



Offer #2024-07861

PhD Position F/M Exploring low-precision arithmetic for continual learning tasks on edge devices

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

About the research centre or Inria department

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 a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

Assignment

While machine learning models have achieved impressive results in recent years on many individual tasks (e.g. object recognition, classification, language models), they are obtained with static models that are not capable of adapting their behavior over time. In a dynamic environment, adapting the behavior of a model would require restarting the training process each time new data becomes available, quickly becoming impractical due to constraints such as storage and privacy issues.

Continual learning [2] studies such problems stemming from an infinite/incremental stream of data and the need to extend their behavior to additional tasks. The major challenge is to learn without significant degradation in accuracy for previously learned tasks, a problem known as catastrophic forgetting. An added complication is the use of low precision arithmetic (e.g. sub 16-bit floating-point formats such as FP8), that can also have an impact on the performance of the model and on the choice of continual learning approach. This aspect of the impact of arithmetic on task performance in a continual learning scenario seems to have so far received little to no attention [3, 4], although if continual learning systems are to be deployed in the real world, especially on embedded or edge devices, such considerations will become paramount.

The goal of this thesis is therefore to investigate the performance impact of using low precision arithmetic in the context of training and deploying continual learning systems on edge devices and propose task-aware number format precision switching strategies and custom hardware architectures for continual learning tasks. The starting point will be implementing, testing, and adapting various low precision variants of continual learning methods (replay, regularization and parameter isolation). To do so, we envision using the Avalanche [5] continual learning library, which will integrate the mptorch [1, 6] framework developed in the TARAN team for doing custom precision computations during DNN training and inference.

The second and main objective of the PhD thesis will then be to validate the developed techniques through a prototype of an accelerator for training in the context of low-precision continual learning. Synthesis of the specialized architecture on a target hardware platform will demonstrate the gains in performance and energy of the automatically generated accelerators. This parallel architecture will include configurable arithmetic operators implementing various precision and number representations as defined by an exploration methodology. Concretely, the architecture will first be validated in an FPGA accelerator for training based on previous work in the team [1, 6]. Second, an ASIC prototype will be designed to reach the highest energy efficiency. The team has such experience in designing custom chips, evaluating performance and power consumption in advanced technology, and even going down to a silicon prototype (even if hardly reachable in the frame of a PhD thesis). Our main focus is on energy-efficient embedded systems, such as autonomous vehicles, or on ultra-low-power IoT (Internet of Things) devices. A heterogeneous host-accelerator model of computation in which a fast accelerator (e.g., FPGA or ASIC) with support for low-precision arithmetic is connected to a slower general purpose host device (e.g., a RISC-V CPU) that can perform high-precision arithmetic.

References

[1] S. B. Ali, S.-I. Filip, and O. Sentieys. A stochastic rounding-enabled low-precision floating-point mac for dnn training. In IEEE/ACM Design, Automation & Test in Europe Conference & Exhibition (DATE), pages 1–6, 2024.

[2] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuyte-laars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[3] C. F. S. Leite and Y. Xiao. Resource-Efficient Continual Learning for Sensor-Based Human Activity Recognition. *ACM Transactions on Embedded Computing Systems*, 21(6):1–25, 2022.

[4] Y. Li, W. Zhang, X. Xu, Y. He, D. Dong, N. Jiang, F. Wang, Z. Guo, S. Wang, C. Dou, et al. Mixed-Precision Continual Learning Based on Computational Resistance Random Access Memory. *Advanced Intelligent Systems*, page 2200026, 2022.

[5] V. Lomonaco, L. Pellegrini, A. Cossu, A. Carta, G. Graffieti, T. L. Hayes, M. De Lange, M. Masana, J. Pomponi, G. M. Van de Ven, et al. Avalanche: an end-to-end library for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3600–3610, 2021.

[6] M. Tatsumi, Y. Xie, C. White, S.-I. Filip, O. Sentieys, and G. Lemieux. MPTorch and MPArchimedes: Open Source Frameworks to Explore Custom Mixed-Precision Operations for DNN Training on Edge Devices. In *ROAD4NN 2021-2nd ROAD4NN Workshop: Research Open Automatic Design for Neural Networks*, 2021.

Main activities

Context: The position is within the TARAN team, based in Inria Rennes. Within the context of the FAIRe project (started in late 2023), the candidate will be involved in common initiatives with other members of the project, in particular the DFKI RIC and AV teams.

When: The desired starting date is October 1st 2024.

Application: Informal inquiries are strongly encouraged and the interested candidates can contact us for any extra information. Applications are accepted until the position is filled. The application file should include: CV, motivation letter, transcripts for the courses taken in the last two years of study, contact information and letters of support from two references (title, name, organization, e-mail).

Skills

The successful candidate should be highly motivated and creative and be familiar with writing and analyzing numerical code. The position requires a strong background in computer science and in particular hardware design, and modern deep learning techniques applied to continual learning tasks. Additionally, a good understanding of continuous optimization algorithms is a plus. Good programming skills in Python/C++ are also required as well as an excellent grasp of hardware design languages (e.g. VHDL or Verilog).

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Possibility of teleworking (90 days per year) and flexible organization of working hours
- Partial payment of insurance costs

Remuneration

Monthly gross salary amounting to 2100 euros for the first and second years and 2200 euros for the third year

General Information

- **Theme/Domain :** Architecture, Languages and Compilation
- **Town/city :** Rennes
- **Inria Center :** [Centre Inria de l'Université de Rennes](#)
- **Starting date :** 2024-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-08-15

Contacts

- Inria Team : [TARAN](#)
- PhD Supervisor :
Sentieys Olivier / Olivier.Sentieys@irisa.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Please submit online : your resume, cover letter and letters of recommendation eventually

Defence Security :

This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :

As part of its diversity policy, all Inria positions are accessible to people with disabilities.