Recently, several frameworks such as TensorFlow [1] and PyTorch [2] emerged and represent the DL network as a directed graph whose nodes represent convolution operations and edges represent data dependences between them. The goal of this PhD thesis is to work on how to allocate the convolution operations and how to schedule them to achieve a better efficiency, typically in the context of platforms consisting of heterogeneous resources such as GPUs and multicore nodes.

The goal of this PhD Thesis is to improve the scheduling and resource allocation strategies along several directions. First, the resource allocation algorithm does not take into account the specificities of the application. Indeed, it is for instance close to the default StarPU scheduling algorithm [3] used for general task graphs.

Second, it has been proved that for specific applications such as linear algebra kernels, injecting some static knowledge based on a more sophisticated scheduling algorithm can strongly improve the performance of greedy algorithm [4]. Third, in the context of DL, the same graph of convolution layers is used many times on different input data along the execution of the DL algorithm, what is close to the context of steady state scheduling [5], that has been proved to be more tractable than general scheduling. At last, another opportunity is to develop high level simulation techniques, that could be used in particular to detect bottlenecks with respect to a DL network and to a parallel architecture. This possibility could more speculatively be especially interesting in the context of DL, since it may help to redesign the network itself to cope with bottlenecks. We will first concentrate on classical layers (Fully Connected Layers, Convolutional Layers, Recurrent Layers) before considering Pl@ntNet [6] as a target network.

About the research centre or Inria department

Our aim is to develop tight formulations for combinatorial problems by combining the latest reformulation techniques, such as Lagrangian and polyhedral approach, non-linear programming tools and graph theoretics tools. Through industrial partnerships, the team targets large scale problems such as those arising in logistics (routing problems), in planning and scheduling, in network design and control, and in placement problems (cutting stock problems).

International and industrial relations

We have an associated team in Brazil through which we collaborate with Artur Pessoa and Eduardo Uchoa (Universidade Federal Fluminense) and Marcus Poggi (PUC-Rio) Our Industrial partners are Pascale Bendotti and Marc Porcheron (EDF, R&D Dpt OSIRIS), and Fabien Rodes (société Exeo Solutions).

Context

Within the framework of a partnership

- collaboration between 4 Inria teams: RealOpt, Zenith, Storm and Tau (IPL HPC-BigData)

Assignment

Regarding the context of HPC and AI platforms, many research challenges arise. The main one is to develop efficient algorithms and software to exploit the power of these systems. We are interested in the design of efficient algorithms and software for the following specific tasks:

- scheduling and resource allocation
- online optimization
- compilers and runtime systems

The keys to success

Technical skills and level required: The candidate will be required to have a solid background in Combinatorial Optimization (scheduling, resource allocation, online algorithms) and/or in Deep Learning (TensorFlow, PyTorch) and a taste for both domains.

Conditions for application

Thank you to send:

- updated CV
- cover letter
- letters of recommendation eventually
- Master Degree Transcripts

Defence Security:

This position is likely to be situated in a restricted area (ZRR), as defined in Decrease 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.

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.
Main activities

These research directions require the joint knowledge of experts in deep learning algorithms, dynamic runtime scheduling and scheduling theory and will benefit in particular to Pl@ntNet application.

The PhD student will be localized in Bordeaux and will be co-supervised by Olivier Beaumont (RealOpt) and Alexis Joly (Zenith), in close collaboration with Guillaume Charpiat (Tau) and Samuel Thibault (Storm). Several stays (1 week) in Saclay and Montpellier will be scheduled during the PhD Thesis.

Benefits package

- Subsidised catering service
- Partially-reimbursed public transport

Remuneration

1982€ / month (before taxes) during the first 2 years, 2085€ / month (before taxes) during the third year.