Global Deconvolutional Networks for Semantic Segmentation

Vladimir Nekrasov
nekrasowladimir@unist.ac.kr
Janghoon Ju
janghoon.ju@unist.ac.kr
Jaesik Choi
jaesik@unist.ac.kr

Ulsan National Institute of Science and Technology
50 UNIST, Ulju, Ulsan, 44919 Korea

Motivation. Semantic segmentation is a crucial computer vision task, solving which would enable a thorough scene understanding of the environment. The areas that already benefit from the automatic semantic segmentation include biomedical imaging [2], autonomous driving [4]. Further enhancement of current models will necessarily increase the number of possible applications, as well as quality of performance.
The transfer learning of deep convolutional networks pre-trained for image classification on ImageNet has proven to be successful in semantic segmentation. In these models, last fully-connected layers are replaced by convolutional ones followed by a learnable deconvolution or fixed interpolation to acquire the output of the same spatial size as the input. Usually, the segmented mask is coarse. Several ways to deal with this have been proposed, including the 'skip'-layer architecture [3] and post-processing with probabilistic graphical models [1].
Algorithm. The combination of graphical models with deep networks requires carefully designed differentiable operations to mimic approximate inference, while traditional upsampling approaches tend to operate only locally. To overcome these issues, we propose an alternative novel architecture aimed to perform an upsampling globally, as well as enforce the correct label recognition. For the first task, we propose the equivalent of deconvolution, which we call *'global interpolation'*. We denote the decoded information of the RGB-image $\mathbf{I} : \mathbf{I} \in \mathbb{R}^{3\times H\times W}$, as $\mathbf{x} : \mathbf{x} \in \mathbb{R}^{C\times h\times w}$, where C represents the number of channels, h and w define the reduced height H and width W, respectively. To acquire $\mathbf{y} : \mathbf{y} \in \mathbb{R}^{C\times H\times W}$, an upsampled signal, we apply the following formula:

$$\mathbf{y_c} = \mathbf{K_h}\mathbf{x_c}\mathbf{K_w^T}, \forall \mathbf{c} \in \mathbf{C} \qquad (1)$$

where the matrices $\mathbf{K_h} \in \mathbb{R}^{H\times h}$ and $\mathbf{K_w} \in \mathbb{R}^{W\times w}$ are interpolating each feature map of $\mathbf{x}$ through the corresponding spatial dimensions. Contrary to a simple bilinear interpolation, which operates only on the closest four points, Eq. (1) allows to include much more information on the rectangular grid. Besides that, we append an additional *multi-label classification loss* to correct the wrong predictions of the network. The complete architecture can be seen in Figure 1.

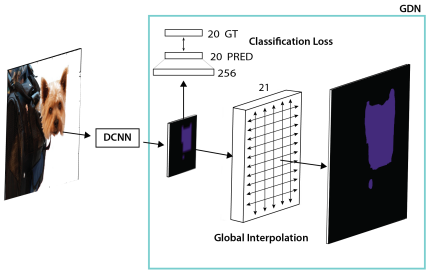

Figure 1: **Global Deconvolutional Network**. Our adaptation of FCN-32s [3]. We upsample the reduced signal with the help of global interpolation and append the multi-label classification loss to increase the recognition accuracy.

Results. We evaluate the proposed approach extending two publicly available models: FCN [3] and DeepLab [1]. We show the superior performance over them and achieve 74.02% mean IoU on the test set of the PASCAL VOC benchmark, which is close to the state-of-the-art performance without exploiting larger datasets.

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[2] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *NIPS*, 2012.

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

[4] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009.