Offre n°2025-09139

PhD Position F/M Memory minimization for 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

Niveau d'expérience souhaité : Jeune diplômé

A propos du centre ou de la direction fonctionnelle

The Centre Inria de l'Université de Grenoble groups together almost 600 people in 22 research teams and 7 research support departments.

Staff is present on three campuses in Grenoble, in close collaboration with other research and higher education institutions (Université Grenoble Alpes, CNRS, CEA, INRAE, …), but also with key economic players in the area.

The Centre Inria de l'Université Grenoble Alpe is active in the fields of high-performance computing, verification and embedded systems, modeling of the environment at multiple levels, and data science and artificial intelligence. The center is a top-level scientific institute with an extensive network of international collaborations in Europe and the rest of the world.

Mission confiée

Context on memory peak minimization

In this proposal we want to tackle the problem of memory minimization when sequentially executing tasks (representing processes, programs, blocks of code, instructions, . . .) with data dependencies, represented as a dataflow task graph. This is critical for applications such as large neural networks that require huge amounts of memory space. To execute a set of tasks on a given system, the highest memory demand, i.e., the memory peak, of the tasks must be smaller than the memory available on the system. The memory peak is the maximum amount of live data at any time; this is the size of data produced by the tasks so far, and which are required for the computations of further tasks. Three techniques can be used and combined to meet such memory constraint:

- **scheduling**: to schedule tasks according to their impact on memory (for example, tasks consuming data and decreasing the amount of live memory should be executed as soon as possible);
- **offloading**: to move live data to a slower but larger upper-level memory (e.g., cache, RAM, disk), and reloading them when required, which gives the opportunity to execute another data intensive task in between;

- **recomputing (aka rematerialization)**: to erase and recompute data, by re-executing the tasks having produced more data than they consumed, and keeping their (smaller) input data live instead of the (larger) output one, which also gives the opportunity to execute another data intensive task in between.

Considering all three techniques, scheduling, offloading, and recomputing, gives rise to trade-offs between the minimization of the memory peak and the execution time; this problem is challenging and PSPACE-complete.

Principales activités

Description and objectives of the PhD

During the previous years, we have addressed the memory peak minimization problem of general task graphs by using only the scheduling technique [1, 2]. Our approach, based on original graph transformations, finds the optimal sequential schedule in terms of memory peak for a wide class of task graphs. This technique is able to optimally solve the problem on some large dataflow task
graphs, up to 50, 000 tasks in our experiments.

These results encourage us to study memory minimization for neural networks. This entails to take into account the particular shape and tasks used by neural networks. Indeed, graphs representing neural network have specific shapes (linear, U-shaped, etc.) and their most common tasks are specific operations (matrix multiplications, convolutions, activations, pooling, etc.). Moreover the other two techniques, offloading and recomputing, should also be considered to extend and apply our previous work to neural networks.
Offloading consists of data movement from a size-limited memory (RAM or GPU global memory) to a bigger but slower one (disk or RAM, respectively). Recomputing is useful for the training phase of neural networks, decomposed in two passes: forward and back propagation. It can be used to store only a part of all neurons' outputs during the forward pass, so that the missing ones will
be recomputed during the back propagation pass.

The overall objective of the PhD is, by taking into account the specificities of a given neural network and by using the three techniques mentioned above, to minimize the execution time overhead while fitting in a given memory budget (i.e., optimization under constraint). This represents an opposite viewpoint compared to

our previous work were the memory peak was minimized for a constant time budget.

Compétences

Interested candidates are expected to have a good background in formal methods, machine learning
and compilation. Good relational and English skills are also important for the project

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 (90 days per year)
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Informations générales

- **Thème/Domaine :** Systèmes embarqués et temps réel
  Systèmes d'information (BAP E)
- **Ville :** Montbonnot
- **Centre Inria :** Centre Inria de l'Université Grenoble Alpes
- **Date de prise de fonction souhaitée :** 2025-10-01
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2025-09-30

Contacts

- **Équipe Inria :** SPADES
- **Directeur de thèse :**
  Fradet Pascal / pascal.fradet@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'e?orce 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

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.