FID3025 Architecting Efficient AI Hardware with technology and architectural design space exploration 7.5 credits

Konstruktion av effektiv AI-maskinvara genom utforskning av teknologi- och arkitekturdesignrymden

This is a translation of the Swedish, legally binding, course syllabus.

If the course is discontinued, students may request to be examined during the following two academic years

Establishment

Course syllabus for FID3025 valid from Autumn 2020

Grading scale

P, F

Education cycle

Third cycle

Language of instruction

The language of instruction is specified in the course offering information in the course catalogue.

Intended learning outcomes

After passing this course, the students will be able to

1. Analyze the requirements of a real-life machine learning problem in terms of storage, computation and power,
2. Make informed decisions based on available technology, architectural options, accurate estimates of area, performance and energy that would best meet the targets for the machine learning problem,
3. Create low-energy custom AI solutions that would contribute to a sustainable development,
4. Evaluate major research trends and understand what are the open challenges that the community is focusing on.

Course contents

The course consists of the following two modules:

Requirements Analysis

In this module, we study how to systematically extract requirements in terms of computational operations, their types, interconnect and storage. These requirements are logical and are independent of the implementation style. Many real-life examples will be discussed in class and students assigned problems for hands-on experience.

Being able to understand the energy requirements is the first step in creating low-energy and thus sustainable solutions.

Architecting AI Hardware and Understanding technology and architectural trade-offs

In this module, we study what are the architectural trade-offs when implementing AI hardware. We go into the details of memory hierarchy and their technology options. Memory is the most dominant cost-component and we study how to exploit temporal locality to minimize the cost of memory storage and memory access.

Next to memory, interconnect is the biggest challenge. Wires are the worst scaling aspect of technology today. For instance, moving data by 1 mm on a chip is comparable in energy cost to a single precision floating point. Besides energy cost, interconnect plays a strong role in architectural decisions as well. For instance, it is a common mistake to increase parallelism in computation without increasing the parallelism in access to memory. We show how we can architect designs that allow increase in computation with matching increase in bandwidth to memory.

Finally, we also study what are the options to implement the arithmetic operations in Neural Networks. We also study how to do trade-offs in terms of accuracy vs. implementation cost with the help of a concrete case study from the field of bacterial genome recognition.

Knowing these architectural and technological options to reduce energy will contribute to sustainable AI solutions.

Specific prerequisites

Enrolled as a doctoral student.

Examination

- EXA1 - Written examination, 7.5 credits, grading scale: P, F

Based on recommendation from KTH's coordinator for disabilities, the examiner will decide how to adapt an examination for students with documented disability.

The examiner may apply another examination format when re-examining individual students.

Students will be assigned papers to read and present to the class. Their presentation will be used to judge how well they have grasped the contents of the paper and relate it to their problems.

Small projects will be defined keeping in mind the research interests of the students or groups of students.

Other requirements for final grade

To get a passing grade, the students must attend at least 80% of classes, submit all assignments and give the final presentation on their project.

Ethical approach

- All members of a group are responsible for the group's work.
- In any assessment, every student shall honestly disclose any help received and sources used.
- In an oral assessment, every student shall be able to present and answer questions about the entire assignment and solution.