The Feasibility of Deep Feature Pyramid for Semantic Segmentation

Wen Chen[1, a], Shuhao Ma[2], Jin Wang[1], Qing Zhu[1]

[1]Beijing University of Technology Beijing, China

[2]Dalian Maritime University Dalian,China

[a]cw_pro@163.com

Keywords: deep convolutional neural networks, semantic segmentation, image pyramid, multi-scale feature fusion, deep feature pyramid, pascal voc dataset

Abstract: Image pyramid is an significant approach to improve the performance of the convolutional neural networks (CNNs) on the computer vision tasks(e.g., image classification, object detection, image segmentation), but it takes huge computation time and occupies large memory. In this paper, image pyramid has been improved. We first point out that different tasks and different data sets have different requirements for feature fusion, and then we propose a more flexible approach, called Deep Feature Pyramid (DFP), which can alleviate the large memory requirements and a lot of computation time of image pyramid to some extent. Our approach enables the model to achieve good performance with less memory overhead and computation time. Compared with image pyramid, our method greatly reduces the memory occupancy and computation time, especially when the image scale is large. We validate the performance of DFP in semantic segmentation, which is a very demanding task for feature fusion, and the experimental data set is Pascal VOC2012.

1. Introduction

CNNS [1] have achieved significant performance on computer vision tasks that is attributed to the excellent ability to extract feature. Designing a strong extractor of feature is essential to computer vision task, fortunately, the previous works like [2, 3], which can capture image information in a variety of forms [4]: low -level edges, mid-level edge junctions, high-level object parts and complete objects, has provided the basis for us. In particular, the architecture of the CNNs has achieved remarkable achievement[5, 6, 7, 8, 9, 10] in image classification, object detection and image segmentation, compared with traditional approach [11]. A good performance of a model is often determined by many factors. One of the factors is mutil-scale feature fusion[12, 13, 14], which is a significant technology to improve the performance of the model. Multi-scale feature plays an important role not only in deep learning before rising, but also in deep learning. Employing multi-scale feature can further enhance the performance of the model with bottleneck state, in the tasks of image classification, semantic segmentation and target detection. In the domain of computer vision, image pyramid increases the depth of images by stacking different scales of images. In the end to end architecture, the images are first resized to different scales, before being feed into the network, and then these images are inferred and learned by the neural network through a shared net [15]. Shared net learns global information on small scale image, learns local information on large scale image. Combining global information and local information enables model has a comprehensive understanding of the image. However, image pyramid technology has great limitations in practical applications. The training model requires a lot of equipment, because the image pyramid technology needs to store numerous gradients in memory than single-scale input in error back-propagation process. Moreover the image pyramid also increases the computation time in the forward propagation process, which becomes more unfeasible for real-time tasks. This is an inherent drawback of the image pyramid technology, which is fatal for the application in practice.

2. Ease of Use

For the current characteristics of deep learning, we consider three challenges in computer vision tasks, all of which are currently possible using the methods shown in figure. 1 to enhance the performance of models. These approaches are successful examples of using multi-scale feature to solve these challenges.

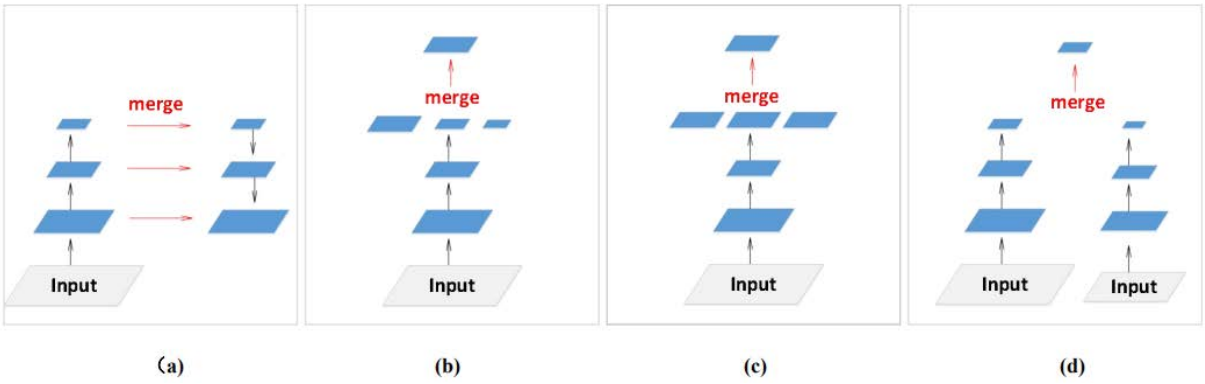

Figure 1. (a) Combination of encoding-decoding structure and skip connection approach. (b) This is the spatial pyramid pooling, which uses different down sampling steps to achieve different receptive field . (c) Atrous spatial pyramid pooling. Capturing different scales of information by inserting zero into the convolution kernel. (d) The typical image pyramid structure can be combined with the first three feature fusion methods.

The first challenge is how to improve the accuracy of image classification. Academia has been exploring the field of image classification for many years, and have made great progress. In particular, CNN has outperformed hunman beings. Although the accuracy of image classification is very high, however, this work [14] has shown that multi-scale feature fusion will improve the performance of CNN. Especially pyramid pooling technique not only has demonstrated the importance of mult-scale feature, but also enabled the model to accept arbitrary size of image, so that the model can accept multi-scale training. PSPnet is a variant of spatial pyramid pooling and has achieved good performance in semantic segmentation tasks.

The second problem is how to solve the small target loss problem in consecutive down-sampling operations, which allows CNNs to learn increasingly abstract feature representations. By enlarging the convolution kernel, the network can get a larger receptive field and have a excellent ability to extract features. However, the larger convolution kernel will greatly increase the memory consumption. Therefore, the commonly used network adopt the convolution kernel size is 3x3 and 1x1. [3]Work has proved that the use of small convolution kernel can increase the depth of the network with a certain memory limit, so that the network achieves a higher degree of non-linearity. However, with the increase of the depth of the network, the number of down-sampling increases, excessively small objects in the image may disappear in the high level feature map, which makes it difficult to detect small objects. This work[19] has shown multi-scale fusion technology will further improve the loss of small targets.The third difficulty comes from the spatial information loss in semantic segmentation by the invariance of CNNs.

This invariance is an important way to improve the robustness of the model in image classification tasks, but is an obstacle in semantic segmentation. Image semantic segmentation is a dense prediction task, and every pixel must be correctly classified. Since 2015, deep convolution neural network has made great progress in semantic segmentation, which benefits from FCN[20]. CNNs can achieve more robust features through consecutive down-sampling. However, in image semantic segmentation tasks, not only robust features but also precise spatial information are needed. However, consecutive down-sampling operations constantly lose spatial information, which makes it difficult for image semantic segmentation tasks requiring precise spatial information. At present, a series of multi-scale feature fusion methods appear in image semantic segmentation tasks, and have solved the contradiction to a great extent between classification and segmentation, especially ASPP

network, which has excellent feature extraction ability and avoids down-sampling operation. [21, 22, 23, 24, 25] These works have shown that multi-scale technology is great significance in semantic segmentation technology.

3. Prepare Your Paper Before Styling

3.1 DFP for Avoiding large memory consumption

Our goal is to introduce the mind of DFP, which enables models achieve more flexibility, and build a feature pyramid at different scales, for more computer tasks and different data sets. We advocate the use of DFP to avoid the waste of memory space and computation time. Firstly, we define depth, which can be applied to any architecture. In principle, we define the depth according to the number of network down-sampling. As shown in the figure 2, from the bottom to the top, the defined depth changes with the number of down-sampling. For example, the number of down-sampling is 5, the depth is 5, the number of down-sampling is 0, the depth is 0.

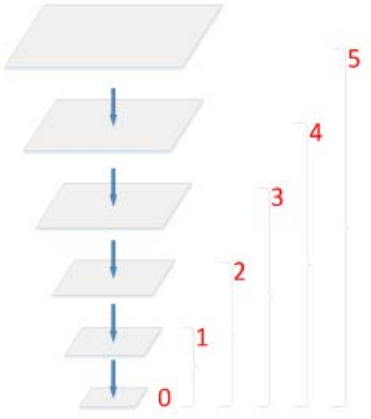

Figure 2. Illustrating the concept of depth

Figure. 3 is an architecture that combines the idea of image pyramid with the concept of depth. Two special cases, the network with depth of 5 will degenerate into the original image pyramid network, and the network with depth of 0 will degenerate into the network without image pyramid characteristics. The purpose of DFP is to explore a balance between resource consumption and model accuracy by avoiding these two special cases with the concept of depth.

There are 2 steps in the stage of extracting feature for any network using DFP idea.The first stage, network accepts a single scale input and does not have shared network strategy. This strategy avoids wasting memory and computing time on different scales for the low level convolution layer when image pyramids are adopted for multi-scale feature fusion. Generally speaking, the feature maps generated by the low level layer in the network are not often used in image classification and target detection. When adopt image pyramid strategy, with the increase of different scales, the low level layer will generate more feature maps, which occupy large memory and consume more computing time.

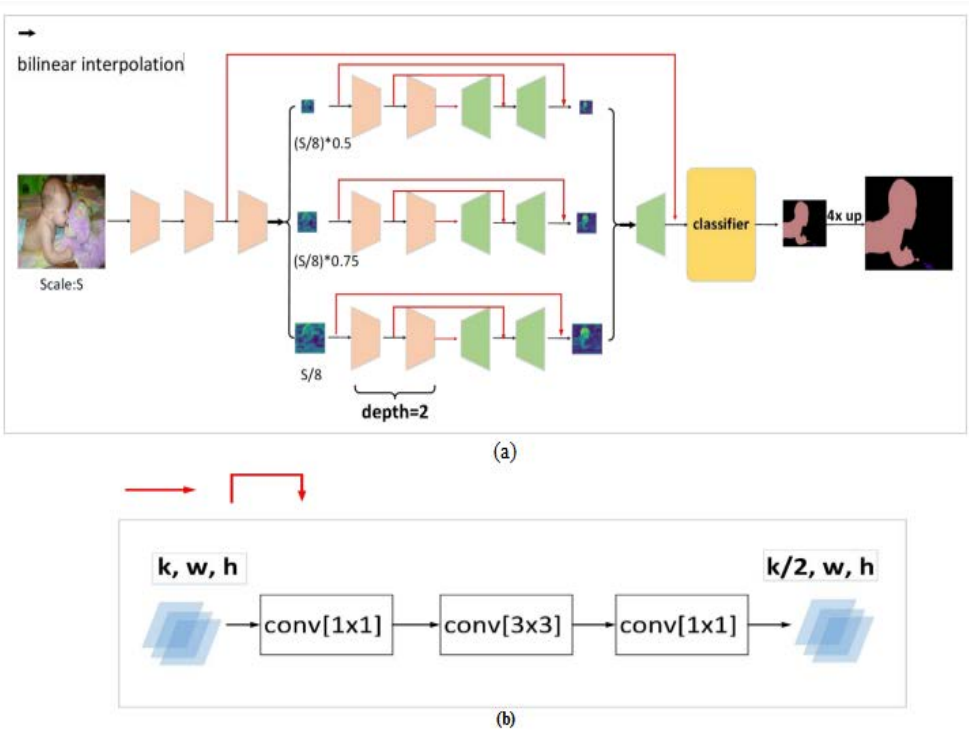

Figure 3. (a)Illustrating the model of FCN adopt DPF idea. In this illustration, the depth of the model is 2, and the bilinear interpolation is used to scale the feature map. (b) This is an encoder that compresses the feature map to k/2 channels.

3.2 DFP for Semantic Segmentation

Fully convolutional network is a success of deep convolutional neural networks models has enabled remarkable progress in pixel-wise semantic segmentation by end-to-end training. Recently work Mask RCNN implements the instance segmentation by adding FCN. In the FCN design, in order to retain the spatial information of the feature map, the fully connection layer is removed, and deconvolution is adopted in the up-sampling stage. But in our experiment, bilinear interpolation is used instead of deconvolution in the up-sampling stage, and the convolution kernel of 1x1 and 3x3 is used. In practice, using deeper resenet-152 and resnet-101 model generally yields better performance than vgg modes. In our experiment, we extract feature maps from multiple branch and up-sample feature map by parallel branch.

3.3 Experimental details

In our experiment, we combine the FCN with our method by removing the full connection layer of the resnet50 pre-trained on the Image-net as a network feature extractor. Our loss function is the sum of the cross entropy of each pixel. In our experiment, 30 epochs were iterated. Firstly, 10 epochs were trained, then the BN [26] parameters were fixed and the other 20 epochs were iterated. We use the stochastic gradient descent algorithm (SGD) [27] to optimize the objective function, the weight decay coefficient is 0.0005, the initial learning rate is 0.0005, the learning rate change strategy is $\mathrm{lr} = \mathrm{lr} * (\frac{total\_epoch - epoch}{total\_epoch})$, due to memory constraints, we set the batch to be 3, and crop patch to be 512x512. Random horizontal flip was adopted as the data augmentation strategy, and what we found is that the larger padding weakens the convolution kernel. Our proposed approach are implemented by pytorch and experiments run on devices GTX1060 with 6GB memory. Our evaluation criteria is mIoU:

$$\frac{1}{C}\sum_{i=0}^{C}\frac{p_{ii}}{\sum_{i=0}^{C}p_{ij}+\sum_{j=0}^{C}p_{ji}-p_{ii}}$$

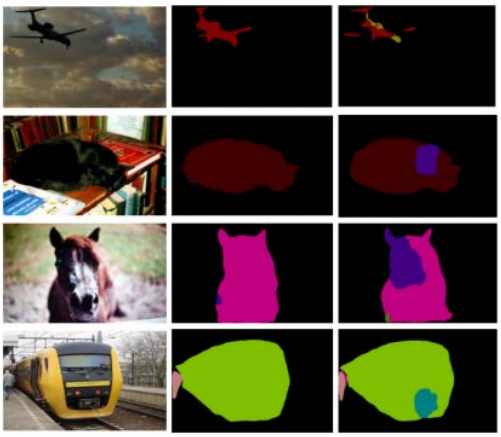

Figure 4. Comparison of the model's performance using the DFP method with the model's performance without the DFP method

3.4 Result analysis

As can be seen from tableI, the precision of the model is significantly higher than that of the non-pyramid (depth=0) method when the depth is 1. When the depth is 5, the model is equivalent to using the traditional pyramid method, compared with the depth of 4, 3 and 2, the precision of the model is not significantly improved, but the memory requirement and computation time increases sharply.Thus it can be seen that the traditional pyramid method is too rigid and not flexible enough, resulting in high demand for resources, while our work has explored the flexibility of the DFP thought and the ability to improve model accuracy no less than the pyramid method.

Table 1. Performance on PASCAL VOC 2012 test set. “training memory” is the consumption of memory during the training stage. “testing time” is the consumption of computing time for each batch during the test stage. We set batch=3. Scaling factors are 1.0, 0.75 and 0.5.

network	depth	training memory (MB)	testing time (sec)	mIoU
Resnet50	5	4677	0.098	69.87
	4	4459	0.095	69.66
	3	3994	0.086	69.72
	2	3609	0.063	69.69
	1	3349	0.059	68.26
	0	3275	0.050	63.79

4. Conclusion

Our proposed DFP is an open and flexible thought that applies multi-scale fusion technology. We define the concept of depth, which can be adjusted according to different tasks requirements to avoid the use of image pyramids caused excessive waste and difficult application of image pyramids technology. We have proven that out approach is not only improvement for performance of model ,but also with much less resource consumption.

References

[1] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition, in Proceedings of the IEEE, 1998.

[2] K. He, X.Zhang, S.Ren, and J.sun. Deep Residual Learning for Image Recognition. arXiv preprint arXive: 1512.03385, 2015.

[3] K. simonyan and A. Zisserman. Very deep convolutional networksfor large-scale image recognition. arXiv preprint arxiv:1409.1556, 2014.

[4] M. D. Zeiler, G. W. Taylor and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 2018 - 2025.

[5] A. krizhevsky, I. Sutskever,and G.E.Hinton. Imagenet classification with deep convolutional neural networks in NIPS, 2013.

[6] G. Papandreou, I.Kokkinos, and P-A. Savalle. Modeling local multiple instance learning, and sliding window detection, in CVPR, 2014.

[7] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248 - 255.

[8] T. Y. Lin, M.Maire,S.Blongie, J. Hays. Microsoft coco: Common objects in context,In ECCV, 2014.

[9] R.Girshick.Fast R-CNN. In ICCV, 2015. 1, 2, 3, 4.

[10] S.Ren.K.He,R.Girshick,and J.Sun. Faster R-CNN: To-wards real-time object detection with region proposal networks. In NIPS, 2015.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627 - 1645, Sept. 2010.

[12] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6230-6239.

[13] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834 - 848, 1 April 2018.

[14] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015.

[15] L. Chen, Y. Yang, J. Wang, W. Xu and A. L. Yuille, "Attention to Scale: Scale-Aware Semantic Image Segmentation," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 3640 - 3649.

[16] O.Ronneberger, P. Fischer, T.Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv preprint arxiv :1505.04597.

[17] V. Badrinarayanan, A. Handa, R. Cipolla. SegNet:A Deep Convolutional Encoder-Decoder Architecture for Robust SemanticPixel-WiseLabelling. arXiv preprint arxiv :1505.0729 3.

[18] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6517 - 6525.

[19] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 936 - 944.