Deep Neural Networks for Object Detection

Jiahao You[1, *]

[1] School of Computer and Information, Anhui Normal University, Wuhu 241002, China

* Corresponding Author Email: YJH@AHNU.EDU.CN

Abstract. Object detection is one of the most basic and challenging tasks in image and video domains. The research on computer vision tasks is getting more and more attention, such as some tasks: object classification, object monitoring, etc. This paper provides a literature review that summarizes the detailed algorithms and application scenarios for object detection. Analysing and summarizing the latest research results in the current Object detection field, and summarize the relevant data sets and evaluation indicators, and based on this, summarize the current research problems in the Object detection field, and discuss the future research challenges in the Object detection field. possible future research directions.

Keywords: Computer vision. Object detection. Deep learning. Convolution Neural Network.

1. Introduction

In recent years, computer vision has been widely studied in various fields and developed at an unprecedented speed. Object detection is the most basic task. Image segmentation can be performed according to the characteristics of the target. It is not only necessary to judge the type of the image, but also to mark the actual position of the image. There have also been many significant achievements in the field of object detection. With the continuous development of technology, many application scenarios of object detection have also been developed, such as face detection [1], robot navigation [2], medical diagnosis [3], automatic driving [4], etc., all of which have huge practical applications potential. Nowadays, there are more and more practical applications based on object detection. How to improve the actual effect of object detection and improve the efficiency of the algorithm has great research value [5,6].

The object detection algorithm mainly uses deep learning technology. For example, Shang et al. [7] proposed the MANet for the semantic segmentation of remote sensing images. The model embeds the channel attention mechanism and fuses semantic features. The results are better than the other six most advanced networks.

However, these methods have some general problems: first, when detecting multiple object classes at the same time, the detection speed is slow and the accuracy is low. Second, some object detection methods or models can only object a specific view of a single object, and cannot detect the same object from different perspectives or changes [8]. The object detection algorithm requires high resource consumption, such as GPU and CPU computing resources, but the computational resources of systems are limited [9]. Only trained objects can be detected by the detection model, and some similar objects will be mistaken for objects in the trained model.

Although object detection is widely used in various tasks and solves many practical problems, at this stage, there are very few specific reviews on this task, beginners do not have a suitable starting article to explore this research field, limited to the previous articles, unable to understand the latest research trends and methods. In addition, the current review can also be considered as a simple summary of the object detection field, which can be beneficial to the development of subsequent specific research work. Therefore, the latest object detection algorithms at the current stage are summarized by us, and the common data sets and evaluation indicators for tasks are introduced. In addition, we also explain some mainstream object detection algorithms. Finally, we introduce some state-of-the-art methods for the task and the potential challenges in the task.

2. Background

2.1. Task definition

Object detection is a popular research field, mainly to detect target instances, for pictures or videos, such as dogs, humans and other instance objects. It is the work of pre-processing for subsequent tasks such as image and video recognition, positioning, etc., so as to have a favourable impact on subsequent tasks. However, when performing the task of object detection, there are many difficult processes, such as various practical conditions such as occlusion, darkness, etc. in the image. This work has practical applications in video surveillance, autonomous driving, face recognition, computer vision, etc.

2.2. Datasets

2.2.1. Microsoft common objects in context (MS COCO)

MS-COCO is one of the most challenging datasets in the current object detection field, with fewer objects but more instances. After the data set appeared, after 2015, related data competitions will be held. Compared with the classic dataset ImageNet, the objects of this dataset are significantly smaller and denser, and this dataset helps the object detection model to understand the real world, and the trained model will be better[10].

2.2.2. ImageNet large scale visual recognition challenge (ILSVRC)

ILSVRC is not the name of a dataset, but a computer vision challenge, held from 2010-2017, after the competition, all datasets are made public. Images are collected based on ImageNet for practical detection challenges. Beginning in 2010, the dataset was small, with only 1000 categories. In subsequent technology and the development of the competition, the dataset contains a total of 30 fully labeled categories based on video clutter, average number of objects, etc. factor distinction. ILSVRC contains more than 15 million images, which is a great challenge to the subsequently developed object detection algorithms[11].

2.2.3. Pascal VOC

PASCAL VOC assists in the evaluation of newly developed object detection algorithms that can perform various tasks based on this dataset, which was collected based on challenges held in 2005-2012. Every year, different datasets are used for real-world testing, among which the most popular versions are VOC07 and VOC12. Each year's datasets are developed from existing datasets. The same as ILSVRC is that in the first year, the dataset has only four objects, but in VOC12, the dataset consists of 11530 training objects and 27450 annotation objects, and the object class is also increased to 20 This dataset annotates and reproduces common objects in real life to help the model better reconstruct the real world [12].

2.2.4 Open image dataset (OID)

Open Images dataset V4, which is mainly for three image tasks, image classification, object detection and visual relationship detection, all images are downloaded from Flicker, a total of 9.2 million images, and all images are marked by professional annotators, so The excellent point of this data set is that all the labeled images in this data set are of high quality, which provides convenience for the subsequent training process, and also helps the model to understand the world more accurately, providing practical object detection tasks. Good validation dataset [12].

2.3. Metrics

In current research, average precision (AP) is most commonly used to evaluate the evaluation criteria for object detection tasks. First, let's briefly introduce some basic concepts before AP, such as true negative (TN), false negative (FN), true positive (TP), and false positive (FP). TP is the correct prediction for true values and TN is the prediction for negative values. Correctly predict the class.

However, since many objects cannot be detected in the object detection task, the concepts of FP and FN are not suitable for object detection.

In object detection, Intersection over Union (IoU) is used to evaluate the accuracy of the actual predicted bounding box. IoU can evaluate the overlapping area of the ground truth BBg and the predicted BBp, a bounding box is divided by the predicted bounding box and the ground-truth bounding box. Equation 1 shows the importance of IOU in object detection, and how its actual detection works. Equation 1 shows the actual IoU calculation method.

$$Io\,U\left(BB_g, BB_p\right) = \frac{\text{area}\left(BB_g \cap BB_p\right)}{\text{area}\left(BB_g \text{U} BB_p\right)} \tag{1}$$

For the object detection task, to judge whether the detection of an object is correct, the threshold t is given first, and then the overlap between t and IoU is found for actual classification. Compare t and IoU, if IOU $\geq$ t; otherwise, it is incorrect. If there are different detectors on a bounding box in the same region, only one detector can be considered correct.

For the object detection method, the concept of longitude and recall can be used for evaluation, and the concept definition uses TP and FP. Precision P is calculated by using TP and FP. The longitude calculated by a model is calculated using Equation 2.

$$P = TP/(TP + FP) \tag{2}$$

The recall rate is a part of the actual detection of the correct instance for the current object detection model. The concept of recall rate is defined by using TP and FN. The third equation is the calculation formula for the recall.

$$R = TP/(TP + FN) \tag{3}$$

In object detection, the detector identifies relevant objects to find ground truth objects. If a detector has relatively high precision and recall values during testing, then it can be said that the detector is a Nice detector. The average precision, on the other hand, is calculated for the average over the recall interval. The higher the AP value, the better the performance of the method, and conversely, the smaller the value, the worse the effect. mAP, which is an indicator to measure the accuracy of all categories of detectors, is the actual indicator used to measure the detection accuracy. The calculation of mAP can be calculated by Equation 4

$$mAP = \frac{1}{N}\sum_{i=0}^{N} AP_i \tag{4}$$

where N is the total number of classes and APi is the ith class of AP. Another metric, the F1 score, is the harmonic mean of precision and recall. Sensitivity (SEN) and specificity (SPE) metrics are used for evaluation.

$$\text{SEN} = \text{True Positive}/\text{True Positive} + \text{False Negative} \tag{5}$$

$$\text{SPE} = \text{True Negative}/\text{True Negative} + \text{False Positive} \tag{6}$$

Some studies have used a combination of these metrics to assess outcomes. However, recall and precision are the most widely used combinations.

3. Recent researches

Dai et al. [13] addressed the key issue of how to improve the performance of object detection heads in object detection algorithms. The main challenges in developing a good object detection are: the scale awareness of the head, and the spatial awareness that the head has. And recent research has only focused on addressing one of the above problems in various ways. This paper proposes a new detection head, called dynamic head, which unifies scale-awareness, space-awareness, and task-awareness. The attention mechanism is deployed separately on each specific dimension of the feature, i.e. horizontal, spatial and channel. The scale-aware attention module is only deployed in the

horizontal dimension. It learns the relative importance of different semantic levels to enhance features at appropriate levels according to the scale of a single object. The spatially aware attention module is deployed in the spatial dimension (i.e. height × width). It learns coherent discriminative representations at spatial locations. The task-aware attention module is deployed on the channel. It guides different feature channels to support different object-based kernel responses for different tasks (e.g., classification, box regression, and center/keypoint learning). Final Experimental Results Extensive experiments on the MS-COCO benchmark demonstrate the effectiveness of the method. Compared to EfficientDet (24) and SpineNet (8), the dynamic head uses 1/20 the training time but performs better.

Xie et al. [14] aimed at the unsupervised pre-training method designed for object detection deficiencies in image classification. A simple and effective self-supervised object detection method DetCo is proposed. Its advantages come from the multi-level supervision of the representations of different stages in the middle, for the contrastive learning between the global image and local patches. These two designs facilitate discriminative and consistent global and local representations at each level of the feature pyramid, optimize backbone network features, improve object detectors through multi-dimensional, multi-scale approaches, and simultaneously improve detection and classification, to have better results. Extensive experiments on VOC, COCO, Cityscapes, and ImageNet show that DetCo not only outperforms state-of-the-art methods on a range of 2D and 3D instance-level detection tasks, but is also competitive in image classification.

Sun et al. [15] have always relied on dense priors for Object detection, but there are many problems with the current dense priors, the pipeline will produce some redundant or repeated results, the network is very sensitive to heuristic rules, and the final performance Largely affected by the size of the anchor box. In this paper, sparse R-CNN is proposed, which is a pure sparse method without enumerating object candidates for the image grid, and making a fixed sparse set of learned object proposals for the object selection head, i.e. The proposed feature, which is a high-dimensional (e.g. 256) latent vector. It is able to encode rich instance features. In particular, the proposed feature generates a series of custom parameters for its unique object recognition header. Sparse R-CNN demonstrates its accuracy, runtime, and training convergence performance on the challenging COCO dataset, comparable to well-established detectors.

Dai et al [16] address the training and optimization challenges of DETR, which require large-scale training data and extremely long training programs. The existing task cannot be directly applied to pre-train DETR's morphers. The main reason is that DETR is mainly concerned with spatial localization learning. In this paper, we propose a pretext task called random query patch detection for unsupervised pre-training of DETR (UP-DETR) for object detection. Specifically, we randomly crop patches from a given image and then provide them to the decoder as queries. With a pre-trained model, detection can be made from raw images, in the query patch, UP-DETR is introduced, and, extending this operation, it is possible to have query of object groups as well as multi-query patches of attention masks. In experiments, UP-DETR outperforms DETR on PASCAL VOC and COCO Object detection with faster convergence and better average accuracy.

Sun et al [17], found that DETR, which treats Object detection as a set prediction problem, achieves state-of-the-art performance but requires a long training time to converge. To analyze the reasons for the difficulty of DETR optimization, extensive experiments were conducted and it was found that the cross-attention module of the Transformer decoder to obtain the target information from the image is the main reason for the slow convergence. To speed up the convergence speed, the encoder-only version of DETR was further investigated by removing the cross-attention module. It is found that the encoder-only DETR shows great improvement in detecting small objects, but, performance of detecting large objects is poor. This paper analyzes the convergence problem of DETR, and the Hungarian loss of the model has an important impact on the instability of the two-part matching. Based on this, in order to speed up the training process based on the transformer method, this paper proposes two models, both of which are modified versions of DETR, and in subsequent tests show that the proposed method not only converges faster than the original in the evaluation of

the COCO 2017 detection baseline DETR is much faster and also significantly outperforms DETR in terms of detection accuracy and other baselines.

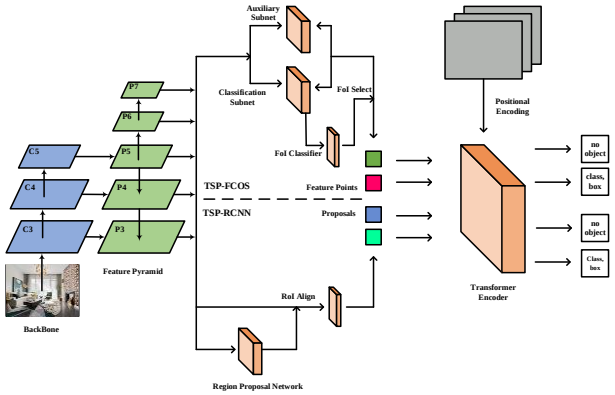

Figure 1 The network architectures of TSP-FCOS and TSP-RCNN

Xu et al [18]. studied the problem of labeling large datasets. The current dataset labeling process is time-consuming and expensive. Current mainstream methods are trained on unlabeled networks, compared to previous complex methods. This paper proposes an end-to-end semi-supervised method for efficient annotation. The training can effectively improve the labeling quality of the data set, and for the new object detection algorithm, a more accurate data set can promote the training of object detection. At the same time, two effective and simple techniques are also proposed, the first is that the unclassified classification loss can be score weighted. The second is for selecting box jittering methods for box regression learning. This method is obviously better than the previous method, and on the new target detector, after labeling with this algorithm, the detection accuracy can be improved by 1.5mAP, which promotes the development of the algorithm in the field of object detection.

4. Challenges

Detection efficiency is a major challenge for object detection models. On the basis of ensuring accurate classification and positioning, object detection also requires extremely fast detection speed. For example, in the actual application process, if there is a demand for real-time detection, the detection speed Just put forward a very high demand. Although some researches have been aimed at this, the speed of the current algorithm is far from meeting the demand in actual production scenarios.

Scalability is also a major challenge for object detection. The current data is increasing at an unprecedented rate. For the current large data set, manual labeling is very difficult, so it is necessary to rely on algorithms for labeling, mostly supervised methods, so the object detection algorithm is scalable. Therefore, there is no better way to deal with data objects that have not been seen before in the actual environment.

For the input data, people can quickly distinguish the different perspectives of the same object, but for the object detection algorithm, this is a completely different feature or object. How can we propose an object for different perspectives, and quickly and correctly identify it, which is also a very challenging research direction for current algorithms.

With the development of current wearable devices and Internet of Things technology, people's needs are gradually changing. Before, a lot of computing power and storage space were required for object detection, but now limited by space and resource constraints, it is necessary to propose more efficient The algorithm can perform real-time detection on the basis of limited resources, which is also a new application. Moreover, the position and size of pictures in the real-time environment are constantly changing, and the computational complexity is high, which also poses a higher challenge to the algorithm.

The design of the current object detection algorithm model is becoming more and more complex. For example, networks such as AlexNet to ResNet and ResNext often require a large amount of data and many GPUs for training, which often results in waste of resources, and in actual deployment,

there will also be It is a great difficulty. Based on this, how to design a lightweight and efficient network is what current researchers need to consider.

At present, most of the object detection algorithms are only suitable for 2D images, and the detection effects for other 3D, lidar, etc. are poor, and object detection is also highly correlated with drones, robots and other fields. Therefore, the detection of 3D images, It poses new challenges for current object detection.

5. Conclusion

In this paper, we summarize that Object detection is receiving more and more attention in the field of computer vision. There can be more opportunities as well as applications as well as techniques to implement Object detection. This paper first introduces a simple definition of Object detection and the techniques required for Object detection, which allows a brief understanding of the Object detection task. This is followed by an introduction of the Object detection application area, which allows to clarify the practical scenarios for the algorithm. In order to test the algorithms for future Object detection tasks, we first introduce four different datasets in this field, namely MS COCO, ILSVRC, Pascal VOC, OID, and propose some common test metrics for these tasks. Then, we briefly outline the recent application scenarios and, based on them, propose some current problems and challenges in the field of Object detection algorithms, which can be briefly advanced for future research and applications.

References

[1] X. Ming, F. Wei, T. Zhang, D. Chen, N. Zheng, and F. Wen, 'Group Sampling for Scale Invariant Face Detection', IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 2, pp. 985–1001, Feb. 2022, doi: 10.1109/TPAMI.2020.3012414.

[2] I. Muhammad, K. Ying, M. Nithish, J. Xin, Z. Xinge, and C. C. Cheah, 'Robot-Assisted Object Detection for Construction Automation: Data and Information-Driven Approach', IEEEASME Trans. Mechatron., vol. 26, no. 6, pp. 2845–2856, Dec. 2021, doi: 10.1109/TMECH.2021.3100306.

[3] R. Yang and Y. Yu, 'Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis', Front. Oncol., vol. 11, p. 638182, Mar. 2021, doi: 10.3389/fonc.2021.638182.

[4] Y. Xu, H. Wang, X. Liu, H. He, Q. Gu, and W. Sun, 'Learning to See the Hidden Part of the Vehicle in the Autopilot Scene', Electronics, vol. 8, no. 3, p. 331, Mar. 2019, doi: 10.3390/electronics8030331.

[5] L. Fan, T. Zhang, and W. Du, 'Optical-flow-based framework to boost video object detection performance with object enhancement', Expert Syst. Appl., vol. 170, p. 114544, May 2021, doi: 10.1016/j.eswa.2020.114544.

[6] C. B. Murthy, M. F. Hashmi, N. D. Bokde, and Z. W. Geem, 'Investigations of Object Detection in Images/Videos Using Various Deep Learning Techniques and Embedded Platforms—A Comprehensive Review', Appl. Sci., vol. 10, no. 9, p. 3280, May 2020, doi: 10.3390/app10093280.

[7] R. Shang, J. Zhang, L. Jiao, Y. Li, N. Marturi, and R. Stolkin, 'Multi-scale Adaptive Feature Fusion Network for Semantic Segmentation in Remote Sensing Images', Remote Sens., vol. 12, no. 5, p. 872, Mar. 2020, doi: 10.3390/rs12050872.

[8] B. F. Klare et al., 'Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A', in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, Jun. 2015, pp. 1931–1939. doi: 10.1109/CVPR.2015.7298803.

[9] L. Liu et al., 'Deep Learning for Generic Object Detection: A Survey', Int. J. Comput. Vis., vol. 128, no. 2, pp. 261–318, Feb. 2020, doi: 10.1007/s11263-019-01247-4.

[10] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, 'COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images', arXiv, arXiv:1601.07140, Jun. 2016. doi: 10.48550/arXiv.1601.07140.

[11] O. Russakovsky et al., 'ImageNet Large Scale Visual Recognition Challenge', Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[12] J. Kaur and W. Singh, 'Tools, techniques, datasets and application areas for object detection in an image: a review', Multimed. Tools Appl., Apr. 2022, doi: 10.1007/s11042-022-13153-y.

[13] X. Dai et al., 'Dynamic Head: Unifying Object Detection Heads with Attentions', in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, Jun. 2021, pp. 7369–7378. doi: 10.1109/CVPR46437.2021.00729.

[14] E. Xie et al., 'DetCo: Unsupervised Contrastive Learning for Object Detection', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 8372–8381. doi: 10.1109/ICCV48922.2021.00828.

[15] P. Sun et al., 'Sparse R-CNN: End-to-End Object Detection with Learnable Proposals', in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, Jun. 2021, pp. 14449–14458. doi: 10.1109/CVPR46437.2021.01422.

[16] Z. Dai, B. Cai, Y. Lin, and J. Chen, 'UP-DETR: Unsupervised Pre-training for Object Detection with Transformers', in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, Jun. 2021, pp. 1601–1610. doi: 10.1109/CVPR46437.2021.00165.

[17] Z. Sun, S. Cao, Y. Yang, and K. Kitani, 'Rethinking Transformer-based Set Prediction for Object Detection', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 3591–3600. doi: 10.1109/ICCV48922.2021.00359.

[18] M. Xu et al., 'End-to-End Semi-Supervised Object Detection with Soft Teacher', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 3040–3049. doi: 10.1109/ICCV48922.2021.00305.