Two thesis proposals within Labex Cominlabs LEANAI project

Dynamic Precision Training of Deep Neural Network Models on the Edge

at IRISA (Rennes) and/or LS2N (Nantes)
Contacts: **Silviu Filip** (silviu.filip@inria.fr), **Anastasia Volkova** (anastasia.volkova@univ-nantes.fr), **Elisa Riccetti** (elisa.riccietti@ens-lyon.fr)

Recent developments in deep learning (DL) are putting a lot of pressure and pushing the demand for intelligent edge devices capable of on-site learning. The realization of such systems is, however, a massive challenge due to the limited resources available in an embedded context and the massive training costs for state-of-the-art deep neural networks (DNNs). In order to realize the full potential of deep learning, it is imperative to improve existing network training methodologies and the hardware being used.

In the context of the upcoming LeanAI project, funded through the Labex CominLabs initiative, we are looking for two excellent PhD candidates to work on these problems. We will be focusing on both the *arithmetic* and *algorithmic* levels with the end goal of designing new mixed numerical precision *hardware architectures* for DNN training that are at the same time more energy-efficient and capable of improved performance in a resource-restricted environment. The expected outcomes include new mixed-precision algorithms for neural network training, together with open-source tools for hardware and software training acceleration at the arithmetic level on edge devices. Such tools are important to help democratize access to AI technologies.

The first PhD will work on the algorithmic/optimization aspects of reduced/mixed precision DNN training. A starting point will be the investigation of gradient-based backpropagation methods and the development of mixed-precision variants, with a main focus of using dynamic and adaptive number format precision choices in order to enhance energy efficiency [9, 10]. In a second stage, the candidate will also study alternative training strategies in a dynamic/mixed precision context that might show enhanced convergence guarantees and/or are more tolerant to low precision numerical effects. In particular, we will study and develop dynamic mixed-precision methods based on second-order optimization algorithms [11] and on hierarchical neural network modelling [3].

The second PhD will be concerned with the arithmetic and hardware generation aspects of the mixed-precision training approaches being devised in the context of the other PhD. The main task will revolve around building a toolkit of software/hardware co-design methods related to arithmetic operations in the context of training, which will rely on existing high level synthesis (HLS) tools from established vendors (such as those from Xilinx) for deployment. A large part of the work will revolve around the design of custom mixed-precision operators for the various operations performed during training and inference (*e.g.* dot product / matrix multiply operations and activation function evaluation).

An important aspect of both PhDs will be the efficient implementation and the practical evaluation, through extensive experiments and analysis of the new mixed-precision training algorithms and HW operators/accelerators, respectively. A crucial brick in this edifice will be the development of a mixed-precision training simulation framework (we intend it to be an extension to PyTorch) that supports an exhaustive list of number formats and numerical precisions. The main programming languages will therefore be Python (for coding the interface of the framework and performing of experiments with the developed algorithms) and C++ (for adding support to PyTorch for the number formats and precisions we will explore and for the HLS-based development at the core of the second PhD thesis).

Context: The successful candidates will be members of either the TARAN (at IRISA) or OGRE (at LS2N) teams, based in Rennes and Nantes, respectively. Extended stays/visits to the OCKHAM (at LIP) team in Lyon are also envisioned, especially in the context of the first PhD. The two theses are part of the upcoming LeanAI project, funded

through the CominLabs initiative (https://cominlabs.inria.fr/en/), which will start later this year. They will be working directly with Silviu Filip (https://people.irisa.fr/Silviu-Ioan.Filip/), Elisa Riccietti (http://perso.ens-lyon.fr/elisa.riccietti/), Anastasia Volkova (https://avolkova.org), Rémi Gribonval (http://perso.ens-lyon.fr/remi.gribonval) and Olivier Sentieys (http://people.rennes.inria.fr/Olivier.Sentieys/).

When: The desired starting date is October 1st 2021.

Who: The successful candidates should be highly motivated and creative. Both positions require a strong background in applied mathematics and/or computer science, with knowledge of DL techniques. Additionally, a strong understanding of continuous optimization algorithms is required for the first PhD, whereas good knowledge of hardware design is a must for the second position. Good programming skills are also required in both cases.

Application: Informal inquiries are strongly encouraged and the interested candidates can contact any member of the project for additional details and information. Applications are accepted until the positions are filled. The application should be sent by email to Silviu Filip (silviu.filip@inria.fr), Anastasia Volkova (anastasia.volkova@univ-nantes.fr) and Elisa Riccetti (elisa.riccietti@ens-lyon.fr) and it should include:

- motivation letter
- CV
- transcripts for the courses of the last two years of study
- contact information of two references (title, name, organization, e-mail)

References

[1] J.-M. Muller, N. Brunie, F. de Dinechin, C.-P. Jeannerod, M. Joldes, V. Lefèvre, G. Melquiond, N. Revol, and S. Torres. *Handbook of Floating-Point Arithmetic*. Birkhäuser Boston, 2nd edition, 2018.

[2] C. Lauter and A. Volkova. A framework for semi-automatic precision and accuracy analysis for fast and rigorous deep learning. In *2020 IEEE 27th Symposium on Computer Arithmetic*, pages 103–110, 2020.

[3] H. Calandra, S. Gratton, E. Riccietti, and X. Vasseur. On a multilevel Levenberg–Marquardt method for the training of artificial neural networks and its application to the solution of partial differential equations. *Optimization Methods and Software*, pages 1–26, 2020.

[4] N. Brisebarre, G. Constantinides, M. Ercegovac, S.-I. Filip, M. Istoan, and J.-M. Muller. A high throughput polynomial and rational function approximations evaluator. In *2018 IEEE 25th Symposium on Computer Arithmetic (ARITH)*, pages 99–106. IEEE, 2018.

[5] A. Fan, P. Stock, B. Graham, E. Grave, R. Gribonval, H. Jegou, and A. Joulin. Training with quantization noise for extreme fixed-point compression. *arXiv preprint arXiv:2004.07320*, 2020.

[6] B. Barrois and O. Sentieys. Customizing Fixed-Point and Floating-Point Arithmetic – A Case Study in K-Means Clustering. In *2017 IEEE International Workshop on Signal Processing Systems (SiPS)*, pages 1–6. IEEE, 2017.

[7] T. Zhang, Z. Lin, G. Yang, and C. De Sa. QPyTorch: A low-precision arithmetic simulation framework. *arXiv preprint arXiv:1910.04540*, 2019.

[8] S. Fox, J. Faraone, D. Boland, K. Vissers, and P. H. Leong. Training deep neural networks in low-precision with high accuracy using FPGAs. In *2019 International Conference on Field-Programmable Technology (ICFPT)*, pages 1–9. IEEE, 2019.

[9] A. Rajagopal, D. Vink, S. Venieris, and C.-S. Bouganis. Multi-Precision Policy Enforced Training (MuPPET): A precision-switching strategy for quantised fixed-point training of CNNs. In *International Conference on Machine Learning*, pages 7943–7952, 2020.

[10] Y. Fu, H. You, Y. Zhao, Y. Wang, C. Li, K. Gopalakrishnan, Z. Wang, and Y. Lin. FracTrain: Fractionally Squeezing Bit Savings Both Temporally And Spatially For Efficient DNN Training. *arXiv preprint arXiv:2012.13113*, 2020.

[11] A. S. Berahas and M. Takáč. A robust multi-batch L-BFGS method for machine learning. *Optimization Methods and Software*, 35(1):191–219, 2020.