Offer #2024-07376

Post-Doctoral Research Visit F/M Efficient Deep Learning (IDP 2024)

Contract type : Fixed-term contract

Level of qualifications required : PhD or equivalent

Other valued qualifications : PhD degree in Computer Science/ Mathematics/ Machine Learning/other technical field

Fonction : Post-Doctoral Research Visit

Context

Objective

Optimise the training and inference of modern neural networks to create large-scale AI models for science. Develop theoretical approaches and corresponding software.

Work environment

You will be a part ofTopal INRIA team in Bordeaux, which includes experts in both HPC and AI fields.

Particularly, for the last several years members of the Topal team have been working on optimizing the training of neural networks by applying techniques from high performance computing, linear and tensor algebra (please, see papers from ICML'23, ICML'23, IJCAI'22, NeurIPS'21 and workshop WANT@NeurIPS'23). You will work closely with Julia Gusak, Lionel Eyraud-Dubois, and Olivier Beaumont.

Is regular travel foreseen for this post?

Short-term visits to conferences and collaborative laboratories. In particular, the team is involved with a tight collaboration with Caltech within the framework of Associated Team ELF.

Assignment

Scientific Research context:

The unprecedented availability of data, computation, and algorithms has enabled a new era in AI, as evidenced by breakthroughs like Transformers and LLMs, diffusion models, etc., leading to groundbreaking applications such as ChatGPT, generative AI, and AI for scientific research. However, all these applications share a common challenge: they keep getting bigger, which makes training models harder. This can be a bottleneck for the advancement of science, both at industry scale and for smaller research teams that may not have access to very large training infrastructure. While there already exists a series of effective techniques (e.g., see the overview [2]), recent ones either still rely on manual hyperparameter settings or lack automatic joint optimization of orthogonal approaches (e.g., pipelining and advanced re-materialization).

Work description:

Concerning the training phase, one group of methods proposes advanced parallelization techniques, such as model and pipelined parallelism, for which the members of Topal already contributed [1, 3, 4]. They are used to split models across devices. Another group of methods considers effective optimizers. For example, ZeRO optimizer proposes optimizer state/gradients partitioning to reduce memory footprint during the optimization step. Additionally, to reduce the required per-GPU memory allocation, offloading and checkpointing (or re-materialization) techniques can be used. Offloading to CPU saves memory at the price of an overhead on communications, while activation checkpointing recomputes parts of the computational graph when applied, thus saving memory at the price of an overhead on computations. All types of techniques can be combined to achieve better throughput. Recent papers consider a combination of pipeline parallelism with activation checkpointing techniques [5, 6].

An important point is that algorithms with theoretically better time/memory complexity in practice might provide fewer benefits as it could be expected from analytical derivations. The reason is the overhead caused by specific hardware we use to train or execute neural networks. To make deep learning algorithms efficient in real life it is important to combine software and hardware optimization when

creating new deep learning algorithms.

During the Postdoc contract we plan to propose novel approaches to improve efficiency (memory/time/communication costs) of neural network training and inference. Particularly, by finding best model execution schedule which allows using different types of techniques, including but not limited to parallelisms, re-materialization, offloading, low-bit computations. Along with theoretical contribution to the field, there will be developed software to automatically optimize the training and inference of modern deep learning architectures.
Potential applications will include, but not be limited to, computer vision, natural language processing, climate, etc.

References:

[1] Zhao, X., Le Hellard, T., Eyraud-Dubois, L., Gusak, J. & Beaumont, O. (2023). Rockmate: an Efficient, Fast, Automatic and Generic Tool for Re-materialization in PyTorch. Proceedings of the 40th International Conference on Machine Learning
[2] Gusak, J., Cherniuk, D., Shilova, A., Katrutsa, A., Bershatsky, D., Zhao, X., Eyraud-Dubois, L., Shlyazhko, O., Dimitrov, D., Oseledets, I. & Beaumont, O. (2022, July). Survey on Large Scale Neural Network Training. In IJCAI-ECAI 2022-31st International Joint Conference on Artificial Intelligence (pp. 5494-5501). International Joint Conferences on Artificial Intelligence Organization.
[3] Beaumont, O., Eyraud-Dubois, L., Shilova, A., & Zhao, X. (2022). Weight Offloading Strategies for Training Large DNN Models.
[4] Beaumont, O., Eyraud-Dubois, L., & Shilova, A. (2021). Efficient combination of rematerialization and offloading for training dnns. Advances in Neural Information Processing Systems, 34, 23844-23857.
[5] Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V. and Zhang, E., 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990.
[6] Li, S., & Hoefler, T. (2021, November). Chimera: efficiently training large-scale neural networks with bidirectional pipelines. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-14).

Main activities

Activities:

- Implement different techniques for efficient multi-GPU training and inference.
- Propose new approaches for efficient deep learning (based on pipelining, checkpointing, offloading, and other optimization techniques).
- Develope software to automatically optimise the training and inference of modern deep learning architectures.
- Perform experiments with modern neural networks, including GPT-like models and Neural Operators. Potential applications will include, but not be limited to, computer vision, natural language processing, climate, etc.
- Analyze the performance of models using profiling tools.
- Write scientific papers
- Collaborate with Topal colleagues in Europe and US

Skills

Technical skills and level required:

- Good knowledge in Machine Learning and Deep Learning
- Basic knowledge in Linear algebra, Optimization, Probability Theory, Calculus
- Experience with Python, PyTorch, LaTeX, Linux, Git (will be a plus: Docker, Singularity, Slurm)
- Publications at A/A* conferences in Machine Learning will be a plus.

Languages: English

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Remuneration

2788€ / month (before taxs)

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
  Scientific computing (BAP E)
- **Town/city :** Talence

- **Inria Center :** Centre Inria de l'université de Bordeaux
- **Starting date :** 2024-10-01
- **Duration of contract :** 2 years
- **Deadline to apply :** 2024-05-03

Contacts

- **Inria Team :** TOPAL
- **Recruiter :**
  Beaumont Olivier / Olivier.Beaumont@inria.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.

The keys to success

Passionate about AI and HPC, taste for the design of algorithm, their implementation, and experimental validation.

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

Thank you to send:
- CV
- Cover letter
- Support letters (mandatory)
- List of publication

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.