CSCI 1390, Spring 2025: Systems for Machine Learning

Class Meetings: Tuesdays, Thursdays 09:00 - 10:20, CIT 368

Canvas

EdStem

Staff

Course logistics are also available on this calendar.

Description

Many applications, across industries varying from ecommerce to education, rely on data processing and machine learning systems for data analytics tasks. Deep learning techniques are now being applied to problems such as search, coding assistants, and chip placement. Due to how widely used these applications are, performance, specifically latency, throughput, and hardware efficiency, is very important. However, achieving high performance in these systems can be challenging.

This class will explore systems-related challenges related to building, training, deploying and managing large-scale data processing and machine learning systems. The learning goals of the class are for students to deeply understand the systems techniques that allow ML systems to be deployed efficiently, at scale. Topics include efficient training and inference, understanding how to build hardware ML algorithms, understanding ML algorithm performance, GPU programming and cuda, recent transformer architectures, efficient retrieval, and more.

This is the first offering of this class; students should anticipate technical difficulties, and the course components and schedule are subject to change.

Registration

Due to constrained TA resources, this offering of CS 1390 will be capped at 56 students. Students requesting an override must (a) have fulfilled the prerequisites (CS 300, 330, or 1310, and one of (CS 410, 1410, 1470, 2470, 1420) and (b) complete course assignments during shopping period. All registered students will be responsible for completing all course assignments regardless of when they register for the class. We will be managing the waitlist through the waitlist form some of you may have filled out. The class is *already full*, so unless registered

students drop the class, we will not be letting students off the waitlist. *If you are planning to drop the class*, please do so ASAP so we can let students off the waitlist.

New waitlist form: Waitlistform Even if you filled out the old form, please fill out the new form; so we can keep track of folks who are still interested in the class. The TAs will also be posting it on EdStem.

Auditing

If you wish to audit (or "vagabond") the class, note that the aforementioned limits on TA resources still apply, so we won't be able to offer auditors grading support or support in office hours.

Components and Grading

Course Projects: 65 %

The largest component of this class is a series of four course projects. Each project will involve an implementation task and writeup that answers conceptual questions. Course staff will grade projects based on the correctness and performance of the implementation, and clarity of the writeup. Projects may either be graded offline by course staff, or through live meetings with course staff. During these gradings meetings, students should be able to answer technical questions about the project. Projects in this course offering are in both Python (using Pytorch) and Cuda (to program GPUs). The breakdown of how much each project contributes to the final grade is below (percentages add to 65).

- Project 1: 18 %
- Project 2: 18 %
- Project 3: 18 %
- Project 4: 11 %

Written Homeworks: 20 %

There will be two written homeworks, where students read pieces of technical writing (either a research paper or a tech report), and students must respond to conceptual questions. Responses will be graded on clarity. Late hours cannot be used on the written homeworks.

Post Lecture Quizzes: 10 %

After almost all the lectures, the teaching staff will post a short quiz that reinforces lecture content, which will be graded on completeness and correctness. If you have attended the lecture (or watched the lecture video posted to canvas), you should be able to answer the quiz questions. Students *cannot submit PLQ responses late*, but they can *miss up to 5 PLQs across the entire semester*, no

questions asked. If there is a PLQ, it will be posted right after each lecture and the response is due at the start of the *next* lecture (responses will not be accepted after this).

Class Participation: 5 %

The final component of the grade is class participation. Students can participate in a number of ways: answering each other's questions on EdStem, asking particularly helpful questions in EdStem, lecture, or office hours, and helping each other in office hours.

Late Policy (for Course Projects)

Each student has 144 total late hours for projects. These hours are designed to help you cope with unexpected emergencies, and only apply to project submissions. In other words, you can hand in your projects late, but the total amount of lateness summed over all the deadlines must not exceed 144 hours. For the purpose of calculating late hours, we exclude hours between midnight and 7am. This means that you can get a good night's sleep before you continue working on your project without losing extra late hours while you sleep.

If you don't hand in a project by the last day of classes, we'll give the assignment a zero (i.e., no credit whatsoever). However, if you hand an assignment in late, and your total late time (including the late time for that assignment) exceeds 144 hours, and you hand it in by the last day of classes, then we'll give it a D or an F depending on whether it seems to basically work. Therefore, it's better to complete and hand in a project even if you have already used your late hours.

You can divide up your 144 hours among the projects however you like; you don't have to ask or tell us, but you *cannot use more than 72 hours per project.* Finally, can only use the 144 hours for the projects, and not for post-lecture quizzes or for the written homework. If you want an exception to these rules, you will need a Dean's note.

Finally, if you find yourself struggling, please contact the instructors as soon as you feel able; and if you have a health condition that affects your learning or classroom experience, please contact us so we can seek accommodations. We do not generally make exceptions to the late policy, but we work hard to help all our students complete the course without overwhelming stress.

Time Breakdown

Students should expect to spend 3 hours per week in lecture, and approximately 10 hours a week working on the course projects and/or written homeworks, for a total of 180 hours over the entire week.

Collaboration Policy

Students should understand and follow the Brown Academic Code and the Code of Student Conduct.

Additionally, specific to this course, we *encourage working with other students to build conceptual understanding and debug software issues*. However, each student is responsible for their own project implementation and writeup, and their own written homework response.

Students must understand their submissions; instructors will interview and quiz students about their answers as part of the grading process to determine this. For all class assignments, students must cite all sources (people, websites, papers, etc.) that they consult as a part of their work. External sources include but are not limited to previously published articles, blog posts, Stackoverflow or similar sites, conversations with other people, etc. This policy is not meant to discourage the use of external sources, but rather to codify a standard academic practice. Be generous with citations.

A note about generative AI: students should not use generative AI to fill in code snippets for any of the course projects, or write answers to the conceptual questions or written assignments. Remember that instructors always reserve the right to ask students about their implementations and writeups, so it is important for students to have worked through the assignments.

Finally, by taking this class, you agree to never post solutions for any assignments publicly.

Schedule

Homeworks and Projects

Note that all dates are tentative and we may change the schedule as the course goes on. Homeworks and projects will be released latest by 5 PM on the release date, and will be due at *6 PM on the due date*.

Date	Assignment	Due Date
01/28/2025	Project 1: Parallelism	02/19/2025 (with mid-project checkin due on 02/04/2025)
02/19/2025	Written HW 1	02/27/2025
02/25/2025	Project 2: Attention & KV Caching	03/18/2025
03/18/2025	Project 3: Cuda Kernels	04/15/2025
04/15/2025	Project 4: Vector Databases	04/24/2025

Date	Assignment	Due Date
04/24/2025	Written HW 2	05/06/2025 (end of reading period)

Lectures (PLQs Posted Here)

Lecture recordings are available here (accessible to anyone at Brown).

Date	Topic	Notes	PLQ Link
01/23/2025	Introduction to the Course	Intro Slides	No PLQ
01/28/2025	Introduction to Deep Learning Systems	Notes	No PLQ
01/28/2025	Overview of DL Stack, Data Parallelism and All-Reduce	Notes	PLQ
02/04/2025	ZeRO Redundancy, Tensor Model Parallelism, Pipeline Parallelism	Notes	PLQ
02/06/2025	Attention, Transformers Overview	Notes	PLQ
02/11/2025	Intro to Hardware and Arithmetic Intensity	Notes	PLQ
02/13/2025	Transformers Flops Analysis	Notes	No PLQ
02/20/2025	Transformers Training Flops and Opportunities for Hardware Optimization	Notes	PLQ
02/25/2025	Linear Attention Algorithms and Flash Attention	Notes	No PLQ
02/25/2025	Flash Attention cont. and Intro to Inference Optimization	Notes	PLQ

Credits

Website designed using pandoc. The class content borrows heavily from cs229s at Stanford University, created by Azalia Mirhoseini and Simran Arora, and cs 15-849, taught by Zhihao Jia. We have also drawn on work from Simon Boehm, Deepak Narayanan, and cs149 at Stanford. We thank Brown T-Staff for managing the Hydra cluster, without which it would be difficult to complete projects 2 and 3. The collaboration policy is taken from cs1675 at Brown. The late hours policy is taken from Brown's cs300.