Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Object Detection Based on the Improved Single Shot MultiBox Detector

To cite this article: Songmin Jia *et al* 2019 *J. Phys.: Conf. Ser.* **1187** 042041

Object Detection Based on the Improved Single Shot MultiBox Detector

Songmin Jia[1,2], Chentao Diao[1,2*], Guoliang Zhang[1,2], Ao Dun[1], Yanjun Sun[1,2], Xiuzhi Li[1,2] and Xiangyin Zhang[1,2]

[1]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;

[2]Beijing Key Laboratory of Computational Intelligence and Intelligent System.

Email:diaochentao@163.com

Abstract. Aiming at the poor effect of deep learning algorithm on small objects detection, the SSD object detection method based on feature fusion is proposed. The reasons for low detection rate and poor robustness of classical SSD object detection methods are analysed; and through the theoretical analysis and comparative experiments, the characteristic fusion layer was proposed. The shallow layers with high resolution and deep layers with strong semantics are fused with the feature fusion structure; finally, a complete feature fusion structure is designed with the residual block to increase the width and depth of the network. The contrast experiment on the PASCAL VOC dataset was conducted for detection capability and detection accuracy, and experimental result indicates that when the confidence is set to 0.5, the mAP of the SSD method based on feature fusion is 78.04%, which is 0.8% higher than the classical SSD algorithm and 4.8% higher than the Faster RCNN algorithm. Obviously, the proposed algorithm improves the ability of small objects, and verifies the effectiveness of the proposed algorithm.

1. Introduction

In the field of computer vision, object detection is an important research topic. In recent years, various algorithms based on convolutional neural network (CNN) have been applied to object detection tasks, and the detection accuracy and efficiency have been effectively improved [1-5]. However, detecting object at different scales is still a challenging research task. Aiming at the problem of poor performance of multi-scale object detection in current algorithms, the solutions proposed by relevant scholars can be divided into two main categories. One is to extract object features at different scales based on image pyramids to complete multi-scale object detection; the other is to calculate the corresponding feature map of the original image, and then build a feature pyramid on the feature map to complete multi-scale object detection.

Building image feature Pyramid is the basic solution to realize multi-scale object detection at present [6]. These pyramids are scale invariant because the scale changes of objects are canceled out at different levels of the pyramids. Before the deep learning model, manual design of image features Pyramid feature is a popular method [7-9]. This method is particularly important for DPM [10] object detector which relies on a large number of scale sampling. At present, in the task of object recognition, the method of feature extraction using the CNN has replaced the method of artificial design feature

[11-12]. In the field of object detection, the CNN has better detection effect than traditional manual design features, but the detection performance still needs to be improved [13, 14].

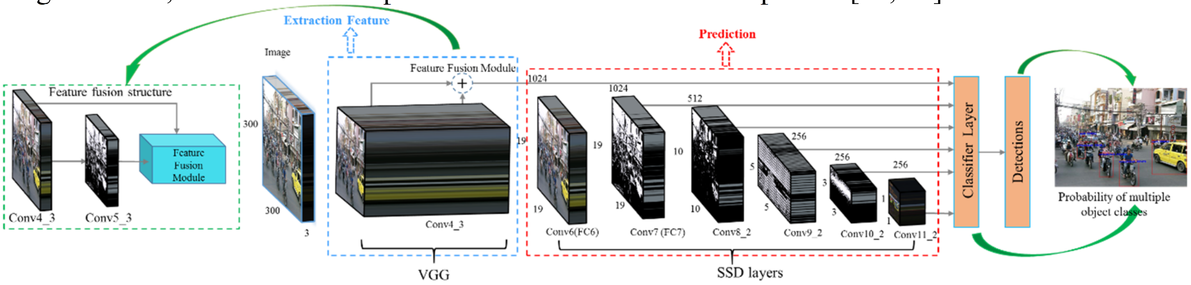


Figure 1. SSD based on feature fusion Network Model

In this paper, a method of SSD object recognition based on feature fusion is proposed (shown in figure 1) for object recognition. In Section 2, the proposed algorithm is introduced about the structure of the method and the processing features. The experiment results and analyses are described in Section 3. Finally, conclusions are drawn in Section 5.

2. Method analysis

The SSD algorithm takes VGG16 as the basic network and several user-defined layers as the functional layers to construct an efficient object detection framework. SSD uses the characteristic pyramid in convolution network to replace the original method of multi-scale object detection based on full connection layer. However, SSD algorithm adopts a non-discriminatory approach to different levels of features, which makes it unable to take into account local details and texture features and global semantic features, thus affecting the detection efficiency of the system for small-scale objects. Figure 2 show that the SSD algorithm is less effective in detecting smaller objects (such as people, sheep and cars in figure 2).

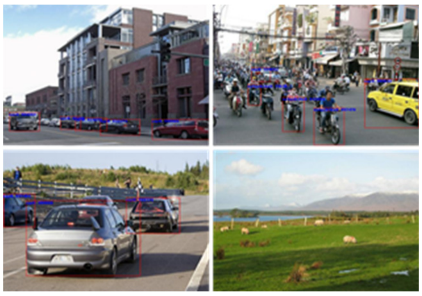


Figure 2. The sketch of detect result of SSD for small objects

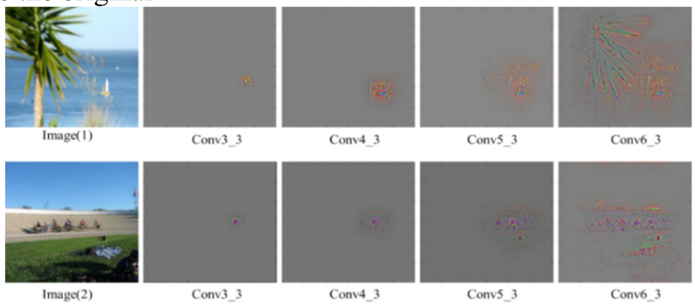


Figure 3. The receptive filed of the partial layer in the SSD

It is necessary to determine which level of texture information and semantic information is significantly suitable for feature fusion. Figure 3 shows the receptive field of the SSD algorithm in different layers. Obviously, the receptive field of the object image feature is small when Conv3_3 is used, and the receptive field of Conv6_3 is too large, which leads to a large amount of background noise. In Conv4_3 and Conv5_3, the size of the receptive field is moderate relative to the small-scale object, and the object feature information can be obtained completely.

In the CNN, the receptive field denotes the mapping region of the pixels of the output characteristic image of each layer on the original image. Figure 4 shows the receptive field intention of different layers. It can be seen from figure 4 that the size of the receptive field is related to the size and step size of the convolution nucleus of all the network layers before the layer. Firstly, the sliding step size is calculated. In the neural network, the step value of each layer is the product of the step value of all previous layers, so the step value formula of the first layer is shown in equation (1).

$$strides(i) = \prod_{j=1}^{i-1} stride_j \tag{1}$$

The size of the receptive field is calculated from top to bottom, that is, the deepest receptive field on the first layer is first calculated, and then the receptive field is regressed to the first layer of the network, as shown in equation (2).

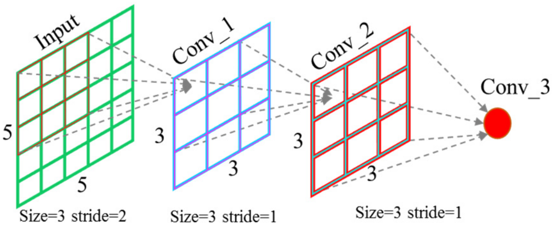


Figure 4. The illustration of the receptive filed

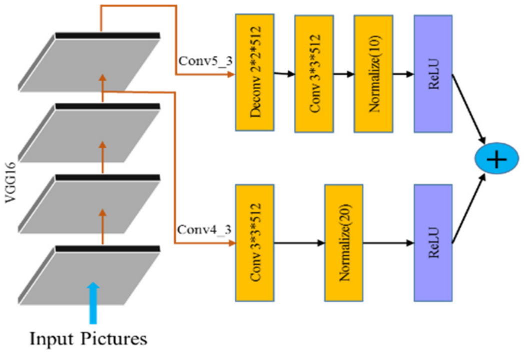


Figure 5. The structure of feature fusion module

$$V'_{rf} = ((V_{rf} - 1) * stride(i)) + Size_{conv} \tag{2}$$

Where V'_{rf} represents the current receptive field, V_{rf} is the current receptive filed in the first layer, $Size_{conv}$ is the size of convolution nucleus.

The receptive field area of Conv4_x to Conv5_x in the SSD network model is calculated by equation (2). The receptive field from Conv4_1 to Conv5_3 are 44, 76, 92, 132, 164 and 196, respectively. Conv4_3 receptive field has a moderate size relative to the small-scale object, and can extract the feature information of the small-scale object completely. In addition, compared with Conv4_3, Conv5_3 has three convolutions and one pooling operation, and the semantic information has been enhanced to avoid the impact of a large number of noises on small-scale object detection. So, this paper proposes a feature fusion framework based on Conv4_3 and Conv5_3.

The aim of feature fusion structure is to design an effective framework to fuse the features of different layers in convolutional neural network, and then send the features of this layer into the detector to complete target detection. The design process is as figure 5.

Assuming that $X_i, i \in Q$ is the extracted depth feature, the fusion architecture should logically satisfy the following equation (3):

$$X_{output} = \Gamma_1\{\chi_i(X_i)\} \qquad i \in Q \tag{3}$$

Where χ_i is the feature layer for preprocessing, which is used to fuse all feature maps in the current layer; Γ_1 is the feature fusion framework; Q is the number of channel of the feature map of the current layer. The structure of Γ_1 is shown in figure 5. The feature of Conv4_3 and Conv5_3 is fused by 3*3 convolution layer, normalization layer and activation layer. Considering the size of Conv5_3, it is necessary to be handled by 2*2 deconvolution layer.

Based on the above improvements, the complete object detection framework proposed in this paper can be shown in figure 1. The proposed framework of feature fusion can effectively reduce the possibility of introducing a large amount of background noise into the receptive field to affect small-scale object detection. Meanwhile, fusing feature maps of different layers directly enhances the relevant semantic information, and makes use of the information extracted from each layer of the structure, which makes the proposed features have stronger discriminant performance and effectively improves the object detection accuracy of the proposed algorithm.

3. Experiments

3.1. Experimental setup

In order to evaluate the performance of the algorithm proposed, experiments were carried out on PASCAL VOC 2007 and 2012 datasets. In this section, the test results of SSD algorithm based on feature fusion are compared with those of SSD algorithm and Faster RCNN algorithm. And the performance of the proposed method for small object detection is improved compared with the original algorithm. Finally, the real-time performance of the algorithm is compared and analysed.

The VGG16 network pre-training model used in this paper can be trained based on ImageNet data set. The hyper-parameters are set as follows: batch size = 32, gamma = 0.1, momentum = 0.9, input size 300 x300, optimization type SGD. In addition, the newly added layer is initialized using Xavier [22]. The initial learning rate is set to 10^{-3}, and then adjusted to 10^{-4}, 10^{-5} and 10^{-6} respectively when the number of iterations is 60k, 80K and 100K.

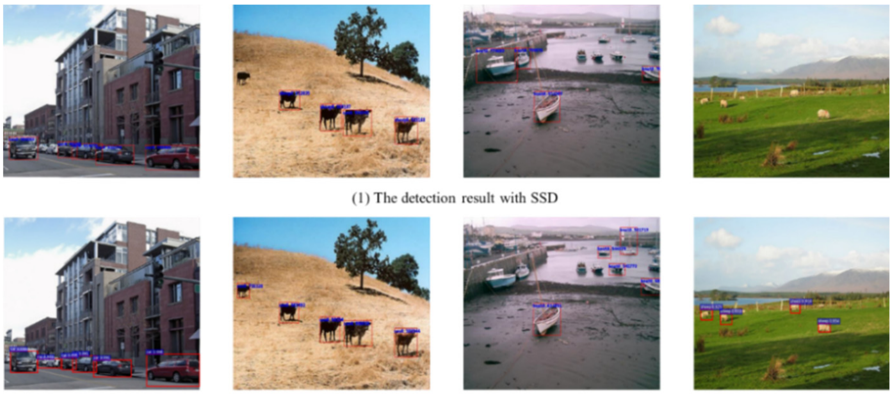
3.2. Performance evaluation algorithm based on dataset

In this paper, the detection model of the proposed algorithm is trained on the dataset of PASCAL VOC2007 and PASCAL VOC2012. The VOC2007 and VOC2012 datasets contain 9963 and 22531 images with 20 targets respectively.

After experiments, the influence of different layer feature fusion on the overall detection performance is analysed on PASCAL VOC2007 test set. The results of experiment are that Conv4_3 and Conv5_3 achieves 78.04% mAP, Conv4_3 and Conv6 achieves 79.95% mAP, and Conv4_3, Conv6 and Conv7 achieves 77.84% mAP. As shown in the result, using Conv4_3 and Conv5_3 for feature fusion can significantly improve the detection accuracy. Conv6 has larger sensing field than Conv5_3, which leads to more background noise being introduced, thus reducing the detection accuracy. In order to evaluate the overall performance of the proposed algorithm, the detection results of 20 kinds of objects are compared and analysed on PASCAL dataset, as shown in table 1. The average accuracy of the proposed algorithm is 78.33%, which is 1.13% and 5.13% higher than SSD and Faster RCNN, respectively. In addition, as shown in figure 6, the detection effect of the first behaviour SSD algorithm and the second behaviour SSD algorithm proposed in this paper are shown. Obviously, the proposed algorithm has been significantly improved than the original SSD algorithm in small-scale object detection. Obviously, the algorithm is superior to the above two methods in object detection accuracy, thus verifying the effectiveness of the proposed algorithm.

4. Conclusion

In this paper, the SSD object detection algorithm based on feature fusion was proposed. By constructing a feature fusion architecture, the object detection performance of the algorithm is effectively improved, and especially for the detection effect of small-scale objects. First, a framework for efficiently integrating image texture features and global feature depth models is designed. Then, the feature information extracted from each layer of the structure is transformed to the detector of the network, and further mining the semantic information of network features. Finally, using the PASCAL VOC dataset, experiments are made for testing and verifying the performance of the proposed algorithm. The experimental results show that the mAP of the proposed SSD object detection algorithm based on feature fusion is 78.0%, which is both higher than that of the SSD and Faster RCNN algorithms, and further verifies the advancement and effectiveness of the algorithm. And the future work will optimize the network parameters, increase the number of training samples, and improve the robustness and adaptability of the model to achieve better detection performance.



(1) The detection result with SSD

(2) The detection result with Ours algorithm

Figure 6. Comparison of results between SSD algorithm and feature fusion based on SSD algorithm

Table 1. The test result on PASCAL VOC2007 test dataset (with IOU=0.5)

Methods	MAP (%)	Air	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
SSD300	77.2	78.5	85.9	75.7	71.0	49.2	85.3	86.5	87.7	60.7	82.3
Faster RCNN	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9
Ours	78.0	80.3	86.3	76.8	72.1	51.4	86.1	86.7	88.0	61.4	82.6

Methods	MAP (%)	Desk	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	Train	TV
SSD300	77.2	76.8	84.3	86.7	84.5	79.1	51.7	77.4	78.8	86.6	76.7
Faster RCNN	73.2	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
Ours	78.0	77.1	85.6	87.7	86.4	79.6	53.8	78.9	79.2	87.8	78.1

Acknowledgments

This research is financially supported by the 2017 BJUT United Grand Scientific Research Program on Intelligent Manufacturing (No. 040000546317552) and the National Natural Science Foundation of China (No. 61703012).

References

[1] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. arXiv preprint.

[2] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[3] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

[4] Lee, K., Choi, J., Jeong, J., & Kwak, N. (2017). Residual features and unified prediction network for single stage detection. arXiv preprint arXiv:1707.05031.

[5] Wang, R. J., Li, X., Ao, S., & Ling, C. X. (2018). Pelee: A Real-Time Object Detection System on Mobile Devices. arXiv preprint arXiv:1804.06882.

[6] Liu, W., Anguelov, D., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[8] Yang, F., Choi, W., & Lin, Y. (2016). Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2129-2137).

[9] Bansal, A., Chen, X., Russell, B., Gupta, A., & Ramanan, D. (2016). Pixelnet: Towards a general pixel-level architecture. arXiv preprint arXiv:1609.06694.

[10] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[11] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[12] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9), 1627-1645.

[13] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[14] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.