Offer #2025-09139

PhD Position F/M Memory minimization for neural networks

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Level of experience : Recently graduated

About the research centre or Inria department

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.

Assignment

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.

Main activities

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.

Skills

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

Benefits package

- 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

General Information

- **Theme/Domain :** Embedded and Real-time Systems
  Information system (BAP E)
- **Town/city :** Montbonnot
- **Inria Center :** Centre Inria de l'Université Grenoble Alpes
- **Starting date :** 2025-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-09-30

Contacts

- **Inria Team :** SPADES
- **PhD Supervisor :**
  Fradet Pascal / pascal.fradet@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.

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

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.