Recent Advances of Generic Object Detection with Deep Learning: A Review

Xin Li[1,2], Yingying Li[1,2(✉)], and Shushu Li[1,2]

[1] Anhui Agricultural University, Hefei, China
yyhgd@ahjzu.edu.cn
[2] School of Electronics and Information Engineering, Anhui Jianzhu University, Hefei 230601, China

Abstract. Object detection is an important and challenging problem in computer vision. It has been widely applied in many vision tasks, such as object tracking, image segmentation, action recognition, etc. With the rapid development of deep learning, more state-of-the-art object detection methods based on deep learning with some modifications have effectively improved the detection performance. This paper comprehensively reviews object detection methods in the recent five years based on deep learning from object detection framework, including significant advances of the backbone network, multi-scale learning, data augmentation. Finally, we investigate the performance of typical object detection algorithms on popular datasets MS-COCO, PASCAL-VOC, and point out the existing problem for further research.

Keywords: Object detection · Deep learning · Backbone network · Multi-scale learning

1 Introduction

Object detection is one of the most fundamental tasks in computer vision. It plays an important role in many applications, such as object tracking [1], image segmentation [2], action recognition [3], etc. In recent years, object detection has been pushed forward by the success of deep learning techniques to a research highlight. Numerous research progresses on object detection have endlessly been achieved. It is necessary to provide researchers with timely reviews to guide future research on object detection.

Many reviews about object detection have been published. These reviews sum up all kinds of object detection methods from different research perspectives and under specific application scenarios. Wu [4] systematically reviews recent advances in object detection with deep learning, including detection components, learning strategies, applications, and benchmarks. Zhao [5] pays more attention to the typical generic architectures of object detection with progress and useful tricks. Their work also reviews several specific applications, such as salient object detection, face detection, and pedestrian detection. Li [6] provides

X. Yuan et al. (Eds.): ICUIA 2020, CCIS 1319, pp. 185–193, 2020.
https://doi.org/10.1007/978-981-33-4601-7_19

a comprehensive review of generic object detection from 300 research contributions, from the aspects of detection frameworks, object proposal generation, feature representation, context modeling, etc. Jiao [7] provides an overview of the traditional and new applications and some new branches of object detection.

This paper focuses on the new advances of generic object detection in the recent five years, reviewing the research works of deep learning-based object detection. The paper aims to give a comprehensive review in various aspects of object detection, including object detection framework, significant advances of the backbone network, multi-scale learning, data augmentation. In addition to this, we investigate datasets and evaluation of classical object detection algorithms in recent years and we thoroughly analyze their performance.

The rest of the paper is organized as follows: The object detection framework is listed in Sect. 2. Then significant advances in various aspects of object detection are in Sect. 3. The evaluation method of object detection and the comparison of various performances are in Sect. 4. Finally, we conclude and discuss future directions in Sect. 5.

2 Object Detection Framework

Deep learning-based object detection frameworks usually can be divided into two categories: two-stage detectors and single-stage detectors. Two-stage detectors first generate a sparse set of proposals locations and then region classifiers as the next step. Single-stage detectors directly make a categorical prediction of objects at each location along with cascaded region classification as the same step.

2.1 Two-Stage Detector

The Two-stage detectors include the following two processes: one is to propose the candidate boxes, and the other is to make the decision of classifications using multiple feature maps at the top of the network. The most representative two-stage object detectors are the R-CNN [8] series, including fast R-CNN [9], faster R-CNN [10], and Libra R-CNN [11].

R-CNN applies CNNs to bottom-up region proposals in order to localize objects, generate a rich hierarchy of image features by supervised pre-training and domain-specific fine-tuning. Fast R-CNN employs a new training algorithm that fixes the disadvantages of R-CNN and SPPnet to improve training and testing speed while also increasing detection accuracy. Faster R-CNN presents Region Proposal Networks (RPNs) for more efficient and accurate region proposal generation, for RPNs can generate higher quality region proposals than Fast R-CNN for detection. Libra R-CNN integrates IoU-balanced sampling, balanced feature pyramid, and balanced L1 loss. Thanks to its overall balanced design, Libra R-CNN significantly improves the detection performance.

2.2 Single-Stage Detector

The single-stage target detection networks are integrating the two tasks of generating candidate boxes and providing the final classification of the input image as a whole process. The advantage of this framework is that it greatly improves the detection speed. The representative networks of single-stage are SSD [12] and YOLO [13], YOLO9000 [14], YOLOv3 [15], YOLOv4 [16].

SSD is a fast single-stage object detector for multiple categories, which discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales, and uses multiple feature maps at the top of the network to achieve improved performance. YOLO [13] takes object detection as a regression problem to predict bounding boxes and class probabilities directly from input images in one evaluation. YOLO pushes the application of object detection in real-time. But the first Yolo has poor position accuracy in small object detection. The later YOLO [14–16] make improvements on YOLO in positioning accuracy and detection speed, not only to general goals but also to small object detection. YOLOv4 [16] develops the previous object detection model and summarizes the influence of state-of-the-art Bag-of-Freebies and Bag-of-Specials methods of object detection during the detector training. So it is faster and more accurate than other detectors.

3 Review of Significant Advances

3.1 Backbone Networks

Deep learning networks bring a revolutionary breakthrough in object detection rather than just obvious improvements in performance on large databases. Their success results from training an effective backbone network on large labeled images. The most representative backbone networks used in object detection tasks are as follows.

VGG [17] modifies some parameters of the ConvNet architecture and increases the depth of the network by adding more convolutional layers with using very small (3×3) convolution filters in all layers. With the emergence of the convolutional neural network, image recognition has developed rapidly.

ResNet [18] uses a residual learning framework to lighten the training networks, which rebuilds the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. These residual networks are easier to optimize and can gain accuracy from considerably increased depth. ResNet has lower complexity even if it has deeper layers, compared with VGG.

SpineNet [19] is the scale-permuted model instead of the scale decreased model, which provide two major improvements on backbone architecture: One is that SpineNet can retain spatial information as it grows deeper, the other, the connections between feature maps should be able to go across feature scales to facilitate multi-scale feature fusion. It is a good backbone architecture design for tasks requiring simultaneous recognition and localization.

EfficientNets [20] is a family of models with a new baseline, obtained by neural architecture search. This model firstly uses a simple and effective coefficient to quantify the relationship among all three dimensions of network width, depth, and resolution. Benefited from this balancing network depth, width, and resolution, EfficientNets achieve much better accuracy and efficiency than the previous backbone network.

CSPNet (Cross Stage Partial Network) [21] splits the gradient flow to make propagated gradient information through different network paths have a large correlation difference. CSPNet can greatly reduce the amount of computation, and improve inference speed as well as accuracy, so it can relieve the problem that previous networks require heavy computations, help people with cheap devices to enjoy the result of the backbone networks.

3.2 Multi-scale Learning

Neck refers to the fusion of features of the above different scales, with the purpose of generating multi-scale features with both high semantic information and accurate location information, and improving the ability of the model to detect targets of different scales.

FPN Network. The traditional method of extracting multi-layer features is the image pyramid, which is an effective but conceptually simple structure to interpret images with multi-resolution. By changing the scale of the image, the image layer by layer is compared to a pyramid. The higher the level, the smaller the image, and the lower the resolution. We can also extract features from the feature pyramid by using the convolutional network, but this will greatly increase the operation time and require more memory for operation. Therefore FPN in 2017 was put forward, the author through the bottom-up, namely network to process before, top-down, upsampling is used, the results of the sampling on the transverse connection is will and bottom-up generation feature of the same size of the map to merge, and the characteristics of different resolutions is a figure, FPN today is still used in many networks, such as Faster RCNN, Mask RCNN, DSSD [22], etc.

There are also some problems with the classical FPN network. For example, multi-scale characterization improves the detection effect of the deMultiscale learning network, but at the same time makes it impossible for the multiscale features to be fully utilized by the network. Therefore, AugFPN [23] is an improvement on the classical FPN structure. AugFPN innovation lies in three of the components, respectively is Consistent Supervision: narrow the features before fusion of different scale, the semantic gap between Residual Feature Augmentation: analysis on the characteristics of Residual enhance extraction rate constant of context information, reduce the information loss Feature map is on the highest pyramid level, Soft ROI Selection: adaptively learn better ROI. AugFPN's innovations make up for FPN's shortcomings. At present, a new feature pyramid structure called NAS-FPNnas [24] is proposed. The author makes

full use of Neural architecture search with reinforcement learning and trains a controller to select the best model structure in a given search space through intensive learning.

SPP and ASPP Networks. The traditional neural network requires the input of a fixed size of the image, which requires the resize of the image before it is introduced into the network. As a result, the image information changes. To solve this problem, SPP [25] and ASPP [26] are proposed.

SPP extracts features from blocks of different sizes, respectively 4×4, 2×2, and 1×1. Put these three grids on the following feature map, and you can get 21 different Spatial bins. From these 21 blocks, each block extracts a feature, so as to extract the 21-dimensional feature vector. The entire process is completely independent of the size of the input, so you can handle candidate boxes of any size.

The ASPP parallel samples a given input by atrous convolution at different sampling rates. Compared to the conventional convolution operator, atrous convolution can obtain a larger size of the receiving field without increasing the number of kernel parameters. ASPP proposed to connect the feature maps generated by atrous convolution under different expansion rates in series, so that the neurons in the output feature map contain multiple accepting field sizes, encoding the multi-scale information, and finally improving the performance.

3.3 Data Augmentation

Training for a neural network often requires the support of thousands of pictures, and the more data, the better the experimental effect. However, this often does not occur in large data sets, which leads to a new field of data enhancement. MixUp [27] multiplies and superimposes two images at different coefficient ratios, and then adjusts the label using those superimposed ratios. With CutMix [28], it is to overlay the cropped image onto a rectangular area of other images and resize the label according to the size of the mixed area. The random erase [29] and CutOut [30] can randomly select rectangular areas in the image and populate them with a random or complementary value of zero. In addition, style transfer GAN [31] is often used for data enhancement, which can effectively reduce the texture deviation of CNN learning.

4 Evaluation and Databases

Average Precision (AP) is the common metric to evaluate the detection precision, defined as the average detection precision under different recalls, usually evaluated in one class. The mean Average Precision (mAP) refers to the average score of AP across all classes, which is used as an evaluation metric for many object detection datasets.

A number of well-known datasets for object detection have been provided in the past years [32] MS-COCO and PASCAL-VOC [32] are the most representative datasets for generic object detection. We investigate the performance of typical object detection algorithms with different backbone network on popular datasets MS-COCO, PASCAL-VOC. The results are shown Table 1 and Table 2.

Table 1. Detection results on the VOC 2007 test-dev dataset of some typical methods

Model	Backbone	mAP (%)	# of stage	Detection speed (fps)
RCNN [8]	VGG16	66	Two	0.5
Fast RCNN [9]	VGG16	70	Two	7
Faster RCNN [10]	VGG16	73.2	Two	7
Faster RCNN [10]	Resnet101	76.4	Two	5
YOLO [13]	Darcknet19	66.4	One	45
SSD [12]	VGG16	77.1	One	46
YOLOv2 [14]	Darcknet19	78.6	One	40
YOLOv3 [15]	Darcknet53	33	One	51
DSSD321 [22]	Resnet169	78.6	One	9.5
DSSD513 [22]	Resnet169	81.5	One	5.5
Soft Sampling [33]	VGG16	79.3	Two	–
R-FCN-3000 [34]	Resnet101	80.5	Two	30

Table 2. Detection results on the MS-COCO test-dev dataset of some typical methods

Model	Backbone	mAP (%)	# of stage	Detection speed (fps)
Mask RCNN [11]	Resnet101	33.1	Two	4.8
YOLO9000 [14]	Darcknet19	78.6	One	40
FPN [35]	Resnet50	35.8	Two	5.8
NAS-FPN [23]	Resnet50	44.2	Two	92.1
Cascade RCNN [23]	Resnet101	42.8	Two	–
D-RFCN + SNIP [36]	DPN-98	48.3	Two	2
TridentNet [37]	Resnet101	48.4	One	–

5 Conclusion

In recent years, deep learning-based object detection has developed rapidly. Detection accuracy and high precision in real-time systems are the ultimate goals of object detection. This paper provides a detailed review of object detection in

recent five years, covering object detection framework, significant advances of the backbone network, multi-scale learning, activation function, data augmentation. Although significant advances in this domain have been achieved recently, there is still much room for further development. Finally, we propose several promising future directions, such as the interpretability of convolution, the combination of the actual mobile terminal, the balance of accuracy and speed.

Acknowledgements. This work was supported in part by the Natural Science Research Project of Educational Commission of Anhui Province of China under Grant (KJ2018A0521) and the Visiting Scholar Researcher Program through the Young Talents Foreign Visiting and Training Program of Educational Commission of Anhui Province of China under Grant (gxgwfx2018047).

References

1. Kang, H., et al.: T-CNN: tubelets with convolutional neural networks for object detection from videos. IEEE Trans. Circ. Syst. Video Tech. **28**(10), 2896–2907 (2018)
2. Wang, Y., Wang, L., Lu, H., et al.: Segmentation based rotated bounding boxes prediction and image synthesizing for object detection of high resolution aerial images. Neurocomputing (2020)
3. Herath, S., Harandi, M., Porikli, F.: Going deeper into action recognition: a survey. Image Vis. Comput. **60**, 4–21 (2017)
4. Wu, X., Sahoo, D., Hoi, S.C.H.: Recent advances in deep learning for object detection. Neurocomputing (2020)
5. Zhao, Z.Q., Zheng, P., Xu, S., et al.: Object detection with deep learning: a review. IEEE Trans. Neural Netw. Learn. Syst. **30**(11), 3212–3232 (2019)
6. Liu, L., Ouyang, W., Wang, X., et al.: Deep learning for generic object detection: a survey. Int. J. Comput. Vis. **128**(2), 261–318 (2020)
7. Jiao, L., Zhang, F., Liu, F., et al.: A survey of deep learning-based object detection. IEEE Access **7**, 128837–128868 (2019)
8. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
9. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
10. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
11. Pang, J., et al.: Libra R-CNN: towards balanced learning for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
12. Pang, J., Chen, K., Shi, J., et al.: Libra R-CNN: towards balanced learning for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2019)
13. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)

14. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
15. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement, arXiv preprint arXiv:1804.02767 (2018)
16. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934 (2020)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014)
18. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Du, X., Lin, T.Y., Jin, P., et al.: SpineNet: learning scale-permuted backbone for recognition and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11592–11601 (2020)
20. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks, arXiv preprint arXiv:1905.11946 (2019)
21. Wang, C.Y., Mark Liao, H.Y., Wu, Y.H., et al.: CSPNet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391 (2020)
22. Fu, C.Y., Liu, W., Ranga, A., et al.: DSSD: deconvolutional single shot detector, arXiv preprint arXiv:1701.06659 (2017)
23. Guo, C., Fan, B., Zhang, Q., et al.: AugFPN: improving multi-scale feature learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12595–12604 (2020)
24. Ghiasi, G., Lin, T.Y., Le, Q.V.: NAS-FPN: learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7036–7045 (2019)
25. He, K., Zhang, X., Ren, S., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
26. Yang, M., Yu, K., Zhang, C., et al.: DenseASPP for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3684–3692 (2018)
27. Misra, D.: Mish: a self regularized non-monotonic neural activation function, arXiv preprint arXiv:1908.08681 (2019)
28. Yun, S., Han, D., Oh, S.J., et al.: CutMix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6023–6032 (2019)
29. Zhong, Z., Zheng, L., Kang, G., et al.: Random erasing data augmentation. In: AAAI, pp. 13001–13008 (2020)
30. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552 (2017)
31. Geirhos, R., Rubisch, P., Michaelis, C., et al.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, arXiv preprint arXiv:1811.12231 (2018)
32. Everingham, M., Van Gool, L., Williams, C.K.I., et al.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
33. Wu, Z., Bodla, N., Singh, B., et al.: Soft sampling for robust object detection, arXiv preprint arXiv:1806.06986 (2018)

34. Singh, B., Li, H., Sharma, A., et al.: R-FCN-3000 at 30fps: decoupling detection and classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1081–1090 (2018)
35. Seferbekov, S.S., Iglovikov, V., Buslaev, A., et al.: Feature pyramid network for multi-class land segmentation. In: CVPR Workshops, pp. 272–275 (2018)
36. Singh, B., Davis, L.S.: An analysis of scale invariance in object detection SNIP. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3578–3587 (2018)
37. Li, Y., Chen, Y., Wang, N., et al.: Scale-aware trident networks for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6054–6063 (2019)