ECE 498NSU/598NSG - Deep Learning in Hardware

Instructor: Naresh Shanbhag **TAs:** Charbel Sakr and Hassan Dbouk

Prerequisites: ECE 313 and ECE 342

Text: List of papers and instructor notes; **Lecture**: Tu and Th 11:00-12:20, ECEB 3015

Office Hours: Mondays 3PM to 4PM, CSL 414

TAs Office Hours: Fridays 12PM to 1PM, ECEB 5034 (Charbel)
Tuesdays 2PM to 3PM, ECEB 4034 (Hassan)

Course Description: This course will present challenges in implementing deep learning algorithms on resource-constrained hardware platforms at the Edge such as wearables, IoTs, autonomous vehicles, and biomedical devices. Fixed-point requirements of deep for deep neural networks and convolutional neural networks including the back-prop based training will be studied. 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 in just-in-time manner throughout the course. Case studies of hardware (architecture and circuit) realizations of deep learning systems will be presented. Homeworks will include a mix of analysis and programming exercises in Python and Verilog leading up to a term project.

Syllabus

Being a first-time offering, this outline has been designed to provide for some flexibility by allowing the instructor to choose a large subset of the listed topics. The list will become more precise in the second offering.

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 2-5):** algorithmic view of DNNs and CNNs including training via back-propagation. Popular networks (LeNet, AlexNet, VGGNet, ResNet, MobileNet), benchmark datasets (MNIST, CIFAR, ImageNet), and metrics. Understanding the dynamics of learning behavior by drawing analogies between the backpropagation algorithm and LMS. Methods for inference and training in fixed-point. Estimating computational and representational (storage) costs.
3. **Digital Accelerators (Weeks 6-10):** data-flow models of fixed-point deep learning algorithms. Efficient algorithm-to-architecture mapping techniques including data reuse, output, weight and row stationary architectures. Neuromorphic architectures. Energy and latency models to estimate and compare associated costs of various mapping techniques and explore trade-offs. Case studies of digital deep learning architectures (Eyeriss, DianNao series, TPU, Cambricon, TrueNorth), and practical IC realizations.
4. **In- and Near Memory Architectures (Weeks 11-14):** DRAM-based (e-DRAM), 3D architectures (HMC, HBM), SRAM-based deep in-memory architectures, architectures based on non-volatile resistive memories (RRAM PCM, CBM crossbars). Energy, latency and accuracy trade-offs in analog computation. Case studies of architectures (ISAAC, PRIME) and practical IC realizations.
5. **The Future (Week 15):** challenges and opportunities in deep learning hardware – designing programmable architectures, Shannon-inspired models of computation, developing CAD design methodologies, enabling emerging beyond CMOS fabrics, obtaining fundamental limits, and others.

Grading: Course grade will be based on homeworks (50%) involving Python and Verilog programming well as design and analysis problems, a term design project (40%), and scribing assignments (10%). Since this course will be taken by ECE 598NS students as well, each homework will have extra problems specifically for the graduate students. These problems will be optional for ECE 498NS students.