



Parallel Connecting Deep and Shallow CNNs for Simultaneous Detection of Big and Small Objects

Canlong Zhang¹(✉), Dongcheng He¹, Zhixin Li¹, and Zhiwen Wang²

¹ Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University, Guilin, China
zcltyp@163.com

² College of Computer Science and Communication Engineering,
Guangxi University of Science and Technology, Liuzhou, China

Abstract. In order to improve the real-time and accuracy of Faster R-CNN (Region based Convolutional Neural Networks) for detecting small object, a novel object detection model is proposed in this paper. Our model not only keeps the detection accuracy for big object, but also improves significantly the accuracy for small object, and with very little reduction in term of detection speed. Firstly, a shallow CNN is designed and connected with an improved deep CNN by using skip-layers connection method, which makes full use of the convolution characteristics with different layers to improve the detection ability for small object; Secondly, the detection accuracy of our model is improved further by incorporating the region proposal mechanism in Faster R-CNN, and using 12 kinds of anchors to generate object candidates; Finally, a dimensional reducer is designed by connecting ROI-Pool layer and 1×1 convolutional layer, which accelerates the detection of overall network. The test results on image datasets PASCAL VOC and MS COCO show that the detection accuracy of our model is higher than some current advanced models, and small objects is significantly improved.

Keywords: Object detection · Convolutional neural networks
Region proposal · Skip-layers connection

1 Introduction

Object detection technology has been widely used in intelligent transportation, road detection and military target detection. With the advent of deep learning and large-scale visual identification datasets, object detection has developed rapidly, among which the two-step object detection framework based on R-CNN [1–4] and one-step object detection framework based on regression [5–7] are the most representative.

Object detection framework based on R-CNN mainly consists of image convolution, region proposal, classification and regression of the region. In 2014,

Girshick et al. [1] proposed the object detection framework based R-CNN by combining region proposal with CNN, which opened a new era for object detection with deep learning. After that, around the detection accuracy and speed, many improved versions of the R-CNN model have been proposed, such as, the Fast R-CNN [2] incorporating the ROI-Pool layer and multi-task loss function, the Faster R-CNN [3] incorporating Region Proposal Network (RPN), and the Mask R-CNN [4] cooperating instance segmentation for multitask collaboration. Compared with the traditional method of target detection, the R-CNN methods avoid the subjectivity of the manual feature extraction, and realize the end-to-end object feature extraction and classification.

In view of the two-step object detection be very slow, the one-step object detection method avoids the process of region proposal, and performs the object detection by using regression method directly to output category and bounding box of each regions in the image. The YOLO [5] divides the image into grids, and performs regression computation on those grids to gain category and bounding box of the objects, which boosts the detection speed to 45 fps. The SSD [6] introduces the anchor representation of Faster R-CNN into YOLO to general multi-scale regions at each location in the image, which not only improves greatly the accuracy of detector but also makes the detection speed be up to 58 fps. DSSD [7] fuses the deep convolution layers and the shallow convolution layers by using encode-decode network, which can leverage high-level semantic and low-level image feature, so boosts the detection performance on small object and dense object. Above methods do not require region proposal, so compared to the methods based on R-CNN, their detection speed are faster but the accuracy be lower.

The shortage of above two methods is that the detection accuracy of small objects is poor. Therefore, in this paper, an object detection model based on parallel connection of Deep and Shallow CNN (DS-CNN) is designed via innovative use of skip-layers connection, region proposal and anchors. This model not only keeps the detection accuracy for big object, but also improves significantly the accuracy for small object, and with very little reduction in term of detection speed.

2 DS-CNN

The framework of our detection model is shown in Fig. 1, and it consists of four parts: the first part is the feature extraction network, including deep CNN and shallow CNN; the second part is the region proposal network, which is used to generate object candidates; the third part is dimensional reducer, which is used to reduce the dimension of feature of object candidate; the fourth part is fully-connected (FC) layer, classification and regression network.

2.1 The Design of Deep and Shallow CNN

In general, the deeper the level of convolution is, the more obvious the semantic characteristic is, and the easier it is to classify the object, but the more

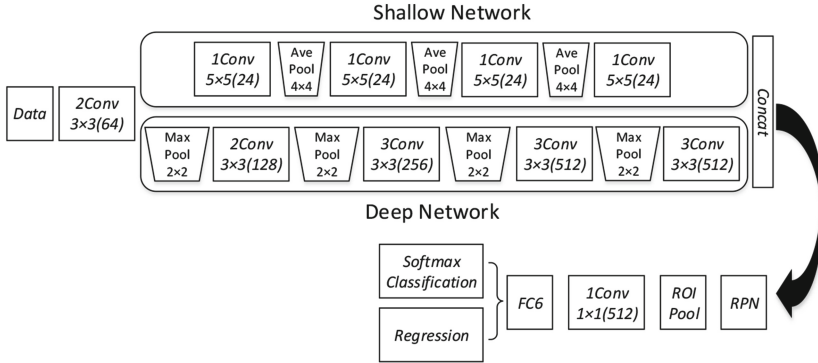


Fig. 1. The framework of DS-CNN. Given an input image, after dealing with two layers of convolutional layers, we use shallow network and deep network to process the feature map into the same size and combine them in the concat layer. The RPN generates 500 region proposals, and then the feature map is processed by the dimensional reducer and fully-connected layer. Finally, we use softmax for classification and multi-task loss function for regression.

information lost. For large-scale objects, this loss is not enough to affect their classification and identification, but it is not the same for small objects. Taking Fast R-CNN [2] as an example, the feature map of last convolutional layer conv5-3 has been reduced by 16 times. For a 500×300 image, the size of the small object is about 32×32 , and it becomes 2×2 after conv5-3. Although the upsampling can expand the image to 7×7 , the loss of information is irreversible. This is the reason why the series method of R-CNN model has a relatively poor detection accuracy on small objects.

For this purpose, we designed deep CNN and shallow CNN based on the VGG16 network. The deep network is used to capture the high-level semantics of large objects, while the shallow network is used to hold the low-level image features of small objects. In the deep network, the parameters of conv1-1 to pool4 are the same as those of VGG16, but the conv5-1 to conv5-3 layers are all modified using dilated convolution with a pad of 2, a kernel size of 3×3 , a stride of 1, and a dilation of 2. Dilated convolution [8] is a common method in the image segmentation, it can expand the receptive field without changing the size of the feature map, and thus contains more global information. The principle is shown in Fig. 2, where (a) is the normal feature map, and (b) is a dilated convolutional map with a dilatation factor of 2. For the 7×7 feature area, the actual convolution kernel size is 3×3 and the hole is 1, that is, the weights of other points except 9 black points are 0. Although there is no change in kernel size compared to (a), the receptive field of this convolution has increased to 7×7 , which allows each convolution output to contain more global information.

In the shallow network, it is no longer necessary to capture high-level semantic features of the image, but rather to obtain the low-level image features, so we don't need a very deep network, that is, we don't need to use a large number of

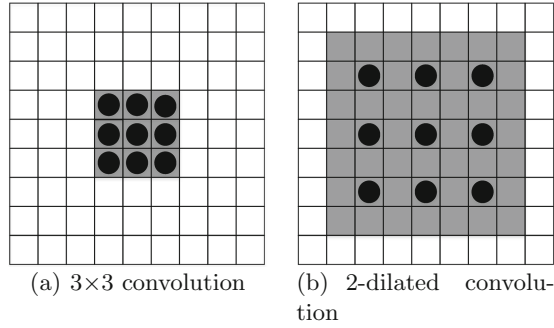


Fig. 2. Principle of dilated convolution.

convolutional layers. In order to achieve better results for the parallel structure, we use the skip-layers connection method to share the parameters of conv1-1 and conv1-2. Starting from conv2-1, only 4 convolutional layers are used, each of these layers has 24 filters with a kernel size of 5×5 . In order to make the final deep network and shallow network have the same spatial resolution, we design an average pooling layer with a kernel size of 4×4 and a stride of 2 after each convolutional layer in the shallow network. Using average pooling instead of maximum pooling in this model ensures that no excessive image information is lost.

After extracting image features, we need to combine the feature maps of the deep network and the shallow network, and to integrate them into a unified space. In this paper, we employ the concat layer to do it, and the dimension of joint features is 536-d.

2.2 Object Candidate Generation

The number and quality of object candidate region proposals affects directly the speed and accuracy of the object detection. RPN [3] directly generates candidate regions on the convolutional map by using the “anchor”. Although RPN is still in the way of window sliding in essence, the detection speed of the whole network is greatly improved because of its regional recommendation, classification and regression sharing the feature of convolution map, so we refer the RPN in the proposed DS-CNN.

RPN scans and convolves the feature maps by using a 3×3 sliding window. At the center of 3×3 sliding window, we give 4 scales(64, 128, 256, 512) and 3 aspect ratios(1:1, 1:2, 2:1), which can generate 12 different region proposal boxes, i.e., 12 kind of anchor. Thus, for the 14×14 feature map, there are about 2300 ($14 \times 14 \times 12$) region proposal boxes. After that, all region proposal boxes are sent to the fully connection classification layer and regression layer to classify and refine the region. The classification layer contains two elements for calculating the probability of the target or non-target. The regression layer contains four coordinate elements (x, y, w, h) for determining the target position. In order to

obtain valid region proposals, we adjusted some of the parameters and using the non-maximum suppression method to preserve the candidate region whose overlap rate with truth region is greater than or equal to 0.5 as a positive sample, and less than 0.3 as a negative sample. Finally, the first 500 positive samples with highest overlap rate are selected as the final region proposals for object detection.

In RPN, the input image can scale up to 1000×600 , but the maximum scale in 12 kinds of anchors is 1024×512 , resulting to the 1024 be beyond 1000, so the parts beyond the border will be cut out. As a result, the maximum size of the anchor is 1000×512 , and this size is large enough to cover the big object in image. Similarly, 256 and 128 scales can be used to deal with medium-size objects. Because each anchor is single-label detection, large object with obvious feature will cover small object with obscure feature. However, by using the minimum scale 64, we can avoid the small objects to be covered by large objects when detecting small objects, so improve the detection accuracy of small objects.

2.3 Dimensional Reducer

The FC layer can integrate the extracted image features, and plays the role of classification in CNN. Because the FC layer is easy to cause parameter redundancy, many classical methods choose to use other types of layers instead of the FC layer. For example, fully convolutional network uses a convolutional layer instead of FC layer, and ResNet [9] and GoogLeNet [10] all use the global average pooling instead of FC layer. Because our model draws on the classification and regression layer of Fast R-CNN, it cannot completely remove the FC layer. Therefore, in our model, a dimensional reducer is designed to replace one FC layer of VGG16 to reduce parameter redundancy. The dimensional reducer consists of a ROI-Pool layer and a 1×1 convolutional layer. The ROI-Pool layer is able to output a fixed size feature map after the RPN, which is used to compress feature maps in this paper. The convolutional layer with a kernel size of 1×1 and a step size of 1 is behind the ROI-Pool layer, which can not only make the structure more compact, but also reduce the dimension of the feature map. We use dimensional reducer to fix the size of feature maps to 7×7 , and to reduce dimension of features from 536 to 512. The compressed features are then input the FC layer. The experimental results show that our structure of dimensional reducer+FC6+Loss is faster than the one of FC6+FC7+Loss in VGG16, and the detection accuracy is also slightly improved.

Similar to the series method of R-CNN, in FC layer, we uses SoftmaxWithLoss for classification and SmoothL1Loss for regression when training, and uses Softmax for classification when testing.

2.4 Joint Training

Like some advanced models, the DS-CNN can also accept end-to-end training and testing. However, by comparison, we find that the alternate training method can obtain better mAP than end-to-end training method on our model. The main

steps of the alternate optimization training are as follows: Firstly, we initialize the feature extraction network with the pre-trained model of ImageNet [11], and gain candidate regions by training the RPN alone on PASCAL VOC. Secondly, we reinitialize the feature extraction network with the pre-trained model of ImageNet, and add the candidate regions generated in the first step. Meanwhile, a separate detection network is trained on the PASCAL VOC dataset using DS-CNN so as to obtain the parameters of convolutional layer through the loss values of the fully-connected layer and the candidate regions of the RPN. Thirdly, we retrain DS-CNN, and use the model obtained in the second step to initialize and fix the parameters of the convolutional layer so that the convolutional layer does not participate the back propagation, and using the RPN model trained in the first step to initialize and fix the parameters of the RPN in the DS-CNN so that the RPN isn't involved in the back propagation. The total purpose of this step is to connect the feature extraction network with the RPN. Finally, we use the parameters of both convolutional layer and RPN obtained in the third step to reinitialize and fix the DS-CNN model so that both the convolutional layer and the RPN isn't involved in the back propagation. The purpose of training in this step is to fine-tune the fully-connected layer and get the most optimized results.

3 Experimental Evaluation

PASCAL VOC [12] and MS COCO [17] are two used widely datasets in the object detection field, and are used to evaluate our DS-CNN. The mAP is used as the main evaluation criterion, and the convergence and detect speed of model are used as two auxiliary evaluation criteria. We also compare our model with state-of-the-art models, and they all use VGG16. All experimental results are obtained by running these models on a PC with Intel Core i7-7700K 4.20 GHz CPU, GeForce GTX 1080Ti GPU, and 16 GB RAM.

3.1 Experiments on PASCAL VOC

The PASCAL VOC 2007 dataset includes 20 object categories, about 5k training images and 5k testing images, and the PASCAL VOC 2012 dataset is similar to PASCAL VOC 2007, but the volume of data has doubled. Small objects of PASCAL VOC dataset are mostly indoor, including bottle, chair, dining table, potted plant, sofa, and tv.

In the first experiment, we use alternate training method to train our DS-CNN on the training dataset of PASCAL VOC 2007, and test the model on the testing dataset of PASCAL VOC 2007. Experimental results are shown in Table 1, where the bold fonts, such as **bottle**, **chair**, are small objects. From the table, we can observe that the mAP of DS-CNN is 72.1%, which is higher than other models. For small objects, the detection accuracy of our model is significantly improved, where the bottle and plant is the most obvious. Although the accuracy of tv is lower than OHEM+FRCN [14], it is also 3.2% better

than Faster R-CNN. However, the detection results on larger objects seem to be unstable, but we notice that most of them can maintain a high accuracy. In order to express object detection results more intuitively, Fig. 3 shows some examples of results on the PASCAL VOC 2007 dataset.

Table 1. The average detection precision (%) of all models on PASCAL VOC 2007, where the training dataset is from PASCAL VOC 2007, and the best score is highlighted in red color.

	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow
FRCN [2]	66.9	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7
Faster R-CNN	69.9	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3
locNET [13]	65.4	70.7	74.2	62.2	48.1	45.2	84.0	74.7	78.6	42.9	73.0
OHEM+FRCN	69.9	71.2	78.3	69.2	57.9	46.5	81.8	79.1	83.2	47.9	76.2
DS-CNN	72.1	71.8	82.3	71.2	58.4	51.3	80.1	82.8	82.5	53.1	78.8
		table	dog	horse	motor	person	plant	sheep	sofa	train	tv
FRCN [2]		67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8
Faster R-CNN		67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6
locNET [13]		67.0	75.4	77.9	66.9	58.1	30.9	65.5	69.3	73.6	69.0
OHEM+FRCN		68.9	83.2	80.8	75.8	72.7	39.9	67.5	66.2	75.6	75.9
DS-CNN		69.1	80.5	82.2	79.9	81.5	45.1	71.2	70.2	79.9	70.8

In order to eliminate the interference caused by the insufficiency of the dataset, we designed the second experiment. Similarly to the first experiment, we still used the testing dataset of PASCAL VOC 2007 for testing, but the training dataset were from PASCAL VOC 2007+2012, by which the volume of training dataset was expanded to three times of the first experiment. The experimental results are shown in Table 2. It is easy to see that the mAP of DS-CNN is 75.8%. Similar to our method, SSD500 also parallel connects the convolution features from different layers. However, its features all are from high-level instead of low-level layers, so the features of small objects cannot be effectively extracted and trained. On the contrary, in our model, the shallow network and the average pooling layer are used to preserve the information of small objects, and the scale 64 is used to enhance the detection of small objects in the RPN, so the DS-CNN performs better than SSD on the detection of small objects. However, SSD enhances the combination of different convolutional layers, uses data augments, and abandons the fully-connected layers and candidate region generating, so the overall performance of object detection is better than DS-CNN. In structure, DS-CNN is similar to Faster R-CNN, and also draws on RPN of Faster R-CNN, so there is a high comparability between them. The accuracy of DS-CNN is higher than that of Faster R-CNN on all objects except boat, where the detection accuracy on small objects is increased significantly, which demonstrates the effectiveness of DS-CNN.

In order to illustrate that our model can also achieve good results in different datasets, we design the third experiment. In this experiment, the training dataset

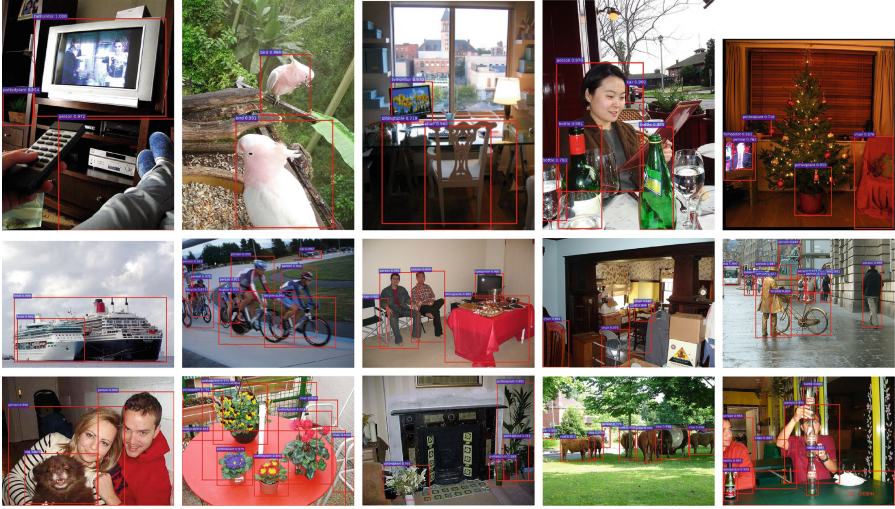


Fig. 3. Some elected examples of object detection results on the PASCAL VOC 2007

Table 2. The average detection precision (%) of all models on testing dataset of PASCAL VOC 2007, where the training dataset is from PASCAL VOC 2007+2012, and the best score is highlighted in red color.

	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow
FRCN	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8
Faster R-CNN	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9
Noc [15]	73.3	76.3	81.4	74.4	61.7	60.8	84.7	78.2	82.9	53.0	79.2
SSD500	75.1	79.8	79.5	74.5	63.4	51.9	84.9	85.6	87.2	56.6	80.1
DS-CNN	75.8	76.9	80.4	75.3	65.3	60.8	85.4	85.7	88.2	60.4	82.7
		table	dog	horse	motor	person	plant	sheep	sofa	train	tv
FRCN		68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster R-CNN		65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
Noc [15]		69.2	83.2	83.2	78.5	68.0	45.0	71.6	76.7	82.2	75.7
SSD500		70.0	85.4	84.9	80.9	78.2	49.0	78.4	72.4	84.6	75.5
DS-CNN		70.3	85.1	85.8	78.8	79.2	45.6	76.1	76.8	84.2	74.3

consists of training dataset of PASCAL VOC 2007+2012 and testing dataset of PASCAL VOC 2007, and testing dataset is from testing dataset of PASCAL VOC 2012. We also compare DS-CNN with FRCN+YOLO [5] and HyperNet [16], and the experimental results are illustrated in Table 3. It is easy to see that our model not only keeps the high detection accuracy for big object, but also improves significantly the detection accuracy for small object, and with very little reduction in detection speed.

Table 3. The average detection precision (%) of all models on testing dataset of PASCAL VOC 2012, where the training dataset is from PASCAL VOC 2007+2012, and the best score is highlighted in red color.

	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow
Faster R-CNN	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1
FRCN+YOLO	70.4	83.0	78.5	73.7	55.8	43.1	78.3	73.0	89.2	49.1	74.3
HyperNet	71.4	84.2	78.5	73.6	55.6	53.7	78.7	79.8	87.7	49.6	74.9
OHEM+FRCN	71.9	83.0	81.3	72.5	55.6	49.0	78.9	74.7	89.5	52.3	75.0
DS-CNN	73.1	82.7	81.1	73.9	55.9	53.7	80.0	76.1	89.8	54.9	76.6
		table	dog	horse	motor	person	plant	sheep	sofa	train	tv
Faster R-CNN		55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
FRCN+YOLO		56.6	87.2	80.5	80.5	74.7	42.1	70.8	68.3	81.5	67.0
HyperNet		52.1	86.0	81.7	83.3	81.8	48.6	73.5	59.4	79.9	65.7
OHEM+FRCN		61.0	87.9	80.9	82.4	76.3	47.1	72.5	67.3	80.6	71.2
DS-CNN		62.2	89.5	81.8	83.9	79.8	48.6	73.6	68.5	81.5	69.4

3.2 Experiments on MS COCO

The MS COCO dataset is more complex than PASCAL VOC, and contains 80 object categories, about 80k images on the training set and 40k images on the validation set. Especially, the dataset has many small objects, so is very suitable for evaluating DS-CNN. We use the end-to-end training method, and set the basic learning rate be 0.001 and the learning strategy be ‘step’. The total iteration step is 490k, and the learning rate is reduced to 0.0001 after 350k iterations. We calculate the mAP@IoU \in [0.5:0.05:0.95] (COCO’s standard metric) and mAP@0.50 (PASCAL VOC’s metric). Experimental results are shown in Table 4. It can be seen that our model has 23.1% mAP on the COCO metric and 43.6% mAP on the VOC metric. It is also interesting to notice that our model performs well on the detection of small and medium objects, and its mAP reaches 6.3% and 25.4% respectively. However, its performance on the large objects seem to be mediocre.

Table 4. MS COCO 2015 test-dev detection average precision (%). All methods use VGG16, and area infers to the size of object.

	Area	FRCN	OHEM	DS-CNN
mAP@[0.50:0.95]	All	19.7	22.6	23.1
mAP@0.50	All	35.9	42.5	43.6
mAP@[0.50:0.95]	Small	3.5	5.0	6.3
mAP@[0.50:0.95]	Medium	18.8	23.7	25.4
mAP@[0.50:0.95]	Large	34.9	37.9	36.3

3.3 Combine from Which Layers?

When using the skip-layers connection method, we need to consider which layers be combined can get the best detection result. For example, the combination of conv3+conv4 +conv5 is the best in ION [18], while the combination of conv1+conv3+conv5 is the best in HyperNet [16]. We give different combinations of cov1, cov2, cov3 and cov5, and use the end-to-end method to train and test each combination on the PASCAL VOC 2007 dataset. The experimental results are shown in Table 5. We found it is no true that the more the number of layers is, the higher the accuracy is, and the best combination is conv1+conv5.

Table 5. The detection accuracy of different combinations of multiple layers.

layers	data+5	1 + 5	2+5	3+5	1+2+5	1+3+5
mAP	69.7%	71.4%	70.8%	69.4%	70.2%	68.6%

3.4 The Evaluation of Speed

Detection speed and convergence speed are two important indexes for evaluating the performance of an object detection model. We compare the DS-CNN with Faster R-CNN on PASCAL VOC 2007. For fair comparison, we also set the number of final candidate regions in Faster R-CNN to 500, and run two models on our PC. We collected each detection time of the model, and averaged all detection times. The detection speed of DS-CNN was about 12 fps, while Faster R-CNN was about 14 fps. In fact, this is an expected result, because DS-CNN consumes more time than Faster R-CNN in feature extraction. However, the difference between 12 and 14 is very little, so the speed has also met our standard: the detection speed has little reduction.

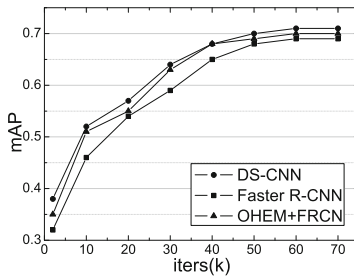


Fig. 4. The mAP at different iterations.

We also used the end-to-end training and testing method to evaluate the convergence speeds of two models on PASCAL VOC 2007 dataset, and recorded

the mAPs of intermediate models with different iterations before generating the final models. The comparison result is shown in Fig. 4. It is easy to see that the two models all had converged when iterating 70k times, and the mAP of DS-CNN is 71.4% while Faster R-CNN is 69.5%. DS-CNN has a faster convergence speed than Faster R-CNN, because it is about 6% higher than Faster R-CNN when iterating 2k times, and it starts to converge after 50k iterations.

4 Conclusions

We designed a new object detection model based on R-CNN. Firstly, we used dilated convolution to design deep neural networks and shallow neural networks, and used skip-layers connection method to connect the two networks. Secondly, we used the RPN to generate object candidates. Thirdly, we designed a dimensional reducer to reduce the dimension of feature maps. Finally, the model output the results of classification and regression. The experimental results illustrated that our model not only keeps the detection accuracy for big object, but also improves significantly the detection accuracy for small object, and with very little reduction in detection speed. However, many more advanced structures cannot be applied to our model due to the limitations of the VGG16 and Fast R-CNN frameworks. In the future, we will research more advanced image feature extraction methods to further improve the accuracy and speed of object detection.

Acknowledgments. The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Natural Science Foundation of China (Grant nos. 61866004, 61663004, 61462008, 61751213), the Natural Science Foundation of Guangxi Province (Grant nos. 2017GXNSFAA198365, 2016GXNSFAA380146), the Scientific Research and Technology Development Project of Liuzhou (Grant no. 2016C050205), and Guangxi Collaborative Innovation Center of Multisource Information Integration and Intelligent Processing.

References

1. Girshick, R., Donahue, J., Darrell, T, Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
2. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: International Conference on Neural Information Processing Systems, pp. 91–99 (2015)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
5. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)

6. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
7. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: DSSD : Deconvolutional single shot detector (2017)
8. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, pp. 770–778 (2015)
10. Szegedy, C., et al.: Going deeper with convolutions, pp. 1–9 (2014)
11. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2014)
12. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. *Int. J. Comput. Vis.* **111**(1), 98–136 (2015)
13. Gidaris, S., Komodakis, N.: LocNet: improving localization accuracy for object detection, vol. 766–767, no. 121, pp. 789–798 (2015)
14. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining, pp. 761–769 (2016)
15. Ren, S., He, K., Girshick, R., Zhang, X., Sun, J.: Object detection networks on convolutional feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(7), 1476–1481 (2016)
16. Kong, T., Yao, A., Chen, Y., Sun, F.: Hypernet: towards accurate region proposal generation and joint object detection. In: *Computer Vision and Pattern Recognition*, pp. 845–853 (2016)
17. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
18. Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside-outside net: detecting objects in con-text with skip pooling and recurrent neural networks, pp. 2874–2883 (2015)