Offre n°2024-07861

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

Le descriptif de l'offre ci-dessous est en Anglais

Type de contrat : CDD

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 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.

Mission confiée

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 preci- sion 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 archi- tectures 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 iso- lation). 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 trough 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 per- formance 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 accelera- tor 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, eval- uating 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) de- vices. 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.

Principales activités

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).

Compétences

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).

Avantages

- 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

Rémunération

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

Informations générales

- **Thème/Domaine :** Architecture, langages et compilation
- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :**2024-10-01
- **Durée de contrat :**3 ans

- **Date limite pour postuler :** 2024-08-15

Contacts

- **Équipe Inria :** TARAN
- **Directeur de thèse :**
  Sentieys Olivier / Olivier.Sentieys@irisa.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

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

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.