Internship proposal in machine learning and signal processing

Deep neural network compression using tensor methods

Context Deep Neural Networks (DNNs) demonstrate good prediction performances in numerous applications. However, the architectures of neural networks are very large, reaching several million parameters, and running them on systems with limited computing capacity (embedded systems) becomes a difficult task. For this reason, we will focus in this internship project on the compression of DNNs by *tensor* methods.

Summary This internship project deals with the study of new compression techniques for deep neural networks, by resorting to tensor decompositions to model and factorize the DNN weights. Recent studies [1] show that DNN weight matrices are often redundant, and by restricting their ranks, it is possible to significantly reduce the number of parameters without a significant drop in performance. In this project, we propose to convert these matrices to a *tensorial* format and to use multidimensional data processing methods to compress them. The goal of this internship is to study different tensor representations, such as the canonical polyadic decomposition (CPD) or Tucker decomposition (TD) [2], for the compression of the converted multidimensional weights. Specifically, we will study the compactness of these representations and their impact on the predictive accuracy of DNNs. In a first stage, the intern student will review the existing state-of-the-art tensor-based compression techniques [1], [3] and will get familiar with the tensor decompositions. Then, we will compare different representations with the goal to improve them and propose new tensor-based scheme for DNN compression.

This internship can be followed by a Ph.D research project
starting October, 2022, at LIS, Toulon

Profile and requirements M2R or engineering school students with major in signal processing, machine learning or applied mathematics. Good python programming skills are required. The knowledge of deep learning frameworks is a desirable plus. The candidate should have good writing and oral communication skills.

Supervision and environment The intern student will join the Signal and Image (SIIM) research team at the LIS laboratory, Toulon. The internship will be supervised by Yassine Zniyed (Associate Professor at Université de Toulon) and Thanh Phuong Nguyen (Associate Professor/HDR at Université de Toulon).

Contact and application procedure Please submit your application, including a CV and the list of your academic records (exam grades) to Yassine Zniyed (zniyed@univ-tln.fr) and Thanh Phuong Nguyen (tpnguyen@univ-tln.fr).

References

[1] A. Novikov, D. Podoprikhin, A. Osokin, *et al.*, "Tensorizing neural networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15, Montreal, Canada: MIT Press, 2015, pp. 442–450.

[2] Y. Zniyed, R. Boyer, A. L. de Almeida, *et al.*, "High-order tensor estimation via trains of coupled third-order cp and tucker decompositions," *Linear Algebra and its Applications*, vol. 588, pp. 304–337, 2020.

[3] V. Lebedev, Y. Ganin, M. Rakhuba, *et al.*, *Speeding-up convolutional neural networks using fine-tuned cp-decomposition*, 2015. arXiv: `1412.6553 [cs.CV]`.