Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/104/20241062

Overview of object detection based on deep learning

Dantong Zhang

University of Birmingham Joint Institute at Jinan University, Jinan University, Guangzhou, China

dxz247@student.bham.ac.uk

Abstract. In recent years, computer vision technology has developed rapidly, as one of its important research directions, object detection has received widespread attention due to its high accuracy. Meanwhile, object detection has many application fields, such as intelligent transportation, medical and health, security systems, etc. Traditional object detection methods have limitations when applied to complex real-world scenarios. To improve the shortcomings of conventional methods, deep learning based object detection algorithms have significantly improved the efficiency of object detection and become a research hotspot in object detection. This article summarizes the algorithms into two-stage and one-stage object detection algorithms based on the technical processes and structural differences in handling object detection tasks. Firstly, several common two-stage and one-stage object detection algorithms and their applications in real-world scenarios are introduced. Then their data sets and algorithm performance are analyzed and compared. Finally, according to the results of the comparison, the existing problems of the two-stage object detection algorithm and the one-stage object detection method are discussed, and their future development directions are pointed out.

Keywords: object detection, deep learning, two-stage object detection algorithms, one-stage object detection algorithms.

1. Introduction

With the rapid development of computer vision, the efficiency of deep learning is also constantly improving. As an important branch of computer vision, Object detection has wide applications in human-computer interaction, security monitoring, biometric recognition, and other fields. Object detection is used to detect pre-defined input objects, and its main categories include traditional object detection and object detection based on deep learning convolutional neural networks. Traditional object detection mainly relies on manually designed features and machine learning algorithms, such as the Histogram of Oriented Gradients (HOG) algorithm proposed by Navneet Dalal et al. in 2005 [1], which achieves object detection and localization in images through three core steps: first, feature extraction is performed, and manually designed feature extractors are used to extract useful information for detection from the original image; second, classifier design is carried out using algorithms such as Support Vector Machine (SVM) and AdaBoost to learn the mapping relationship between features and target categories; finally, the detection method generates candidate regions through strategies such as sliding windows, selective search, or EdgeBoxes, and combines classifier judgment to accurately locate and label the target position [2]. Overall, traditional object detection methods have low feature extraction efficiency,

DOI: 10.54254/2755-2721/104/20241062

insufficient feature generalization ability, and weak robustness, which increases the difficulty of porting between different applications and datasets, while also increasing complexity and maintenance costs.

With the advancement of deep learning, AlexNet significantly improved image recognition performance through the ReLU activation function in 2012, and object detection algorithms based on deep learning have become a research hotspot in this field [3]. Subsequently, Karen Simonyan et al. proposed VGGNet in 2014, which constructs deep networks by stacking small convolutional kernels, demonstrating the importance of network depth in improving recognition accuracy [4].

Based on the technical process and structural differences in handling object detection tasks, deep learning-based object detection methods can be classified into two categories: two-stage and one-stage object detection algorithms. The two-stage object detection algorithm decomposes the detection task into two parts: region proposal and region classification. Representative algorithms include Region-CNN (R-CNN), Faster Regions with CNN features (faster R-CNN), and Feature Pyramid Network (FPN), which perform well in the accuracy of detection tasks. One-stage object detection algorithms directly use object categories and positions as regression tasks, with typical representatives including You Only Look Once (YOLO), Single Shot MultiBox Detector (SSD), CornerNet, and ExtremeNet algorithms.

Regarding the current research status mentioned above, this article first introduces several common two-stage object detection algorithms and one-stage object detection algorithms, summarizes the applications of the two object detection algorithms in various fields, then summarizes the selection and performance of the two methods on datasets and compares them. Finally, based on the results, the existing problems and future research directions of the two algorithms are summarized.

2. Two-stage object detection algorithm

2.1. R-CNN series

R-CNN is one of the earliest methods to apply CNN to object detection. It first uses methods such as Selective Search to generate Region Proposals in the image, then performs CNN feature extraction on each candidate region, and uses Support Vector Machines (SVM) for classification. Finally, it fine-tunes the target position through bounding box regression.

2.2. Fast R-CNN

Fast R-CNN has been improved based on R-CNN, mainly solving the problem of repeated feature extraction in R-CNN. It introduces the Region of Interest (RoI) pooling layer, which allows the entire image to undergo only one feature extraction. Then, each candidate region is subjected to RoI pooling to obtain a fixed-size feature map, followed by classification and bounding box regression.

The innovation of this algorithm lies in the design of a RoI pooling layer to solve the feature extraction problem of candidate regions of different sizes and proportions, enabling Fast R-CNN to handle input images and candidate regions of different sizes. The end-to-end training process from feature extraction to object detection and classification has been achieved, simplifying the training process and improving the overall performance of the model [2].

2.3. Faster R-CNN

Faster R-CNN further improves the generation of candidate regions by introducing a Region Proposal Network (RPN), which is one of the core innovations of Faster R-CNN. It uses CNN to slide windows on feature maps, generating multiple candidate boxes of different scales and aspect ratios for each position, and calculating object scores for each candidate box. Shared Convolutional Features: Faster R-CNN combines the feature extraction parts of RPN and Fast R-CNN into a shared convolutional neural network [5].

DOI: 10.54254/2755-2721/104/20241062

2.4. FPN

The workflow of FPN is as follows. Firstly, the multi-scale problem in object detection is solved by designing the structure of the feature pyramid. Secondly, by combining bottom-up feature extraction and top-down feature fusion, bottom-up and top-down feature fusion, as well as lateral connections, are performed to fuse the rich semantic information of high-level feature maps with the high-resolution information of low-level feature maps. This fusion method of multi-scale features enables the network to simultaneously utilize the advantages of high-level and low-level features, improving the accuracy and robustness of detection [6].

2.5. Algorithm application

The two-stage object detection algorithm has been widely used in multiple fields due to its high detection accuracy and stability. The following are some main application scenarios:

(1) In the field of security, two-stage object detection algorithms can be used for tasks such as facial recognition, vehicle detection, and behavior analysis. By capturing video images through surveillance cameras, algorithms can recognize targets such as faces and vehicles in real-time and issue warnings for abnormal behavior. This technology is of great significance for maintaining public safety and preventing crime.

(2) In the field of industrial quality inspection, two-stage object detection algorithms can be used for tasks such as product defect detection and part recognition. By collecting and analyzing images of products on the production line, algorithms can automatically detect surface defects or identify specific parts, thereby improving production efficiency and product quality.

(3) In intelligent transportation systems, two-stage object detection algorithms can be used for tasks such as traffic sign recognition and vehicle tracking. By analyzing road images, algorithms can recognize traffic signs, vehicles, and other targets in real-time, and provide important data support for traffic management.

3. One-stage object detection algorithms

3.1. YOLO series

To address the real-time limitations of two-stage detectors, the Facebook AI Research Institute, in collaboration with the Allen Institute for Artificial Intelligence and a research team from the University of Washington, proposed the first one-stage object detection YOLO model in the field of deep learning to achieve real-time performance at the 2016 International Conference on Computer Vision and Pattern Recognition [7]. This model transforms the object detection problem into a regression problem and achieves object classification and localization through a single forward propagation.

In 2016, Joseph Redmon and his team pioneered YOLO v1, which revolutionized object detection tasks into direct regression problems, achieving real-time prediction of bounding box coordinates and class probabilities. Although the speed is significant, there is still room for improvement in detecting small objects. Subsequently, YOLO v2 emerged, which not only accelerated the detection process but also significantly improved the accuracy of detection by integrating batch normalization, optimizing classifier resolution, and introducing anchor boxes. YOLO v3 further improves the network structure by using DarkNet-53 as the backbone network and cleverly integrating FPN, achieving precise capture of cross-scale targets and significantly enhancing the ability to recognize small targets. YOLO v4 introduces the Best Practice Set (BoS) and Best Feature Set (BoF) strategies, which optimize the network architecture and training process while maintaining detection speed, promoting a dual leap in detection accuracy and generalization ability [8].

YOLO v5 has shifted its development environment to the Pytorch platform for the first time, which greatly accelerates the inference process with the help of efficient tools such as AutoAnchor. YOLO v6, developed by the Meituan Visual AI team, focuses on the innovation of quantitative technology. Through in-depth analysis of PTQ and QAT, it explores efficient quantitative paths and opens up new avenues for model deployment and performance optimization. YOLO v7 introduces the E-ELAN

DOI: 10.54254/2755-2721/104/20241062

network architecture and a cascaded model scaling strategy for the first time, which finely adjusts the depth and width of the model to maintain optimal structural balance. The YOLO v10 developed by Tsinghua University is the latest iteration of the YOLO series, a dual allocation strategy was proposed to address the issue of nonmaximum suppression of post-processing dependencies hindering end-to-end deployment of YOLO, setting a new milestone for the YOLO series in terms of real-time performance and end-to-end deployment [9].

3.2. SSD algorithm

The SSD algorithm combines the regression idea of YOLO and the anchor mechanism of Faster R-CNN to predict the bounding box of the target by setting anchor points of different scales and aspect ratios on feature maps of different scales. The SSD algorithm is a one-stage object detection algorithm proposed by Wei Liu et al. in ECCV 2016 [10]. It sets default boxes with different aspect ratios and scale ratios at each position of the feature map and then predicts each default box to obtain the final detection result. The SSD object detection model significantly improves performance through three key innovations: 1) fusing multi-layer feature maps to enhance small object detection. 2) Presetting multi-scale prior boxes enhances robustness in complex scenes. 3) Directly predicting in the convolutional layer simplifies the process and reduces computational complexity [10].

3.3. Algorithm application

The application of one-stage object detection algorithms is very extensive, covering multiple fields and industries. Here are some main application scenarios:

(1) In the field of autonomous driving, one-stage object detection algorithms are used to identify key targets in the driving environment of vehicles. Through the accurate detection of these targets, the auto-drive system can obtain road information in real-time, to make correct driving decisions, this application greatly improves the safety and reliability of the auto-drive system.

(2) In the field of smart homes, one-stage object detection algorithms are used to achieve intelligent recognition and control of smart home devices. For example, by combining cameras and one-stage object detection algorithms, it is possible to recognize and analyze the behavior of family members, thereby automatically adjusting the home environment to meet the needs of different family members.

(3) In the field of drones, one-stage object detection algorithms are used to achieve autonomous flight and intelligent monitoring of drones. By real-time detection and recognition of target objects, drones can autonomously adjust their flight attitude and altitude to track targets or take photos.

4. Comparison of dataset and algorithm performance

4.1. Dataset selection and comparison

4.1.1. Dataset selection. The dataset plays an important role as a benchmark in model training and evaluation. Training models on diverse and widely covered image datasets can help them gain stronger abilities in tasks such as image recognition, classification, and segmentation. In the field of object detection, annotating images, including target location and category information, helps models accurately identify and locate various objects in the image. Some typical datasets in object detection tasks, including Pascal VOC, ImageNet dataset, Google Open Image dataset, MSCOCO dataset, and DOTA dataset, provide rich annotation information in different scenarios.

4.1.2. Dataset Comparison. When comparing the one-stage object detection algorithms YOLO, SSD, and CornerNet with the two-stage object detection algorithms for dataset selection, it is possible to analyze them by comparing the bias of dataset selection and the specific datasets applicable. By analyzing Table 1, the following conclusions can be drawn:

(1) The MS COCO dataset, as a widely used dataset in the field of computer vision, is suitable for the above four situations and is an important dataset in object detection learning under deep learning.

DOI: 10.54254/2755-2721/104/20241062

(2) All four algorithms exhibit different requirements for scale and diversity when selecting datasets. YOLO tends to choose datasets with better real-time performance evaluation, such as PASCAL VOC and MS COCO, which demonstrates its emphasis on fast processing capabilities. In contrast, SSD and two-stage object detection algorithms emphasize more on the richness and diversity of the dataset, such as MS COCO and ImageNet, which help improve the model's generalization ability in different scenarios and categories. CornerNet focuses on datasets with precise pixel-level annotations to meet its requirements for keypoint detection.

(3) The accuracy of annotation has a significant impact on the performance of object detection methods. CornerNet has the highest demand for precise pixel-level annotation, which reflects that in keypoint detection, the accuracy of annotation directly determines the training effect and detection accuracy of the model. In contrast, although YOLO and SSD also rely on high-quality labeling, their performance may be slightly less sensitive to labeling accuracy. The two-stage object detection algorithms partially alleviate the impact of annotation accuracy on final performance through multi-stage processing.

Table 1. Comparison of datasets with the same object detection algorithm

	Specific considerations for dataset selection	Example dataset	Significant features
YOLO	Tend to choose datasets with better real-time performance evaluation	PASCAL VOC, MS COCO	Simplify the detection process and optimize the network structure
SSD	Tend to choose datasets with rich target categories and diverse scenarios	MS COCO, PASCAL VOC	Using multi-scale feature maps
CornerNet	Tend to choose datasets with precise pixel-level annotations	MS COCO	Perform keypoint detection
two-stage object detection	Tend to choose datasets that contain complex scenes and diverse targets	MS COCO, ImageNet, PASCAL VOC	Staged processing

4.2. Comparison of algorithm performance

A complete evaluation system has been established in the field of object detection to comprehensively measure the performance of algorithms, including backbone network, frames per second (FPS), mean average precision (mAP), etc. This article compares and analyzes the object detection algorithm by selecting the above indicators. The following conclusions can be drawn from Table 2:

The application of one-stage object detection algorithms is very extensive, covering multiple fields and industries. Here are some main application scenarios:

(1) The significant improvement in FPS from YOLOV to YOLOV3 is mainly due to the optimization of network structure, the improvement of computational efficiency, and more efficient feature extraction methods. The improvement of mAP reflects the progress of the model in detection accuracy, mainly due to the application of more complex network structures, finer anchor box mechanisms, multi-scale detection strategies, and data augmentation techniques. The YOLO series has gradually improved detection accuracy by continuously optimizing these aspects.

(2) Early versions of YOLO used VGG16 as the base network, but with the introduction of DarkNet, the model reduced computational complexity and parameter count while maintaining high performance. DarkNet's design is more compact and efficient, which helps to improve FPS.

(3) The performance of the SSD series on VOC datasets is relatively stable, mainly due to its fixed network structure and detection process. However, due to its relatively simple network structure and fewer feature pyramid layers, SSD may be slightly inadequate in handling complex scenes. Compared to the YOLO series, the SSD series has a lower FPS, which may be related to its network architecture design. SSD sacrifices a certain processing speed while pursuing detection accuracy.

DOI: 10.54254/2755-2721/104/20241062

(4) The two-stage detector achieves high detection accuracy by first generating candidate regions and then performing fine classification and position regression. However, this step-by-step processing approach also increases computational complexity, resulting in lower FPS. The two-stage detector typically requires more computing resources to support its complex processing flow, which to some extent limits its application in scenarios with high real-time requirements.

(5) The YOLO series has gradually improved detection accuracy by continuously optimizing these aspects. Single-stage detectors represented by the YOLO series and SSD series have achieved high processing speed through a one-step detection method. However, while pursuing speed, its detection accuracy may be affected to some extent. Single-stage detectors have been widely used in real-time video surveillance and autonomous driving scenarios due to their high FPS and relatively good mAP performance.

Table 2. Performance comparison of object detection algorithms

	Backbone network	*FPS(framesper second)*	Stage division	mAP(%)		
				VOC2007	VOC2012	COCO(mAP@[0.50,0.95])
YOLO[11]	VGG-16	45.0	one	63.4	57.9	-
YOLOv2[11]	DarkNet-19	40.0	one	78.6	73.4	21.6
YOLOv3[11]	DarkNet-53	51.0	one	-	-	33.0
SSD[2]	VGG-16	19.3	one	79.8	78.5	28.8
R-SSD[2]	ResNet	35.0	one	78.5	80.8	-
	VGGG-16	16.6		80.8	-	-
DSSD321[2]	ResNet-101	9.5	one	78.6	76.3	33.2
F-SSD300[2]	VGGNet	65.8	one	82.7	82.0	27.1
R-CNN[2]	Alex Net	0.03	two	58.5	-	-
	VGG-16	0.50		66.0	53.3	-
Fast R-CNN[2]	VGG-16	7.0	two	70.0	68.4	19.7
Faster R-CNN[2]	ResNet-101	5.0	two	76.4	73.8	34.9

5. Conclusion

Object detection technology based on deep learning has made significant progress and has become a hot research topic in the field of computer vision. This article provides an overview of two-stage and one-stage object detection algorithms and analyzes different algorithms. With the continuous advancement of deep learning, existing methods still have the following problems: 1) One-stage algorithms directly output detection results through one forward propagation, although faster, their detection accuracy is often slightly inferior compared to two-stage algorithms. 2) Due to the small number of pixels occupied by small targets in the image and the lack of rich feature information, one-stage algorithms are prone to missed or false detections when processing such targets, resulting in poor detection performance for small targets. 3) The number of objects of different categories in the dataset of one-stage detection algorithms may vary greatly, and single-stage algorithms may not be able to effectively learn the features of all categories when processing such imbalanced data, resulting in the problem of imbalanced detection categories. 4) The two-stage algorithm divides the detection task into two stages: region recommendation and classification regression, which improves detection accuracy but has high computational complexity, leading to relatively slow detection speed. 5) The quality of regional recommendations generated in the first stage directly affects the detection performance in the second stage. If the regional recommendations are inaccurate, it will lead to difficulties in subsequent classification and regression. 6) The two-stage algorithm usually requires a more complex network structure and more parameters, which increases the training difficulty and computational cost of the model.

DOI: 10.54254/2755-2721/104/20241062

In response to these issues, future research on object detection can be carried out from the following aspects: 1) Designing more refined network structures to optimize traditional structures, such as introducing Feature Pyramid Networks to enhance the detection ability of objects of different scales. 2) Adopting a more reasonable loss function, such as Focal Loss, to balance the weights of difficult and easy samples and improve the detection ability for small targets and difficult-to-distinguish samples. 3) By using data augmentation techniques to increase the number of small targets and difficult-to-distinguish samples, and using methods such as Online Hard Case Mining during the training process to balance the class distribution. 4) Adopt more efficient and accurate region proposal algorithms, such as RPN (Region Proposal Network), to generate high-quality region proposals. 5) On the premise of ensuring detection accuracy, design a lighter network structure to reduce the number of model parameters and computational complexity, and improve detection speed.6) The use of feature pyramids and other techniques to fuse features of different scales can improve the model's detection ability for targets of different scales, while also helping to increase detection speed.

References

[1] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 886-893.

[2] Guo, Q., Liu, N., & Wang, Z. (2023). A review of deep learning-based object detection algorithms. Journal of Detection and Control, 45(06), 10-20+26.

[3] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097-1105.

[4] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International CoRR, 2014,abs/1409. 1556.

[5] Ren, S., He, K., & Girshick, R. (2015). Faster r-cnn:Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 91-99.

[6] Xie, X., Cheng, C., & Yao, Y. (2022). Remote sensing image object detection using dynamic feature fusion. Journal of Computer Science, 45(4), 735-747.

[7] Renomn, J., Divvala, S., & Girshick, R. (2016). You only look once: Unified, real-time object detection. Pro- ceedings of the IEEE conference on computer vision and pattern recognition, 2016, 779-788.

[8] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M.(2020). YOLOv4:optimal speed and accuracy of object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 1544-52.

[9] Wang, A., Chen, H., & Liu, L. (2024). YOLOv10: Real-Time End-to-End Object Detection. arxiv preprint arxiv: 2405.14458.

[10] Liu, W., Anguelov, D., & Erhan, D. (2016). SSD: single shot multibox detector. Computer Vision-ECCV 2016:14th European Conference on Computer Vision, Amsterdam: Springer, 21-37.

[11] Zhou, J., & Wang, J. (2023). A review of YOLO object detection algorithm research. Journal of Changzhou Institute of Technology, 36 (01): 18-23+88.