PhD subject: *Low-rank Tensor Representations of Convolutional Neural Networks*
Advisors: Suraj Kumar and Loris Marchal
Location: ROMA team, Inria Lyon, France

Context

Convolutional neural networks (CNNs) are currently the state-of-the-art models to classify objects in several domains, such as computer vision, speech recognition, text processing etc. Thanks to improved computational capability, we witness several popular complex and deeper CNNs. For example, AlexNet is 8 layers deep, while ResNet employs short connections and is represented with 152 layers. Both have about 60M parameters. CNNs have intensive computational requirements due to their huge complexity and large number of parameters.

Tensors are a natural way to represent high dimensional data for numerous applications in computational science and data science [1]. CP, Tucker and Tensor Train are the widely used tensor decomposition methods in the literature. These decompositions represent a high dimensional object with a small set of low dimensional objects.

Representing a high dimensional tensor with a set of smaller dimensional objects drastically reduces the overall number of parameters. This led to the use of low-rank tensor representations at different layers of CNNs. For example, it has been shown that replacing convolution kernels of ResNet with their low-rank approximations in Tucker tensor representations significantly reduces the number of parameters and improves the overall performance [3]. In a separate work, contributions have been made to replace dense weight matrices of the fully connected layers of AlexNet by their approximations in Tensor-train format [2]. This approach also significantly reduces the number of parameters while achieving the similar accuracy. The above contributions strongly advocate to employ the low-rank tensor representations in CNNs. We view the full CNN as a large tensor and aim to replace it with a set of smaller tensors.

Assignment

We view CNN models as large tensors and plan to represent them with their low-rank tensor representations. The main goal of this PhD thesis is to take advantage of parallel work on tensor computations and various methods to iteratively train tensor based frameworks for the efficient training and prediction with popular CNN models.

This PhD thesis will be held in the ROMA Inria team at LIP, ENS Lyon under the supervision of Suraj Kumar and Loris Marchal.

Main Activities

The candidate is expected to perform the following activities:

- Analyze existing training methods for CNNs and adapt them for tensor based models
- Represent popular CNN models with low-rank tensor representations
- Evaluate proposed models for MNSIT, CIFAR and ImageNet datasets
- Design parallel algorithms for the proposed models

Skills

The candidate must have a Master's degree in Computer Science, Computational Sciences, Applied Mathematics, or a related technical field.
Familiarity with Linear Algebra computations and Neural Networks will be much appreciated.

References

[1] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009. [Online]. Available: https://doi.org/10.1137/07070111X

[2] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov, "Tensorizing neural networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/6855456e2fe46a9d49d3d3af4f57443d-Paper.pdf

[3] A.-H. Phan, K. Sobolev, K. Sozykin, D. Ermilov, J. Gusak, P. Tichavský, V. Glukhov, I. Oseledets, and A. Cichocki, "Stable low-rank tensor decomposition for compression of convolutional neural network," in *Computer Vision – ECCV 2020*, pp. 522–539. [Online]. Available: https://doi.org/10.1007/978-3-030-58526-6_31