CSC-591/791/ ECE 591-019
Software-Hardware Co-design for Intelligent Systems

Instructor(s): Wujie Wen, Associate Professor, Department oof Computer Science, wwen@ncsu.edu

Website: https://wenwujie.github.io/

Objective or Description:

Fueled by the proliferation of sensing data and the advancement of machine learning (ML), intelligence is becoming a household brand in many Cyber-Physical Systems and Internet-of-Things applications. This course will offer in-depth coverage of the latest software-hardware co-design methodologies for developing energy-efficient, low-latency, reliable and trustworthy intelligent systems. Techniques that are widely investigated and adopted in industry companies and academic communities will be discussed and practiced. Topics include but are not limited to: ML basics and system performance evaluation metrics, hardware platforms and software frameworks, modern neural networks, training and testing, design paradigms of algorithm-hardware co-design for ML acceleration like hardware-oriented model compression, pruning and quantization, sparsity, modern ML hardware architectures, near memory and in memory computing, trustworthy and private computing by design, machine vision guided image compression etc. Hands-on projects will be utilized with real-world applications and datasets. The final project is expected to be deployed and evaluated on hardware platforms (such as embedded GPU, embedded TPU) when possible. Classes will be running by combining lecture sessions, presentation sessions, and labs/project.

Prerequisites:
Basic programming experience (e.g., C++, Python) and knowledge of linear algebra, probability etc. Knowledge of computer hardware is a plus.

Textbook:
There are no required textbooks for this course. The class will be mostly relying on related reading materials, including but not limited to the latest research papers from top conferences, tutorials, workshops, and webpages etc.

Topics: (tentative)--modern machine Learning basics/architecture, training and inference, model compression, sparsity, quantization, federated learning, ML hardware architectures for cloud and edge, near memory and in memory computing, trustworthy AI (adversarial machine learning, fault injection attacks etc.), privacy-preserving ML (differential privacy, encrypted inference etc.), machine vision guided image compression etc.

Grading: Lab Assignments (2): 40% (20%+20%)
b) Paper Presentation/Discussion: 20%
c) Project: 40% +10% bonus

Cross-listing in other departments: CSC591/791