



Offre n°2024-07927

PhD Position F/M Efficient Training of Neural Networks

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 center at the University of Bordeaux is one of the nine Inria centers in France and has about twenty research teams. The Inria centre 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 SMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute...

Contexte et atouts du poste

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).

Keywords

Efficient Deep Learning
Artificial Intelligence
High Performance Computing
Resource Optimization
Re-materialization and Offloading
Memory Optimization

Mission confiée

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 PhD 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., Prabhume, 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).

Principales activités

- Implement baselines
- Develop new theoretical approaches
- Write scientific papers
- Develop software

Compétences

Required Knowledge and background

- 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)

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- 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

Rémunération

The gross monthly salary will be 2100€ the first year and 2190€ the second and third year.

Informations générales

- **Thème/Domaine** : Calcul distribué et à haute performance
Calcul Scientifique (BAP E)
- **Ville** : Talence
- **Centre Inria** : [Centre Inria de l'université de Bordeaux](#)
- **Date de prise de fonction souhaitée** : 2024-10-01
- **Durée de contrat** : 3 ans
- **Date limite pour postuler** : 2024-07-31

Contacts

- **Équipe Inria** : [TOPAL](#)

- **Directeur de thèse :**
Gusak Yulia / yulia.gusak@inria.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.

L'essentiel pour réussir

Passionate about Deep Learning and High Performance Computing

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

If you are interested by this job, please could you apply on website jobs.inria with the following documents :

-cv

-cover letter

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.