USC Viterbi
School of Engineering

Course ID and Title: [EE508, Hardware Foundations of Machine Learning]

Units: 4

Term—Day—Time: [Spring 2025] — [Lecture Saturday 12:30-4:10pm – Discussion: TBD]

Location: TBD

Instructor: Arash Saifhashemi

Office: TBD
Office Hours: TBD
Contact Info: saifhash@usc.edu

Teaching Assistant: TBD

Office: TBD
Office Hours: TBD
Contact Info: TBD

Catalogue Description

ML kernels: convolutions, transformers, embeddings. Accelerators: GPUs, input/weight/output stationary accelerators. Distributed ML: data, model and hybrid parallel. Private ML: homomorphic encryption and multi-party computing accelerators.

Course Description

This course presents a unique perspective to ECE (Electrical and Computer Engineering) students who are interested in building Machine Learning (ML) hardware and systems, such as graphics processing units (GPUs) and accelerators, and in designing scalable ML systems such as cloud based ML training and inference pipelines. This course introduces students to computations and memory access kernels that are commonly seen in ML models, including convolutions, transformers, and embedding tables. Students will learn how to transform convolutions to matrix operations and how to accelerate these matrix operations on hardware accelerators. It presents 3 different hardware design paradigms for ML accelerators: input, output and weight stationary accelerators. It provides an in-depth understanding of ML hardware accelerators in the market place, such as GPUs and Tensor Processing Units (TPUs). The course also presents how to scale ML systems using parallelization approaches such as model, data and hybrid parallelism. The course will enable students to understand the basics of privacy in machine learning and how to accelerate private ML systems using homomorphic encryption and multi-party computing.

Learning Objectives

By the end of this course, students will be able to

- Transform convolutional neural network pipelines into matrix computations, map these computations to GPUs, TPUs and ML accelerators and measure the per iteration training runtime.
- Build hardware accelerator blocks for matrix computations using input/output/weight stationary methods for ML inference and training.
- Quantify performance bottlenecks in ML systems through code instrumentation and system monitoring statistics

- Train large ML models using model or data parallel computations and map them to run on a collection of GPUs.
- Design privacy preserving ML models using multi-party computing algorithms and then map them to existing GPU hardware using PyTorch Crypten programming framework.
- Quantify the latency impact of using private ML and identify the execution bottlenecks that lead to privacy latency penalties.

Prerequisite(s): None.
Co-Requisite(s): None.
Concurrent Enrollment: None.
Recommended Preparation: Coding matrix algebraic algorithms, Understanding of gradient descent, PyTorch or Tensorflow programming experience, Memory systems of CPUs and GPUs.

Course Notes

Letter Grading on a 4 point scale.
Course will be offered as in-person course (including DEN based participation).
All lecture materials will be posted on the D2L Brightspaceand any student discussion will be moderated on the Piazza platform.
All the software programming exercises will be done through Github.

Technological Proficiency and Hardware/Software Required

All labs associated with GPUs will be done using USC's HPC CARC GPUs using cloud based virtual instance reservations. If CARC resources are busy we will also use Google Colab facilties that provide educational access to GPUs. A Colab and CARC usage tutorial will be posted online and a hands-on demonstration will be done during the discussion sections.

Required Readings and Supplementary Materials

REQUIRED:

- Sze, Chen, Yang and Emer: "Efficient Processing of DNNs," Morgan&Claypool Press. 2021. ISBN: 9781681738321
- Aggarwal, IBM: "Recommender Systems," Springer, 2016, https://doi.org/10.1007/978-3-319-29659-3_1
- PyTorch Tutorials and Manuals, https://pytorch.org/tutorials/
- A set of links to required papers is provided in the course schedule and paper will also be posted on the class website.

Optional Readings and Supplementary Materials

Recommended: Heidari: "Deep Learning for Robot Perception and Cognition: Chapter 4 GCNs," Elsevier, ISBN: 978-0-323-85787-1

Description of Assignments and How They Will Be Assessed

6 Homework/Programming Assignments.
2 Reading Assignments.
One written midterm
One written final exam.

Assignment#1 will focus on introducing students to Colab and CARC resources and provides a hands-on labs on accessing these resources through cloud.

Assignment#2 will focus on introducing students to CNN models by building an image classification system and demonstrating the convolutions as matrix operations that are accelerated on a GPU hardware. Students will implement convolutions as Toeplitz matrices and then do matrix tiling based acceleration on GPUs, first using Cuda cores and then using Tensor cores.

Assignment#3 will focus on measuring the performance bottlenecks of the CNN execution on GPUs. It will include measuring the GPU tensor and Cuda core utilization, memory bandwidth measurements and scalability bottlenecks. They will also use techniques like quantization and pruning to understand how the model size impacts the latency and model accuracy tradeoffs.

Assignment#4 will focus on implementing the CNN models using input, output and weight stationary approaches and comparing the performance across all three different approaches.

Assignment#5 will ask students to design machine learning pipelines.

Assignment#6 will focus on designing transformers.

Reading assignments will ask students to read and provide a summary report of their understanding on at least 4 ML hardware accelerator papers, such as Google TPU, DaDianNao accelerators.

Participation
No participation credit.

Grading Breakdown

Assessment Tool (assignments)	% of Grade
Assignments 1-6	48%
Midterm	20
Reading Assignment 1-2	10
Final	22%
TOTAL	100%

Grading Scale
Grading scale is unchanged.

Assignment Submission Policy
All assignments will be assigned through Piazza and the completed submissions will also be uploaded to the Piazza website.

Course-Specific Policies
No late assignment submission is allowed.

Attendance
Attendance is required.

Academic Integrity
USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. Scampus, the Student Guidebook, contains the

Student Conduct Code in Section 11.00, while the recommended sanctions are located in Appendix A: http://www.usc.edu/dept/publications/SCAMPUS/gov/. Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty. The Review process can be found at: http://www.usc.edu/student-affairs/SJACS/.

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

You may not record this class without the express permission of the instructor and all other students in the class. Distribution of any notes, recordings, exams, or other materials from a university class or lectures — other than for individual or class group study — is prohibited without the express permission of the instructor.

Use of Generative AI in this Course

Generative AI is encouraged: You are expected to use Generative AI (e.g., ChatGPT and image generation tools) in this class. Learning to use Generative AI is an emerging skill; this is an opportunity for you to discuss with the instructor appropriate use of these tools. Keep in mind the following:

- Generative AI tools are permitted to help you brainstorm topics or revise work you have already written.
- If you provide minimum-effort prompts, you will get low-quality results. You will need to refine your prompts to get good outcomes. This will take work.
- Proceed with caution when using Generative AI tools and do not assume the information provided is accurate or trustworthy If it gives you a number or fact: assume it is incorrect unless you either know the correct answer or can verify its accuracy with another source. You will be responsible for any errors or omissions provided by the tool. It works best for topics you understand.
- Generative AI is a tool, but one that you need to acknowledge using. Please *include a paragraph at the end of any assignment explaining if, how, and why you used AI and indicate/specify the prompts you used to obtain the results*. Failure to do so is a violation of academic integrity policies.

Course Evaluations

Course evaluation will follow the university-wide practice.

Course Schedule

	Topics	Preparation	Deliverables
Week1	Introduction to ML training and inference including topics such as loss computations using L2 norm, cross entropy, SGD and backpropagation algorithms.	Refresh Linear Algebra	Colab and CARC resource Usage
Week2	Deep neural networks based on convolutions such as ResNet, ResNext including topics such as vanishing gradients, skip connections etc.,	Introduction to PyTorch or Tensorflow tutorial online (PyTorch Tutorials: Learn the basics chapter)	HW 1: Building image classification CNN using PyTorch or Tensorflow on Colab/CARC
Week 3	Toeplitz matrices and their usage in CNNs	Review the ML training and inference background materials (Sze et al. Chap 1 and Chap 2)	Implement CNNs using Topelitz matrices on GPUs and CPUs
Week 4	Systolic arrays and design of input and output stationary ML accelerators using systolic computations	Review systolic arracy computing (Posted papers)	HW 2: Implement Matrix tiling and use cuda cores and tensor cores on GPUs to refine the image classification CNN
Week 5	Weight stationary accelerators including TPUs	Review DaDianNao paper (Posted paper from Micro 2014) by the same title)	Measure performance bottlenecks using NVProf and PyTorch profiling framework
Week 6	Row stationary accelerators such as Eyeriss	Review Google TPU (Posted from ISCA 2019 paper by the same title)	HW 3: Implement CNN using weight, input and output stationary schemes. Compare the performance of the 3 schemes.
Week 7	Model reduction schemes such as pruning and quantization	Review Eyeriss papers (Sze et al. Chap 5 and 6)	Midterm#1 Concept and Sample Exam Review
Week 8	Recommender systems in industry including concepts such as embedding tables, two-tower networks	Midterm preparation	Midterm#1
Week 9	Distributed ML with data parallel training using parameter servers and ring-all-reduce	Review DLRM papers from Meta (Posted paper also available at https://arxiv.org/abs/1906.00091)	Run the open source DLRM on GPUs and measure their memory bandwidth and execution unit utilization demands for embedding accesses and MLPs
Week 10	Distributed ML using model parallelism and hybrid parallelism	Review the NCCL library from Nvidia (From 2017 GTC conference and NCCL Manual Chapter 3 NCCL API, available https://docs.nvidia.com/deeplearning/nccl/user-guide/docs/index.html	HW 4: Distributed training of CNNs with parameter servers and NCCL communication library
Week 11	Transformer computational kernels and their use in NLP	Review of OpenAI ChatGPT paper (https://arxiv.org/abs/2303.08774)	Explore scalability of distributed training with data parallelism
Week 12	Transformer parallelism	Review of the Attention network paper (From NeurIPS 2017 paper with the same title)	Explore scalability of distributed training with model parallelization
Week 13	Machine learning pipelines		HW 5: Transformer design
Week 14	Machine learning pipelines		HW 6: Machine learning pipelines.
Week 15	Private ML using federated learning	Review of Federated Learning paper from Google	

		(https://www.nowpublishers.com/article/Details/MAL-083)	
FINAL			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.

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).

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.