CS 492 Special Topics in Computer Science
Systems for Machine Learning
Spring 2021

Instructor: Jongse Park
Time: TU/TH 10:30am – noon
Credits: 3-0-3
Room: E3 #2445

Course description & goals

Welcome! This course is a special topic course where you will learn how the systems for Machine Learning (ML) work. Not only does this course involve the "learn" part, but also you will have to get your hands dirty to "build" an actually-working system for ML.

While the algorithmic aspect of ML is not the focus of this course, you will learn the basics since the ML systems and their computing stack are heavily algorithm-oriented. Given the algorithms, we will look into the entire computing stack--from domain-specific programming interface and platforms (e.g., Tensorflow) to compiler (e.g., TVM) and accelerators (e.g., TPU/GPU). Most contents treated in this course will be based on the state-of-the-art software system and hardware accelerators for ML, which you would find from products and papers that field-leading companies and schools have released and/or published in recent years.

As is always the case, the best way to learn systems is, you build one on your own. You will build your own DNN platform (just like Tensorflow but much simpler) in C/C++, which exclusively and limitedly supports DNN inference. If you don't know what I mean by "inference", don't worry, I will teach you! Later in the semester, you will optimize your own system's performance using various parallelization methods (e.g., AVX, MKL, BLAS, , CUDA, and cuDNN). I will make the projects even more fun by setting these projects as competition where you would be able to brag the performance you achieved from your implementation and beat the fellow students.

The details of lectures and projects are tentative and subject to be updated but the main theme will be more and less the same.

Communication

Discussion: The class discussion board is on Piazza. You may sign up at this link: https://piazza.com/kaist.ac.kr/spring2021/cs492

I will post course-related announcements and information on the board. You must read the discussion board at least once per day, and you should post course-related questions and responses there. I expect you to make good use of the discussion group and of TA support when you have technical or administrative questions or problems. You are responsible for any and all information posted to Piazza by any of the course staff. We will pin any important announcements to the top of the Piazza feed. You are expected to read all announcements within twelve hours of their being posted.

Email to the Instructor or TAs: Emails to course staff should begin with "CS492:" in the subject line, followed by a brief description of the purpose of your email. If you follow this rule, we will be better able to address your questions in a timely manner. If you do not, a response to your email may be delayed indefinitely.

Prerequisites

This class has two prerequisites: CS230 and CS311. The system programming skills you learned from CS230 will be primarily needed for you to deliver your projects. In addition, CS230 offers the basic system software knowledge, which is essential for you to understand the software-related discussions in this course. The undergraduate-level computer architecture background offered by CS311 will be required for you to follow the hardware-related discussions. Without the prerequisite background knowledge and implementation skills, it will be challenging for you to successfully follow the materials.

Grading

This course will include mid-term and final exams, projects, and class participation.
Projects: 70%
Paper Critiques: 30%
Class participation: 10%

Due to the COVID situation, we will not have midterm and final exams. Instead, there will be multiple projects that require a significant amount of time and effort. You will also need to read several papers and write critiques that discuss strengths/weaknesses of the paper, technical insights, etc. Attendance, attitude, and any other miscellaneous evaluation factors will be considered in the 10% of class participation.

Cheating

Instructor and TAs will perform several forms of cheating detections under the hood.
Examples of cheating include: looking at someone else's program, writing your program while talking to someone else about it, talking another student through the solution code, allowing others to look at your solution code, and looking on the Internet for code to solve your programming assignments, including stackoverflow and other such sites. If you have any doubts about what is allowed, ask the instructor.

Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Because such dishonesty harms the individual, all students, and the integrity of the University, policies on scholastic dishonesty will be strictly enforced.

The penalty for cheating on projects and critiques is as follows:

First time: a letter-grade down (e.g., A0 □ B0, B+ □ C+) in the course.

Next: the F grade