Hardware-Software Co-Design for Machine Learning

Course Objectives

The advancement in AI can be attributed to the synergistic advancements in big data sets, machine learning (ML) algorithms, and the hardware and systems used to deploy these models. Specifically, deep neural networks (DNNs) have showcased highly promising results in tasks across vision, speech and natural language processing. Unfortunately, DNNs come with significant computational and memory demands -- which can be Zeta (10^{21}) FLOPs and Tera (10^{12}) Bytes respectively for in Large Language Models such as those driving ChatGPT. Efficient processing of these DNNs necessitates HW-SW co-design. Such co-design efforts have led to the emergence of (i) specialized **hardware accelerators** designed for DNNs (e.g., Google's TPU, Meta's MTIA, Amazon's Inferentia & Trainium, and so on) and (ii) specialized **distributed systems** comprising hundreds to thousands of these accelerators connected via specialized fabrics (e.g., . Furthermore, GPUs and FPGA architectures and libraries have also evolved to accelerate DNNs.

This course aims to present recent advancements that strive to achieve efficient processing of DNNs. Specifically, it will offer an overview of DNNs, delve into techniques to distribute the workload, dive into various architectures and systems that support DNNs, and highlight key trends in recent techniques for efficient processing. These techniques aim to reduce the computational and communication costs associated with DNNs, either through hardware and system optimizations. The course will also provide a summary of various development resources to help researchers and practitioners initiate DNN deployments swiftly. Additionally, it will emphasize crucial benchmarking metrics and design considerations for evaluating the rapidly expanding array of DNN hardware designs, system optimizations, proposed in both 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; understand the need and means to distributed ML; evaluate the utility of various DNN strategies for end-to-end efficient execution; and understand future trends and opportunities from ML algorithms, system innovations, down to emerging technologies (such as ReRAM).

Course Structure

The course will involve a mix of lectures interspersed with guest lectures from industry practioners, heavy paper reading, and discussions.

Course Text

The material for this course will be derived from papers from recent computer architecture conferences (ISCA, MICRO, HPCA, ASPLOS) on hardware acceleration, systems conferences (SOSP, MLSys) for distributing ML, ML conferences (ICML, NeurIPS, ICLR) focusing on future trends, and blog articles from industry (Google, Microsoft, Meta, NVIDIA, Baidu, Intel, Arm).

Syllabus and Outline:

1. **Review of Machine Learning Concepts**
    a. Key principles of Machine Learning - Neural Networks
    b. Overview of Deep Learning – CNNs, Transformers
    c. Training vs Inference - Feed-Forward Networks, Backpropagation
2. **Introduction to Hardware Acceleration Technologies: GPUs, TPUs, and Beyond**
    a. Importance of hardware acceleration in Machine Learning
    b. Overview of Dataflows
        i. Data Reuse
        ii. Dataflow Taxonomy
        iii. Dataflow Analysis
    c. Hardware-specific optimizations for machine learning
        i. Systolic Arrays
        ii. Spatial Accelerators
    d. Detailed discussion of GPUs, TPUs, and emerging hardware technologies
3. **Basics of Distributed Machine Learning**
    a. Gradient descent and stochastic gradient descent
    b. Communication Collectives for distributed training - Broadcast, Reduce, All-reduce, Scatter, All-Gather
4. **Modes of Distributed Training**
    a. Introduction to data parallelism
    b. Introduction to tensor parallelism
    c. Introduction to pipeline parallelism
5. **Benchmarking**
    a. MLPerf
    b. DAWNbench
6. **Emerging Technologies**
    a. ReRAM
    b. Analog Accelerators
7. **Case study and Future trends**
    a. Sustainable Computing for Machine Learning
    b. Other Models – Recommendation models and MOEs
    c. Federated Learning

Class participation: 5%
Assignments (2 assignments): 40%

- [1, 20%]
- [2, 20%]

Midterm: 20%
Presentation on Paper/Case Study: 10%
Project: 25% (Milestones 15%, Report 10%)