Multi-Scale Context Aggregation by Dilated Convolutions
Machine Learning - Project

Dren Gashi, Mike Pereira, Valeriia Vterkovska
Master in Information and Computer Sciences
University of Luxembourg

December 10, 2017

1 Introduction

Computer vision, the idea of enabling computers ”to see” as we do, has caught the interest of many researchers in the past decade. Many problems in computer vision are dedicated to dense or pixel-wise prediction, the task of predicting a label for each pixel in an image. This leads to *semantic segmentation*, where we aim to partition the image into semantically meaningful pieces or segments and to classify each part into one of predefined classes of its enclosing object or region such that multiple pixels with the same classification share the same characteristics. This enables identifying boundaries and objects such as animals, backgrounds, vehicles or human beings. Furthermore, some real world use cases are detecting road signs, tumors, classifying land use and land cover or detecting medical instruments in operations [4].

Convolutional neural networks (CNN) had a big success on segmentation problems and were mainly used [3]. However, they came with compromises. Fully connected layers increased the complexity and *pooling layers* or down sampling layers, that are being used for increasing the receptive field, however, decrease the resolution. The receptive field only grows linearly for CNN. This is where dilated convolution layers [2] are proposed. They enable an exponential expansion of the receptive field while the parameters are growing linearly as one can observe in figure 2. The goal is to make use of local pixel-level accuracy but also to gather global knowledge of a bigger window / context.

2 Proposed Solution

Dilated convolution is a convolution with a stride in the filter. In the past, the dilated convolution has been referred to as ”convolution with a dilated filter”. Dilating the filter means expanding its size and filling the empty positions with zeros. The dilated convolution operator can be applied with the same filter at different ranges using different dilation factors [2] . Dilations create new lattice points between each of the existing points and assign zero values since it doesn’t change the inner product (figure 1). Therefore, dilated convolutions are a way to upscale the filter. A so called front-end module is used that replace the last two pooling layers from classification networks with dilated convolutions. A separate context module is trained separately using the outputs of front-end module as inputs. This module aggregates multi-scale contextual information in order to increase the performance of dense predictions.

2.1 Exponential Expansion

Dilated convolutions have the advantage that they enable an exponential expansion of the receptive fields without loss of resolution or coverage. For instance, in figure 2, we can see that we can obtain very large receptive fields using just a few layers [5]. The receptive field of non-dilated CNN is growing linearly in the other hand. For example, let C_n be a convolutional network of $n \times n$ convolutions. Then, the receptive field size will be $i(n-1)+n$ where i is an index pointing to the layer, which shows a linear expansion [1]. Dilation rate $d = 1$ gives a standard convolution, $d = 2$ means skipping one pixel per input and $d = 4$ means skipping 3 pixels, et cetera. In figure 3, the red dots represent the inputs of a 3×3 filter, and the green area depicts the receptive field captured by each of the inputs. A so-called *front-end* module is developed where all subsequent layers are dilated by a factor of 2 for each omitted pooling layer. Convolutions in the final layer have dilation 4. Hence, dense predictions are obtained without any increase in number of parameters but rather using the same as of the original classification network. The receptive field is the implicit area captured on the initial input by each input (unit) to the next layer. The dilated convolution between signal f and kernel k and dilution factor $l\tau$ is

$$(k *_l f)_t = \sum_{\tau=-\infty}^{\infty} k_t \cdot f_{t-l\tau}$$

as described in [1]. For simple convolution (rate $D = 1$) this would become $f_{t-\tau}$. In the dilated convolution, the kernel only touches the signal at every l-th entry [2].

2.2 Classification using feature maps

A so-called *context module* is designed to increase the performance of dense prediction architectures by aggregating multi-scale contextual information. It preserves the resolution/dimensions of data at the output layer. This is because the layers are dilated instead of pooled, hence the name dilated causal convolutions. The context module allows one to have larger receptive fields with same computation and memory. The module takes C number of feature maps as input and produces C feature maps as output. The input and output have the same form, thus the module can be plugged into existing dense prediction architectures. The more important point is that the architecture is based on the fact that dilated convolutions support exponential expansion of the receptive field without loss of resolution or coverage. Standard initialization procedures do not support the training as observed during experiments [2], hence convolutional networks are usually initialized using random samples.

3 Results and Conclusion

For our tests, the Cityscapes dataset has been used. The resulting output images can be found in the appendix. In terms of accuracy, the proposed implementation beats any prior state-of-the-art implementation. The use of dilated convolution shows a significant advantage over traditional semantic segmentation systems, since it conserves both the resolution and coverage of the input image.

References

[1] F. Huszr. "Dilated Convolutions and Kronecker Factored Convolutions". [Online] Accessed: Dec 5, 2017. Available under
`http://www.inference.vc/dilated-convolutions-and-kronecker-factorisation/`.

[2] F. Yu, V Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions". Conference paper at ICLR 2016.

[3] J. Long *et al.* "Fully Convolutional Networks for Semantic Segmentation". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7-12 June 2015.

[4] M. Thoma. "*A Survey of Semantic Segmentation*". pp. 116, 2016. Available: `http://arxiv.org/abs/1602.06541`

[5] O. Aaron van den *et al.* "WaveNet: A Generative Model for Raw Audio." *CoRR abs/1609.03499 (2016).*

Appendix

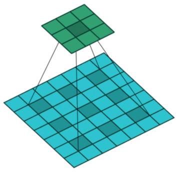


Figure 1: A $k \times k$ filter has been dilated by internally padding with zeros.

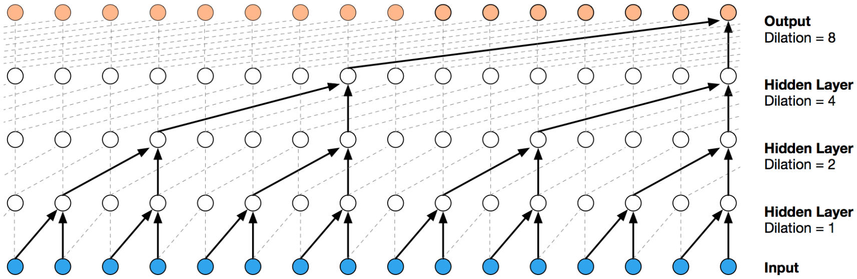


Figure 2: Visualization of a stack of dilated convolutional layers.

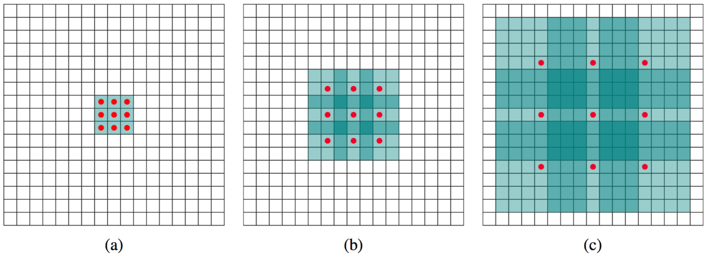


Figure 3: (a) F_1 is produced using a 1-dilated convolution of F_0. Each element in F_1 has a receptive field of 3×3. (b) F_2 is obtained using a 2-dilated convolution of F_1. Each element in F_2 has then a 7×7 large receptive field. (c) F_3 is obtained by a 4-dilated convolution of F_2, resulting in a receptive field size of 15×15. The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly.

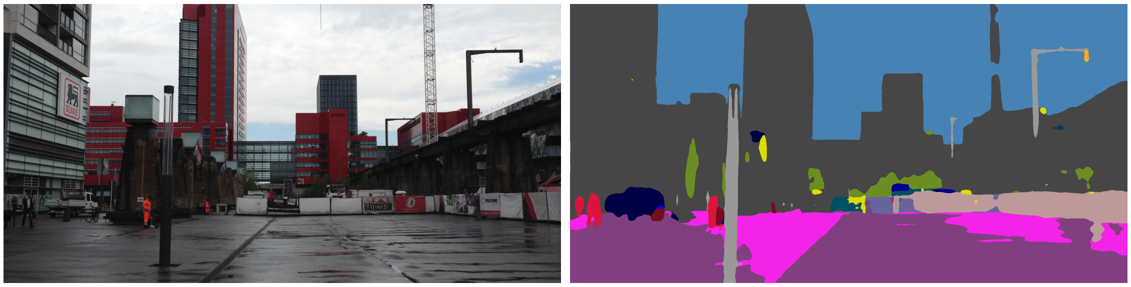


Figure 4: Luxembourg - Belval

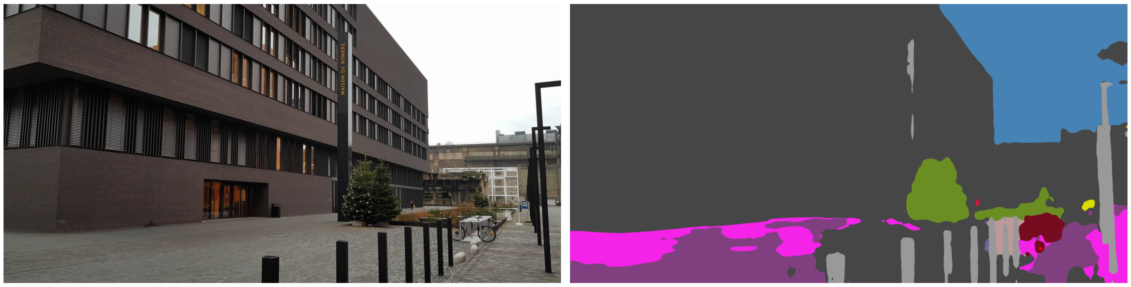


Figure 5: University of Luxembourg - Building MNO

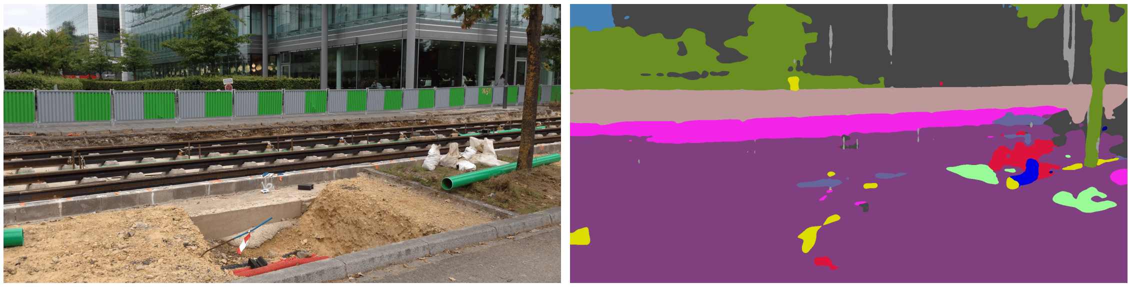


Figure 6: Luxembourg - Kirchberg - Tram rail construction

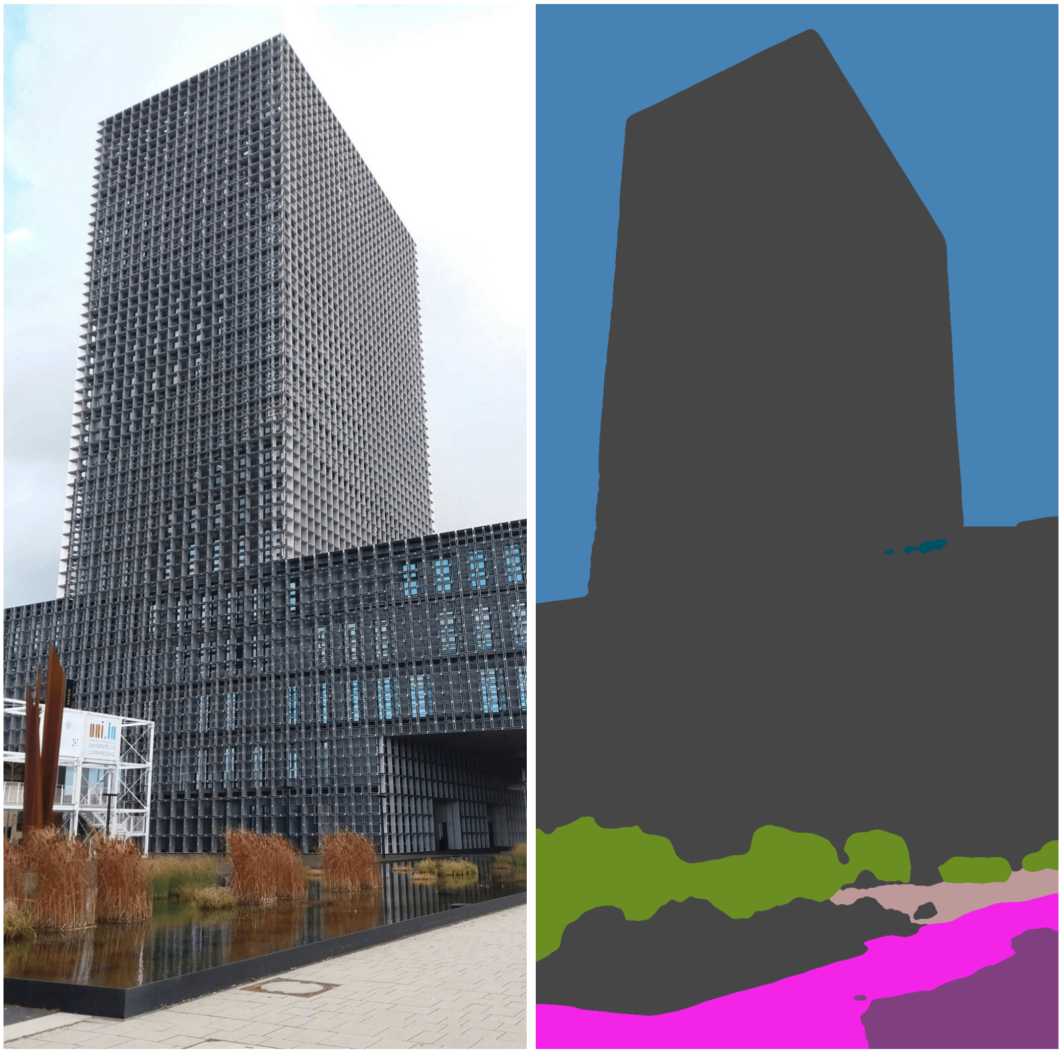
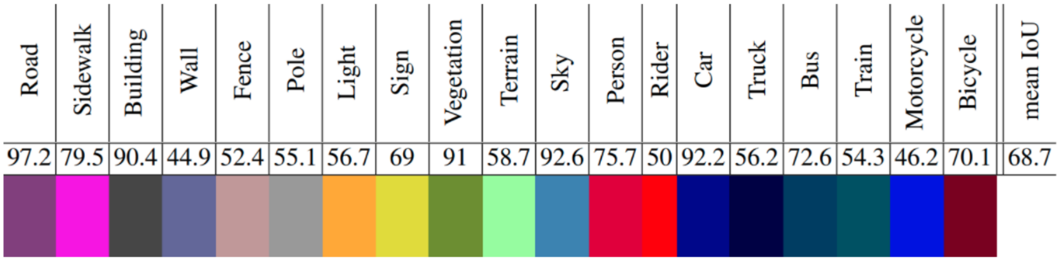


Figure 7: University of Luxembourg - Building MSA



Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mean IoU
97.2	79.5	90.4	44.9	52.4	55.1	56.7	69	91	58.7	92.6	75.7	50	92.2	56.2	72.6	54.3	46.2	70.1	68.7
🟪	🟪	⬛	🟪	🟫	⬜	🟧	🟨	🟩	🟩	🟦	🟥	🟥	🟦	⬛	🟦	🟦	🟦	🟥	

Figure 8: Cityscapes classification, detection accuracy and used color palette