Attention to Refine Through Multi Scales for Semantic Segmentation

Shiqi Yang and Gang Peng(✉)

Key Laboratory of Ministry of Education for Image Processing and Intelligence Control, School of Automation, Huazhong University of Science and Technology, Wuhan, China
{albert_yang,penggang}@hust.edu.cn

Abstract. This paper proposes a novel attention model for semantic segmentation, which aggregates multi-scale and context features to refine prediction. Specifically, the skeleton convolutional neural network framework takes in multiple different scales inputs, by which means the CNN can get representations in different scales. The proposed attention model will handle the features from different scale streams respectively and integrate them. Then location attention branch of the model learns to softly weight the multi-scale features at each pixel location. Moreover, we add an recalibrating branch, parallel to where location attention comes out, to recalibrate the score map per class. We achieve quite competitive results on PASCAL VOC 2012 and ADE20K datasets, which surpass baseline and related works.

Keywords: Semantic segmentation · Attention model · Multi-scale Context

1 Introduction

With the booming of deep learning, many visual tasks have made significant progress. For instance, semantic segmentation, also known as image labeling or scene parsing which aims at giving label for each pixel, has made great breakthroughs in recent years. Efficient semantic segmentation can facilitate plenty of other missions such as image editing.

Recent approaches for semantic segmentation are all almost based on Fully Convolutional Network (FCN) [13], which outperforms the traditional methods by replacing the fully connected layers with convolutional layers in classification network. The follow-up works have extended the FCN from several points of view. Some works [2,14] have introduced the coarse-to-fine structure with upsample modules like deconvolution to give the final mask prediction. And due to the usage of pooling layer, spatial size has decreased largely, for which dilated (or atrous) convolution [6,20] has been employed to increase the resolution of intermediate features and hold the same receptive field simultaneously.

R. Hong et al. (Eds.): PCM 2018, LNCS 11165, pp. 232–241, 2018.
https://doi.org/10.1007/978-3-030-00767-6_22

Other works mainly focus on two directions. One is to post-process the prediction from the CNN through Conditional Random Field (CRF) to get smooth output. These works [1,6,22] are actually ameliorating the localizing ability of the framework. Another direction is to ensemble multi-scale features. Because features from lower layers in CNN have more spatial information and ones from deeper layers have more semantic meaning and less location information, it is rational to integrate representations from various positions since location information is important for semantic segmentation. The first type method for multi-scale combines features from different stages with skip connection to get fused features for mask prediction, such as [6,13,18]. And another type is to resize input to several scales and pass each one with a shared network, it will produce final prediction using the fusion of multi-stream resulting features. There are also methods trying to exploit the capability of global context information, like ParseNet [12] which adds a global pooling branch to extract contextual features. And PSPNet [21] adopts a pyramid pooling module to embed global context information to achieve accurate scene perception.

Attention model has been all the rage in natural language processing area, such as [3], and it has also shown its effectiveness in computer vision and multimedia community recently [4,16,17,19]. It allows model to focus on specific relevant features. Attention-to-scale [7] is the first approach to introduce attention model into semantic segmentation for multi-scale. It takes in different scale inputs. For each scale, the attention model produces a weight map to weight features at each location, and the weighted sum of score maps across all scales is then used for mask prediction. But it only utilizes the feature from specific layer to generate attention, which may omit many contextual details, and this can not ensure that the attention model can guide network to get precise results.

Referring to attention-to-scale, we propose a new attention model in this paper, which also takes in multi-scale inputs but integrates features from different layers, similar to hypercolumns [10]. The attention model has two branch outputs, *i.e.*, one for location attention through which it drives network to focus on large objects or regions for small scale input and pay attention to small targets for large scale just like attention-to-scale, another branch is to recalibrate the score map per class since resulting features from several stages carry contextual information. The outputs from attention model will be applied to multi-scale stream predictions, and final mask prediction is a weighted sum of all these streams.

Our contributions are two aspects as follows:

(1) We introduce a novel attention model into multi-scale streams semantic segmentation framework, the final mask prediction is produced by merging the predictions from multiple streams.
(2) The attention model utilizes fused features from different positions of CNN, which carry more contextual information, and has two branch outputs, where one is for location attention and another is for recalibrating.

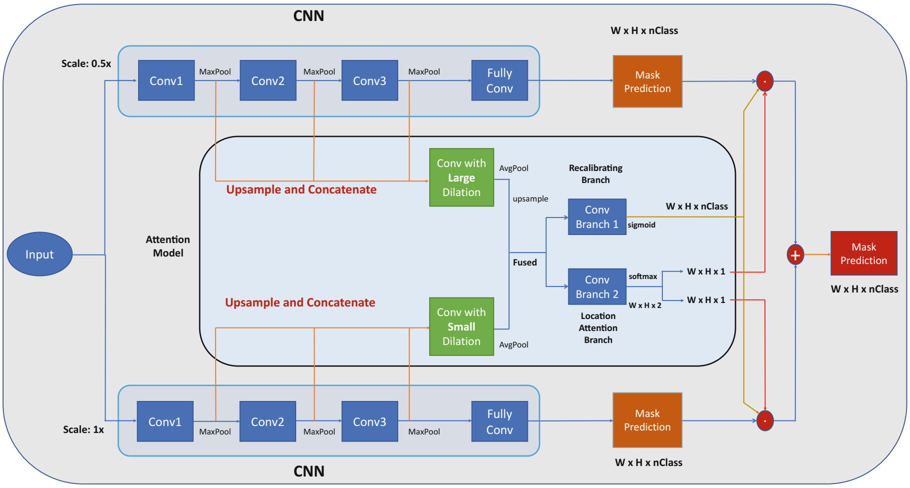

Fig. 1. Architecture of semantic segmentation framework with the proposed attention model. The attention model takes in features from different stages in CNN just like hypercolumns [10], and then it adopts convolutional layer with different dilation to process features for each scale respectively. Attention model produces two kinds of weight maps which are applied to multiple streams predictions. The final mask prediction is a sum of all streams.

2 Proposed Methods

2.1 Attention Model with Multi-scales

Like we mention before higher-layer features contain more semantic information and lower ones carry more location information. Fusion of information from several spatial scales will improve the accuracy of prediction in semantic segmentation. In addition, multi-scale aggregation also catch more contextual representations since some operations like pooling will dispose of the global context information, leading to local ambiguities which will be discussed later. It is the reason why multi-scale fusion gained a lot of popularity.

Since our work is extended from attention-to-scale [7], here we give a brief review on it. In attention-to-scale, the images are resized to several scales which will be fed to a weight-shared CNN, and the attention model takes as input the directly concatenating features from penultimate layer in each scale stream. The attention model consists of two convolutional layers and will produce n channels scores map, where n means the number of input scales. The attention model is expected to adaptively find the best weights on scales. But there exists some problems. The features from penultimate layer surely contain semantic representations, but they lack essential localization and global information fed to the attention model to achieve precise prediction. And we also posit that simply concatenating features from certain position is not conducive to lead the attention model to learn soft weight across scales. Seeing that the attention model is to put large weights on the large object or region in small-scale stream and gives large weights to the small targets in large-scale stream, we think it is rational to handle features from different scales respectively before integrating them.

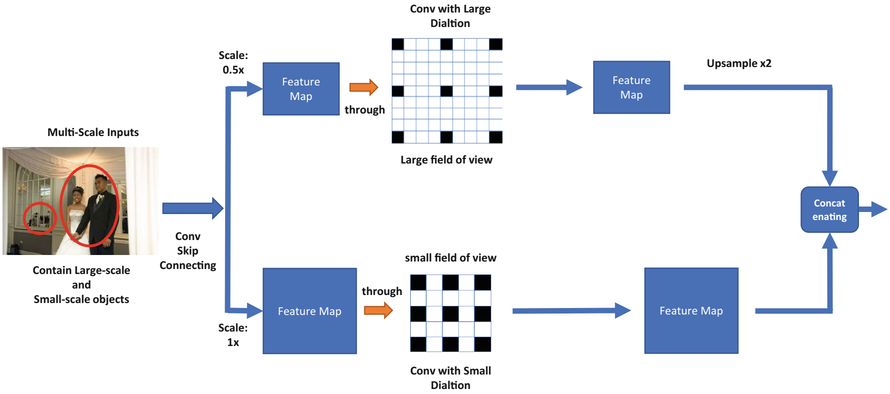

Fig. 2. Convolution with different dilation for different scale. Convolution with large dilation has large field of view while convolution with small dilation has small field of view.

Inspired by hypercolumns, we adopt the philosophy of it. Like depicted in Fig. 1, features from different stages in CNN get upsampled to same size and then we concatenate them all. To keep computation cost at bay, we choose the size of features after two pooling operation as the appointed resolution to do upsampling by bilinear interpolation. Through this way, the acquired features carry more localization and context information.

It is well-known that the structure of network has an impact on the range of pixels of the input image which correspond to a pixel of the feature map. In other words, filters will implicitly learn to detect features at specific scales due to the fixed receptive field. To accomplish our motivation of attention model which is to adaptively put weights on corresponding scale, we add a unique convolutional layer with unequal dilation for each scale. This process is demonstrated in Fig. 2. Convolution with large dilation has large field of view (FOV) and is expected to catch the long-span interlink of pixels for large scale object or region in small scale stream, and small dilation convolution is deployed to encode target of small scale in large scale stream. After the dilated convolution, the features will be concatenated, resulting one contains much more abundant and context information.

By the way, the two-stream CNNs in Fig. 1 are actually the same one when implemented in practice, just like Siamese Network.

2.2 Two Branch Outputs of Attention Model

The concatenated features will go through two parallel convolutional branches: location attention branch and recalibrating branch.

In common with attention-to-scale, the attention model will produce soft weights for multiple scales (we refer to it as location attention). Assuming the number of input scale is n, and the size of mask prediction, which is denoted as P^s for scale s, is W × H, *nClass* means the class number of the objects. The location attention output by the model is shared across all channels. After the refinement of local attention, the mask predictions, denoted as M_i^s, are described as:

$$M_{i,c}^s = \sum_{s=1}^{n} l_i^s \cdot P_{i,c}^s \tag{1}$$

The l_i^s is computed by:

$$l_i^s = \frac{\exp\left(wl_i^s\right)}{\sum_{j=1}^{n} \exp\left(wl_i^j\right)} \tag{2}$$

where wl_i^s is the score map produced by the location attention branch at position $i \in [0, W * H - 1]$ for scale s, before the softmax layer of course.

And since the fused features fed to the attention model contain context information, we want to make full use of them to eliminate some degrees of class ambiguity, *i.e.*, to utilize contextual relationship to enhance the ability of classification. The lack of ability to collect contextual information may increase the chance of misclassification in certain circumstances. To take an example, neural network sometimes tends to take apart a large-scale object into several regions of different classes [11], or maybe classify a boat on the river as a car and so on in scene parsing [21] (these can be observed among visualization results in Sect. 3.1). To deal with these issues, we add a recalibrating branch parallel to location attention. It has the same architecture as location attention branch which means containing two convolutional layers, except that output channel changes to *nClass* and *sigmoid* activation is deployed instead of softmax. This branch aims to find the interdependencies between adjacent objects or regions using the integrating features, and its output is used for recalibrating the score maps before the location attention refinement. Because the contextual relationship stay the same in different scale, the recalibrating outputs are shared across all scales. So the final mask prediction for each stream can be described as:

$$M_{i,c}^s = \sum_{s=1}^{n} l_i^s \cdot [P_{i,c}^s \otimes wr_{i,c}] \tag{3}$$

where the $\otimes$ means element-wise multiplication and $wr_{i,c}$ means output in position i in channel $c \in [0, n-1]$ produced by recalibrating branch. Another choice for recalibrating branch is to predict bias per position in each channel instead of multiplication. But it will bring around 1% performance decrease according to our experiment.

And the ultimate mask prediction is as below, where M^s is the mask prediction of scale s:

$$M_{final} = \sum_{s=1}^{n} M^s \tag{4}$$

As for the loss function, we follow the setting of attention-to-scale, *i.e.*, the total loss function is sum of $1+S$ cross entropy loss functions for segmentation, where S symbolizes number of scales and one for final prediction.

3 Experimental Results

We experiment our method on two benchmark datasets: PASCAL VOC 2012 [8] and ImageNet scene parsing challenge 2016 dataset [23] (it is from ADE20K [24], hereinafter we refer to it as ADE20k).

For all training, we only train the network with 2 scales, *i.e.*, 1x upsample and 0.5x upsample. As for the different dilation, we set it to 2 for small scale and 12 for large scale. And we use the poly learning rate policy [12], meaning current learning rate is computed by multiplying $(1 - \frac{iter}{max\_iter})^{power}$ to base learning rate, where the power is set to 0.9. We refer to the layers in the last stage where gives mask prediction as decoder, layers previous to decoder are encoder. Learning rate of decoder is 10 times that of encoder. All experiments are implemented using PyTorch on a NVIDIA TITAN Xp GPU.

3.1 PASCAL VOC 2012

The PASCAL VOC 2012 [8] segmentation dataset consists of 20 foreground object classes and a background class. The PASCAL VOC 2012 dataset we use is augmented with extra annotation by Hariharan *et al.* [9], resulting in 10582 training images. In experiment we report performance results on original PASCAL VOC 2012 validation set.

DeepLab-LargeFOV [5] is chosen as base model. Since our work is extended from attention-to-scale, in order to compare fairly, we reproduce the DeepLab-LargeFOV and attention-to-scale based on it by ourselves, following the set of

Table 1. Results on PASCAL VOC 2012 validation set. There exists 2 scale streams: 1x and 0.5x. The mIoU means mean intersection of union [13].

Method	mIoU
Baseline (DeepLab-LargeFOV)	61.40%
Merged with MaxPooling	63.88%
Merged with AvgPooling	64.07%
Attention-to-Scale	64.74%
Our method	**67.98%**

Table 2. Ablation study for proposed method on PASCAL VOC 2012. The multi-stage means hypercolumns-like feature integration from different positions. Diverse dilation means utilizing different dilated convolution for multi-scale features. Extra branch means adding recalibrating branch. *-The base model is actually attention-to-scale. †-No diverse dilations means using standard convolution instead.

Method	Multi-stage	Diverse dilations†	Location attention	Extra branch	mIoU
Base model*			√		64.74%
Base model+	√		√		65.80%
Base model++	√	√	√		66.83%
Our method	√	√	√	√	**67.98%**

attention-to-scale [7]. All these experiments use VGG16 [15] as skeleton CNN, which is pretrained on ImageNet. Our reproduction of them yields performance of 61.40% and 64.74% on the validation set respectively. The performance of attention-to-scale is lower than original paper, but the follow-up experiments still can verify effectiveness of our proposed method since ours is directly built on attention-to-scale. Noted that both of attention-to-scale and our work adopt extra supervision, meaning adding softmax loss function for each scale stream. The results of experiment are demonstrated in Table 1.

Merged with Pooling in Table 1 means adopting pooling operation as fusion approach for multi-scale stream instead of attention model. It can be seen that our method surpasses baseline and attention-to-scale by 6.58% and 3.24% respec-

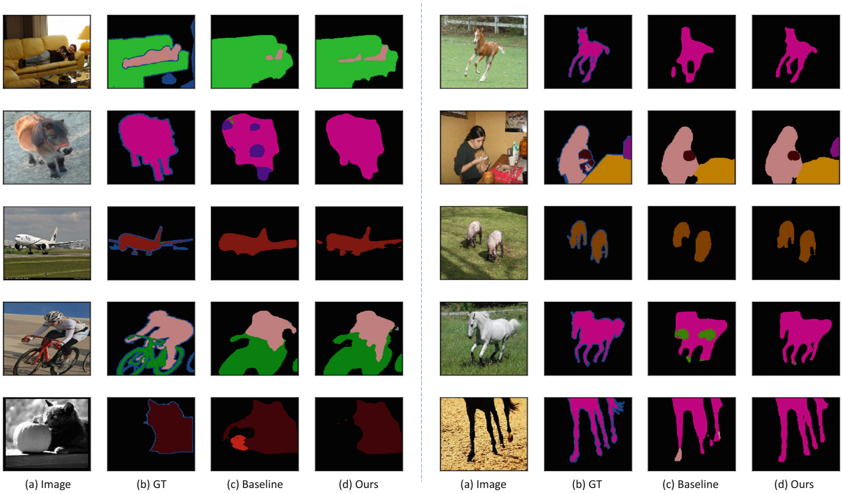

Fig. 3. Representative visual segmentation results on PASCAL VOC 2012 dataset. Images are from train and val set. GT means ground truth, and baseline means attention-to-scale approach. Our proposed method produces more accurate and detailed results.

tively. Furthermore, we conduct additional experiments for ablation study of each module in our method. We cut off certain modules from our proposed method, re-train and report the performance of remainder, which is shown in Table 2. Please noted that base model without all these modules is actually attention-to-scale approach. As you can see, the modules we design indeed take effect on segmentation task.

Since the attention-to-scale has verified the motivation which we share with by visualizing weight maps produced by the attention model, we don't replicate this experiment on our proposed model. Turning to qualitative results, some representative visual comparisons are provided between attention-to-scale and our method in Fig. 3. We observe that unlike attention-to-scale, our method can get finer contour in some cases and probability of breaking down a large-scale object into several pieces decreases. Our results contain much more detailed structure and more accurate pixel-level categorization, which we posit it comes from the utilization of multi-scale and context information as well as the extra branch.

3.2 ADE20K

ADE20K dataset first shows up in ImageNet scene parsing challenge 2016. It is much more challenging since it has 150 labeled classes for both objects and background scene parsing. It contains around 20K and 2K images in the training and validation sets respectively.

We deploy ResNet34-dilated8 [20] (not resnet50 because of limited GPU memory) as base CNN to investigate several different methods. Besides applying attention-to-scale and our proposed attention model, we also experiment on Pyramid Scene Parsing (PSP) [21] module as a comparison, which is a state-of-the-art approach on ADE20K dataset to the best of our knowledge. The experiment results are presented in Table 3. The PSP here doesn't contain auxiliary loss in original paper. We can see that our proposed attention model outperforms other methods, and achieves 4.40% improvement on mIoU over baseline. Besides, we also embed both the PSP module and proposed attention module in baseline and it obtains further performance improvement.

Table 3. Results on ADE20K validation set. *- Two multi-scale attention methods take as input two scale streams: 1x and 0.5x.

Method	mIoU	Pixel accuracy
ResNet34-dilated8 (Baseline)	32.67%	76.41%
Baseline + attention-to-scale*	35.11%	76.82%
Baseline + PSP	36.43%	78.01%
Baseline + our attention model*	**37.07%**	78.57%
Baseline + our attention model* + PSP	**38.21%**	79.29%

4 Conclusion

In this paper, we propose a novel attention model for semantic segmentation. The whole CNN framework takes in multi-scale streams as input. Features from different stage of CNN are fused, then resulting one in each scale goes through convolutional layers with different dilation, which are expected to catch distinctive context relationship for different scales. After that, all these features get concatenated and resulting one is fed into two parallel convolution output branches of the attention model. One of the branches is location attention, aiming to pay soft attention to each location across channels. Another one is designed to fully utilize contextual information to deal with class ambiguity by recalibrating the prediction per location for each class. Experiments on PASCAL VOC 2012 and ADE20K show that proposed method make a significant improvement.

References

1. Arnab, A., Jayasumana, S., Zheng, S., Torr, P.H.S.: Higher order conditional random fields in deep neural networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 524–540. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_33
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Patt. Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Chen, J., Zhang, H., He, X., Nie, L., Liu, W., Chua, T.S.: Attentive collaborative filtering: multimedia recommendation with item-and component-level attention. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335–344. ACM (2017)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062 (2014)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Patt. Anal. Mach. Intell. **40**(4), 834–848 (2018)
7. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2016)
8. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
9. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 991–998. IEEE (2011)
10. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 447–456 (2015)

11. Li, X., et al.: FoveaNet: perspective-aware urban scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 784–792 (2017)
12. Liu, W., Rabinovich, A., Berg, A.C.: ParseNet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
14. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528 (2015)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Song, X., Feng, F., Han, X., Yang, X., Liu, W., Nie, L.: Neural compatibility modeling with attentive knowledge distillation. arXiv preprint arXiv:1805.00313 (2018)
17. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
18. Xia, F., Wang, P., Chen, L.-C., Yuille, A.L.: Zoom better to see clearer: human and object parsing with hierarchical auto-zoom net. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 648–663. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_39
19. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
20. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
21. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
22. Zheng, S., et al.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1537 (2015)
23. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20K dataset. arXiv preprint arXiv:1608.05442 (2016)
24. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)