EE 599: Systems for Machine Learning

Units: 4
Spring 2024
Lecture:
Wednesday: 6:30 - 8:20 pm
Saturday: 9:00 - 10:50 am

Discussion: Monday: 6:30 - 7:20 pm

Location: TBD

Instructor : Arash Saifhashemi
Office: EEB 504B
Office Hours: TBD
Contact Info: saifhash@usc.edu

Teaching Assistant: TBD
Office: TBD
Office Hours: TBD
Contact Info: TBD

Course Description

This course provides an in-depth exploration of systems for machine learning algorithms, focusing on research and recent advancements. The course examines the parallel nature of machine learning algorithms and their compatibility with parallel-everything hardware blocks, such as chip multiprocessors, multithreading, GPUs, and parallel software.

One primary objective of the course is to enable students to understand and leverage the potential of mapping machine learning algorithms to GPUs and TPUs, which has proven to be highly beneficial for enhancing ML performance. While performance improvements remain crucial, the course also emphasizes the significance of power efficiency and other constraints in machine learning systems.

Moreover, the course addresses the wide diversity of ML models and the heterogeneity of hardware accelerators that cater to their computing needs. Students will gain insights into various ML accelerators, both in research and academia, that have emerged to maximize power efficiency and optimize ML performance for different application domains, including Transformer-based natural language processing models. The course places significant emphasis on exploring the diverse range of models and the heterogeneous hardware accelerators utilized to execute these models.

Lastly, the course delves into the critical aspects of security in ML systems. Students will examine the escalating concern regarding security breaches and the importance of data privacy and integrity. The module dedicated to privacy and security aspects equips students with knowledge and tools to ensure trustworthy ML systems, particularly in critical application domains like medicine and transportation.

Learning Objectives

By the end of the course, students will:

- Gain a comprehensive understanding of hardware systems for machine learning algorithms, including chip multiprocessors, multithreading, GPUs, and parallel software.
- Possess the knowledge and skills to effectively map machine learning algorithms on GPUs and massive TPUs to optimize ML performance.
- Understand the importance of power efficiency and other constraints in machine learning systems, enabling them to design and implement ML accelerators that maximize energy efficiency.
- Acquire insights into the diverse computing needs of various ML models, such as Transformer-based natural language processing models, and be able to select appropriate hardware accelerators accordingly.
- Understand the privacy and security challenges associated with ML systems, and be equipped with strategies and tools to ensure data privacy, integrity, and secure microarchitectures.
- Develop a critical awareness of emerging trends and recent developments in hardware systems for machine learning, enabling them to stay updated and adapt to future advancements in the field.

Prerequisite(s): None
Co-Requisite(s): None
Concurrent Enrollment: None
Recommended Preparation:

- Proficiency in programming languages commonly used in machine learning, such as Python, is required.
- Familiarity with linear algebra and calculus concepts is beneficial for understanding machine learning algorithms and their implementation.
- Basic knowledge of computer architecture and parallel computing principles will be helpful in comprehending hardware systems for machine learning.

Course Notes

- This is a Letter grade course.

Technological Proficiency and Hardware/Software Required

Technological Proficiency:

- Proficiency in programming languages commonly used in machine learning, such as Python, is required.

Hardware/Software Requirements:

- A computer with sufficient processing power and memory to handle machine learning algorithms and simulations.
- Access to software libraries and frameworks for machine learning, such as TensorFlow or PyTorch, along with their dependencies.
- Installation of development environments and tools, such as Anaconda or Jupyter Notebook, for coding and experimentation.
- A Github account.

Additional resources:

- USC Computing Center Laptop Loaner Program.
- Zoom information for students

- Blackboard help for students
- Software available to USC Campus.

Required Readings and Supplementary Materials

- REQUIRED READING: Sze, Chen, Yang and Emer: "Efficient Processing of DNNs," Morgan&Claypool Press. 2021. ISBN: 9781681738321
- REQUIRED READING: http://eyeriss.mit.edu/tutorial.html
- A list of required readings and notes will be handed out (detailed schedule below).

Optional Readings and Supplementary Materials

- SUGGESTED READING: Dubois, Annavaram and Stenström: "Parallel Computer Organization and Design" Cambridge University Press, 2012. ISBN: 978-0-521-88675-8. Background material is in this book and some of the planned chip multiprocessor related discussion is in this book. Also, some problems and reading assignments will be picked from the book.

Description and Assessment of Assignments

[Assignments are aligned with the learning objectives, meaning that each assignment serves to measure student performance on at least one learning objective. This section describes each assignment, how it maps onto learning objective(s), and how it will be graded. Faculty can provide detailed descriptions of assignments here, or they can provide descriptions that provide an overview, reserving actual assignment descriptions/prompts for later in the course. Every category of graded work should be briefly described. Note: some programs also require rubrics on the syllabus. Search for examples of rubrics on the CET website.]

Participation

- Class attendance is required.
- We will consider in-class work as participation.

Grading Breakdown

Table 1 Grading Breakdown (Tentative)

Assessment Tool (assignments)	Points	% of Grade
Midterm and Quizzes	20	20
Programming, reading, and homework assignments	40	40
Project	20	20
Final exam	20	20
TOTAL	100	100

Grading Scale

[The following is an example of what a grading scale might look like. Check with your department or school for the recommended grading scale.]

Course final grades will be determined using the following scale:

Table 2 Course Grading Scale

Letter grade	Corresponding numerical point range
A	95-100
A-	90-94
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
D+	67-69
D	63-66
D-	60-62
F	59 and below

Assignment Submission Policy

- Programming assignments require electronic submission.
- Grading will be handled by both automatic and manual means.
- Code submissions may be reviewed and applied to a series of test cases to determine integrity.

Grading Timeline

Students can expect grading and feedback from the instructor in about a week after the submission deadline.

Course Specific Policies

None

Attendance

1. Attendance is expected, and non-attendance may impact the final grade as stated in the syllabus.
2. Accommodations will be made for student athletes and advance notice of religious observations.
3. Missed class sessions may require completion of alternative course work.
4. Students must notify the instructor of anticipated absences or exceptional circumstances.

Classroom norms

Refer to the CET resource, A menu of discussion norms.

Zoom etiquette

Following these Zoom etiquette guidelines promotes effective and respectful online communication.

1. Maintain a respectful and professional tone in all communications.
2. Follow any specific camera requirements set by the instructor.
3. Contact the instructor in advance if unable to keep the camera on during synchronous sessions.
4. Actively participate and engage in discussions.
5. Use hand-raising or chat functions to contribute and ask questions.
6. Avoid multitasking or unrelated activities during Zoom sessions.
7. Respect privacy and refrain from sharing sensitive information without consent.
8. Follow the instructor's guidelines for microphones, video, and screen sharing.
9. Mute your microphone when not speaking to minimize background noise.
10. Listen attentively and seek clarification when needed.
11. Contact the instructor for questions or concerns about Zoom etiquette.

Academic Integrity

The University of Southern California is foremost a learning community committed to fostering successful scholars and researchers dedicated to the pursuit of knowledge and the transmission of ideas. Academic misconduct is in contrast to the university's mission to educate students through a broad array of first-rank academic, professional, and extracurricular programs and includes any act of dishonesty in the submission of academic work (either in draft or final form).

This course will follow the expectations for academic integrity as stated in the USC Student Handbook. All students are expected to submit assignments that are original work and prepared specifically for the course/section in this academic term. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s). Students suspected of engaging in academic misconduct will be reported to the Office of Academic Integrity.

Other violations of academic misconduct include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

- Please ask instructors if you are unsure about what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.
- Collaboration. In this class, you are expected to submit work that demonstrates your individual mastery of the course concepts.
- Group work. Unless specifically designated as a 'group project,' all assignments are expected to be completed individually.
- Computer programs. Plagiarism includes the submission of code written by, or otherwise obtained from someone else.

If found responsible for an academic violation, students may be assigned university outcomes, such as suspension or expulsion from the university, and grade penalties, such as an "F" grade on the assignment, exam, and/or in the course.

Course Content Distribution and Synchronous Session Recordings Policies

USC has policies that prohibit recording and distribution of any synchronous and asynchronous course content outside of the learning environment.

Recording a university class without the express permission of the instructor and announcement to the class, or unless conducted pursuant to an Office of Student Accessibility Services (OSAS) accommodation. Recording can inhibit free discussion in the future, and thus infringe on the academic freedom of other students as well as the instructor. (Living our Unifying Values: The USC Student Handbook, page 13).

Distribution or use of notes, recordings, exams, or other intellectual property, based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study. This includes but is not limited to providing materials for distribution by services publishing course materials. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relationship to the class, whether obtained in class, via email, on the internet, or via any other media. (Living our Unifying Values: The USC Student Handbook, page 13).

Course Evaluations

End-of-Semester Evaluation:
At the end of the semester, students will have the opportunity to provide feedback through a university-wide course evaluation. This feedback helps review their experience and provides valuable insights for instructors.

Mid-Semester Evaluation:
A mid-semester evaluation will be conducted to allow for early course correction. Students can provide feedback on their learning experience up to that point.

Both evaluations are important for continuous improvement and students are encouraged to actively participate.

Course Schedule

Table 3 Course schedule

	Topics/Daily Activities	Readings/Preparation	Deliverables
Week 1	Introduction and applications		
Week 2	An overview of ML algorithms and their computational demands, CNNs ML runtime systems (PyTorch, TensorFlow)		Programming/Homework assignment (PHA)#1 Posted
Week 3-5	Deep dive in DNN accelerators Kernel Computation. Toeplitz matrices Dataflows (weight, input and output stationary systems), loop nests		Programming/Homework assignment (PHA)#2 Posted
Week 6	Quantization		Reading Assignment 1 posted
Week 7	Sparsity in DNN Accelerators		Programming/Homework assignment (PHA)#3 Posted
Week 8	Basics of GPU and TPU microarchitecture. **Midterm Exam**		Reading Assignment 2 posted
Week 9	More on GPU and TPU microarchitecture.		Project topics posted
Week 10	Distributed Learning		PHA#4 posted, Project initial updates
Week 11	Privacy in Deep Learning		
Week 12	Introduction to NLP and Language Models		PHA#5 posted
Week 13	RNNs, LSTM, and Transformers, LLMs		Mid project updates
Week 14	Training and fine-tuning Transformers		PHA#6 posted
Week 15	Generative models		Final project updates
Week 16	Backup for overflow		
FINAL	**Final Exam**		Refer to the final exam schedule in the USC *Schedule of Classes* at classes.usc.edu.

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, comprises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

Other violations of academic integrity include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), collusion, knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Students and Disability Accommodations:

USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline is comprised of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services

(though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.