Efficient Smoothing of Dilated Convolutions for Image Segmentation
Deep Learning - Project Proposal

Konstantin Donhauser, Manuel Fritsche, Lorenz Kuhn, Thomas Ziegler
{donhausk, manuelf, kuhnl, zieglert}@ethz.ch

Abstract
Dilated convolutions have proven highly useful for the task of image segmentation. While they allow the inexpensive capturing of features at different scales, the structure of the dilated convolutional filter leads to a significant loss of information. We hypothesise that inexpensive modifications to Dilated Convolutional Neural Networks, such as additional averaging layers, could overcome this limitation. In this project, we aim to test this hypothesis by evaluating the effect of these modifications for a state of the art image segmentation system.

1 Introduction
The goal in semantic image segmentation is to partition images and label the resulting segments to facilitate the analysis and interpretation of the images. Developing good segmentation algorithms is crucial for many real-world applications such as medical image processing [1], autonomous driving [2] or SLAM systems [3].

Convolutional Neural Networks have proven to be successful in semantic image segmentation [4, 5, 6, 7, 8, 9] and are used in most state of the art algorithms. The DeepLab system, originally presented in [10] and then improved upon in [11, 12, 13], combines several methods and insights to achieve the current state of the art performance on common data sets [14, 15]. They leverage Spatial Pyramid Pooling [16] in combination with dilated convolution (also called atrous convolution) to cheaply and robustly segment objects at different scales. However, dilated convolution introduce a loss of information (described in more detail below).

Recently, several approaches have been developed to reduce the loss of information that follows from the sparseness of the dilated convolution filters. In [17, 18] a strategy is proposed for selecting the rate of sequential dilated convolution layers which reduces the information loss. In [19] additional convolutional layers are used to smooth the input of dilated convolution layers. In contrast, our approach is to do smoothing with inexpensive filtering methods without adding a lot of additional parameters to be trained.

2 Problem Statement
One challenge in image segmentation is that objects may appear at different scales - both within the same image but also between images - which poses a problem for classical convolutional layers. Learning features on different scales is thus both difficult and essential to provide reliable image segmentation. DeepLab [11] overcomes this issue by using dilated convolutions. These convolutions effectively introduce gaps into the filters, which allows for increased filter sizes without introducing more weights. This results in a sparse sampling of the input signal x with rate r. For a filter w this can be written in 1-D as $y[i] = \sum_{k=1}^{K} x[i + r \cdot k]w[k]$.

The method has proven to work well in practice and allows us to deal with object dependencies on different scales without reducing the image resolution. However, because of this sparse sampling only a few points are taken into account for potentially large parts of the image. If these points are noisy or simply bad representatives of the surroundings, the dilated convolution will yield bad results. Moreover, gridding effects occur, which result in uncorrelated neighbouring pixels in the output [17, 19]. This effect introduces a spatial information loss since neighbouring input pixels are usually related to each other.

3 Approach
Seeing these issues, we propose inexpensive modifications to the dilated convolution to make it more robust to local noise and encode more information. Rather than performing dilated convolutions directly on features, we first apply an additional interpolation filter $v[n]$ to capture more of the local information and then compute the dilated convolution on the result. Mathematically, this can be expressed as $y[i] = \sum_{k=1}^{K} \left(\sum_{n=1}^{N} x[i + r \cdot k + n]v[n] \right) w[k]$. Considering that current image segmentation systems already have millions of trainable parameters, a central goal of our project is to find more efficient adaptations than the ones presented in previous works [17, 19]. We will start with evaluating the most obvious interpolation filters, a simple moving average and a Gaussian filter. We intend to compare these to trainable convolutional layers and aggregations of multiple filters.

As a baseline we take the performance of the DeepLabv2 system [12] on the data sets Cityscapes [14] (5000 data points) and PASCAL VOC 2012 [15] (13000 data points). Both data sets contain images with pixel-level annotations. In addition to the baseline provided in the original paper, we will compare our results to the scores achieved by the two "degridding" approaches presented in [17] and [19]. We will use mIOU as our main measure of segmentation quality. Since we are also aiming to be more efficient than existing methods, we will also compare the training times of the different algorithms.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1013–1020, IEEE, 2018.

[3] A. Harati, S. Gächter, and R. Siegwart, "Fast range image segmentation for indoor 3d-slam," *IFAC Proceedings Volumes*, vol. 40, no. 15, pp. 475 – 480, 2007. 6th IFAC Symposium on Intelligent Autonomous Vehicles.

[4] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *European Conference on Computer Vision*, pp. 297–312, Springer, 2014.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[6] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 447–456, 2015.

[7] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3992–4000, 2015.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[9] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658, 2015.

[10] L.-C. 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.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[12] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[13] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[15] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European conference on computer vision*, pp. 346–361, Springer, 2014.

[17] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1451–1460, IEEE, 2018.

[18] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1442–1450, IEEE, 2018.

[19] Z. Wang and S. Ji, "Smoothed dilated convolutions for improved dense prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2486–2495, ACM, 2018.