PEPTIDE FEATURE DETECTION USING CONVOLUTIONAL NEURAL NETWORKS MASTER PROJECT / MASTER THESIS

THORSTEN FALK
COMPUTER VISION GROUP, DEPARTMENT OF COMPUTER SCIENCE
FREIBURG UNIVERSITY
FALK@INFORMATIK.UNI-FREIBURG.DE

LARS NILSE, OLIVER SCHILLING
INSTITUTE OF MOLECULAR MEDICINE AND CELL RESEARCH
FREIBURG UNIVERSITY
LARS.NILSE@MOL-MED.UNI-FREIBURG.DE
OLIVER.SCHILLING@MOL-MED.UNI-FREIBURG.DE

Abstract. Artificial neural networks and especially convolutional neural networks have been widely used for image classification and outperform classical image classification systems. In 2015 they even surpassed human performance in this task. Besides whole image classification, fully convolutional network (FCN) architectures originally used for image compression (auto-encoders), allow efficient semantic segmentation of images. They also allow the design of very efficient and accurate structure detectors. The project aims at using FCNs for the detection of peptide structures in mass-spectrometry based proteomics data. Beside genome sequencing, quantitative proteomics is one of the most useful tools in modern biological and medical research.

Unlike traditional wet lab techniques, mass-spectrometry based proteomics allows complete characterisation of a proteomic sample. The ability to monitor all and not only a few *a-priori* specified proteins opens new routes for biological and medical research. The reliable detection of peptide structures is a key step in the bioinformatic analysis of such data. Such algorithms have much improved over the past decade [1]. But their implementation and maintenance proofs to be labourious. Deep learning strategies appear to be a promising alternative.

Starting with LeNet in the late 1990s, Convolutional Neural Networks (ConvNets) have made great progress in past decades [2]. The extraction of features using machine learning algorithms instead of human-constructed features proves in many use cases the more efficient and easier approach. Public competitions such as the ImageNet - Large Scale Visual Recognition Challenge (ILSVRC) provide a forum for ConvNets continuously to compete and evolve. Image classification networks consist of a feature extraction stage that learns filters optimally tuned to the classification task on several resolution levels. The extracted features are then classified using few fully connected layers. When replacing the fully connected layers for classification by a convolutional image synthesis path, instead of class labels, full resolution images are generated from the learnt feature representation. These image-to-image architectures, called Fully Convolutional Networks (FCN) [3, 4], allow semantic image segmentation or detection within images with extremely high efficiency by exploiting feature dependencies on parallel GPU hardware.

The spectra can be regarded as huge but sparse images, which in principle allow direct use of FCNs to solve the detection task. However, the required large receptive field to identify peptides in raw resolution spectra precludes direct FCN application in practice. Instead smart (ideally loss-less) image downsampling strategies have to be employed. Another promising approach is direct exploitation of the sparsity of the input- and intermediate blobs by reformulating the convolutions for sparse data as in [5]. Additional to the localization of different peptides their bounding box should be estimated, adding a regression task to the pure detection.

For network implementation the caffe [6] framework will be used. Both OpenMS [7] and caffe are BSD-licensed, open-source C++ libraries with Python bindings. The aim of the project is to implement a first prototype of a fully convolutional peptide detector and test its performance on a number of proteomics datasets.

References

[1] L. Nilse, F. C. Sigloch, M. L. Biniossek, and O. Schilling, Toward improved peptide feature detection in quantitative proteomics using stable isotope labeling., Proteomics - Clinical Applications **9**(7-8), 706–714 (Aug. 2015), doi:10.1002/prca.201400173. 1
[2] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature **521**(7553), 436–444 (May 2015), doi:10.1038/nature14539. 1
[3] O. Ronneberger, P.Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241, Springer, 2015, (available on arXiv:1505.04597 [cs.CV]), doi:10.1007/978-3-319-24574-4_28. 1
[4] E. Shelhamer, J. Long, and T. Darrell, Fully Convolutional Networks for Semantic Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2016), doi:10.1109/TPAMI.2016.2572683. 1
[5] B. Graham, Spatially-sparse convolutional neural networks, (2014). 1
[6] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, *Caffe: Convolutional Architecture for Fast Feature Embedding*, Convolutional Architecture for Fast Feature Embedding, ACM, New York, New York, USA, Nov. 2014, doi:10.1145/2647868.2654889. 1
[7] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher, OpenMS: a flexible open-source software platform for mass spectrometry data analysis., Nature Methods **13**(9), 741–748 (Aug. 2016), doi:10.1038/nmeth.3959. 1