Hands-on Machine Learning Training

Session 8 – Semantic Segmentation with Deep Learning

Theoretical Preparation

In the following session, you will implement and try out different techniques for semantic segmentation using deep learning algorithms. The theory of these techniques and algorithms should be understood before the session, so that you can focus on the implementation during the restricted time of the session.

For an overview of semantic segmentation with deep learning, you can find a good review here:

- Y. Guo, Y. Liu, T. Georgiou, M. S. Lew, *A review of semantic segmentation using deep neural networks* in International Journal of Multimedia Information Retrieval, June 2018.

The review can also be found on Moodle. Do not worry, we will not ask for details of the review during the test. However, we recommend going through the review before reading the papers, since you will get a nice introduction to the topic as well as further literature.

The topics covered are explained in the following papers:

- O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation* in MICCAI, 2015.
- J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation* in CVPR, 2015.

The two papers can be found on Moodle. Please read these and understand the basic concepts behind them. As there are inconsistencies in literature regarding nomenclature, let us clarify some points and define the nomenclature used in this session.

- You will come into contact with two types of upsampling layers. The first is *bilinear upsampling* as known from signal processing. The second is *transposed convolution*, also reffered to in literature as *deconvolution* or *fractionally strided convolution*.
- *Dilated convolutions* are synonymous with *atrous convolutions*.
- *Skip-connections* will refer to concatenation operations, whereas *Shortcuts* will refer to summation operations.

Also, animations as given here https://github.com/vdumoulin/conv_arithmetic may prove helpful in understanding the different introduced concepts.

A small test at the beginning of the session will cover the general ideas introduced in these papers. Please note that you need to pass this test in order to participate at the session.

Am I Prepared?

You might find the following remarks and questions helpful to check if you understood the papers and are prepared for the test:

- **U-Net**
  - What are the differences between the U-Net approach vs. the FCN architecture proposed by Shelhamer et al.
  - How can the transposed convolution layer be understood in terms of signal processing
  - Why is it important to utilize both deep and shallow features for segmentation? What are their respective characteristics?
- **Fully Convolutional Networks**
  - How is semantic segmentation defined?
  - What is different in a Fully Convolutional Network compared to e.g. a VGG-Net classifier?
  - Which kind of layers are used?
  - What defined the spatial resolution of the output segmentation map?

Further Reading

If you are interested in learning more about semantic segmentation, we refer to the following work for further reading. Looking at the code may provide you a further feel on how ML-based research in CV can be structured.

- J. Wang et al., *Deep High-Resolution Representation Learning for Visual Recognition* in TPAMI, 2020.
  Compared to the encoder-decoder networks like the U-Net, high-resolution feature maps are maintained throughout the network in HRNet.
- A. Tao et al., *Hierarchical Multi-Scale Attention for Semantic Segmentation* on arXiv, 2020.
  Coupling encoder-decoder networks (DeepLabv3+) with a classical multi-scale segmentation approach and implicit self-attention (known from “transformers”). Code can be found here.