Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Progress of Object Detection Based on Deep Learning

Aohua Zhang *

Department of Material Science and Technology, Beijing Forestry University, Beijing, China

* Corresponding Author Email: ynanaa@connect.ust.com

Abstract. Object detection is commonly utilized in fundamental computer vision research fields: medical, autonomous driving, sensing monitoring, and other fields. With the advancement of high-performance computing technology, the computing speed of hardware has made a significant advancement. Meanwhile, machine learning-related algorithms, especially deep learning-based algorithms, greatly boosted the detection speed and accuracy. In practical applications, engineers have little understanding of the principles, effects, and performance of these algorithms, which seriously hinders the industrial application of object detection. However, the advancement has fueled challenges and competition for better modules and industrialization. Accordingly, the main purpose of this study is to present the algorithms of object detection and summarize it from three aspects: data set, algorithm, and performance, so as to provide a reference for researchers in related fields. In terms of algorithms, the paper mainly introduces anchor-free, single-stage, and two-stage algorithms. Finally, the existing problems and future research directions for target detection are discussed.

Keywords: Deep Learning; 2D Object Detection; Computer Vision

1. Introduction

Object detection is a crucial subtask of a broad range of utilization: commercial, industrial, medical, self-driving, and other fields. The advancement of high-performance computing technology has contributed to breakthroughs in deep learning. Moreover, the end-to-end algorithms have significantly increased object detecting speed and accuracy. Therefore, object detection is increasingly an attractive field of study and it has brought a huge improvement from 20 years ago to now.

In the early stage of research, some academics created object detecting algorithms relying on features of geometry, texture, and key points. Since this method lacks effective feature extraction methods, it needs a hand-crafted feature design depending on the characteristics of the objects. Besides, hand-crafted methods still have limitations in possible scenarios and they have difficulties in detecting objects with varied illuminations, angles, and sizes accurately. However, optimizers and end-to-end methods have promoted a rapid improvement in object detection.

This study primarily introduces the progress of some algorithms and relevant statistics of the object detection field and gives a summary and reference for relevant research. The article structure is divided into three sections: datasets, algorithms, and performance. Concerning datasets, this part introduces PASCAL VOC, ILSVRC, and MS COCO datasets for testing and training modules. In the second part of algorithms, this paper mainly introduces the anchor-free, one-stage, and two-stage algorithms. Regarding performance, comparisons are made from two perspectives: parameters and average precision. Finally, the last section discusses the existing problems and future research directions of target detection.

2. Object Detection Datasets

Large datasets could provide high-quality data for training and testing detection models and some dataset challenges also accelerate algorithms to update and perform better. From 2005 to now, there have been some famous datasets built and released for promoting detection models, including the

dataset of ImageNet (e.g., ILSVRC 2012, ILSVRC 2013), PASCAL Visual Object Classes Challenges (e.g., PASCAL VOC 2007, PASCAL VOC 2012) and Microsoft COCO (e.g., COCO 2015, COCO 2017) [1]. The data of these widely used datasets is shown in Table 1, and it could be seen the development of quantities and varieties of the different dataset versions.

2.1. PASCAL VOC

There are three significant 2D challenging competitions and one of these collections is PASCAL VOC, which is held by the Pattern Analysis, Statical Modeling, and Computational Learning organization. It has two famous datasets: PASCAL VOC 2007 and PASCAL VOC 2012 for many challenging contents: classification, object detection, segmentation, and human layout and they have 20 classes of objects in 4 main classes: vehicles, household, animals, and person.

2.2. ILSVRC

The ImageNet is built to measure and compare the progress of computer vision. The dataset of ILSVRC contains over 14 million labeled images, and ILSVRC uses a sub-dataset for the challenge. Table 1 shows the change of emphasis of ILSVRC's versions.

2.3. MS COCO

The Microsoft COCO is developed to deal with challenges including identifying non-iconic views of items, contextual inference, and 2D localization. In contrast to PASCAL VOC, the COCO dataset has a larger image count and images have more complex backgrounds and a greater number of objects. And the MS COCO dataset has become a new test standard in object detection.

Table 1. Statistics of some well-known datasets.

Dataset	Train		Validation		Test
	Images	Objects	Images	Objects	Images
PASCAL VOC 2012	5717	13609	5823	13841	10991
ILSVRC 2012	1281167	1000	50000		100000
ILSVRC 2013	395909	345854	20121	55502	40152
COCO 2017	118287	860001	5000	36781	40670

3. Milestones of Deep Learning Based Object Detection Methods

Numerous noteworthy advancements have been made in recent decades. The key task of object detection is localization, and detecting location methods could be classified into two kinds depending on anchors. The anchor-free detector uses regression directly on objects, and the anchor-based detector uses anchors for regression. The anchor-based algorithms also have two categories: one-stage detector and two-stage detector that generates region proposals initially for altering detecting targets with anchors.

3.1. Anchor Free Detectors

3.1.1. CornerNet

Hei Law and Jia Deng proposed the CornerNet detector in 2019, which is a one-stage method for object detection without anchor boxes [2]. In contrast to detectors that overlap the ground truth boxes with an abundance of anchor boxes, the CornerNet is designed to use pairs of corners for predicting bounding boxes. Another new part of CornerNet is corner pooling, which is used for localizing corners by generating richer information on the boundaries of objects. For module architecture, the prediction models follows the hourglass network. Each module is followed by a corner pooling and generates sets of corresponding heatmaps, offsets, and embeddings.

3.1.2. CenterNet

A few months after CornerNet was released, X. Zhou et al. presented the CenterNet model, an end-to-end one-stage method [3]. As a more efficient model, the CenterNet only uses bounding boxes' center points to represent the object. For network architecture, the CenterNet uses two Hourglass Modules as its backbone structure and it has three outputs. By feeding images into this convolutional network, it could generate heatmaps, offsets, and height and width data. In the decoding process, this model uses the non-max-suppression (NMS) to find the maximum values in heatmaps as indexes of center points, and then it could find heights, weights, and offsets. Then in the encoding process, the CenterNet maps key points into areas of Gaussian Neighbors in heatmaps and it processes offsets, widths, and heights with similar mapping relations. The model uses the sum of focal loss, offset loss, and height and width loss for regression.

3.1.3. FCOS

Zhi Tian et al. introduced a detection system free of proposals and anchors, the FCOS (Fully Convolutional One-Stage), in 2019 [4]. In architecture, the FCOS uses CNN as the backbone and it trains every location for regression. Then the FCOS uses FPN for multiple layers prediction and it uses classification, Center-ness, and regression for final boxes. Center-ness is the main novel method used to treat those low-quality predicting bounding boxes, and the FCOS uses single-layers paralleling with classification to generate locations of Center-ness, which could finally decrease the weights of these low-quality boxes and drop these boxes by NMS processing.

3.2. One-stage Detectors

3.2.1. OverFeat

OverFeat is proposed by Pierre Sermanet et al. in 2013, which is the first approach using AlexNet for recognition, localization, and detection [5]. Compared to AlexNet, the novel techniques used in OverFeat are three: multiscale images as inputs, offset pooling processing after the fifth convolutional layer, and sliding windows in fully connected layers. For localization, the OverFeat uses the regressor to replace the classifier and the regressor could produce 4 units to represent the coordinates of edges of bounding boxes. Finally, the OverFeat chooses a set of predictions with a higher confidence level and merges these to generate final bounding boxes.

3.2.2. YOLO

Joseph Redmon et al. presented the YOLO in 2015, a single-stage object detector [6]. YOLO simplifies the process of extracting region proposals and regresses whole images to generate classes, locations, and confidences. There are two fully connected layers and 24 convolutional layers in the YOLO network to generate 2 bounding boxes with 5 predictions in each: center point location in grid cells, width and height, and confidence. Profiting from its simple and efficient architecture, YOLO performs faster than two-stage detectors in that time, but also it has disadvantages in accuracy.

3.2.3. YOLOv2 and YOLO9000

In 2016, Joseph Redmon and Ali Farhadi introduced the updated version, YOLOv2. Then they utilized a strategy to train the YOLO9000 system on both the COCO detection and ImageNet classification datasets, which finally could detect about 9000 classes. As an updated version, YOLOv2 uses seven tricks for improvement, and the high-resolution classifier pre-trained on ImageNet promotes its detection performance. In the network, the YOLOv2 adds batch normalization in every convolutional layer and it adds the passthrough layer to get more fine-grained features. Then the YOLOv2 replaces fully connected layers with Anchor Boxes and uses k-means clustering to automatically generate priors. The YOLOv2 also uses offsets for direct location prediction and it resizes images after every 10 batches for multi-scale training. All these tricks have improved the performance of YOLOv2 in different degrees.

3.2.4. YOLOv3

In 2018, Joseph Redmon and Ali Farhadi proposed the third version of YOLO, YOLOv3. In YOLOv3, there are three new tricks for enhancement: using darknet-53 as the new backbone and using layer-wise addition; using FPN for prediction and still keeping the k-means clustering for prior boxes; using the binary logistic classification to replace SoftMax.

3.2.5. SSD

W. Liu et al. introduced the SSD algorithm in 2016 [7]. Compared to YOLOv1, SSD has 2 updates: using the prior box for default scales and aspect ratios and detecting with multi-scale feature maps. For prediction, the SSD uses various scales of feature maps to select anchor boxes, and then it chooses the best performance box for prediction. With these updates, the SSD has improved in detection speed and accuracy.

3.2.6. RetinaNet

T.-Y. Lin et al. introduced the RetinaNet model in 2018, which deals with a much more important problem: the reason why one-stage detectors have inferior performance in accuracy [8]. The paper states that two-stage detectors use prior boxes to filter, alter about 1000-2000 anchors, and fix the ratio of foreground and background to 1:3, while one-stage detectors have to deal with about 100000 samples. This phenomenon makes positive samples not trained well, and they designed the focal loss function to rebalance it.

3.3. Two-stage Detectors

3.3.1. R-CNN

R. Girshick et al. presented the Regions with CNN (R-CNN) in 2014, as a bridge between image classification and object detection tasks. R-CNN has 5 main parts of module design [9]. After inputting images, the selective search algorithm generates thousands of region proposals, and it follows 5 steps to produce different scales of region proposals: over-segmentation, calculating similarities of neighbors, combining these regions, calculating similarities between the merged regions and other neighbors, and repeating the combining and calculating. Then the R-CNN resizes region proposals to the same size and computes feature maps by training Convolutional Neural Networks, which mainly contains the supervised pre-training and fine-tuning. The R-CNN then uses SVM (Support Vector Machines) to classify regions and uses Non-Maximum Suppression for filtering boxes. Finally, the R-CNN uses bounding box regression to revise locations with ground truth boxes.

3.3.2. Fast R-CNN

Ross Girshick presented the updated version in 2015, Fast R-CNN [10]. The novel change is the Region of Interest Pooling layer (RoI) following 5 convolutional layers, which divides regions into H x W cells and uses max pooling on them to generate fixed-size cells. Also, the Fast R-CNN has two changes: using convolution for the whole image and using SoftMax to replace the SVM classifier.

3.3.3. Faster R-CNN

A few months after Fast R-CNN was released, S. Ren et al. proposed the Faster R-CNN [11]. The Faster R-CNN has two main features: submitting region proposals via the RPN (Region Proposal Network) and presenting the concept of anchors. For RPN, it relies on the CNN backbone's feature maps, and RPN computes offsets of bounding box regression of anchors and uses SoftMax to classify the positive and negative anchors. Then the R-CNN puts the region proposals produced by RPN into RoI Pooling for resizing and it uses a classifier to generate probabilities of classes and offsets of proposals. Benefiting from its efficient architecture, the Faster R-CNN could reach the near-real-time detection level.

3.3.4. FPN

T.-Y. Lin et al. introduced a prediction model, the Feature Pyramid Networks, in 2017 [12]. For most traditional convolutional models, classifiers could not use all the information during convolutional processing and they could cause high costs in RAM, losing small objects and mistakes in classifying because of less information on a large scale in the Pyramidal feature hierarchy. The FPN develops a combination of the bottom-up model and the top-down model by using lateral connections, which could generate high-level semantics of all scales for extracting features. In the network, the FPN has two pathways and lateral connections, then the FPN adds element by element to the subsampled results.

3.3.5. Mask R-CNN

K. He et al. presented a blended object detection model, Mask R-CNN, in 2018 [13]. The essential part of Mask R-CNN is using a sub-branch paralleling with classification and regression to predict segmentation masks. In the Mask R-CNN network framework, it uses RPN to generate region proposals, and then it uses RoIAlign to generate features for pixel-wise sigmoid. In RoIAlign, the Mask R-CNN uses RoI Pooling for first quantization, and then it extracts histograms and uses max pooling combining them, which has a large improvement in misalignment and keeps space locations.

4. Comparison of Detectors

4.1. Comparison of Parameters

In object detection design, parameters are an important part affecting the model performance. Parameters impact the space complexity of the model, which determines the usage of GPU memory when a model is training or running, and finally, they impact the performance evaluation. In network architecture, the parameters of convolutional layers are calculated by the formula: $(kernels * kernels) * c\text{hannel}_{inputs} * c\text{hannel}_{outputs}$, and the parameters of fully connected layers are calculated by the formula: $\text{weights}_{\text{in}} * w\text{eights}_{\text{out}}$. But the parameters could not just decide the performance and speed. For example, the YOLOv3-tiny has 8 million parameters and YOLOv5n has 1.8 million parameters, but YOLOv3-tiny is faster than the YOLOv5n. Table 2 shows the number of parameters of some detectors.

Table 2. Parameters of some well-known object detectors.

Algorithm Type	Algorithm	Backbone	Parameters(M)
Ancor Free	CornerNet	Hourglass-104	57.3
	CenterNet	Hourglass-104	52.38
	FCOS	ResNet-101-FPN	53.3
One-Stage	YOLOv3	Darknet-53	40.55
	RetinaNet	ResNet-101-FPN	35.60
	Faster R-CNN		60
	Mask R-CNN	ResNet-101-FPN	63.75

4.2. Comparison of Effects

For object detectors, it is needed to evaluate their performance on datasets and compare them with other models' advantages and disadvantages. For object detection models, there are three aspects of evaluation: classification accuracy with accuracy, precision, recall rate, AP (Average Precision) and mean AP; IoU (Intersection over Union) for localization accuracy; and FPS (Frames Per Second) for detecting speed. FPS usually is used for near-real-time detectors to compare their speed, and IoU usually is used in model design for prediction and calculation loss functions, such as the YOLO series. AP evaluation is first introduced in PASCAL VOC 2007, but it is calculated with the Interpolated AP method with fixed numbers of division parts. PASCAL VOC 2010 changed its AP calculating method

to use integral calculating of the area under the curve, which is a more accurate method. MS COCO still uses the Interpolated AP method, but it has been updated to use 100 points on the PR curve to improve accuracy. Table 3 shows the mAP performance of several object detectors on COCO test-dev, and Table 4 shows their mAP performance on PASCAL VOC 2007 [14]. More detectors have used the COCO dataset for training in recent years because the COCO dataset updates more data of images and classes and overfitting problems appear when models are trained on certain datasets.

Table 3. Comparison of mAP of some well-known object detectors on COCO test-dev.

Algorithm Type	Algorithm	Backbone	mAP
Anchor Free	CornerNet	Hourglass-104	43.2
	CenterNet	Hourglass-104	47.0
	FCOS	ResNet-101-FPN	41.5
One-Stage	YOLOv2	DarkNet-19	21.6
	YOLOv3 + Darknet-53		33.0
	SSD		28.8
	RetinaNet	ResNet-101-FPN	39.1
	RetinaNet	ResNeXt-101-FPN	40.8
Two-Stage	Fast R-CNN		19.7
	Mask R-CNN	ResNet-101-FPN	38.2

Table 4. Comparison of mAP of some well-known object detectors on PASCAL VOC 2007.

Algorithm	Backbone	mAP
CenterNet	DLA34	80.7%
YOLO		63.4%
YOLOv2	DarkNet-19	78.6%
SSD		81.6%
R-CNN		58.5%
Fast R-CNN		70.0%
Faster R-CNN		73.2%

5. Conclusion

This paper has reviewed datasets, some important detectors, technologies, and metrics and it shows the development and achievement made by deep learning-based detectors in the last 10 years. It also could be seen that techniques are not following a certain development way and their attempts are not always promoting the performance of all metrics and datasets. But the tries are the basis and origins of inspiring, merging, and combining techniques for development. Object detection does not only focus on aspects of accuracy and speed, because of its widely used applications, it also focuses on the following directions.

5.1. Generalization

For more applications, detectors should not only orientate certain datasets but also it should be trained for more real-world circumstances, such as self-driving. For instance, object detection on self-driving also should deal with various devices and data formats, such as voxel, mesh, point cloud, and pseudo point cloud. Besides, the generalization of models should also deal with the limitations of overfitting on certain datasets and the relation between robustness and accuracy.

5.2. End-to-end System

For industrialization, an end-to-end object detection system could have more advantages in speed and control. For example, the speed of the YOLO series makes them reach the near-real-time level of detection, but industrialization also needs a balance between accuracy and speed.

5.3. Robustness to Variation

Detectors should also deal with the varied appearance of certain classes. A certain object could appear in different instances and positions, and it could also show in different conditions of view point, occlusion, and clutter. Detectors should have high robustness to these variations for high accuracy.

5.4. High Efficiency

It is needed for detectors to have high efficiency in time, memory, and storage. High efficiency is related to hardware, but algorithms also need to schedule the hardware efficiently to face several different object categories and their possible locations in the real world.

Considering how quickly computing technology is evolving, the evolution of algorithms and models could generate in months, and this paper is expected to provide some references for researchers in related fields

References

[1] Z. Zou, et al. Object detection in 20 years: A survey. Proceedings of the IEEE, 2023, 111(3):257-76.

[2] Law H, and Deng J. Cornernet: Detecting objects as paired keypoints. Proceedings of the European conference on computer vision (ECCV), 2018, 734-750.

[3] X. Zhou, et al. Objects as points. arXiv preprint, 2019, arXiv:1904.07850.

[4] Z. Tian, et al. FCOS: Fully convolutional one-stage object detection. arXiv preprint arXiv:1904.01355. 2019.

[5] Sermanet P, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint, 2013, arXiv:1312.6229

[6] J. Terven, et al. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. Machine Learning and Knowledge Extraction, 2023, 5(4), 1680-1716.

[7] W. Liu, et al. Ssd: Single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, 2016, Part I 14, 21-37.

[8] TY. Lin, et al. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2017, 2980-2988.

[9] Girshick, R., et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, 580-587.

[10] Girshick R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015, 1440-1448.

[11] S. Ren, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6):1137-49.

[12] TY. Lin, et al. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2117-2125.

[13] K. He, et al. Mask r-cnn. Proceedings of the IEEE international conference on computer vision, 2017, 2961-2969.

[14] W. Wang. Internimage: Exploring large-scale vision foundation models with deformable convolutions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 14408-14419.