Job vacancy #2023-06544

PhD Position F/M Hardware-guided compression & fine-tuning of Transformer-based models

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.

Context

Supervisors: Silviu-Ioan Filip (1) and Olivier Sentieys (1)
(1) Univ Rennes, Inria
Place: Campus de Beaulieu, Rennes, France
Contacts: olivier.sentieys@inria.fr, silviu.filip@inria.fr

Assignment

This PhD Thesis will be funded through the HOLIGRAIL project.

Main activities

Transformer-based large language models (LLMs) have been getting significant attention recently with the arrival of networks such as ChatGPT and GPT-4. While they show impressive potential in a variety of tasks (notably in vision and natural language processing), most modern Transformer-based architectures are nevertheless characterized by significant storage (with up to hundreds of billions of parameters and hundreds of GB of required memory/disk space), computational and energy costs, making it challenging to fine-tune, deploy, and use them in many practical settings without a powerful hardware infrastructure. For instance, even state-of-the-art compressed model families, such as LLaMA [6], require at least 7B parameters (LLaMA-7B) and at least 3.5GB of storage (i.e., when using 4-bit integer quantization).

The PhD candidate will be tasked with investigating extreme compression methods for Transformer-based models, combining pruning and quantization methodologies (such as post-training quantization methods [2] and quantization-aware training [3]). We will look at hybrid format mixed precision approaches (e.g., mixing integer and floating-point data [1]), taking hardware constraints into account (e.g., available compute units and supported formats and memory). The goal is to efficiently explore the large design space of available quantization formats and propose compressed models that are optimized for low latency and energy consumption. This will entail extending low precision simulation tools that we have been developing in our team (mptorch, built on top of PyTorch [4,5]), and also working towards FPGA/ASIC hardware accelerator prototypes for small to medium sized Transformer-based models that will be developed with other members in the HOLIGRAIL project.
Context: The successful candidate will be a member of the TARAN team, based in the Inria Research centre at Rennes University and IRISA Lab. in Rennes, France. The thesis is part of the upcoming PEPR HOLIGRAIL project, part of the larger PEPR programme in Artificial Intelligence. It brings together researchers working on machine learning, computer arithmetic, hardware acceleration and compiler optimization for embedded systems and deep learning applications from University of Rennes, Inria, CEA List, INSA Lyon and Grenoble-INP. HOLIGRAIL is a large and competitive project that will fund more than 20 people ranging from PhD students to postdoctoral fellows.

References:


Skills

When: The desired starting date is October 1st 2023 or as soon as possible after.

Who: The successful candidate should be highly motivated and creative. The position requires a strong background in computer arithmetic, computer architecture, with knowledge of Deep Learning models (ideally LLMs) and techniques. Strong proficiency in Python with knowledge of popular deep learning frameworks such as PyTorch or TensorFlow is also required.

Application: Informal inquiries are strongly encouraged and the interested candidates can contact us for additional details and information. Applications are accepted until the positions are filled. The formal application should be sent by email to Silviu Filip (silviu.filip@inria.fr) and Olivier Sentieys (olivier.sentieys@inria.fr) and it should include:

- motivation letter
- CV
- transcripts for the courses undertaken in the last two years of study
- references and recommendation letters
- links to publications or MSc thesis if relevant
- contact information of two references (title, name, organization, e-mail)

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 2082 euros for the first and second years and 2190 euros for the third year

**General Information**

- **Theme/Domain**: Architecture, Languages and Compilation System & Networks (BAP E)
- **Town/city**: Rennes
- **Inria Center**: Centre Inria de l'Université de Rennes
- **Starting date**: 2023-10-01
- **Duration of contract**: 3 years
- **Deadline to apply**: 2023-12-03

**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

For more information, please contact olivier.sentieys@inria.fr

**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.