2018-00523 - [Campagne CORDI-S-CRI Paris] Domain-Specific Compilation and Optimization, a Reinforcement Learning Approach

**Contract type**: Public service fixed-term contract  
**Level of qualifications required**: Graduate degree or equivalent  
**Other valued qualifications**: Master degree with specialization in programming languages and/or machine learning  
**Function**: PhD Position

### Context

This PhD thesis will involve a collaboration with an open source project called Tensor Comprehensions, led by Facebook Artificial Intelligence Research (FAIR), and the researchers and developers in this community, including machine learning and compilation experts at ETH Zürich, MIT and NVIDIA.

https://research.fb.com/announcing-tensor-comprehensions  
http://pytorch.org/2018/03/05/tensor-comprehensions.html

While the Tensor Comprehensions project focuses on programmability and achieving a practical impact for a wide range of machine learning applications, this PhD thesis proposal pursues a more specific, higher risk objective: automating much of the work of high-performance computing engineers in charge of producing the numerical libraries of machine learning operators (e.g., NNPACK for CPU, CUDNN for GPU). While the thesis is expected to have a strong impact on the practice and experience, it raises semantical and combinatorial questions that cannot be addressed in the typical research and development cycles of industry or open source projects. This is the reason why we are applying for longer term academic support for this research. Complementary support will come from the European research grant MNEMOSENE and from our industrial partners to address shorter term objectives, to provide access to state of the art hardware, and to support the travel costs of international collaboration.

### Assignment

The work will place in the PARKAS team of INRIA and the Département d’Informatique of École Normale Supérieure, supervised by Albert Cohen.

Close interactions will take place with Ulysse Beaugnon, 3rd year PhD student, Oleksandr Zinenko, postdoctoral researcher, and Nicolas Tollenaere, expert engineer.

Deep learning models with convolutional and recurrent networks are ubiquitous and analyze massive amounts of audio, image, video, text and graph data, with applications in automatic translation, speech-to-text, scene understanding, ranking user preferences, ad placement, etc. Competing frameworks for building these networks such as TensorFlow, Chainer, CNTK, Torch/PyTorch, Caffe1/2, MXNet and Theano, explore different tradeoffs between usability and expressiveness, research or production orientation and supported hardware. They operate on a DAG of computational operators, wrapping high-performance libraries such as CUDNN (for NVIDIA GPUs) or NNPACK (for various CPUs), and automate memory allocation, synchronization, distribution. The production of the highly optimized, target-specific implementations for these library functions comes at a very high engineering cost, and typical programming languages and compilers, including domain-specific ones, have failed to deliver competing performance with expert-written code. Furthermore, typical library implementations often do not offer optimal performance for a user’s particular dataset, missing optimizations between operators as well as specialization to the size and shape of data.

We will focus on machine learning as a primary application, for the ideal mix of operator diversity, the need for algorithmic exploration, the potential for dataset specialization and context-dependent optimization (the mix of memory-bound and compute-bound kernels), and the maturity of the software environment and expertise available in the PARKAS team to address these challenges [1,2]. Additional

### General Information

- **Theme/Domain**: Architecture, Languages and Compilation  
- **Town/city**: Paris  
- **Inria Center**: CRI de Paris  
- **Starting date**: 2018-09-01  
- **Duration of contract**: 3 years  
- **Deadline to apply**: 2018-04-23

### Contacts

- **Inria Team**: PARKAS  
- **Recruiter**: Cohen Albert / albert.cohen@inria.fr

### The keys to success

The candidate should have prior exposure to programming languages and/or machine learning, preferably both.

The candidate should be prepared to engage into a cross-domain research programme, combining elements of computing systems research, programming languages, and machine learning (specifically reinforcement learning).

The knowledge of the Rust programming language is a plus. Experience with OCaml and C++ programming is expected.

### Conditions for application

**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.

**Warning**: you must enter your
applications to numerical solvers and linear algebra will flow naturally from this primary focus.

The thesis will address the design of a domain-specific language close to the mathematics of deep learning operators and the construction of a domain-specific compiler for this language. Following the pioneering approach of the Telamon algorithm and tool developed by Ulysse Beaugnon [3], we will evaluate operation research strategies combining analytical modeling, constraint programming and branch and bound optimization to navigate the combinatorial search spaces of domain-specific program transformations involved in the acceleration of a selected set of numerical kernels. We will also adapt reinforcement learning techniques, including multi-armed bandit schemes, popular in such domain-specific scenarios [4], and attempt to combine these with the former. The student will prototype these concepts and algorithms in a domain-specific code generator, to deliver a so-called active or built-to-order library [5].

[1] https://research.fb.com/announcing-tensor-comprehensions

Main activities

The typical activities of the student include algorithm design, formal semantics of a domain-specific language, compiler construction and prototyping, experimental evaluation, analytical and statistical studies to drive the research of more suitable algorithms. The work will also involve related work studies, contributions to the production of scientific reports and publications, the presentation and dissemination of these works, the participation to summer schools and conferences, and collaboration and travel within the community of experts and potential users.

Skills

Technical skills and level required:

Languages : International English.

Relational skills : Work in a multicultural, collaborative research environment.

Other valued appreciated : Independence, Open Source, Interest in the construction of robust software artifacts.

Benefits package

- Subsidised catering service
- Partially-reimbursed public transport

Remuneration

Gross Salary per month: 1 982 € the first 2 years and 2 085 € the last year