Fuzzy Systems and Data Mining VIII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220366

Improved Attention Mechanism-Based Object Detection Method

Jing ZHOU [a,1] and Ze CHEN [a]
[a] *School of Artificial Intelligence, Jianghan University, China*

Abstract. Aiming at the problems of low detection accuracy for objects with non-significant features in the FCOS network, a new object detection method based on attention mechanism is proposed to improve the performance of FCOS network, which can effectively guide the network to focus on the detailed features. According to the verification experimental results on KITTI dataset, the AP value of the improved attention mechanism-based method for car and person detection is improved by 1.1% and 4.9% compared with the standard FCOS network, respectively, and the average category accuracy value is improved by 3%. Thus, the experimental results show the effectiveness of the proposed method.

Keywords. object detection, attention mechanism, deep learning network, KITTI

1. Introduction

Recently, the deep neural network simulating human brain intelligence has made revolutionary progress in the field of visual tasks [1-3] and can realize 2D object detection, recognition and tracking [4][5]. The traditional 2D object detection methods based on deep-learning mainly regress object position via anchor mechanism, such as SSD [6], YOLOX [7], RetinaNet [8], Mask RCNN [9], YOLOF [10], etc.

However, the detection methods with anchor mechanism such as FasterRCNN [4], SSD [6], consume too much computation and memory, and are difficult to deal with the objects that vary greatly in scale and the small objects with non-significant features due to the fixed size and aspect ratio of anchor box. Therefore, the anchor-free methods have attracted extensive attention recently. Compared with the anchor-based methods, the anchor-free detectors have no hyper-parameters associated with the anchor box and their frameworks are more simplified. Thus, they will outperform anchor-based methods with respect to detection performance.

As a typical anchor-free algorithm, FCOS [11] network abandons anchor box and performs detection and regression directly on pixels. Meanwhile, combined with feature pyramid network, FCOS network can extract multi-scale object features and achieves more robustness. The weighting coefficient of center-ness is adopted in FCOS network, and the score of the object is product of classification score and center-ness coefficient. The weight of low-quality object can be reduced by center-ness coefficient and then the object can be easily excluded by Non-maximum suppression (NMS) operation, which effectively improves the detection accuracy. However, the weight of detail feature of

[1] Corresponding Author: Jing Zhou, School of Artificial Intelligence, Jianghan University, Wuhan, 430056, Hubei, China. Email: zhj131@jhun.edu.cn.

object is also easily reduced, which damages the detection accuracy. Aiming at this issue, a network is designed to focus on the key information of the object by the attention mechanism in this work, thus, effectively improving feature extraction ability of the network and achieving higher detection accuracy on the testing dataset.

2. FCOS object detection method

2.1. Network structure

FCOS is a full-convolution and pixel-level object detection network, and it is consisted of backbone, Feature Pyramid Network (FPN) and detection head module, as shown in Figure 1.

The residual network ResNet-50 [12] is taken as the backbone for FCOS network. The object is passed through three convolution layers of Conv3, Conv4 and Conv5 in ResNet-50 to extract features, and the extracted multi-scale features $\{C_3, C_4, C_5\}$ are fed into the feature pyramid FPN [13] module.

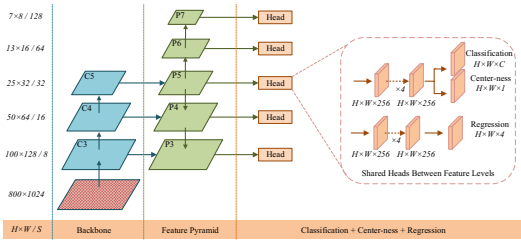


Figure 1. Structure of FCOS model.

The objects with different sizes can be detected based on the feature maps from different layers in FPN. In Fig. 1, the extracted feature maps of five layers are respectively denoted as {P3, P4, P5, P6, P7}, where P3, P4 and P5 are derived from the backbone feature maps of C3, C4 and C5. The feature maps P6 and P7 of the feature pyramid are respectively obtained from P5 and P6 through a convolution layer with the stride 2. Finally, the feature maps of P3, P4, P5, P6 and P7 are obtained with the corresponding stride of 8, 16, 32, 64, 128, respectively. The feature map obtained by FPN is input to the detection head module, which is composed of two branches for object classification and position regression.

2.2. Loss function

Object detection of FCOS network is realized through the classification of foreground objects and the regression for object location, and the loss of the network is consisted of classification and regression loss. By adopting center-ness strategy to the loss function, the FCOS network can accurately filter the negative samples far from the object.

During classification process, the coordinate of each position (x, y) at feature map F_i can be pushed back to $(\lfloor \frac{s}{2} \rfloor + xs, \lfloor \frac{s}{2} \rfloor + ys)$ on the original map, which is near the center of receptive field at position (x, y). If the coordinate falls in any ground-truth bounding box, the feature map will be classified as the positive sample, whose category c^* belongs to B_i. Otherwise, it is classified as a negative sample, whose category c^*

belongs to the background. The symbol F_i represents the feature map of i_{th} layer in CNN, and the s denotes total stride from original image to the i_{th} layer map. The set of bounding boxes labels is $\{B_i\}$, in which each bounding box is described by the coordinate (x_0, y_0) of the top-left corner and (x_1, y_1) of the lower-right corner, that is, $B_i = \left(x_0^{(i)}, y_0^{(i)}, x_1^{(i)}, y_1^{(i)}, c^{(i)}\right) \in \mathbb{R}^4 \times \{1,2 \dots C\}$.

Besides classification label, the regression label can be set as an offset vector $t^*=(l^*, t^*, r^*, b^*)$, where (l^*, t^*, r^*, b^*) respectively represents the distances from the predicted box to the four boundary lines of the bounding box, as denoted in Eq. (1).

$$\begin{aligned} l^* &= x - x_0^{(i)}, \quad t^* = y - y_0^{(i)} \\ r^* &= x_1^{(i)} - x, \quad b^* = y_1^{(i)} - y \end{aligned} \tag{1}$$

Where (x, y) represents the centroid of the predicted box, the (x_0, y_0) and (x_1, y_1) represent the boundary coordinates of ground truth box.

Given regression objectives l^*, t^*, r^*, b^*, the definition of center-ness coefficient is defined in Eq. (2).

$$centerness^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\min(t^*, b^*)}}. \tag{2}$$

The center-ness coefficient is varied from 0 to 1, and binary cross entropy loss is applied for training, then the score of the detected object is product of classification score and center-ness coefficient. The center-ness score of the feature close to the object is higher, whereas the feature far away from the object has a lower score, that is, center-ness coefficient can effectively weaken the weight of object edge and improve the detection performance.

As illustrated in Eq (3), total loss of FCOS network can be obtained based on the classification and regression loss.

$$L\left(\{p_{x,y}\}, \{\{t_{x,y}\}\}\right) = \frac{1}{N_{pos}} \sum_{x,y} L_{cls}(p_{x,y}, c^*_{x,y}) + \frac{\lambda}{N_{pos}} \sum_{x,y} \mathbb{I}_{\{c^*_{x,y}>0\}} L_{reg}(t_{x,y}, t^*_{x,y}) \tag{3}$$

To address sample imbalance, classification loss L_{cls} in Eq. (3) is designed as focal loss, and the regression loss L_{reg} is designed as IOU loss. Meanwhile, N_{pos} represents the positive sample quantity, and λ is given as 1 to balance two types of losses. The sum calculation is carried out on the whole feature map, where $\mathbb{I}_{\{c^*_{x,y}>0\}}$ represents the indication function and equals to one when $c^*_{x,y} > 0$, otherwise equals to zero.

The FCOS network focuses on the object whose center point is in the center of grid, and the object without the grid center will be easily ignored, indicating that the capability of FCOS network for extracting the detailed object features is insufficient.

3. FCOS object detection network

3.1. Attention mechanism

To enhance the feature extraction capability of FCOS network, visual-attention mechanism is applied on the network to focus on the local details of object. The essence of attention mechanism is to locate the key feature regions that represent significant differences between different objects. By learning the weight distribution of image features, the original features are weighted so that more attention can be focused on the details of the image.

In this paper, the attention mechanism is imported to detect the significance of object, and different weights are assigned to each pixel of the feature map. Greater weight is

assigned to detail features, so that more attention is paid to the positive sample pixels, and then the detection accuracy can be effectively improved.

The effective mechanism of channel attention is firstly proposed in Squeeze and Excitation (SE) network to establish the correlation among features [14], and then a corresponding SE block is constructed to perform feature recalibration.

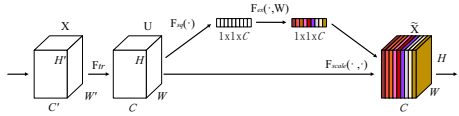


Figure 2. Squeeze-and-Congestion attention structure

The standard structure of SE block is shown in Fig. 2. The feature X is subjected to a given set of convolution transformation $F_{tr}: X \to U, X \in \mathbb{R}^{W' \times H' \times C'}, U \in \mathbb{R}^{W \times H \times C}$ to get the feature U, and then the feature U is aggregated to generate a single-channel 1x1xC feature map through global pooling, which can form the channel weight of the original feature map by the activation function. The channel weight describes the global weight distribution of the channel feature. By learning the weight, the knowledge from the global receptive field of the network is utilized by lower levels and the feature mapping U is re-weighted to obtain the output of the SE block, which is fed directly to subsequent layers, so that the network can focus more attention on the detailed features of object through weighted feature mapping.

The SE attention block can be directly added in the network architecture, and the ability of the feature recalibration performed by SE blocks can be improved by calling the attention module multiple times.

3.2. Improved attention-based object detection network

To promote the performance of FCOS network, the improved detection method is proposed in this work by the attention mechanism.

The architecture of the backbone network ResNet-50 of FCOS network and the basic module bottleneck are shown in Fig. 3. The ResNet-50 structure shows that the Bottleneck module is invoked multiple times.

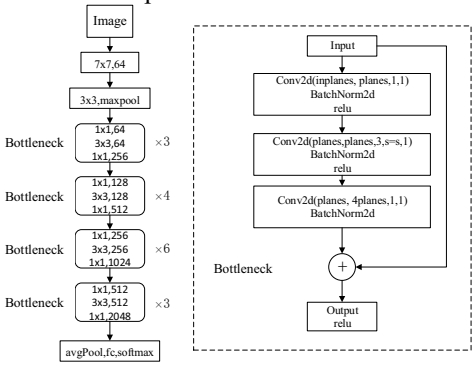


Figure 3. Resnet-50 structure and bottleneck module

The basic module bottleneck can be improved by introducing SE attention block, and the architecture of improved FCOS model is designed as illustrated in Fig. 4, which is combined with the attention mechanism-based SE block and the ResNet-50 network.

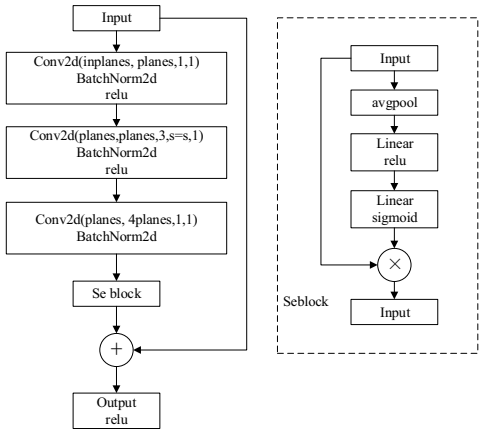


Figure 4. Improved Bottleneck structure

According to the attribute of attention mechanism, the learning ability for significance feature of the network can be enhanced by invoking SE block of bottleneck multiple times.

4. Experiment

4.1. Experiment Setting

The programming language adopted in the experiment is pytorch 1.3.0 and python 3.7.4. The operating environment is based on the GPU of GeForce GTX 1660Ti 6GB, and memory of RAM 16GB, and the CPU model is Intel ® Core ™ i5-9300H CPU@2.40GHz x8.

Meanwhile, the optimizer of the network model is SGD with the momentum of 0.9, and the weight decay coefficient equals to 0.0001, the epoch equals to 60, and batch size is set to 1. The initial learning rate is set as 0.001, and WARMPUP_STEPS parameter is set to fine tune the learning rate, and the parameter of GLOBAL_STEPS is set as the number of trained steps.

The learning rate tuning rules are formulated as follows:

```
if GLOBAL_STEPS < WARMPUP_STEPS:
    lr = float(GLOBAL_STEPS / WARMPUP_STEPS * LR_INIT)
if GLOBAL_STEPS == 20001:
    lr = LR_INIT * 0.1
if GLOBAL_STEPS == 27001:
    lr = LR_INIT * 0.01
```

4.2. Evaluation of results

The experiments mainly focus on detecting person and cars in KITTI dataset to evaluate the performance of the network proposed in this work. KITTI is a popular benchmark dataset in automatic driving field at present. The dataset consists of real images collected from urban, rural and highway scenes and contains 7481 pictures, including 5985 training sets and 748 verification sets.

The evaluation criteria of object detection network is the average precision (AP) value, and the definition of AP value is shown as Eq. (4).

$$AP = \int_0^1 p(r)dr \tag{4}$$

Where p is precision probability of the detected object.

The definition of mean of average precision value, that is mAP value, is shown in Eq. (5).

$$mAP = \frac{1}{N}\sum_{i=1}^{N} Ap_i \tag{5}$$

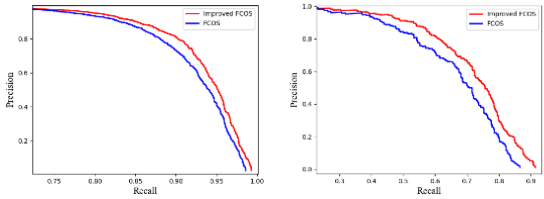
The improved network proposed in this work is tested on KITTI dataset. To verify the detection effect for person and car of the improved network, we compare it with the classical detectors, *e.g.*, Faster R-CNN, SSD and YOLOX. The experimental results are illustrated in Table 1.

Table 1 Comparison of AP value and mAP value with different models

Method	car AP	Person AP	mAP
FCOS	0.917	0.667	0.792
Improved FCOS	0.928	0.716	0.822
Faster-RCNN	0.768	0.609	0.728
YOLOX	0.914	0.722	0.818
SSD	0.865	0.698	0.761

As displayed in Table 1, the AP value of the improved FCOS network for car detection is 0.928, achieving the increasement of 1.1% compared with the standard FCOS. The AP value of person detection is 0.716, achieving the increasement of 4.9% compared to FCOS. And the mAP value is increased by 3%. Meanwhile, the improved FCOS network proposed in this work outperforms other classical detection methods of YOLOX [7], SSD [6] and Faster R-CNN [4] for person and car detection.

The PR curves of person and cars detected by the improved FCOS network and standard FCOS are displayed in Fig. 5. It describes that the improved FCOS method obviously improves the recall rates and accuracy for both person and cars.



a) PR curve of car b) PR curve of person

Figure 5. The PR curves for FCOS models

Meanwhile, the visualization results are displayed in Fig. 6 and Fig. 7, and we observe that the improved FCOS method can detect the truncated object and achieve higher detection precision.

Figure 6. FCOS detection result

Figure 7. Improved FCOS detection result

5. Conclusion

By introducing the attention mechanism-based SE block to the FCOS object detection network, a new attention mechanism-based object detection method is proposed in this work. The improved method can focus on the significant features of the detected object, weaken the subordinate features, and further improve detection performance. Experimental results deliver that the improved FCOS model is superior to YOLOX, SSD and Faster R-CNN detection models, and its performance is significantly improved compared to the standard FCOS method, meanwhile, AP value of each category is obviously promoted.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62106086) and Natural Science Foundation of Hubei Province (No. 2021CFB564).

References

[1] Meerdink S., Bocinsky J., Zare A., et al. Multitarget Multiple-Instance Learning for Hyperspectral Target Detection. IEEE Transactions on Geoscience and Remote Sensing, 2022; 60(1):1-14.
[2] Law H, Deng J. Cornernet: Detecting objects as paired keypoints. Proceedings of the European conference on computer vision (ECCV). 2018. p. 734-750.
[3] Zhou X, Zhuo J, Krahenbuhl P. Bottom-up object detection by grouping extreme and center points. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 850-859.
[4] Ren S., He K., Girshick R., Sun J. Faster RCNN: Towards real time object detection with region proposal networks. IEEE TPAMI, 2017; 39(6): 1137-1149.
[5] Girshick R. Fast R-CNN. In: The IEEE International Conference on Computer Vision (ICCV), 2015. p. 1440-1448.
[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. Computer Vision-ECCV 2016, Cham, 2016. p. 21-37.
[7] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021.
[8] Lin TY, Goyal P, Girshick R., et al. Focal loss for dense object detection. The IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017. p. 2999-3007.
[9] Kaiming H., Georgia G., Piotr D, et al. Mask R-CNN. The IEEE International Conference on Computer Vision (ICCV), 2017: 1-12.
[10] Chen Q, Wang Y, Yang T, et al. You only look one-level feature. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. p. 13039-13048.
[11] Tian Z, Shen C, Chen H., et al. FCOS: Fully Convolutional One-Stage Object Detection, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. p. 1-13.
[12] Kaiming H, Xiangyu Z, Shaoqing R, Jian S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. p. 1-12.
[13] Lin TY, Dollar 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, 2017. p. 2117–2125.
[14] Jie H, Li S, Gang S. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. p. 1-13.