IL2230 Hardware Architectures for Deep Learning 7.5 credits

Hårdvaruarkitekturer för djupinlärning

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

Establishment

Course syllabus for IL2230 valid from Autumn 2021

Grading scale

A, B, C, D, E, FX, F

Education cycle

Second cycle

Main field of study

Electrical Engineering, Computer Science and Engineering

Specific prerequisites

- Knowledge in digital technology, 7,5 credits, corresponding to completed course IE1204.
- Knowledge of microprocessor design and instruction execution 7,5 credits, corresponding to completed course IS1200.
- Knowledge and skills in programming, 6 credits, corresponding to completed course DD1310/DD1311/DD1312/DD1314/DD1315/DD1316/DD1318/DD1331/DD100N/ID1018.
- Knowledge of hardware techniques, 7,5 credits, corresponding to completed course IS2202/IL2225/IL2236.

- Knowledge of digital hardware design in HDL, 7,5 credits, corresponding to completed course IL1331/IL2203.

Active participation in a course offering where the final examination is not yet reported in Ladok is considered equivalent to completion of the course.

Registering for a course is counted as active participation.

The term 'final examination' encompasses both the regular examination and the first re-examination.

Intended learning outcomes

After passing the course, the student should be able to

- describe and explain basic neural networks and deep learning algorithms and their relations
- explain and justify the hardware design space for deep learning algorithms
- choose and apply an appropriate deep learning algorithm, to solve real problems with artificial intelligence in embedded systems
- analyse and evaluate hardware implementation alternatives for deep learning algorithms
- suggest and justify an implementation architecture for applications with deep learning in embedded resource constrained systems
- discuss and comment new hardware implementation architectures for deep learning and new brain-like computer system architectures that utilise new devices and new concepts

in order to

- understand the necessity, importance, and potential of accelerating deep learning algorithms with low power consumption through specialized hardware architecture
- discuss, suggest and evaluate specialised hardware architectures to implement deep learning algorithms and utilise deep learning concepts in resource constrained reliable systems.

Course contents

The course consists of two modules. Module I introduces basic knowledge in machine learning and algorithms for deep learning Module II focuses on specialised hardware implementation architectures for deep learning algorithms and new brain-like computer system architectures. Apart from presenting relevant informative knowledge, the course contains laboratory and project assignments to create understanding of the related algorithms applied to deal with real problems and to contrast and evaluate alternative implementation architectures, in term of performance, cost, and reliability.

Module I: Algorithms for deep learning

Module I introduces basic machine learning algorithms, basic neural network algorithms and algorithms for deep learning. Among a number of machine learning algorithms, this module introduces the algorithms for linear regression, polynomial regression, logistic regression that are fundamental and most relevant for neural networks. For neural networks we consider perceptrons, multi-layer-perceptrons and in particular the back-propagation algorithm. After presenting traditional statistical learning machine learning and neural networks this module further examplifies deep learning algorithms, specifically Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

Module II: Architecture specialization for deep learning

Module II examines specialised hardware based implementation architectures for deep learning algorithms. From a broad spectrum of potential hardware architectures the design alternatives, such as GPGPU:s, domain specific processors, FPGA/ASIC-based accelerators are presented, together with their advantages and disadvantages. In particular limitations and design alternatives for using deep learning algorithms in embedded resource constrained systems will be discussed. Furthermore this module will discuss new architectures in deep learning for computer system design such as brain-like computer system architectures. A case study with analysis, evaluation and application of a deep learing architectures will be carried out.

Examination

- TEN1 - Written exam, 4.5 credits, grading scale: A, B, C, D, E, FX, F
- LAB1 - Laboratory work, 3.0 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.

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

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.