2020-02469 - PhD Position F/M Precision Tuning for Deep Neural Networks Training

Type de contrat : CDI
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Doctorant

A propos du centre ou de la direction fonctionnelle

The Inria Rennes - Bretagne Atlantique Centre is one of Inria's eight centres and has more than thirty research teams. The Inria Center is a major and recognized player in the field of digital sciences. It is at the heart of several European associations, highly innovative PMIs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

Contexte et atouts du poste

Supervisors: Silviu-Ioan Filip (1), Anastasia Volkova (2) and Olivier Sentieys (1)
(1) Univ Rennes, Inria
(2) Univ Nantes, LS2N
Place: Campus de Beaulieu, Rennes, France

Contacts: silviu.filip@inria.fr, anastasia.volkova@univ-nantes.fr, olivier.sentieys@inria.fr

Mission confiée

The PhD Thesis will funded by Inria.

Principales activités

Context of the PhD Thesis

Deep Learning is one of the most intensively and widely used predictive models in the field of Machine Learning. Convolutional Neural Networks (CNNs) [7] have shown to achieve state-of-the-art accuracy in computer vision [6] and have even surpassed the error rate of the human visual cortex. These neural network techniques have quickly spread beyond computer vision to other domains. For instance, deep CNNs have revolutionized tasks such as face recognition, object detection, and medical image processing. Recurrent neural networks (RNNs) achieve state-of-the-art results in speech recognition and natural language translation [3], while ensembles of neural networks already offer superior predictions in financial portfolio management, playing complex games [9] and self-driving cars [20].

Despite the benefits that DL brings to the table, there are still important challenges that remain to be addressed if the computational workloads associated with NNs are to be deployed on embedded edge devices that require improved energy efficiency. For instance, the amazing performance of AlphaZero [9] required 4 to 6 weeks of training executed on 2000 CPUs and 250 GPUs for a total of about 600kW of power consumption (while the human brain of a Go player requires about 20W). Recent work [10] analyzing the carbon footprint of current natural-language processing models shows an alarming trend: training one huge Transformer model [20] for machine translation emits the same amount of CO2 as five cars in their lifetimes (fuel included). Such taxing demands are pushing both industry and academia to concentrate on designing custom platforms for DL algorithms that target improved performance and/or energy efficiency.

One general way to increase the performance and efficiency in computing is through reducing the numerical precision of basic arithmetic operations. In the case of DL systems, there are two main computational tasks: training and inference. Training requires vast quantities of labelled data that are used to optimize the network for the task at hand, usually by way of some form of stochastic gradient descent (SGD) algorithm. Inference, on the other hand, is the actual application of the trained network, which can be replicated onto millions of devices. Between the two, reducing numerical precision during inference has received the most attention from the research community over the last years, with some promising results in certain applications [11][13]. Much less has been done for the training phase, the main reason being that the effects of low-precision arithmetic on training algorithms are not yet well understood. This has by no means stopped major players in the hardware space to start designing architectures that offer increasing support for low-precision arithmetic. In the particular case of training, there are already commercial platforms that mix 32-bit high precision floating-point computing with low precision 16-bit formats for increased performance [4,5,8],

Objectives of the PhD Thesis

With this thesis, we want to conduct a thorough analysis of reduced numerical precision training of DL systems. We plan to do this at two levels: arithmetic (use appropriate numerical formats and bit widths for all the computations used during training) and algorithmic (by attempting to improve the practical convergence properties of the optimization procedures used to train neural networks).

A first objective is for the PhD student to build/augment a deep learning platform with custom precision arithmetic. This will require building customized floating-point operators down to very few bits of exponent and mantissa which offer a desirable balance between accuracy and energy efficiency. In parallel, the plan is to investigate how mixed precision support (i.e. hard-ware support for several numeric formats with varying costs and accuracies) during successive iterations of the SGD training algorithm impacts accuracy and performance.

With respect to existing works that generally consider predefined numeric formats, our aim is to do a more in-depth analytical design space exploration by looking at the entire spectrum of low precision...
floating-point arithmetic formats and how the working precision can be effectively varied in-between training iterations. The student will also have the task of validating the developed techniques through a prototype of an accelerator for CNN training in the context of a collaboration with other researchers in the team.

References


Compétences

The student is expected to develop techniques for the mathematical formulation of quantization and compression models in deep neural networks, and energy-efficient hardware accelerators for machine learning applications. We also expect to have implementations of the developed techniques into a hardware accelerator developed as an FPGA prototype. The designs will primarily be done through High-Level Synthesis tools from C/C++. Ideal candidates should possess the following skills:

- Programming experience in C/C++ and Python.
- Good knowledge of computer architecture, computer arithmetic, hardware design, and embedded systems.
- Good knowledge of deep neural networks and deep learning.
- Experience with hardware acceleration on FPGA boards (especially Xilinx Zynq) would be a plus.
- Experience with deep learning frameworks (especially Pytorch) would be a plus.
- Basic knowledge in linear algebra, optimization, numerical computing, and machine learning in general.

Mostly importantly, we seek highly motivated and active students. A master in Computer Science, Computer Engineering, or Electronics Engineering is required.

Avantages

- Subsidised catering service
- Partially-reimbursed public transport

Rémunération

PhD student: monthly gross salary amounting to 1982 euros for the first and second years and 2085 euros for the third year.