ECE 498NSU/NSG – VLSI in Machine Learning (Fall 2024)

Instructor: Naresh Shanbhag

TAs:
Vignesh Sundaresha: vs49@illinois.edu
Kaining Zhou : kainingz@illinois.edu

Prerequisites: ECE 313 and ECE 342 or instructor's consent

Text: List of papers and instructor notes;

Lecture: M and W 10:00-11:20, ECEB 2022

Instructor Office Hours: Wednesdays 2PM-3PM, CSL 414

TA Office Hours: Thursdays 2pm-4pm, ECEB 2036

Course Description: This course will present challenges in implementing machine learning algorithms in VLSI (silicon) for applications such as wearables, IoTs, autonomous vehicles, and biomedical devices. Simple single-stage classifiers will be discussed first followed by deep neural networks. Finite-precision analysis will be employed to design fixed-point networks to minimize energy, latency, and memory footprint. Training algorithms of both single-stage and deep nets (back-prop) will be presented followed by their fixed-point realizations. Algorithm-to-architecture mapping techniques will be explored to trade-off energy-latency-accuracy in deep learning digital accelerators and analog in-memory architectures. Fundamentals of learning behavior, fixed-point analysis, architectural energy, and delay models will be introduced just-in-time throughout the course. Case studies of hardware (architecture and circuit) realizations of deep learning systems will also be presented. Homeworks will include a mix of analysis and programming exercises in Python and Verilog. NSU section will complete a term project involving the implementation of deep nets on an embedded hardware platform such as an FPGA/MCU. NSG section will write a term paper based on the literature review of a specific topic of their interest, and conduct research project on that topic.

Course Grading: NSU section will be graded on weekly homeworks (30%) involving Python and Verilog programming well as design and analysis problems, and two midterms (30%), and a term design project (40%). NSG section will be graded as: 25% (homeworks), 25% (two midterms), 30% (research project), and 20% (term paper).

Topical Outline

1. **Introduction (Week 1):** modern day applications in human-centric (e.g., biomedical/wearable devices) and autonomous (unmanned vehicles) platforms. Historical overview of AI, connections to neuroscience, early single stage neural networks (ADALINE, perceptron).
2. **Deep Neural Networks (Weeks 2-6):** Survey of popular networks and datasets. Estimating computational and storage costs. Design of fixed-point DNNs and CNNs for inference and training. Interplay between learning behavior, precision of computation, energy, and latency. Reducing network complexity via model compression and lightweight network design.
3. **DNN Accelerators (Weeks 7-10):** Study relationship between memory BW, peak and achievable performance, and energy efficiency. Algorithm-to-architecture mapping techniques. Statistical error compensation to push the limits of energy-latency trade-offs. Case studies of DNN accelerators.
4. **In-Memory Computing (Weeks 10-13):** In-memory architectures for SRAM and embedded non-volatile resistive memories. Energy, latency and accuracy trade-offs and design methods for in-memory computing. Case studies of in-memory architectures.
5. **The Future (Week 14-15):** advanced topics.