Anchor Assignment in Object Detection

Subjects: Computer Science, Artificial Intelligence

Contributor: Ping Han , Xujun Zhuang , Huahong Zuo , Ping Lou , Xiao Chen

Due to the advantages of edge computing in terms of latency, bandwidth, and security, it provides more effective technical support for many applications that require real-time security. A security system is one of the applications for edge computing in Internet of Things (IoT), where the core end device is a camera for surveillance and smart security. Smart security based on object detection is one of the important applications of edge computing in IoT. Anchors in object detection refer to points on the feature map, which can be used to generate anchor boxes and serve as training samples. The training samples in object detection use anchors, which are feature points on the feature map that can be used for classification and regression.

anchor assignment object detection loss aware

1. Introduction

Due to the advantages of edge computing in terms of latency, bandwidth, and security, it provides more effective technical support for many applications that require real-time security. A security system is one of the applications for edge computing in IoT, where the core end device is a camera for surveillance and smart security. Intelligent security systems require accurate and real-time processing of images and video from cameras, which poses a challenge to existing object detection technologies.

The training samples in object detection use anchors, which are feature points on the feature map that can be used for classification and regression. In anchor-based detection models (e.g., RetinaNet [1]), multiple anchor boxes can be tiled with anchors as the center, while in anchor-free models (e.g., FCOS [2]), anchors can be used as anchor points to directly predict regression boxes. In recent years, although great breakthroughs have been made in object detection, there are still some shortcomings in the definition of positive and negative samples in current object detection. As the aspect ratio of the ground-truth (GT) boxes are not fixed, it is difficult for the ordinary center prior to select suitable anchors as positive samples. Moreover, the distribution of objects within a GT box is uncertain, there is often a large amount of background within the labeled GT box, and there may also be interference, such as occlusion. In this case, the traditional approach based on fixed Intersection-over-Union (IoU) assignment [1] does not select better positive samples, which poses a challenge for positive and negative sample assignment. To address the problems posed by fixed IoU, [3] uses the statistical characteristics of IoU between the candidate anchor boxes and a GT box to give the algorithm a different IoU threshold. Ref. [4] proposes a new anchor quality evaluation scheme and uses the quality scores of anchors fitting a probability distribution to achieve a truly dynamic assignment; however, the computational effort is large and reduces the training speed.

2. Anchor Assignment in Object Detection

2.1. Anchor-Based and Anchor-Free Detectors

Object detection is a fundamental yet challenging task in computer vision, requiring the model to predict a bounding box with a category label for each target object in the image. Currently, deep-learning-based object detection algorithms dominate and can be divided into anchor-based and anchor-free approaches according to whether an anchor box is predetermined. Anchor-based object detection algorithms can be classified into two-stage methods (e.g., R-CNN [5], Fast R-CNN [6], Faster R-CNN [7], and R-FCN [8]) and single-stage methods (e.g., SSD [9] and RetinaNet [1]) according to whether candidate boxes are generated or not. These methods first tile a large number of pre-defined anchor boxes on the image, then predict the category, refine the coordinates of these anchor boxes one or more times, and finally output these refined anchor boxes as prediction boxes. The two-stage methods are slow but accurate; however, the advent of FPN [10] and Focal Loss [1] has enabled single-stage methods to achieve the same accuracy as two-stage methods at a faster speed.

Anchor-based object detection methods require artificially setting parameters such as the scale and aspect ratio of the anchor box according to the characteristics of the dataset, resulting in parameter sensitivity and poor generalization of the detector. Therefore, anchor-free detectors have gradually become a research hotspot. Anchor-free detectors can be divided into center-based methods (e.g., FCOS [2]) and keypoint-based methods (e.g., CornerNet [11] and ExtremeNet [12]). Anchor-free detectors can directly predict objects without pre-defined anchor boxes, eliminating the hyperparameters associated with anchor boxes and achieving similar performance to anchor-based detectors, while also giving them more potential in terms of generalization capability.

The anchor point In FCOS is equivalent to the center of the anchor box in RetinaNet. An anchor point and the corresponding anchor box both correspond to the same location on the feature map. Therefore, they are collectively referred to as an anchor in this research when there is no need to distinguish between an anchor point and anchor box.

2.2. Loss Function in Object Detection

Losses in object detection models typically include classification losses and regression losses, corresponding to the classification and regression branches, respectively. Calculating the loss based on the ground truth and the predicted values output by the detection head allows back-propagation to update the model parameters. The main classification losses commonly used in object detection models are OHEM [13], Focal Loss [1], GHM [14], Quality Focal Loss [15], and Varifocal Loss [16], and the main regression losses are IoU Loss [17], Distribution Focal Loss [15], GIOU Loss, and DIoU and CIoU Loss [18]. In addition, some object detection models add a parallel auxiliary branch to the regression branch, such as the centerness branch in [2] and the IoU branch in [4][19][20]. These branches predict a centerness or IoU for each anchor and can also calculate the loss from the true value, which is usually calculated using binary cross entropy (BCE) loss.

2.3. Anchor Assignment in Object Detection

Positive and negative sample assignment refers to the process of determining whether each anchor should be assigned as a positive or negative sample during training, a process also known as training sample selection, anchor assignment, and label assignment. Anchor assignment during training is an important factor affecting the performance of the object detection model. **Figure 1**a shows the anchor inside the GT box. The white box is the GT box and the different colors of anchors represent the high or low quality of the anchors. The red anchors are located in the background rather than the object and should be classified as negative anchors. The yellow anchors contain part of the object information and hence are not optimal positive anchors. Green anchors contain rich object information and should be used as positive anchors. Anchor assignment is used to pick out the green anchors and avoid the red anchors. The white box in **Figure 1**b is the GT, the green box is one of the anchor boxes corresponding to the green anchor, and the red box is their intersection area. By calculating the proportion of the intersection area to the concatenation of the two boxes, the IoU of the anchor box and the GT box can be obtained. People hope that the IoU of the predicted box and the GT box is as large as possible, rather than the IoU of the anchor box and the GT box.

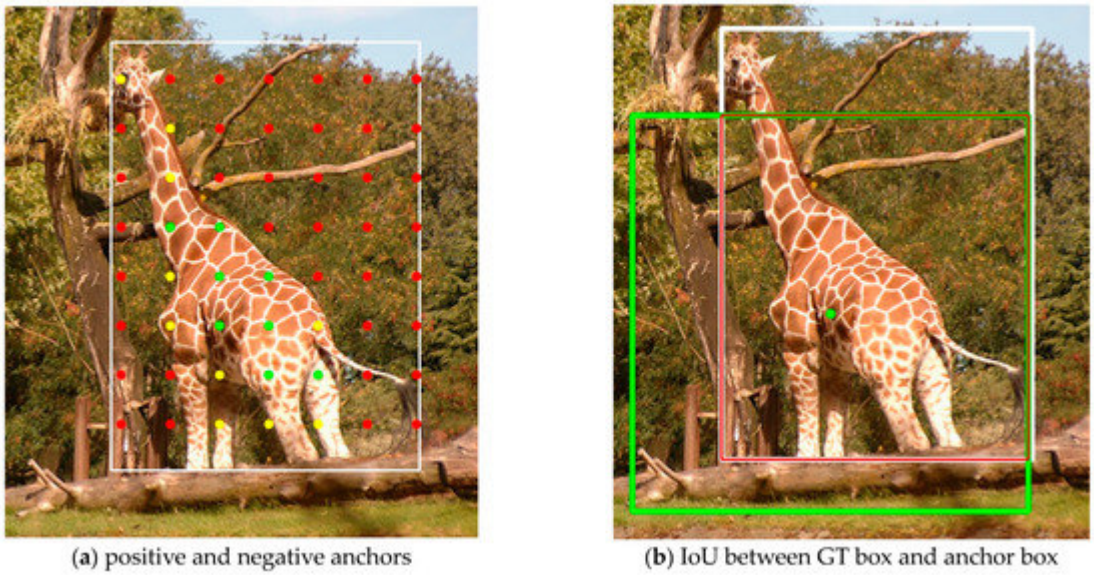

Figure 1. Example of anchors and the IoU of a GT box. (**a**) The white box is the GT box. The red dots are anchors that need to be assigned as negative anchors. The yellow dots are anchors that are better than red anchors, but not optimal positive anchors. The green dots are anchors which are expected to be assigned as positive anchors. (**b**) The white box is the GT box, and the green box is one of the anchor boxes of the green anchor. The red box is the intersection area between these two boxes.

Traditional positive and negative sample assignment strategies use manually designed hard assignments. Anchor-based detection models (e.g., RetinaNet [1]) empirically lay nine anchor boxes of different scales and aspect ratios at each location, and use a fixed IoU threshold to classify the anchor boxes into positive, ignored, and negative samples. This assignment requires multiple anchor boxes to be tiled per location, with many hyperparameters and a large computational cost. Anchor-free detection models (e.g., FCOS [2]) discard the tiling of anchor boxes and assign positive and negative samples using spatial constraints (i.e., restricting positive anchor points to inside the GT boxes) and scale constraints (i.e., setting a fixed maximum regression range for each feature layer). The

subsequent FCOS [21] uses center sampling on top of the original constraints. The shortcoming that the above hard-assignment approaches all have in common is that they do not take into account the fact that the most meaningful anchors are not equally distributed within the GT boxes for objects of different sizes, shapes, or occlusion conditions, and the conditions for dividing positive and negative samples are not the same.

Due to the inadequacy of hard-assignment approaches, many algorithms for adaptively assigning anchors have emerged. For example, MetaAnchor [22] and Guided Anchoring [23] argue that tiled anchor boxes are not optimal. MetaAnchor uses the anchor function to dynamically generate anchor boxes from arbitrary customized prior boxes, thus changing the shape of anchor boxes during training. Guided Anchoring proposes the Guided Anchoring Region Proposal Network (GA-RPN), which generates anchor boxes through additional network modules to better fit the distribution of various objects. Both MetaAnchor and Guided Anchoring modify the structure of the model itself to varying degrees. FreeAnchor [24] defines the training process as one of maximum likelihood estimation based on classification loss and regression loss, and constructs a candidate set of anchors for each GT box. However, when a GT box has many high-quality anchors, this approach does not match the GT box well to the appropriate anchor.

ATSS [3] uses the sum of the mean and standard deviation of IoU values from a set of the closest anchors as the dynamic threshold. However, as the anchor boxes are constant, for the same GT box, this threshold is invariant during the training process. Moreover, using L2 distance is not reasonable for some GT boxes with a large difference in length and width. NoisyAnchors [25] proposes cleanliness scores, using soft labels and re-weighting based on classification and localization losses, which mitigates the effect of noise in the anchors but fixes the number of positive anchors throughout the training process. MAL [26] uses linear scheduling to reduce the number of positive samples as training proceeds but is prone to fall into suboptimal solutions and requires heuristic feature perturbations to mitigate this issue. PAA [4] assumes that the joint loss distribution of positive and negative samples follows the Gaussian distribution, and uses the Gaussian Mixture Model (GMM) to cluster the candidate positive samples to obtain the final positive samples. This approach does not make use of the shape of the GT box. In addition, the GMM requires continuous iteration, which is computationally intensive and extends the training time.

References

1. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
2. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635.
3. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, Virtual, 13–19 June 2020; pp. 9756–9765.

4. Kim, K.; Lee, H.S. Probabilistic Anchor Assignment with IoU Prediction for Object Detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 355–371.
5. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
6. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
8. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
10. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–16 July 2017; pp. 2117–2125.
11. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
12. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 850–859.
13. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
14. Li, B.; Liu, Y.; Wang, X. Gradient harmonized single-stage detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8577–8584.

15. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 21002–21012.
16. Zhang, H.; Wang, Y.; Dayoub, F.; Sunderhauf, N. Varifocalnet: An iou-aware dense object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 20–25 June 2021; pp. 8514–8523.
17. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.
18. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
19. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of localization confidence for accurate object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 18–23 June 2018; pp. 784–799.
20. Goldman, E.; Herzig, R.; Eisenschtat, A.; Goldberger, J.; Hassner, T. Precise detection in densely packed scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5227–5236.
21. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A Simple and Strong Anchor-Free Object Detector. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 44, 1922–1933.
22. Yang, T.; Zhang, X.; Li, Z.; Zhang, W.; Sun, J. MetaAnchor: Learning to detect objects with customized anchors. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 2–8 December 2018; pp. 318–328.
23. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region proposal by guided anchoring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2965–2974.
24. Zhang, X.; Wan, F.; Liu, C.; Ji, R.; Ye, Q. FreeAnchor: Learning to match anchors for visual object detection. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 147–155.
25. Li, H.; Wu, Z.; Zhu, C.; Xiong, C.; Socher, R.; Davis, L.S. Learning from noisy anchors for one-stage object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 13–19 June 2020; pp. 10588–10597.
26. Ke, W.; Zhang, T.; Huang, Z.; Ye, Q.; Liu, J.; Huang, D. Multiple anchor learning for visual object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition, Virtual, 13–19 June 2020; pp. 10206–10215.