EE 382V HARDWARE ARCHITECTURES FOR MACHINE LEARNING

Instructor: Dr. Lizy K. John
Semester: Spring 2020

Course Description:

Machine learning (ML) has become a dominant computing workload. This course provides coverage of architectural techniques to design hardware for training and inference in machine learning systems. Hardware choices for machine learning include CPUs, GPUs, GPU+DSPs, FPGAs, and ASICs. Currently CPUs are used in inferencing tasks while most training is done mostly using GPUs. Tradeoffs in implementing training and inference workloads using these different compute paradigms will be explored. Emerging ML accelerators will be studied. Students will read research papers and complete a major project. Projects can be hardware design projects or characterization/benchmarking/optimization projects.

Course contents:

Basics of Machine Learning and Neural Networks;
Computing need for machine learning;
Overview of hardware platforms for training and inference (CPU, GPU, GPU+DSP, FPGAs, ASIC);
Systolic Arrays
Architectures for ML in the cloud and at the edge;
Memory systems for ML;
In-memory or near-memory computing for ML;
Temporal and spatial parallelism for machine learning;
Energy aware architectures for ML;
ASIC design for machine learning;
GPU based acceleration for ML;
FPGA based acceleration for ML;
Hardware-software co-optimization for machine learning;
Energy, area, delay trade-offs;
Case study of ML chips: Google TPU, MIT Eyeriss, emerging AI chips;
ML benchmarking (MLPerf);
HW/SW Co-design of AI Compute Systems;

Prerequisites:

EE 460N/382N.1 Computer Architecture course
Good programming skills (C, C++, Unix).

Materials:
A collection of papers from IEEE/ACM Conferences

Grading:

Class Participation, Paper Readings, Critiques, Presentations: 35%
Assignments: 30%
Project: 35%