Systematic Analysis of CNN and Its Optimization Algorithms Based on Object Detection

Shuoguo Zhang*

Beijing Lize International Academy, Beijing, 100020, China

*Corresponding author: Zsg20080708@outlook.com

Abstract. Object detection is a very important research field in computer vision. In recent years, because of the development of Convolutional Neural Networks (CNNs), object detection methods are also more and more advanced. From the earliest Region-based Convolutional Neural Network (R-CNN), researchers proposed a two-stage detection idea, to You Only Look Once (YOLO) single-stage detection, which greatly improved the speed of object detection. There are also Transformer models like the Detection Transformer (DETR), which no longer require the complex post-processing steps of the original model, making detection more straightforward. This paper mainly sorts out how these CNN-based detection methods are developed step by step, analyzes their advantages and disadvantages, such as how the algorithm model achieves a significant improvement in accuracy and efficiency through multi-scale feature fusion, the challenges in detecting small objects and realizing real-time detection, and looks forward to possible future research directions, providing important theoretical and practical guidance for further improving model performance and expanding real-world applications.

Keywords: CNN; Object Detection; R-CNN series; Faster R-CNN; YOLO series.

1. Introduction

Object detection occupies a key position in the field of computer vision research, and its core task is to identify the category of a specific object in the image and accurately determine the specific location of the object in the image. In recent years, deep learning technology has made rapid development, in this context, object detection algorithm based on CNN has achieved great progress. This development has completely reversed the research landscape in the field. From the original two-stage approach R-CNN, through Faster R-CNN and YOLO, to the transformer architecture, object detection technology has evolved in terms of the balance between accuracy and speed, as Arkin et al. pointed out [1]. The development process from CNN to Transformer has improved the detection efficiency and greatly improved the performance of multi-scale object recognition by relying on the global attention mechanism.

This paper comprehensively sorts out the development process of CNN-based object detection technology, and deeply analyzes the core innovation points and architecture evolution process in key literatures. From the two-stage framework of the R-CNN family, to the one-stage real-time detection of YOLO, to the Transformer fusion paradigm of DETR. This survey comprehensively evaluates the performance metrics, advantages and limitations of these methods, and discusses the challenges of current research and the future development trends of object detection technology. Based on this analysis, this review provides a comprehensive reference for researchers and helps the field move towards a more efficient direction.

2. Analysis of the theoretical basis

2.1. Definition principle and development history of CNN

CNN is a kind of deep learning model specially for processing grid-like topology data such as images. The core principle of the model is that with the help of convolutional layer, pooling layer and activation function structure, Layer by layer, visual features are extracted from low level to high level. The operational principle of CNN is shown in Fig. 1.

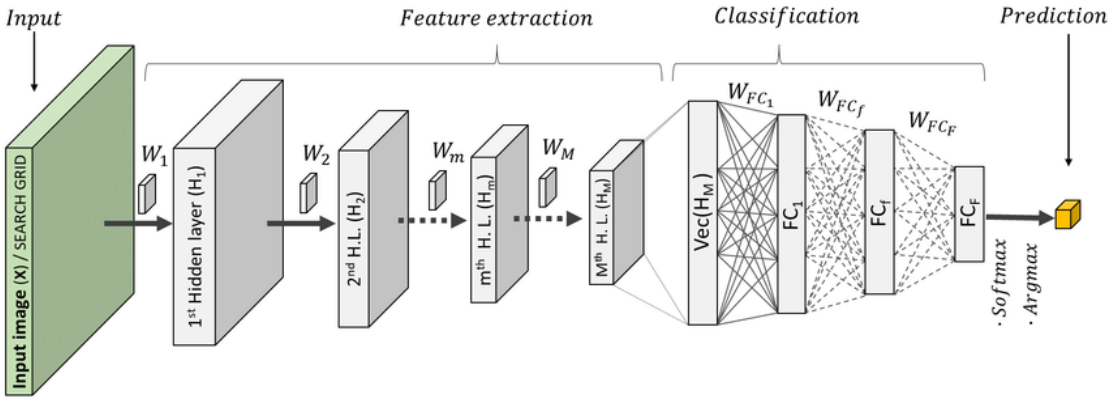

Fig.1 The operational principle of CNNs (Picture credit: Original)

In the field of object detection, the development of CNN, also known as CNN, is closely related to the development of detection frameworks. In the early days, R-CNN was used as a feature extractor to replace traditional visual features. The Region Proposal Network (RPN) is integrated with the detection network to achieve end-to-end training. Single-stage detectors such as YOLO simplify the process and treat the entire detection task as a single regression problem, which improves the detection speed.

2.2. Object Detection

The key task of object detection is to identify the objects in the image and classify the objects in the image. Bounding boxes are a technique that is commonly used to accurately label the position of an object in an image. Through the combination of these technologies, the joint recognition of object identity and space can be realized. CNN-based object detection methods can be divided into two major categories: two-stage algorithms and one-stage algorithms. The two-stage algorithm first generates candidate regions by region proposal network technology, and then classifies and regresses these regions. The one-stage algorithm performs object classification and location prediction directly on the image. The two-stage algorithm has the advantage of high accuracy and the disadvantage of slower speed. Single-stage algorithm has the advantage of high speed and the disadvantage of low accuracy. In practical applications, the object detection method should be selected according to the specific needs and specific analysis.

The process of the two-stage algorithm is to generate a series of candidate regions that may contain objects, and then extract features with the help of CNN, and then use SVM for classification and regression refinement of the bounding box. The representative algorithms are: R-CNN, Fast R-CNN. These methods are more accurate, but relatively slow to compute.

Compared with the two-stage detection algorithm, the one-stage algorithm omitted the step of generating candidate regions in the first stage. Predict the bounding box after detecting the class of the object directly in the image grid. YOLO treats image detection as a single regression task on an S×S grid. The representative algorithms in the one-stage algorithm include: YOLO, DETR. Compared with the two-stage algorithm, the single-stage algorithm has the advantage of faster detection speed, but the accuracy of the early single-stage algorithm model is lower than that of the two-stage algorithm.

3. Comprehensive analysis of the research

3.1. R-CNN: The Beginning of Deep Detection Based on Region Proposal

The architecture development of R-CNN series shows the progress and breakthrough made by deep learning in the field of object detection. Initially, R-CNN proposed a three-stage algorithm framework including "region proposal + CNN feature extraction + classification and regression", and SPPnet introduced a spatial pyramid pooling structure [2]. The problem of low efficiency caused by

repeated convolution calculation in R-CNN is effectively solved. The SPP layer is located after the convolutional layer, which can perform parallel pooling of proposals after a single full image forward propagation, avoiding the distortion of each proposal and the redundancy of convolution operation. DeepMask extends the R-CNN paradigm from the perspective of object detection proposal generation [3]. At its core, the discriminative CNN is used to learn category-independent segmentation masks directly from raw pixels. This innovation emphasizes the data-driven proposal generation approach, breaking through the limitation of R-CNN's dependence on external proposals, and the generalization ability to unseen categories.

Although the R-CNN family has made certain achievements, it also faces some challenges. First, the generation of candidate regions for object detection relies on external algorithms such as selective search, which may cause error accumulation. And the computing process is highly dependent on GPU, which limits its application in real-time.

3.2. Faster R-CNN: The Formation of an End-to-End Detection Framework

Compared with the previous object detection algorithm, the advantages of Faster R-CNN algorithm are "fast" and "real-time". The most important innovation of Fast R-CNN is the proposed region proposal network. The region proposal network is a fully convolutional network module that can generate high-quality region proposals on convolutional feature maps. This innovation greatly improves the detection speed of object detection algorithms, and effectively promotes the development process of object detection. The subsequent R-FCN algorithm constructed a pure fully convolutional region detection architecture, which completely abandoned the non-shared design pattern in the Faster R-CNN algorithm and achieved a trade-off between translation invariance and translation variance. The feature pyramid network uses the pyramid hierarchy structure of CNN itself to build an architecture with top-down paths and lateral connections [4]. This architecture can generate high-resolution feature maps with richer semantics at various scales without the need for image resizing operations.

The advantage of these object detection algorithms is that they are sub-second fast, and the accuracy of these algorithms is improved through shared computation and specific optimization. However, the advantages of these algorithms are different: Faster R-CNN/R-FCN focuses more on detection speed, and FPN/Mask R-CNN focuses more on multi-task fields [5].

Future research directions for object algorithm detection include fusing with one-stage detection algorithms, and integrating Transformer backbone. Through these ways to improve the detection speed and reduce the delay. These fast frameworks solve problems such as insufficient computing power, and more importantly, improve the scalability of object detection by means of modular design. Moving object Detection from the lab to real-time applications: Transitioning to autonomous driving, surveillance, and more. Subsequent research needs to focus on edge device deployment and privacy protection to achieve a wider range of applications.

3.3. YOLO: A unified real-time detection framework

The YOLO family of algorithms is a representative framework in single-stage object detection algorithms. The innovation of the YOLO series of algorithms is to reformulate the object detection task, transforming the object detection task into a regression problem of a single CNN. This innovation can predict the bounding box directly based on the full image pixels and can predict the class of the object in the image. This innovation significantly optimizes the process of object detection, enabling previously intractable end-to-end optimization and further enabling real-time inference. In 2015, Redmon et al. proposed the YOLO algorithm for the first time, which transformed the object detection task into a regression problem, and used a single CNN to predict bounding boxes and class probabilities directly from image pixels, successfully achieving end-to-end training and inference [6]. In 2018, Redmon et al. proposed YOLOv3 algorithm. V3 algorithm innovatively introduced a multi-scale prediction mechanism, and drew on the idea of feature pyramid network to improve detection performance [7]. This significantly improves YOLO's ability to detect small objects. In 2020,

Bochkovskiy et al. proposed YOLOv4 algorithm [8]. V4 algorithm incorporates excellent technologies and training strategies in the field of object detection in these years, uses CSPDarknet53 as the backbone network, and introduces spatial pyramid pooling structure and path aggregation network PAN. The fusion of these excellent techniques and training strategies effectively improves the perception and feature fusion ability of V4 network, and improves the detection accuracy of the model. In 2021, Wang et al. proposed the Scaled-YOLOv4 algorithm, innovatively proposed the model scaling method, and systematically discussed the influence of network depth, width, input resolution and structure on detection performance [9]. A lightweight model YOLOv4-tiny and a high-precision model YOLOv4-large are designed respectively. The development of YOLO series algorithms shows the research trend in the field of object detection: From the early simple framework, it has gradually developed towards multi-scale, multi-task and multi-scene applications, and the collaborative optimization of training strategy and network structure has become a key means to improve performance.

3.4. DETR: A new paradigm of end-to-end detection based on Transformer

Traditional object detection is generally carried out by means of "candidate box + non-maximum suppression": a large number of regions that are likely to contain the object are generated at the beginning, and then filtered gradually, which complicates the model process and makes it difficult to achieve end-to-end optimization. In order to solve the above problems, Carion et al. proposed the DETR algorithm model, which pioneered the addition of Transformer architecture to the object detection task [10]. Different from other object detection algorithms, DETR redefines the detection process as a "set prediction" problem.With the help of a set of learnable queries, the model can directly predict a fixed number of target boxes and determine the category of each detected object. DETR's innovation completely eliminates the anchor box and post-processing step, enabling a revolutionary detection paradigm.

Initially, DETR algorithm has some shortcomings, its convergence speed is relatively slow, and its calculation is relatively large. And DETR is less accurate in identifying small targets. After that, Deformable DETR is improved to solve the above problems by introducing a deformable attention mechanism, so that the DETR model can only focus on the key feature regions [11]. Because of this feature, the computational complexity of the DETR model is effectively reduced. The deformable attention mechanism fuses multi-scale features, which greatly improves the ability of DETR model for small object detection. With these improvements, the Transformer detector has been greatly optimized in terms of training speed, detection accuracy, and real-time performance, which makes the DETR model gradually have practical value.

An analysis of the overall development of DETR reveals a transition from “proof of concept” to “practical implementation” of DETR series models. It uses global attention modeling to replace the local feature extraction method of traditional convolution, which enables the model to understand the overall structure and detail relationships of the image at the same time. With the continuous development of technologies such as deformable attention, multi-scale fusion, and iterative optimization, DETR is expected to become a key model for general visual modeling in the future.

4. Challenges and Outlook

4.1. Challenges

Although there have been some achievements in object detection based on CNNs, such as R-CNN introducing the concept of region proposals, YOLO series enabling real-time detection, and DETR proposing an end-to-end Transformer framework, there are still some key challenges.

The recognition of small objects in multi-scale object detection is still a major difficulty. The fixed input size of CNN limits its ability to adapt to different scales. Although FPN relies on multi-layer fusion to alleviate this problem to a certain extent, there are still problems that high-resolution feature maps are computationally intensive and small objects are susceptible to noise. For example, the AP

performance of DETR for small objects on COCO dataset is lower than the baseline, and YOLOv3 has high localization error in dense scenes.

Lightweight deployment is also a challenging task, because the number of parameters in deep networks is quite large, which makes it difficult to adapt to edge devices. Although SPP-net reduces the dependence on the fully connected layer to a certain extent, the repeated calculation process of the convolutional layer is still time-consuming. Even though Deformable DETR optimizes the attention mechanism, it still needs to be adapted for mobile platforms.

4.2. Outlook

Based on a comprehensive analysis of the existing literature, it can be found that although the object detection technology based on CNN has made certain achievements, it still faces many challenges. As mentioned in the review paper, multi-scale object detection and effective recognition of small objects are still one of the difficulties in current research. In addition, how to achieve a better balance between accuracy and real-time performance, and how to design a more lightweight model to adapt to resource-constrained devices are also key research directions in the future.

Combined with the global modeling ability of Transformer and the multi-scale feature extraction ability of CNN, YOLOv3 and its variants can be optimized with the help of knowledge distillation and neural architecture search technology, making it more suitable for the deployment of Internet of Things devices. The feature pyramid network is extended to a dynamic structure, which can improve the generalization ability of the model in cross-domain scenarios. Promote the realization of applications such as autonomous driving.

5. Conclusion

This article provides a comprehensive review of object detection techniques based on CNNs, covering from the pioneering two-stage paradigm of R-CNN, to the end-to-end training achieved by Faster R-CNN, to the breakthrough of YOLO series in real-time detection. CNN architectures have strongly promoted the development of this field. These methods improve the detection accuracy and efficiency by introducing region proposals, shared convolutional features and one-stage regression.

Innovative measures such as feature pyramid network and spatial pyramid pooling optimize the multi-scale feature representation, solve the problem of traditional CNN dependence on fixed input size, and improve the detection ability of small objects and deformed objects.

However, the traditional CNN detector still relies on manual components such as non-maximum suppression in terms of process, and has certain limitations in modeling global information, which challenges the balance between small object detection and real-time performance. It simplifies the detection process by using the idea of set prediction, and shows a strong global modeling ability. However, its slow convergence speed and high computational complexity still need to be optimized.

Future research will continue to develop in integrating the advantages of different architectures, solving existing challenges such as small object detection, and pursuing higher efficiency. In particular, hybrid models of CNN and Transformer can be explored, such as Deformable attention extended by Deformable DETR to enhance multi-scale fusion. Knowledge distillation and neural architecture search, namely NAS, are introduced to optimize YOLO variants to achieve edge device deployment, and cross-domain generalization is also needed to adapt to emerging scenarios such as medical imaging and satellite remote sensing. With the development of computing resources and the advancement of algorithm innovation, object detection technology is expected to reach the Pareto optimum of accuracy and efficiency after 2025, which will promote the wider application of artificial intelligence in the field of visual understanding.

References

[1] Arkin E, Yadikar N, Xu X, et al. A survey: object detection methods from CNN to transformer. Multimedia Tools and Applications, 2022, 82(14): 21353-21383.

[2] Bhatti U A, Tang H, Wu G, et al. Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence. International Journal of Intelligent Systems, 2023, 2023(1): 8342104.

[3] Gao M, Zheng F, Yu J J Q, et al. Deep learning for video object segmentation: a review. Artificial Intelligence Review, 2023, 56(1): 457-531.

[4] Zhu L, Lee F, Cai J, et al. An improved feature pyramid network for object detection. Neurocomputing, 2022, 483: 127-139.

[5] Xu X, Zhao M, Shi P, et al. Crack detection and comparison study based on faster R-CNN and mask R-CNN. Sensors, 2022, 22(3): 1215.

[6] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 779-788.

[7] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. 2018. arXiv:1804.02767.

[8] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection. 2020. arXiv:2004.10934.

[9] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-YOLOv4: Scaling Cross Stage Partial Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021.

[10] Carion N, Massa F, Synnaeve G, et al. End-to-End Object Detection with Transformers. European Conference on Computer Vision (ECCV). Cham: Springer, 2020: 213-229.

[11] Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection. 2020. arXiv:2010.04159.