Small Object Detection with Multiple Receptive Fields

Yongjun Zhang\textsuperscript{a}, Tao Shen\textsuperscript{b}
Beijing University of Posts and Telecommunications, Beijing 100876, China
\textsuperscript{a}yjzhang@bupt.edu.cn, \textsuperscript{b}shentao@bupt.edu.cn

Abstract. Small object detection has been a problem in deep learning convolutional neural network models. A multi-rate dilated convolution module is proposed to form a feature map to locate small objects. Benefiting from the fact that the dilated convolution does not add extra complexity while maintaining the characteristics of the high-resolution feature map, this paper replaces the traditional convolution network by setting the dilated convolution with the ratio of 1, 2, and 5, and fuses the different convolution rate convolutions. The feature map branch, extracting high-level features, using the FPN algorithm as the basic network framework, combining a variety of receptive field feature maps, enhancing the detection ability of the algorithm for small objects, and the convergence accuracy is greatly improved. Experiments show that the proposed method achieves 81.9\% MAP in the PASCAL VOC2007 dataset, which exceeds the traditional object detection algorithm.

1. Introduction
In deep learning, small object detection is a field of great concern. There are two definitions of small objects. One is relative size. If the length and width of the object size is 0.1 of the original image size, it can be regarded as a small object, and the other is the definition of absolute size, that is, the size is less than 32*32. The object of a pixel can be considered a small object. In recent years, the rapid development of convolutional neural networks, algorithm frameworks such as Faster RCNN [1] and SSD [2] have been well used for general purpose object detection, but there are still some shortcomings and difficulties for small object detection.

Small object detection has certain difficulty, which is caused by the background of small object pixels and high computing power requirements. At present, there are many algorithms with superior performance for small object detection. For example, the FPN algorithm [3] improves the detection effect of small objects by introducing a Top-Down structure. The Mask R-CNN algorithm [4] predicts the object in parallel by adding branches and RCNN. In particular, the problem of Misalignment is solved by using RoIAlign. However, the network structure of these algorithms is designed based on the classification task. It is characterized by a large number of Pooling layers, and the final feature map will be compressed very small. The partial shift of the feature caused by the Pooling layer is actually less friendly to the inspection task where location information is important. The part of the classification task is often occupied in the image. In the end, the feature layer is small, but it has no effect, and the sensory field becomes large, so that the unit feature can contain more information. However, for the detection task, the proportion of the object in the figure is not fixed, and the feature of the low resolution and high receptive field will lose too much detail information, which makes the detection of small objects difficult. Therefore, the network suitable for detection should have a higher
resolution. In this paper, the dilated convolution module with multi-void rate is designed to avoid the problem of the reduction of the receptive field while maintaining the calculation amount.

2. Related Work

2.1. Generally object detection network
Object detection is to find the exact location of an object in a given picture and mark the category of the object. Therefore, the problem to be solved by the object detection is the entire process problem of where and what the object is. The R-CNN series algorithm is a general purpose object detection framework. The R-CNN algorithm divides the object detection into two stages: the algorithm first generates a series of candidate frames as samples, and then uses the convolutional neural network to classify the samples. Compared with the object detection algorithm of one stage, the R-CNN framework is more dominant in detection accuracy and positioning accuracy. A typical R-CNN network is illustrated in Figure 1.

![R-CNN Network structure diagram](image)

**Fig. 1 R-CNN Network structure diagram**

The traditional R-CNN algorithm uses the Selective Search algorithm to evaluate the feature similarity of adjacent image sub-blocks. By scoring the merged similar image regions, the candidate frame of the region of interest is selected as the sample input to the convolution. Inside the neural network structure, the positive and negative sample features consisting of the network learning candidate frame and the calibration frame form the corresponding feature vector, and then the support vector machine design classifier classifies the feature vector, and finally completes the border regression operation for the candidate frame and the calibration frame. Achieve the purpose of objecting detection. Fast R-CNN [5] designed a pooling layer structure of ROI pooling, which effectively solves the problem that the R-CNN algorithm must crop and scale the image area to the same size. The Faster R-CNN introduces the RPN (Region Proposal Networks) network, and divides the algorithm structure into two parts. The RPN network first determines whether the candidate frame is the object, and then determines the object type by the multi-task loss of the classified positioning. This algorithm is flexible and robust for many subsequent improvements.
2.2. Small object detection network

Small object detection has always been a problem in deep learning convolutional neural network models. Most of the existing object detection documents are for general purpose detection, such as the classic single-stage method yolo [6] and SSD, the two-stage method faster-RCNN, etc. These methods are mainly for general object data. The solution is designed to be designed so that the detection effect is not ideal for small objects in the image.

The traditional method of detecting small objects generally adopts an image pyramid, that is, a multi-scale image pyramid is sampled on the training picture. The up-sampling can enhance the fine-grained features of small objects, and theoretically optimize the positioning and recognition of small objects. However, the convolutional neural network model based on image pyramid training has very high requirements on computer power and memory. The development of computer hardware is still difficult to compete. Therefore, this method is rare in practical applications. In the past two years, the method of using multi-layer feature maps (characteristic pyramid, RNN idea, layer-by-layer prediction) has been proposed, which has significantly improved the effect of small object detection. The FPN algorithm proposes a multi-scale object detection algorithm. The top-level features are fused by up-sampling and low-level features. At the same time, the high-level semantics of low-level features and high-level features are utilized, and the prediction effects are achieved by fusing the features of these different layers. The ION algorithm [7] mainly focuses on the context information and multi-scale information in the object detection process. The structure of skipping pooling and IRNN is used to realize the fusion of multi-scale features and context information.

In this paper, a dilated convolution module with multi-void rate is introduced in the FPN algorithm framework, which eliminates the resolution loss caused by feature fusion to some extent.

3. Network Structure

Our goal is to improve the accuracy of small object detection while minimizing the complexity of the model. Inspired by feature pyramids and dilated convolution, our network architecture is improved based on FPN networks. The network framework we propose is shown as Figure 2.

![Network structure diagram](image-url)
In particular, our algorithm framework uses a single-scale image as input, and multiple feature maps are obtained through multi-void rate convolution module (MDC) for feature pyramid fusion, and finally the object category and bounding box in the image are obtained.

3.1. Multi-rate dilated convolution module (MDC)

The dilated convolution was first proposed in DeepLab [8] and is widely used in image segmentation. Dilated convolution has the following advantages over traditional convolutional networks: 1. Expanding the receptive field: The spatial resolution is reduced when down-sampling (pooling or s2/conv). In order to not lose resolution and still expand the receptive field, dilated convolution can be used. On the one hand, when you feel the wild, you can detect the large object, and on the other hand, the resolution is high and you can accurately locate the object. 2. Capture multi-scale context information: the dilated convolution has a parameter to set the dilation rate, the specific meaning is to fill the convolution kernel with dilation rate. Therefore, when setting different dilation rate, the receptive field will be different. That is, multi-scale information is obtained.

The use of dilated convolution can preserve the feature map high-scoring rate information without increasing the computational complexity, which is undoubtedly very important for small object detection. This has also been demonstrated in some small object detection networks in recent years. For example, DetNet [9] effectively expands the receptive field by introducing a low-complexity bottleneck.

RFBNNet [10] integrates the RFB module into the top convolutional layer of the SSD network structure, and the accuracy is improved in the case of control loss. But a single increase in dilated convolution tends to cause a checkerboard effect: local information loss: since the calculation of the dilated convolution is similar to the checkerboard format, the convolution result obtained by one layer, from the independent set of the previous layer, does not depend on each other. Therefore, there is no correlation between the convolution results of this layer, that is, local information is lost. By introducing a multi-cavity rate convolution module, we can make the basic network learn more scale information, so as to improve the accuracy without increasing the complexity. The multi-cavity rate convolution module is shown as Figure 3.

Before the dilated convolution, the feature map is up-sampling by bilinear interpolation, and then the dilated convolution is performed, so that more global information can be extracted while improving the pixels of the feature map. Then use the dilated convolution with a void ratio of 1, 2, and 5. The larger the void rate, the larger the receptive field, and the global and local information can be extracted by combining different void rate information.

The multi-void rate convolution module mitigates the board problem by combining convolution modules with different dilated ratios, so that the feature map obtains more continuous pixel
information, and improves the efficiency of small object detection by increasing the dependence of adjacent pixels.

3.2. Feature fusion module
In order to fuse the multi-scale features of the network layer, we adopted the FPN framework. On the whole, FPN utilizes the Top-Down structure, which combines the high-level and low-level feature information, so that the underlying semantic information is still rich while maintaining a large resolution and improving the detection effect of small objects. The feature fusion method is different from the FPN network. After the dimension of the lower layer is reduced by a 1*1 convolution, it is merged with the high-level features at the pixel level. In order to enhance the effect of small object detection, a weight is added to the lower layer features here. The merged features can be input through a 3*3 convolutional layer for prediction, or the above operations can be repeated to merge with lower-level features.

High-level features are enhanced by sampling the higher-level, more semantic high-level feature maps and then laterally connecting the features to the previous layer. It is worth noting that the two-layer features of the lateral connection are the same in spatial size. This should be done primarily to take advantage of the underlying positioning details. The previous layer can be used after the 1*1 convolution kernel. The purpose is to change the channels, and the channels of the latter layer are the same. The combination method is to add between pixels. Iterate through the process until the finest feature map is generated. At the beginning of the iteration, a 1*1 convolution kernel is added after the C5 layer to produce the coarsest feature map, and the 1*3 and 3*1 convolution kernels are used to process the merged feature map (to eliminate the up-sampling). The aliasing effect) to generate the final required feature map.

4. Experiments

4.1. Implementation Details
The method of RPN is used to generate the bounding box and the fast R-CNN is used for object detection. In order to prove the simplicity and effectiveness of our method, we have minimized the original system of RPN and fast R-CNN to adapt it to our multi-void rate convolution module.

We have adopted a multi-void rate convolution module on the FPN framework. The training strategy includes image enhancement, e.g., smooth L1 loss for localization and softmax loss for classification. In particular, we have adopted a joint training strategy. In the RPN phase, we use a box with an IOU greater than 0.45 for the ground truth box as a positive sample, and a box with an IOU of less than 0.3 for the ground truth box as a negative sample. Obviously, in the scenario of small object detection, it is easy to see the imbalance of positive and negative samples. There are currently three methods for screening negative samples: 1) random sampling; 2) sorting by Score, retaining Top n negative samples (Hard Negative); 3) Half machine sampling, the other half Score sorting. We have adopted a new sample set equalization method: retain Top n difficult negative samples, and add weight to the negative samples. The final loss function is:

\[ L(k, k^*, t, t^*) = L_{cls}(k, k^*) + \lambda L_{reg}(t, t^*) \]

(1)

Where \( L_{cls}(k, k^*) \) is the softmax loss between the two categories, and \( L_{reg}(t, t^*) \) is the L_1 loss of the predicted bounding box and the ground truth box. In particular, the weight of the negative sample is introduced in \( L_{cls}(k, k^*) \), specifically:

\[ L_{cls}(k, k^*) = L_{cls}(k, k^*) \ast weight \]

(2)
weight = 1 + \frac{1}{1 + rank(k)}

(3)

Where \( rank(k) \) is the result of sorting according to Score.

In the detection phase, we used the parameter settings of R-CNN. Let the prediction result \( t = (t_x, t_y, t_w, t_h) \), label \( \mathbf{t} = (t_x^*, t_y^*, t_w^*, t_h^*) \). Including:

\[
\begin{align*}
    t_x &= \frac{G_x - P_x}{P_x} \\
    t_y &= \frac{G_y - P_y}{P_y} \\
    t_w &= \log\left(\frac{G_w}{P_w}\right) \\
    t_h &= \log\left(\frac{G_h}{P_h}\right)
\end{align*}
\]

\( P_i = (P_x, P_y, P_w, P_h) \) represents the center point of the ground truth box, as well as its length and height.

In the end, we can get our end-to-end object detection network and back-propagation through joint training. In practice, we follow the steps below to train.

Step 1: Initialize the CNN network
Step 2: Train the RPN network to generate a region proposal
Step 3: Train the network and use the region proposal in step 2 for object detection.
Step 4: Fine-tune the RPN network parameters
Step 5: Enter the object detection network fine-tuned for the region proposal obtained in step 4.
Step 6: Output the network fine-tuned in steps 4 and 5 as the final network.

Before the fourth step, the RPN network and the detection network are independently trained. In the end, we combined these two networks to get a unified object detection network. The training process sets the input image size to 300 × 300 pixels, the training batch size is set to 1, and the epoch is set to 150,000. During the training process, the training loss is guaranteed to decrease smoothly. The initial learning rate of the first 100k epoch is 0.005, and the last 50k. The epoch is 0.0005 and the momentum is 0.9. In an epoch, 64 RoIs were sampled as samples. All network layers are initialized with "Xavier".

In order to prove the validity of our multi-cavity rate convolution module, we performed experiments on the PASCAL VOC2007 training set with different cavity size convolutions. The results are as follows:
It can be seen that the small object category and object frame can be extracted better and faster using the dilated convolution rate with a combination of 1, 2, and 5.

In addition, in order to verify the impact of learning rate on training loss, we use three strategies of learning rate to train:

- **Strategy 1**: Fixed learning rate is 0.005
- **Strategy 2**: Fixed learning rate is 0.0005
- **Strategy 3**: The initial learning rate of the first 100k epoch is 0.005, the latter 50k epoch is 0.0005, and the momentum is 0.9.

The training results of the three strategies are as follows:

It can be seen that setting a large learning rate is beneficial to the fast convergence of the loss function under the premise that the initial learning rate is fixed. When the loss function converges, using a smaller learning rate facilitates the convergence of the curve to the global minimum. Therefore, the learning rate set in Strategy 3 also makes the final training network work best under the premise of ensuring fast convergence of the network.
4.2. Main Results

We did experiments on the PASCAL VOC2007 training set and compared it with popular object detection algorithms. At the same time, we also conducted an in-depth analysis of object frame extraction and detection performance.

The PASCAL VOC2007 data set includes 9,963 images for a total of 20 objects. Among them, there are 5011 training sets and 4952 test sets. All algorithm models were trained on the VOC2007 training set and tested on the VOC2007 test set.

Table 1 shows the results of the average accuracy of the different models of the network model in the PASCAL VOC2007 test set. It can be seen from the table that the average detection accuracy of the method in the small object category is significantly higher than other network models. The average accuracy of the network proposed in this paper is 1% higher than other networks, such as bird, sheep, plant.

The other small objects are 4.5%, 7%, and 5.3% higher than other networks, respectively, which proves the effectiveness of the proposed network.

Table 1. Test Results for Different Categories of VOC2007

|       | Faster RCNN | MR-CNN | SSD   | FPN   | MDC-FPN |
|-------|-------------|--------|-------|-------|---------|
| aero  | 76.5        | 80.3   | 79.5  | 86.1  | 88.5    |
| bike  | 79.2        | 84.1   | 83.9  | 89.3  | 86.4    |
| bird  | 70.9        | 78.5   | 76.3  | 79.5  | 84      |
| table | 65.7        | 73.7   | 77.2  | 79.5  | 73.2    |
| dog   | 84.8        | 87.2   | 86.1  | 86.4  | 87.6    |
| horse | 84.6        | 86.5   | 87.5  | 89.2  | 88.2    |
| plant | 38.8        | 48.5   | 52.3  | 53.3  | 58.3    |
| sheep | 73.6        | 76.3   | 77.9  | 77.9  | 84.9    |

In addition, in order to illustrate the feature effects extracted by each part in multi-scale detection, we visualize the network feature maps of the following figure, as shown below:

Fig. 6 Training set sample
Fig. 7 Feature map visualization. From left to right, from top to bottom, it is C2, C3, C4, C5

It can be seen that the underlying feature map mainly extracts edge and texture features, and the high-level feature map is more abstract and extracts semantic features. For Figure 5, in order to detect the car, it is impossible to distinguish the information of the car only in the high-level feature map. The visualization that incorporates the underlying features into the higher layers is shown below:

Fig. 8 Feature map visualization. From left to right, from top to bottom, P2, P3, P4, P5
It can be seen that after combining the two parts, the upper layer feature map can extract the high-level semantic information, and the P2 layer corresponding to the C2 layer can distinguish the contour and semantics of the car, which can enhance the detection effect of the small object.

5. Conclusion
This paper proposes a new type of high-efficiency small object detection algorithm. Unlike most object detection algorithms that rely on ultra-deep backbone networks, our algorithm is based on lightweight backbone networks and is integrated by adding multi-rate dilated convolution modules. The characteristics of different size receptive fields make full use of the feature information extracted by the dilated convolution, which improves the overall detection effect of the algorithm and further improves the detection effect of the algorithm on small objects. Through comparison experiments, it can be found that the overall performance of the proposed algorithm on the Pascal VOC2007 data set is better than the traditional RCNN algorithm and most of the mainstream object detection algorithms, meeting the requirements of real-time detection, and the effect of small object detection is significantly improved.

References
[1] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. 2015.
[2] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [J]. 2015.
[3] Lin T Y, Dollár, Piotr, Girshick R, et al. Feature Pyramid Networks for Object Detection [J]. 2016.
[4] Kaiming H, Georgia G, Piotr D, et al. Mask R-CNN [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018:1-1.
[5] Girshick R. Fast R-CNN[C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2016.
[6] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. 2015.
[7] Bell S, Zitnick C L, Bala K, et al. Inside-Outsi de Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2016.
[8] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 40(4): 834-848..
[9] Li Z, Peng C, Yu G, et al. DetNet: A Backbone network for Object Detection [J]. 2018.
[10] Liu S, Huang D, Wang Y. Receptive Field Block Net for Accurate and Fast Object Detection [J]. 2017.