CS249r: Special Topics in Edge Computing

Special Topic: Machine Learning Runtimes

**Special Topics:** The course covers a broad range of hardware and software topics in the context of smart/intelligent embedded systems. Traditional embedded systems are passive electronic devices that perform a single task and operate in isolation. In contrast, modern embedded systems are intelligent devices that involve complex hardware and software to perform a multitude of cognitive functions collaboratively. Designing such systems requires us to have a deep understanding of the target application domains, as well as an appreciation for the coupling between the hardware and the software subsystems. To this end, the course is designed to help students come up to speed on various aspects of edge computing hardware and software, including interactions with remote cloud computing systems and microcontroller edge endpoints. Each year the course focuses on one or more emerging areas, including but not limited to autonomous vehicles, robotics, mobile, etc. The course is ideal for students interested in cross-layer and interdisciplinary work. It is a seminar-style course so students are expected to discuss and interact after reading the assigned research papers. Finally, there are programming assignments that provide the foundation for a course project.

**Current Topic:** The Fall 2019 CS249r course will focus on **Machine Learning Runtimes.** Machine learning, specifically deep learning, is shaping the world of computer science and engineering. There are over 50 new machine learning (ML) models released each day (based on arXiv). As of last year, there are over 45 new hardware startups set on accelerating ML. To cope with the growing diversity in models and hardware, the number of ML runtimes/frameworks and libraries is also swelling. Software development of ML runtimes is a fast-paced and ruthless to shave off seconds of executions time in model training time and performing inferences. The course focuses on how to optimize ML runtimes (PyTorch, TensorFlow, MXNet, etc.) for performance and power efficiency (in the context of warehouse-scale machines), and size (in the context of embedded systems).

**Teaching Fellow:** TBD

**Course Meeting Times:** TBD

**Office Hours:** TBD

**Webpage:** Canvas
Semester: Fall 2019  
Instructor: Prof. Vijay Janapa Reddi  
Office: Maxwell Dworkin 147  
Email: vj@eecs.harvard.edu

Prerequisites: CS141 and/or basic computer architecture and digital design, CS 161 or some other course where the student has had some level of systems programming experience

Topics: The course will cover a range of topics that include but are not limited the following:
- System design, architecture and implementation for autonomous machines/systems (simulators, physics engines, robot model, hardware, etc.)
- Runtime systems for swarm intelligence, and power and performance efficiency
- Algorithms for decision making (reinforcement learning, supervised learning, etc.)

The course will require students to read a variety of different papers from top conferences.

Grade Formula:
- Homework – 25% (3 assignments including programming exercises)
- Exam – 15% (30 minute oral exam, to be scheduled)
- Class Participation – 25%, Including:
  - Discussion Participation (throughout semester during class)
  - Discussion Leadership (you will be assigned certain papers to lead the class through the in-class discussions)
  - Paper Reviews (Highlight the key positives/negatives in papers)
- Project (including final project presentation) – 35%
  - Students will undertake an in-depth course project on a topic of their choosing. The project will introduce the students to research in the field of computer architecture. Project proposals will be due about midway through the class and regular meetings with the teaching staff will help refine the goals and milestones. Final output of the project is a presentation to the class and a project write up in the form of a research style paper.