1 Research project objectives

The rapid development of deep learning has resulted in a massive increase in the use of deep neural networks in various fields of science and life. They are currently one of the most common tools, e.g. in image recognition and segmentation, text translation, drug discovery, recommendation systems, and speech recognition. With the dynamic development of deep learning, the depth of neural networks is increasing, which translates into their quality, but also to their much larger size and computational cost. This issue is especially problematic for mobile devices such as smartphones and for modules of embedded systems used, among others, in Internet of Things systems. To solve this problem, we can either improve hardware, compress networks using all sorts of methods, or improve the computational/memory performance of neural network modules such as the self-attention mechanism used in transformers. This project deals with a compression approach and improving computational/memory performance in deep neural networks using low-order approximation methods.

2 Research project methodology

This project aims to significantly contribute to the development of advanced methods of compression and acceleration of convolutional and capsule neural networks and to propose new effective alternatives to the self-attention mechanism. The project will use advanced methods of linear and multilinear algebra based on matrix and tensor factorization methods. The convolutional layer weights tensor can be efficiently approximated as a product of a sequence of low-rank tensors. In turn, the self-attention module can be replaced with an effective differentiable matrix or tensor factorization layer. In order to go beyond the schemes of linear tensor factorizations, the project will develop new tensor factorization methods with non-linear transformations, which will improve the quality of the estimated factors, and it will be possible to reduce the level of compression even lower. In addition, new alternatives to self-attention layers based on projective non-negative matrix factorization and non-negative tensor factorization will be proposed. As part of the experiments, networks of various sizes will be used, from medium networks on the CIFAR-10 set, to large networks trained on the ImageNet, Pascal-VOC or WikiText-103 set. The obtained results will be compared with state-of-the-art competing methods using metrics that calculate parameter compression and floating-point compression. In addition, the inference time and memory consumption will be measured and compared with the testing time of the core network on various computing platforms.

3 Expected results

As a result of the project, new methods of compression and acceleration of deep neural networks based on low-rank approximation methods will be created, which will certainly contribute to the further development of this field. The developed tools can noticeably improve the performance of modern mobile devices and embedded system modules by compressing the target networks that are to operate on them. The expected scientific effect will be the preparation of a series of four publications, which are planned for publication at the best conferences related to artificial intelligence (e.g. NeurIPS, CVPR, ICCV, ECCV, ICLR, IJCAI) and in journals such as IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, Knowledge Based Systems or Neurocomputing.