High Perspective Faster R-CNN

Quankai Liu[1,a], Kefeng Li[1,b] and Li Liu[1,c,*]

[1]School of Information Science and Electrical Engineering,
Shandong Jiaotong University Jinan, China
a. 17864196668@163.com, b. 535521861@qq.com, c. 89041604@qq.com
**corresponding author: Li Liu*

Keywords: Faster R-CNN, attention, RPN, respective field.

Abstract: In this paper the original Faster R-CNN model structure was improved to solve the problem of limited receptive fields. The RPN sub-module was combined with an attention layer to get a higher perspective, so that the new model can better analyze the whole picture information and highlight meaningful information. Experiments validate that our method achieves a new improvement for the object detection on PASCAL VOC2007 dataset. And when the change ratio of the middle multi-layer perceptron in the channel attention was set to 1, result reached the highest correct rate of 71%, more than the original model.

1. Introduction

Nowadays, with the development of object detection, there is a growing demand for detector that deal with noise, occlusion and blur. Although the computer vision community has shared many excellent datasets, it is still possible that images have many defects. If we want to improve the performance of our model, the defects of the images themselves are problem we must solve.

At present, object detection algorithms based on deep learning can be roughly divided into two categories: one-stage detector and two-stage detector. One-stage detector, represented by YOLO V3[1], includes SSD[2], RetinaNet[3], etc. The advantage of this kind of detector is its fast operation speed, which is to meet the real-time detection. Two-stage detector, represented by Faster R-CNN[4], includes SSP-Net[5], Fast R-CNN[6], Mask R-CNN[7], etc. High accuracy is the advantage of this detector, which is to meet the high accuracy detection.

In this paper, in order to achieve a better result and avoid the influence above factors, we improve RPN module, to enable it to have a better ability of analyzing problem comprehensively. The new sub-module called “HP-RPN”, which is no longer limited to the part receptive field. This improved model is called “high perspective Faster R-CNN”, (“HPF R-CNN”).

2. Structure of HPF R-CNN

As a two-stage classification model, the Faster R-CNN model is improved with modifying the Region Proposal Network (RPN) sub-module while maintaining the sec-detector module[4]. The overall structure of HPF R-CNN is displayed in Fig.1. Image features are extracted by Vgg16[8]. The feature maps are sent to HP-RPN to get Region Of Interest (ROI). The final classification and the final correction coordinates are obtained in second detector.

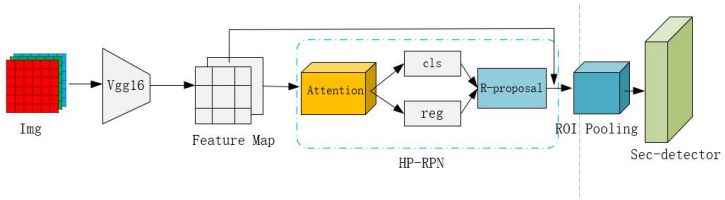

Figure 1: The overview of HPF R-CNN.

Firstly, in HPF R-CNN the images are convoluted and pooled by the Vgg16 to get the feature maps. Vgg16 is a feature extraction network, in which there are 16 convolution layers and full connection layers. The convolution kernel size is 3 * 3, and the stride is 2. A feature map whose size is 1/16 of the original length and width is obtained after Vgg16[8]. Then the feature maps are sent to HP-RPN to be processed using the attention mechanism[9]. These reinforced feature maps have a higher perspective, which can acquire more information and find the key information of the feature maps.

For the HP-RPN module, the feature maps are added attention anchor and they are classified and regressed respectively. The returned feature vectors are operated with Non-Maximum Suppressions (NMS)[10] to reduce invalid vectors, and the results are received by R-proposals. Two new loss values are obtained after the classification and regression, which participate in the calculation of the loss value of the final model later. It will complete the mapping at feature maps to get ROI. When the above operations are completed, the first task of detecting a target object is completed.

Next, R-proposals are sent to ROI Pooling and Sec-detector. In ROI Pooling, feature maps of different scales are sent in for pooling operation, which can ensure that fixed size feature maps are output. The advantage of ROI Pooling is to speed up the training and testing procedure and realize end-to-end training. Lastly, the second detection is carried out by the Sec-detector, in which the feature maps are fully connected and followed by being classified and regressed respectively again. The corresponding category indexes are returned, and the coordinates of the proposal box are further correct through the second regression. When the second coordinate correction is finished, the final coordinate of the proposal box is close to the real value. After one iteration of training, the model returns the final coordinates of the object category index and proposal box. The second task of detecting target category is achieved.

3. HP-RPN

RPN module is one of the critical parts of Faster R-CNN, for the emergence of anchor which greatly improves the recognition accuracy and completes the end-to-end training[4]. According to the features extracted by model, it lays different proportions of anchor on the original image to generate candidate frames matching objects of various scales, which greatly reduces the calculation cost of regional proposal. When the feature is extracted by the extraction network, nine different sizes of anchors are generated on the feature maps with each point as the center. Then 3 * 3 window is slided on the feature map to do convolution operation. Each sliding window generates a 512 dimensional feature vector, which is sent to the classification and regression operations respectively. The confidence of the feature vector as the foreground and the offset from the four coordinates of the dimension box is returned by these two operations. If the confidence of the vector is higher than the threshold, it means foreground. Otherwise, it means background.

However, the disadvantages of the original RPN sub-module are obvious. When judging the anchor, only the information of the receptive field is concerned without including the semantic information of the whole image. This makes it impossible to analyze feature information more

comprehensively in the case of blur or occlusion. In practical application, when the number of picture features is large or the difference is very slight, it will cause inaccurate detection or coordinate positioning deviation.

Therefore, the attention mechanism is involved to solve this problem. The improved RPN sub-module is shown in Figure 2.

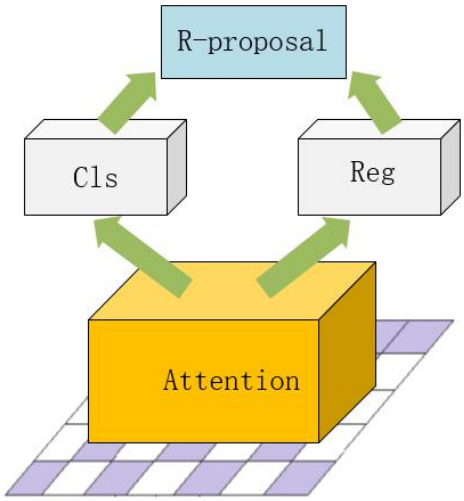

Figure 2: The overview of HP-RPN.

The attention mechanism can effectively avoid the limitation of receptive field and can encourages discrimination between object classes[9]. An attention layer is added to alleviate the conflicts of the diversity of features and find the optimal trade-off between discrimination and diversity[9]. Finally, the improved Convolutional Block Attention Module (CBAM)[11] is added to the HP-RPN sub-module. 9 different shape and size anchors have been generated for each center point on the reinforced feature map. After calculating the estimated coordinates and judgement of whether it is foreground and NMS, about 2000 anchors are selected. They are outputted as R-proposal. The attention model has two sub-modules: channel attention and spatial attention.

When the feature maps are sent to attention layer, they are added channel attention and spatial attention successively to reinforce feature maps. Due to using both average-pooling and max-pooling operations in channel attention, two different spatial context descriptors are generated and combined. Connection with surrounding information is strengthened and meaningful information is highlighted. The two vectors with different weights are added to get the channel attention vector, which has more surrounding information and strengthens meaningful information. Two redistributed vectors are sent to the multi-layer perceptron. In channel attention of original CBAM feature vector is convoluted by the multi-layer perceptron, reduced to 1/2 of the original shape and then restored to the original shape[11]. We adjust the parameters of the multi-layer perceptron to keep the shape unchanged during the convolution, which can reduce information loss during shape transformation. Spatial attention as complementary to the channel attention, corresponding regions of meaningful information are shown. Channel attention focus on more semantic information and spatial attention focus on more position information. Both of channel attention and spatial attention can compute complementary.

Reinforced feature maps are outputted by attention layer, with more surrounding area information and highlighted meaningful information. The main body information gains great weight while other information gains small weights, which help the subsequent classification and coordinate correction.

Reinforced feature maps are sent to classification to judge whether the object is the foreground or background and sent to regression to make the first coordinate correction. The number of overlapping frames is reduced with NMS to the greatest extent, not only greatly reduces the calculation burden, but also improves the accuracy of coordinates. For the training process, 2000 anchors still need to be filtered to 256, calculating coordinate offset and adding positive sample weight again. The coordinates of anchors are normalized for subsequent R-CNN classification and regression. When R-proposal is generated, it is sent to the subsequent module for the second detection.

The HPF R-CNN can be trained end-to-end by optimizing the total loss function, total loss is composed of hp-rpn loss and r-cnn loss. They provide their own classification loss value and regression loss value respectively. The formula is as follows.

$$L_{hp-rpn} = L_{hp-rpn\_cls} + L_{hp-rpn\_reg} \tag{1}$$

$$L_{r-cnn} = L_{cls} + L_{reg} \tag{2}$$

$$L_{total} = L_{hp-rpn\_cls} + L_{hp-rpn\_reg} + L_{cls} + L_{reg} \tag{3}$$

HP-RPN is trained by L_{hp-rpn} to judge whether there is a foreground and make the first coordinate correction. Sec-detector is optimized by L_{r-cnn} to return the target category index and makes a second coordinate correction.

4. Experiment

The experiment of this paper is carried out on the hardware platform of 1080ti graphics card. The software platform is CUDA10.1 and CUDNN7.4. In order to maintain the same environmental conditions as the original model, Vgg16 is still selected as the feature extraction module. The dataset is the PASCAL VOC2007 as the original model. PASCAL VOC2007 includes 9963 pictures and 20 classes. Its training set and test set respectively contain 5011 pictures and 4952 pictures. The major of dataset is picture related to people, which greatly increases the complexity of dataset pictures[12].

Compared with the original model, HPF R-CNN has a better perception of feature details and a higher perspective to obtain more information for image analysis. Figure 3(a), (b), (c) and (d) obviously reflects the advantages of our model compared with the original model. Such as Fig.3(a), (b) and (c), in a complex scene, if the object is small or has no obvious difference from the surrounding features, the original model is difficult to recognize it. However, because of the addition of attention mechanism, our model strengthens the meaningful feature weight and reduce the invalid feature weight. The reinforced feature is more sensitive to subtle changes, which find designated object in a complex or slightly different picture. Figure 3(d) shows another problem of original module that when the target and the target closely fit and the image itself is fuzzy or noise and other factors, it cannot correct the coordinates accurately or even make a wrong judgment. This problem is improved in our model, with the weight allocation avoiding the detection model only focusing on category information. The meaningful information is strengthened and the invalid information is weakened, to highlight the primary information in a large extent. Compared with the traditional results, our model outputs more accurate target coordinates under the condition of correct recognition.

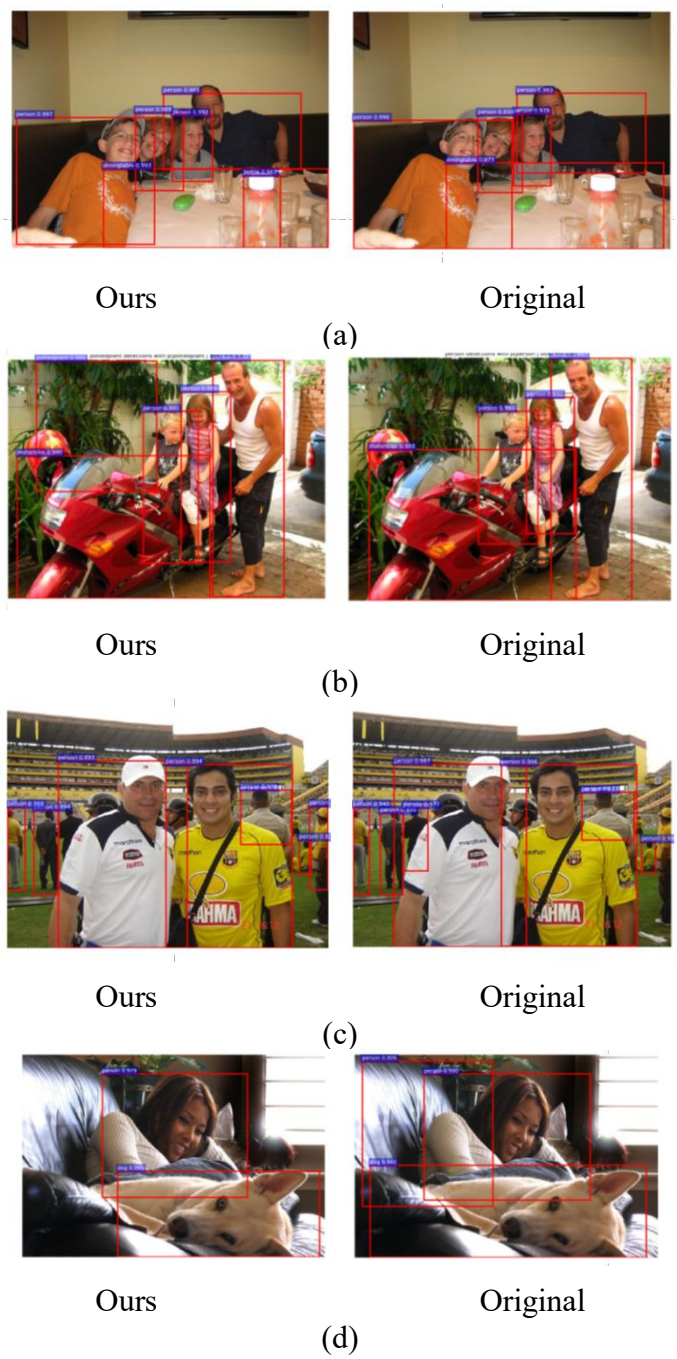

Figure 3: Comparison between HPF R-CNN and the original model.

In order to show the improvement of our model more accurately, the two models are compared on PASCAL VOC2007 test set. The test result is shown in Table 1. Table 1 demonstrates that that the accuracy of detection of small object or object with complex colors is greatly improved. Because the characteristics of large-scale object are very different from those of their surroundings, the promotion is not obvious.

Table 1: Object detection Average Precision(AP).

Method / Class	Faster	HPF	Method / Classes	Faster	HPF
aero	70.0	69.9	dog	80.3	81.5
bike	80.6	78.9	horse	79.8	83.2
bird	70.1	68.7	moto	75.0	76.2
boat	57.3	56.9	person	76.3	77.9
bus	78.2	79.6	plant	39.1	44.0
bottle	49.9	55.7	TV	67.6	72.0
car	80.4	82.0	sheep	68.3	69.7
cat	82.0	84.8	sofa	67.3	68.0
chair	52.2	53.4	train	81.1	75.8
cow	75.3	76.1	mAP	69.9	71.0
table	67.2	65.9			

Our attention mechanism is consisted of channel attention and spatial attention, in which channel attention plays a major role. After feature maps finish average-pooling and max-pooling operations, they are sent to the multi-layer perceptron. There is a ratio in the multi-layer perceptron to control the size of the middle feature. Experiments show that when ratio is taken as 1, the feature information is protected to the greatest extent and the result is the best. The comparison result is shown in Table 2.

Table 2: The results of channel attention with different transformation coefficient.

Ratio / Class	1/2	1	Ratio / Class	1/2	1
aero	70.3	69.9	dog	80.6	81.5
bike	79.8	78.9	horse	83.0	83.2
bird	68.8	68.7	moto	75.7	76.2
boat	58.2	56.9	person	78.1	77.9
bus	77.5	79.6	plant	43.0	44.0
bottle	56.1	55.7	TV	73.9	72.0
car	82.9	82.0	sheep	70.0	69.7
cat	84.1	84.8	sofa	65.3	68.0
chair	53.9	53.4	train	74.9	75.8
cow	74.9	76.1	mAP	70.9	71.0
table	65.8	65.9			

During the experiment, Training times are kept the same as the original 70000 times, and taken 1 to 1/10 of ration for training, and finally gotten the ratio of 1 as the optimal parameter. The attention mechanism in HPF R-CNN not only makes the model focus on that "what" is meaningful given an input image, but also focus on "where" of information. Because of this ability, HPF R-CNN can get more accurate target categories and more detailed and accurate frame coordinates compared with the original model in the face of complex pictures or pictures with indistinct feature differences.

5. Conclusions

HF-Faster R-CNN is presented as a highly efficient model. Compared with traditional Faster R-CNN, it has more powerful performance with an average precision of 71%. The features are weighted by the attention layer, to make the HPF R-CNN more sensitive to meaningful information and not limited by the receptive field. A higher perspective of analyzing images makes our model have better detection results and makes it robust to image defects. Moreover, the proposed model will be investigated for general object detection tasks.

Acknowledgments

This work is supported by National Natural Science Foundation of Shandong Province (ZR2015FL018), ShanDong Key Research and Development Plan (2018GGX101044, 2017GGX201006).

References

[1] Redmon, J.a.F., Ali, Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[2] W. Liu, D.A., D. Erhan, C. Szegedy, S. Reed, C.-Y. and a.A.C.B. Fu, Ssd: Single shot multibox detector. Proceedings of the European conference on computer vision, 2016. 5, 8.

[3] T.-Y. Lin, P.G., R. Girshick, K. He, and P. Dollar., Focal loss for dense object detection. IEEE transactions on pattern analysis and machine intelligence, 2018.

[4] S. Ren, K.H., R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015. 5, 7, 8: p. 91–99.

[5] He Kaiming, Z.X., Ren Shaoqing, Sun Jian SSP-Net：Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE transactions on pattern analysis and machine intelligence, 2015: p. 1904-16.

[6] Girshick, R., Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015. 3, 7: p. 1440-1448.

[7] K. He, G.G., P. Dollar, and R. Girshick, Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on, 2017: p. 2980–2988.

[8] Simonyan, K.a.Z., Andrew, Very deep convolutional networks for large-scale image recognition. Computer Science, 2014.

[9] J. Fu, J.L., H. Tian, Z. Fang, and H. Lu, attention network for scene segmentation. arXiv preprint, 2018.

[10] Neubeck, A.a.G., Luc J. Van, Efficient Non-Maximum Suppression. International Conference on Pattern Recognition (ICPR 2006), 2006: p. 20-24.

[11] S. Woo, J.P., J.-Y. Lee, and I. So Kweon, Cbam: Convolutional block attention module. European Conference on Computer Vision, 2018: p. 3–19.

[12] Everingham, M.E., S. M. Ali， Van Gool, Luc... ThePascalVisual Object Classes Challenge: A Retrospective. International Journal of Computer Vision, 2015: p. 98-136.