課程名稱

人工智慧晶片設計與應用

AI-on-Chip for Machine Learning and Inference

3 學分

開課教師: 陳中和，劉峻誠

開課日期

109 學年第二學期 (2021, Spring semester)

Intended students: Undergraduate senior, graduate students, EE or CSE major.

(大學部4年級與碩士生)

Prerequisites:

Digital logic design

Computer organization

Familiar with C, or C++, and Verilog

Objectives:

To teach students, computer architecture and programming for machine learning and inference, neural net model quantization and optimization, edge AI accelerator hardware design and AI application development, case study on contemporary machines.

Grading Policy:

Exam: 2 x 25% = 50 %

Project: 2 x 25% = 50 %

Class Handout:

AI-on-chip system overview

Machine learning hardware design

1D-PE design for convolution operation

Instruction level and thread level parallelism machines

OOO Processor, multi-threading

Data level parallelism machine, SIMD ISA

RISC-V ISA and vector architecture

GPGPU architecture

Overview of CASLab GPGPU system

Architecture of CASLab GPGPU

Supplemental class materials:

Introduction of OpenCL

Introduction of CASLAB GPGPU Compiler

Reference Textbook:

1. Computer architecture, A Quantitative Approach, by John Hennessy and David Patterson, 5/6th edition
2. Deep Learning-Hardware Design (深度學習-硬件設計) 劉峻誠, 2020.

LABS:

LAB1: ML Tool Introductions and Installations (GPU)

LAB2: Implement Lenet-5 model in Tensor-flow (GPU)

LAB3: Kneron Accelerator Platform (KAP) and SDK (AI Accelerator)

LAB4: AI model on Kneron KAP (AI Accelerator)

LAB5: OpenVINO and Intel Movidius (AI Accelerator)

LAB6: OpenCL Exercises on CASLab GPU (GPU)

Projects:

P1: Implement a Tensor Processing Unit (TPU) in Verilog.

P2: Propose an DL application implementation using Kneron KAP and Intel Movidius.

**Class: Wednesday, 2:10 PM To 5 PM.**

**EEB classroom: 92277**

**2/24 (陳中和)**

Lecture: Overview of AI-on-Chip (2 hours)

Lab1/: ML Tool Introductions and Installations (one hour)

**3/3 (陳中和)**

**Domain specific computing (State of art, CIM)**

LAB2: Implement Lenet-5 model in Tensor-flow (GPU)

**3/10 (陳中和)**

**Domain specific computing: 1D PE design for convolution layer (3 hours)**

**3/17, 3/24 (Kneron陳宇春)**

Total 6 hours

Lecture: Case study, Introduction of Kneron NPU

LAB3/4: Kneron Accelerator Platform (KAP) and SDK (AI Accelerator)

Announcement of AI application implementation project P2.

**3/31 (Intel王宗業)**

Intel OpenVINO and Intel Movidius (2 hours)

LAB 5: OpenVINO and Intel Movidius (AI Accelerator) (one hour)

4/7 (陳中和)

Domain specific computing

TPU, Eyeriss, etc

**Announcement of TPU Project.**

4/14 (陳中和)

Lecture: OOO processor, MIMD processor

Advanced computer architecture for instruction level and thread level parallelism applications

4/21 Exam 1 (陳中和)

4/28 and 5/5 (陳中和)

Total 6 hours

Lecture: Data level parallelism machine, SIMD ISA

Lecture: Vector, GPU architecture

Video for discussion

Accelerating AI: Past, Present, and Future by Krste Asanovic

https://www.youtube.com/watch?v=8n2HLp2gtYs

5/12: (陳中和)

Lecture: CASLAB GPU, overview, architecture (2 hours)

LAB6: OpenCL Exercises on CASLab GPU (GPU) (one hour)

5/19 (陳中和)

Exam 2

**5/26 (劉峻誠)**

Lecture: Industry perspectives of AI-on-Chip (3 hours)

**6/2 (劉峻誠)**

Lecture: AI-on-Chip advanced topics (3 hours)

**6/9 and 6/16 (陳中和)**

Total 6 hours

Student final project demo and presentation

Final exam week (week- 18)