Learning Adaptive Receptive Fields

Hongzhi Liu

Jun 18, 2018

Abstract

In deep neural network, the notion of receptive field refers to the extent of data that are path-connected to a neuron. After the introduction of Fully Convolutional Network (FCN), receptive field has become especially important for deep image parsing network and could significantly affect the network's performance. Today, I read a thesis written by Zhen Wei, who is from State Key Laboratory of Information Security. His team introduce a novel approach to regulate receptive field in deep image parsing network automatically. By end-to-end training, the whole framework is data-driven without laborious manual intervention. Besides, they demonstrate the method's superior regulation ability over manual designs.

1. Overview of Regulating Receptive Field

The introduction of FCN [6] has placed the receptive field in a prominent position. The forward process of FCN to generate dense classification result is equal to a series of inference using sliding windows on input image. However, dilation designs in question are all based on trials or designers' observation on dataset. This is not difficult but rather laborious and time-consuming. This paper written by Zhen Wei is the first trial to replace such process with an automatic way.

Recent works have already accentuated on adapting network's structures to realizing different receptive fields such as [1]. By setting different dilation values, the convolutional kernels could expand its receptive field accordingly. However, there are several main drawbacks in this approach that should be addressed. Firstly, these dilation values are always treated as hyper-parameters in network design. Secondly, such selection results are not generic across different image parsing tasks or even various dataset under the same task. During network transfer, such selection procedure would be repeated again. Thirdly, dilated convolutional kernels only produce discrete values of receptive fields, making it even harder to find a finer receptive field.

In this paper, Dr. Wei proposes a learning based data-driven method for regulating receptive field in deep image parsing network automatically [8]. The main idea is to introduce a novel affine transformation layer before the convolutional layer whose receptive field needs to be regulated. This inflation layer uses derivable interpolation algorithms to enlarge or shrink feature maps. The following layers perform inference on these inflated features and thus receptive fields after the inflation layer are changed. Then inference results will be resized to a fixed size by *'inflation layer'*. To corroborate the method's effectiveness, they conduct experiments on both general image parsing task as well as face parsing task. With proper initialization settings, the proposed method could achieve even superior performance comparing to the best manually selected dilated convolutions.

2. Framework of Method for Regulating Receptive Field

Wei further elaborates on the details of their methods in the paper, including an overview on modified network structure, implementation of the inflation layer and interpolation layer and a loss guidance for multi-path network to realize a multi-scale inference with our data-driven method.

Fig. 1 presents the details of the framework. In the single path network, the inflation layer and the interpolation layer are inserted before fc6 layer and after fc8 layer respectively. The regulation of receptive field is operated on pool5 features.

While in multi-paths version, the initializations of each parallels are the same. In order to break this symmetry and achieve discriminative, multi-scale inference, a loss guidance layer is added to enforce each parallel focus on different scales.

2.1. The Affine Transformation Layers

The affine transformation layers include *the inflation layer* and *interpolation layer*. The inflation layer learns a parameter f, standing for the inflation factor. That is, the feature map will be enlarged by f times before the following convolution operations. There are two steps in the inflation operation, namely coordinate transformation and sampling. To formulate the first process, let (x_i^s, y_i^s) and (x_i^t, y_i^t) to

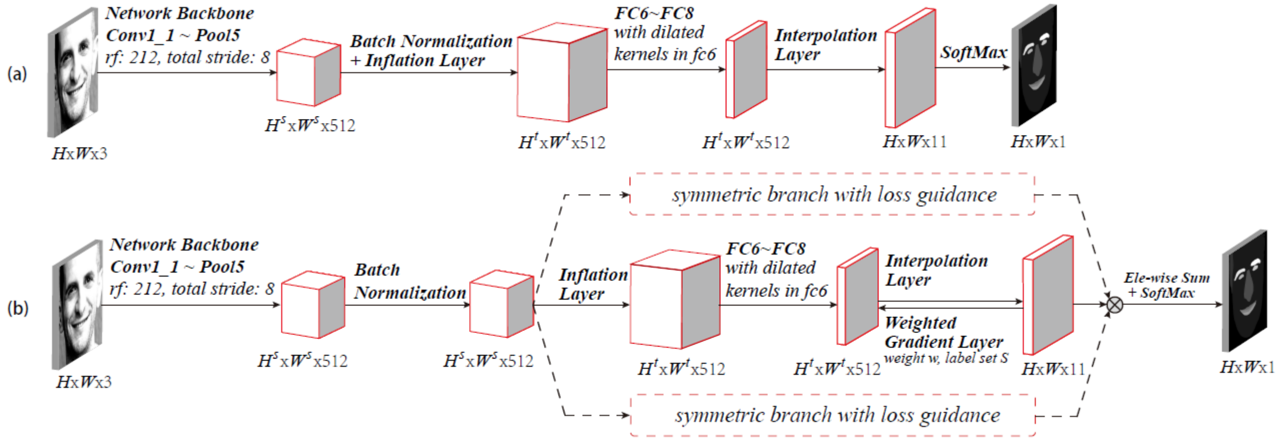

Figure 1. The framework of the method. (a): modified single path network. New layers are inserted before fc6 layer and after fc8 layer. (b): modified multi-paths network where all branches are with the same structure and initialization. *Weighted Gradient Layers* are used to break the symmetry during training.

be the coordinates in the source feature map and the target feature map respectively.

The inflation process builds up an element-wise coordinate projection as Equation 1:

$$x_i^t = f \cdot x_i^s, \quad y_i^t = f \cdot y_i^s. \tag{1}$$

Also, the size of the feature map changes accordingly as Equation 2:

$$H^t = f \cdot (H^s - 1) + 1, \quad W^t = f \cdot (W^s - 1) + 1. \tag{2}$$

where H and W are the height and width of feature maps, superscript t means 'target' and s means 'source'.

Wei normalizes G_{inf} by dividing $H^t \times W^t$, which is the number of pixels in a target feature map as Equation 3:

$$G_{inf} = \frac{1}{H^t W^t} \sum_{c}^{C} \sum_{i}^{H^t \times W^t} \frac{\partial Loss}{\partial V_i^c} \frac{\partial V_i^c}{\partial f}. \tag{3}$$

In practice, Wei and his team simply add these two gradients together to update the inflation factor f as Equation 4:

$$\frac{\partial Loss}{\partial f} = G_{inf} + G_{inter}. \tag{4}$$

In this way, it is possible to learn the inflation factor during the end-to-end training.

2.2. The New Receptive Field

To calculate the range of new receptive fields, the team can transform the question to obtain an equivalent kernel size of fc6 layer while feature maps are unchanged. Denote the original kernel size as k, the new equivalent size is $k' = \lceil (k+1)/f \rceil$. Thus the extent of the new receptive field is $212 + 8 \times (k' - 1)$, where 212 is the receptive field in pool5 layer, 8 is the overall stride from $conv1$ layer to pool5 layer in the network backbone.

2.3. Loss Guidance for Multipaths Network

Deep networks with multi-scale receptive fields have brought performance improvement in image parsing task [2]. This kind of network usually has several slightly different parallels to achieve multiple receptive fields. Our method can be also used in similar structures to realize further improvement and take place of hand-craft dilated convolutional kernels. To achieve this, as shown in Fig. 1 (b), fc6, fc7 and fc8 layers are first copied to make parallels. The output of fc8s are fused by a summation operation. Then, inflation and interpolation layers are inserted before each fc6 layers and after fc8 layers. A shared BN layer is appended after pool5.

To break this symmetry of framework, a weighted gradient layer is added behind each interpolation layers during training. To formulate this process, the team have a method as Equation 5:

$$G^c_{s,i} = wG^c_{t,i}, \quad w = \begin{cases} W, & \text{if} \quad l_i \in S \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

$G^c_{s,i}$ comes from source feature maps while $G^c_{t,i}$ comes from target feature maps. Such weighted gradients will induce each branch to focus on different label, scales and thus lead to obtain discriminative receptive fields.

3. Comparison with Previous Face Parsing Method

Tab. 1 shows a quantitative face parsing comparison between their method and other state-of-the-art methods. Wei and his team use reported results from [4], [7] and [5]. Their method uses the single path network with the initial dilation value of 8. Even without CRF or RNN post-process, their method still achieves the highest accuracy.

Table 1. Quantitative evaluation on Helen dataset [3]. Their method achieves state-of-the-art performance on face parsing task.

Model	F-score
Liu *et al.* [4]	0.738
Smith *et al.* [7]	0.804
Liu *et al.* [5]	0.847
Their method	0.9021

References

[1] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint arXiv:1412.7062*, 2014. 1

[2] L. C. 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. *IEEE TPAMI*, 2018. 2

[3] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012. 3

[4] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE TPAMI*, 2011. 2, 3

[5] S. Liu, J. Yang, C. Huang, and M. H. Yang. Multi-objective convolutional learning for face labeling. In *CVPR*, 2015. 2, 3

[6] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[7] B. Smith, l. Zhang, J. Brandt, Z. Lin, and J. Yang. Exemplar-based face parsing. In *CVPR*, 2013. 2, 3

[8] Z. Wei, Y. Sun, J. Wang, H. Lai, and S. Liu. Learning adaptive receptive fields for deep image parsing network. In *CVPR*, 2017. 1