Hardware Acceleration for Machine Learning

Prerequisite(s): ECE 6100 / CS6290, or equivalent
Instructor: Tushar Krishna

Course Objectives

The recent resurgence of the AI revolution has transpired because of synergistic advancements across big data sets, machine learning algorithms, and hardware. In particular, deep neural networks (DNNs), have demonstrated extremely promising results across image classification and speech recognition tasks, surpassing human accuracies. The high computational demands of DNNs coupled with their pervasiveness across both cloud and IoT platforms has led to a rise in specialized hardware accelerators for DNNs. Examples include Google's TPU, Apple's Neural Engine, Intel's Nervana, ARM's Project Trillium, and many more. In addition, GPUs and FPGA architectures and libraries have also been evolving to accelerate DNNs.

This course will present recent advances towards the goal of enabling efficient processing of DNNs. Specifically, it will provide an overview of DNNs, discuss various platforms and architectures that support DNNs, and highlight key trends in recent efficient processing techniques that reduce the computation and communication cost of DNNs either solely via hardware design changes or via joint hardware design and network algorithm changes. It will also summarize various development resources that can enable researchers and practitioners to quickly get started on DNN design, and highlight important benchmarking metrics and design considerations that should be used for evaluating the rapidly growing number of DNN hardware designs, optionally including algorithmic co-design, being proposed in academia and industry.

Learning Outcomes

As part of this course, students will: understand the key design considerations for efficient DNN processing; understand tradeoffs between various hardware architectures and platforms; learn about micro-architectural knobs such as precision, data reuse, and parallelism to architect DNN accelerators given target area-power-performance metrics; evaluate the utility of various DNN dataflow techniques for efficient processing; and understand future trends and opportunities from ML algorithms down to emerging technologies (such as ReRAM).

Course Structure

The course will involve a mix of lectures interspersed with heavy paper reading and discussions. A semester long programming-heavy project will focus on developing a hardware accelerator for ML, and prototyping it on a FPGA.

Course Text

The material for this course will be derived from papers from recent computer architecture conferences (ISCA, MICRO, HPCA, ASPLOS) on hardware acceleration and ML conferences (ICML, NIPS, ICLR) focusing on hardware friendly optimizations, and from blog articles from industry (Google, Deep Mind, Baidu, Intel, ARM, Facebook).

Syllabus and Outline

1. **Overview of Deep Learning**
   - Traditional Machine Learning
   - Neural Networks
   - Deep Neural Networks - CNNs and RNNs

2. **Training vs Inference**
   - Feed-Forward Networks
   - Backpropogation

3. **Dataflows**
   - Data Reuse
   - Dataflow Taxonomy
   - Dataflow Analysis

4. **Accelerating DNNs in Hardware**
   - GPUs
   - Spatial Accelerators
   - Systolic Arrays
   - FPGAs

5. **HW-SW Co-Design**
   - Binary Neural Networks
   - Bit-Precision
   - Pruning and Sparsity

6. **Benchmarking**
   - MLPerf
   - DAWNbench

7. **Emerging Technologies**
   - ReRAM
   - Analog Accelerators

8. **Future Trends**
   - Unsupervised Learning
   - Reinforcement Learning
   - Federated Learning

Course Grading

Lab Assignments	40% (4 Labs - 10% each)
Midterm	10%
Paper Critiques	10%
Presentation on Paper/Case Study	10%
Project	30% (Milestones 10%, Report 10% Presentation 10%)

Course Policies

Attendance and Absence. Students are expected to attend all lectures and exams. If you have a documented emergency or a university mandated reason because of which you have to miss an exam, get in touch with the instructor before (preferable) or latest by the day of the exam.

Learning Accommodations. If needed, we will make classroom accommodations for students with disabilities. These accommodations should be arranged in advance and in accordance with the office of Disability Services (http://www.adapts.gatech.edu)

Honor Code. Students are expected to abide by the Georgia Tech Academic Honor Code (http://www.policylibrary.gatech.edu/student-affairs/academic-honor-code). Honest and ethical behavior is expected at all times. All incidents of suspected dishonesty will be reported to and handled by the office of student affairs. You will have to do all assignments individually unless explicitly told otherwise. You may discuss with classmates but you may not copy any solution (or any part of a solution).