous improvements in object detection speed and accuracy, making it more suitable for detecting small objects.

With the support of previous YOLO model versions, the YOLOv8 model runs faster and more accurately, it offers a consistent framework for building models that can carry out tasks like object detection, instance segmentation, and image classification [9]. YOLOv8 is also quite effective and adaptable, enabling a variety of export formats and running the model on both the CPU and GPU. There are five models in each category of the YOLOv8 model, working together to complete detection, segmentation, and classification tasks. Among them, YOLOv8 Nano is the fastest and smallest model, while YOLOv8Extra Large (YOLOv8x) is the most accurate but slowest model. YOLOv8 incorporates design features from previously proposed algorithms such as YOLOX, YOLOv6, YOLOv7, and PPYOLOE, reaching a new height in real-time detection.

3.2.2. SSD. SSD is a one-stage object detector that came before YOLO V2 and after YOLO V1. It is based on VGG-16 and adds auxiliary structures to improve performance. SSD uses multi-scale feature maps, as well as the idea of multi-scale feature maps in Darknet53 in YOLO V3. The smaller receptive fields of each cell on the shallow feature map are suitable for detecting smaller objects, while the larger receptive fields of each cell on the deeper feature map are suitable for detecting larger objects. During training, SSD matches each true box with the default box with the best Jaccard overlap and trains the network accordingly. SSD has many advantages, including predicting objects of different sizes at different feature scales, an end-to-end classification and regression implementation similar to VGG16, and true real-time performance [10]. For a network with an input size of 300x300, it achieves 74.3%mAP and 59FPS on the VOC 2007 test set using Nvidia Titan X, and for a network with a size of 512x512, it achieves 76.9%mAP.

4. Conclusion

This article systematically introduces the structural principles and advantages and disadvantages of traditional object detection and deep learning-based single-stage and two-stage object detection algorithms. Single-stage algorithms stand out with the advantage of detection speed and have become the most popular object detection algorithm for a time. However, there are still some difficulties and challenges for object detection based on deep neural network learning. For example, detecting small objects is limited because these objects contain fewer fine-grained features and with the development of network training, it is simple to lose detail features. Modern object detection systems generally require both speed and accuracy, as well as real-time performance. The field of object detection has grown rapidly due to the popularity of deep learning and is sure to bring about a transformation in the market.

References

[1] Papageorgiou C P, Oren M, Poggio T. General framework for object detection, 1998, Comput. Vis., 18.

[2] Lowe DG. Distinctive image features from scale-invariant key points. 2004, Inter. J. of Compu. Vis. 60(2): 91–110.

[3] Kosuke Mizuno,Yosuke Terachi,Kenta. An FPGA Implementation of a HOG-based Object Detection Processor. 2013, Trans. Sys. LSI Des.,6(0).1097–1105.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation.2016 IEEE Conf. Compu. Vis. Patt. Rec. 779-788

[5] Girshick R. Fast R-CNN.2015, Conf. Compu. Vis. 1440–1448.

[6] Ren SQ, He KM, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. 2015, Inter. Conf. Neu. Infor. Proc. Sys., 91–99.

[7] Joseph Redmon, Divvala Santosh, Girshick Ross, et al. You Only Look Once: Unified, Real-Time Object Detection, 2016, IEEE Conf. Compu. Vis. Patt. Rec.: 179-188.

[8] Joseph Redmon, Farhadi Ali. YOLO9000: Better, Faster, Stronger, 2017 IEEE Conf. Comput. Vis. Patt. Recog., 6517-6525

[9] J Joseph Redmon, Farhadi Ali Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018

[10] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multiBox detector. 2016 Euro. Conf. Comput.Vis. 21–37.