Original Research Article

The Object Detection of Anchor Free—A Survey

Rui Huo, Shuili Zhang [*Corresponding Author], *Huaiyuan Sun, Xve Tian*

School of Physics and Electronic Information, Yan'an University, Yan'an, Shaanxi Province, China

Abstract: Since Ross Girshick introduced deep learning to object detection and proposed RCNN networks in 2014, the field of object detection has been blasting ahead. Deep learning-based object detection algorithms typically classify and regress region proposals. In the one-stage detector, these region proposals are the anchor boxes generated by the sliding window approach. In this paper, we first give a brief overview of anchors and compare the advantages and disadvantages of both anchor base and anchor-free object detection. Then, we collect and organize the anchor-free object detection algorithms (such as CenterNet, FCOS, FSAF, etc.) that have received much attention and use in recent years. We also present a detailed description of the anchor-free object detection algorithms in three parts (keypoint-based, Segmentation-based, and YOLO series).

Keywords: Deep learning; Object detection; Image recognition; Anchor

1. Introduction

The term anchor first appears in the paper Faster R-CNN[1], where the authors state that "At each sliding-window location simultaneously predict multiple region proposals, where The k proposals are parameterized relative to k reference boxes, which called anchor". The k proposals are parameterized relative to k reference boxes, which is called anchor". In more general terms, then, an anchor is a set of pre-defined boxes, and during training, the training samples are constructed with the offsets of the real box locations relative to the pre-defined boxes. This is equivalent to roughly "framing" the object with the preset boxes, and then adjusting it on top of these preset boxes. Figure 1 shows the comparison of existing models and anchor-free models, from which it is clear that most of the anchor-free models basically surpass the one-stage and two-stage object detection models in terms of accuracy. So in recent years, it can be said that the object detection model has entered the anchor free era.

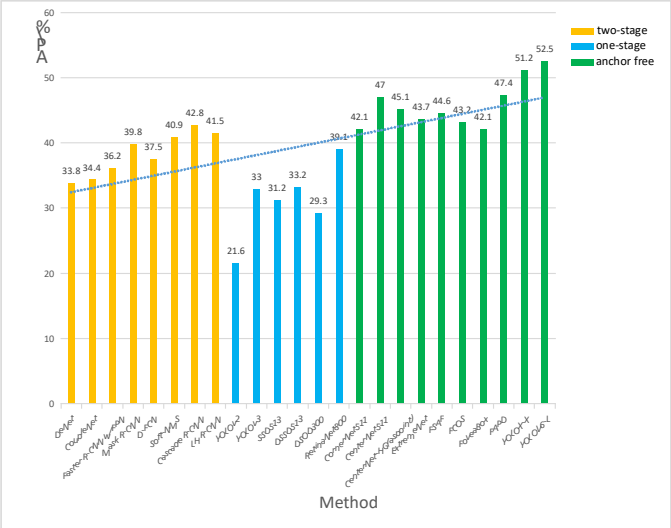

Figure 1. Performance (AP) comparison chart for each type of model.

In fact, anchor free is not a new concept, it can be traced back to the YOLOV1 algorithm, [2]which is the earliest anchor-free model, while recent anchor-free methods are mainly divided into keypoint-based and Segmentation-based[3]methods.

2. Keypoint-Based Object Detection Model for Anchor Free

In the previous section we mentioned that anchor-free object detection models can be divided into two main types, that is keypoint-based and segmentation-based. in this section we mainly introduce the following two keypoint-based anchor-free object detection models

2.1 CenterNet

CenterNet (CenterNet: Keypoint Triplets for Object) proposed by kaiwen duan et al. in CVPR2019[4] can be understood as an improved version of CornerNet. The main idea is to use Hourglass as the backbone to extract the image features. using Cascade Corner Pooling as the module shown in Figure 2(b) to extract the Corner heatmaps of the image. The bounding box of the object is obtained based on the upper left and lower right corner points. All bounding boxes define a central region. Then the center point of the object will further filter the extracted bounding box: if no center point exists in the middle region of the box, this box is considered unreliable; if a center point falls in this center region, this bounding box is retained and the score of its bounding box becomes the average of three points. Using the Center Pooling module as shown in Figure 2(a), the Center heatmap of the image is extracted, and all object centroids are obtained according to the Center map.

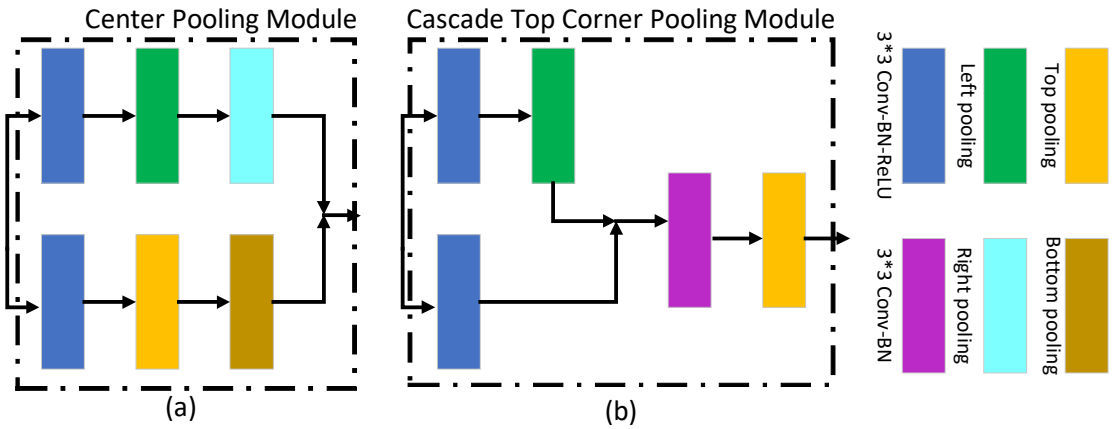

Figure 2. The structures of the center pooling module (a) and the cascade top corner pooling module (b).

Performance: CenterNet has improved a lot in speed and accuracy. On the COCO dataset, CenterNet achieves 47.0% AP, which is higher than any existing one-stage detectors.

2.2 CenterNet (Objects as Points)

CenterNet was proposed by xingyi zhou et al. in CVPR2019[5]. The Centernet network mainly consists of Resnet50[6] to extract image features, followed by the deconvolution module Deconv (three deconvolutions) to upsample the feature map, and finally, three branch convolution networks are used to predict the heatmap, the width and height of the object and the centroid coordinates.

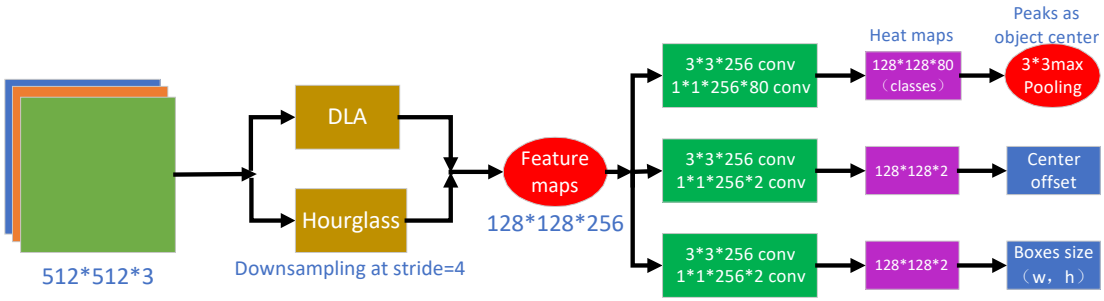

Figure 3. Network architecture of Center Net (Objects as Points).

Figure 3 shows the network structure of CenterNet, it can be seen that CenterNet scales the image to 512x512 size (long side scaled to 512, short side filled with 0), then the scaled 1x3x512x512 image into the network image after resnet50 extract features to get feature1 size 1x2048x16x16, feature1 after deconvolution module Deconv, three times upsampling to get feature2 size 1x64x128x128. feed feature2 into three branches for prediction, predicted heatmap size 1x80x128x128 (indicates 80 categories), predicted length and width size 1x2x128x128 (2 denotes length and width), and the predicted centroid offset size is 1x2x128x128 (2 denotes x, y)

Performance: Using Hourglass-104 as the reference network not only can obtain 40.4% AP but also can obtain 14FPS; the AP and FPS obtained after using DLA-34[7] as the reference network can reach a compromise between accuracy and speed.

3. Segmentation-Based Object Detection Model for Anchor Free

The CenterNet (Objects as Points) mentioned in the previous section already has similarities with the segmentation-based model, which also uses the pixel points of the feature map as the location of a detection box and then predicts them directly using full convolutional branching.

3.1 FSAF

FSAF was proposed by Chenchen Zhu et al. in CVPR 2019[8]. The authors propose an anchor-free mechanism-based feature selection (FSAF) module. As a one-stage component, it can be combined with a feature pyramid embedded in a one-stage detector for online feature selection.

The FSAF module allows each instance to automatically select the most appropriate feature layer in the FPN. In this module, the basis of feature selection is changed from the original instance size to the instance content, which enables the model to automatically learn to select the most appropriate feature layer in the FPN.

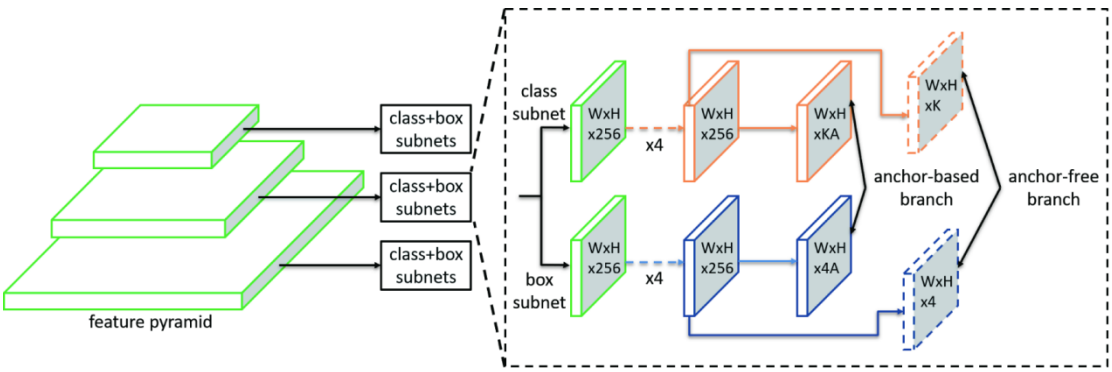

Figure 4. Network architecture of RetinaNet with FSAF module. Reproduced from ref. 28 with permission from the IEEE. copyright 2023.

As shown in Figure 4, FSAF uses RetinaNet[9] as the base structure, and adds a branch of FSAF in parallel with the original classification subnet and regression subnet to achieve full end-to-end training without changing the original structure. The FSAF also contains two branches: classification (using the sigmoid function) and box regression, which are used to predict the class and coordinate values of the object. For regression, it is the prediction of the 4 positions offset mapping. In the inference, FSAF can output the prediction results as a branch alone, or simultaneously with the original anchor-based branch. When both branches are present, the outputs of both branches are combined and the final prediction is obtained using NMS.

Performance: Its accuracy is higher than many one-stage detectors, and it can achieve 44.6% AP on the COCO dataset using the ResNeXt-101 network.

3.2 FCOS

FCOS is a paper by Zhi Tian et al. at CVPR2019[10]. FCOS designs a simpler and more flexible framework by eliminating anchor frames. It solves the object detection problem at the pixel level prediction, using FPN to substantially reduce the ambiguity of the frames of the pixel regression at the overlap, and proposes a new "center-ness" branch to reduce the weight of low-quality detection frames.

The open root is used to slow down the decay of centrality, which ranges from 0 to 1. Therefore, binary cross-entropy (BCE) loss[11] can be used for training. For inference, the final score (used to rank the predicted bounding boxes) is the product of the classification score and the centrality.

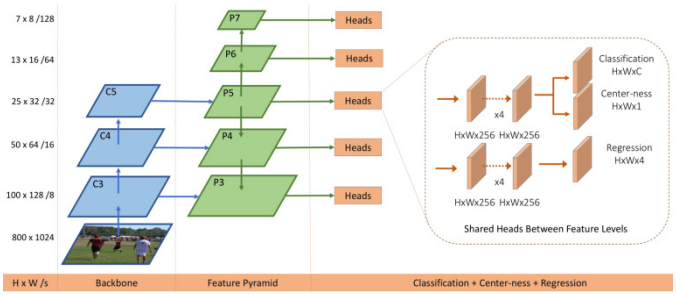

Figure 5. The network architecture of FCOS.Reproduced from ref. 30 with permission from the IEEE. copyright 2023.

As shown in Figure 5, according to the FPN, objects of different sizes are detected on different levels of feature maps. P3, P4, and P5 are obtained from the feature map C3, C4, and C5 of the backbone CNN after a 1×1 convolutional lateral join, and P6, P7 are obtained from P5, P6 after a stride=2 convolutional layer, respectively. So, the final obtained P3, P4, P5, P6, and P7 correspond to stride=8,16,32,64,128, respectively. Then, the regression range of the bounding box is directly restricted on the different feature layers. Finally, the Head network is shared on different feature layers. If the position (x, y) falls inside any real box (x0, y0, x1, y1) (top left and bottom right), it is considered a positive sample, and its category is labeled as the category of this real box; otherwise, it is considered a negative sample.

Performance: FCOS outperforms existing one-stage detectors, while FCOS can also be used as an RPN in the two-stage detector Faster RCNN and is largely superior to the RPN. the improved version of FCOS can achieve an AP of 44.7%.

4. YOLO Series

YOLO series can be said to be the masterpiece of one-stage object detection, as mentioned in the first section, YOLOV1 is an anchor-free model, but its subsequent versions, whether Joseph Redmon or Alexey

Bochkovskiy, still introduced anchor. With the rapid development of the anchor-free model in recent years, recent versions of YOLO such as YOLOv8 do not use anchor.

YOLOv8's anchor box principle is based on predefined bounding boxes that predict the location and size of the target. An anchor box is a set of predefined bounding boxes, each of which corresponds to a position on the feature map of the network output. By predicting the offset and scale of the anchor frame, the accurate position and size of the target in the image can be determined, thus improving the detection accuracy.

5. Conclusion

In this literature survey, we mainly introduce the current novel and popular anchor-free object detection model from three aspects. There are two Keypoint-based anchor-free object detection models. The biggest advantage of this type of anchorless model is that the fast detection speed of the detector greatly reduces the time and computational power required. Then four Segmentation-based anchor free models FSAF and FCOS are introduced. The greatest credit for this class of anchor-free models to rival anchor-based methods in accuracy should be attributed to FPN, followed by Focal Loss. Finally, YOLO models using anchor free are presented. They also do not use anchors and are far ahead of their predecessors in both accuracy and speed. Although these models have performed quite well, they have not yet shown the full potential of object detection.Due to the large number of applications of artificial intelligence in various fields, there is still a long way to go before object detection faces many problems and challenges in the future.

Statements and Declarations

I solemnly declare: I abide by academic ethics, advocating rigorous style of study. The graduation thesis submitted is the result of my independent research under the guidance of my supervisor. This paper does not contain anything that has been published or written by others, except as expressly stated and quoted in the paper. The paper is written by myself and I am responsible for the content written.

Conflict of Interest

There is no conflict of interest

Funding

This research was funded by the National Natural Science Foundation of China (No.62264015);the Foundation of the Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data(Grant No.IPBED17),Teaching Reform Project of Yan'an University, China (Grant No. YDJG23-34).

References

[1] Ren S Q, He K M, Girshick R,Sun J (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39:1137-1149

[2] Redmon J, Divvala S, Girshick R, Farhadi A,Ieee (2016) You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),PP 779-788

[3] Li B, Shi Y, Qi Z Q,Chen Z S (2018) A Survey on Semantic Segmentation. In: 18th IEEE International Conference on Data Mining Workshops (ICDMW),PP 1233-1240

[4] Duan K W, Bai S, Xie L X, Qi H G, Huang Q M, Tian Q,Ieee (2019) CenterNet: Keypoint Triplets for Object Detection. In: IEEE/CVF International Conference on Computer Vision (ICCV),PP 6568-6577

[5] Zhou X, Wang D,Krhenbühl P (2019) Objects as Points. arXiv:1904.07850

[6] He K, Zhang X, Ren S,Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),PP 770-778

[7] Yu F, Wang D, Shelhamer E,Darrell T (2018) Deep Layer Aggregation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,PP 2403-2412

[8] Zhu C C, He Y H, Savvides M,Soc I C (2019) Feature Selective Anchor-Free Module for Single-Shot Object Detection. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),PP 840-849

[9] Lin T Y, Goyal P, Girshick R, He K,Dollár P (2020) Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 42:318-327

[10] Tian Z, Shen C, Chen H,He T (2019) FCOS: Fully Convolutional One-Stage Object Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV),PP 9626-9635

[11] Qin X, Zhang Z, Huang C, Gao C, Dehghan M,Jagersand M (2019) BASNet: Boundary-Aware Salient Object Detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),PP 7471-7481

[12] He K, Gkioxari G, Dollar P,Girshick R (2017) Mask R-CNN. arXiv:1703.06870v3

[13] He X, Zhao K,Chu X (2019) AutoML: A Survey of the State-of-the-Art. arXiv:1908.00709

[14] Zoph B,Le Q V (2016) Neural Architecture Search with Reinforcement Learning. arXiv:1611.01578

[15] Kong T, Sun F, Liu H, Jiang Y, Li L,Shi J (2020) FoveaBox: Beyound Anchor-Based Object Detection. IEEE Transactions on Image Processing 29:7389-7398

[16] Iandola F N, Han S, Moskewicz M W, Ashraf K, Dally W J,Keutzer K (2016) SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360

[17] Zhu C, Chen F, Shen Z,Savvides M (2019) Soft Anchor-Point Object Detection. arXiv:1911.12448