USC
Viterbi
School of Engineering

ITP 450 Advanced Computing in Applied Machine Learning
Units: 2
Term: Fall
Day/time: 50 minutes each on Tuesdays and Thursdays
Location: KAP 267

Office Hours: TBD

Instructor: Byoung-Do Kim, Hao Ji, Iman Rahbari
Contact Info:
Byoung-Do (BD) Kim, **bdkim@usc.edu**
Hao Ji, **haoji@usc.edu**
Iman Rahbari, **irahbari@usc.edu**

Contact Info: carc-support@usc.edu

IT Help: Viterbi IT, CARC Support

Hours of Service:
Monday – Friday, 9:00 a.m. – 5:00 p.m.
Contact Info:
Viterbi IT: DRB 205 (213) 740-0517 engrhelp@usc.edu
CARC Support: carc-support@usc.edu

Last updated: 8/1/24

Course Description

Through lectures, interactive hands-on sessions, and a team project, students will learn how to apply available tools and technologies in advanced computing and deep learning to solve science and engineering problems while working with modern high-performance computing systems.

Students will work on one team project to complete by the end of the semester. For the team project, students will use state-of-the-art computing resources at USC's Center for Advanced Research Computing (CARC) to analyze real-world datasets using the techniques discussed in class, provide insights on the datasets, and present their descriptive or predictive models to the class.

Catalogue Description

Theoretical and practical approaches of various computational methods in high performance computing in applied machine learning.

Learning Objectives

Through attending lectures and engaging in the interactive hands-on sessions, the students will develop a strong foundation in advanced computing tools and learn to apply them in data science and deep learning problems". Upon successful completion of this course, students will demonstrate the ability to apply modern tools and techniques in data analysis and deep learning to solve science and engineering problems, while effectively using the advanced high-performance computing systems at USC. More specifically, the students will be able to:

- Explain the hardware and software components of an HPC cluster and how to use them effectively

- Compare CPU and GPU and identify which one would be a more suitable choice in a given scenario

- Develop codes in Python, using CUDA-enabled libraries, to perform data analysis and machine learning tasks on GPUs

- Explain fundamental concepts in "deep" learning and implement those in python scripts, using appropriate libraries such as PyTorch, using HPC systems to solve real-world science and engineering problems

- Demonstrate an understanding of the cloud infrastructure (in AWS), identify the services required to perform data analysis and deep learning tasks and utilize them to address practical science and engineering problems

Intended Audience

This course is intended for students who *already have a foundational understanding of python programming and data analysis* and are interested in using high-performance computing to solve real-world problems such as autonomous driving, image classification, predictive maintenance, as well as for their research using machine learning with GPU on large scale datasets.

Prerequisite(s): ITP 449 or DSCI 352 or MATH 446 or BUAD 425 or CSCI 467
Recommended Preparation:

Course Notes

1. The lecture format will vary with the topic and will include slides posted prior to class and computational codes as appropriate. Lectures will cover the fundamentals of computational methods essential to understanding the data science approaches.
2. Given the focus of this class on the computational process of data analysis, discussion during lectures and the following hands-on workshops will use example published data; relevant information and the data will be made available before the class.
3. Hands-on workshops will expose students to all steps of data handling, with emphasis on understanding the computational methods, production of basic numeric and graphical outputs, and data interpretation.

Technological Proficiency and Hardware/Software Required:

- Basic Python programming skills
- Students are expected to have their own laptop
- Data science SW stack is available on the High-Performance Computing system offered by the Center for Advanced Research Computing (CARC)s

Recommended Readings

Book Chapters:

- Sterling, T., Brodowicz, M., & Anderson, M. (2017). High performance computing: modern systems and practices. Morgan Kaufmann.
  - Chapter 1 (sections: 1.1, 1.2, 1.3), Chapter 2 (sections: 2.1, 2.2, 2.3, 2.7, 2.8, and 2.9) Chapter 3
- Bandyopadhyay, A. (2019). Hands-On GPU Computing with Python: Explore the capabilities of GPUs for solving high performance computational problems. Packt Publishing Ltd.
  - Chapters 4, 6, and 8
- Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.
  - Chapters 10 and 14
- Raschka, S., Liu, Y. H., Mirjalili, V., & Dzhulgakov, D. (2022). Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python. Packt Publishing Ltd.
  - Chapter 19
- Dabravolski, V. (2022). Accelerate Deep Learning Workloads with Amazon SageMaker. Packt Publishing Ltd.
  - Chapter 1

All these books are available online through USC Libraries.

Online Resources:

- https://www.carc.usc.edu/user-guides
- https://www.carc.usc.edu/user-guides/data-science

Description and Assessment of Assignments, Grading Breakdown

Assignment	**% of grade**
Assignments	**40**
Midterm	**30**
Team Project	**30**
TOTAL	**100**

Team Project:
During week 13, the students will be given a team project on the image classification to work within the groups of 3-4 people. Each team can choose different tasks for their project depending on their interest and expertise. The key deliverables are:

1. Team project presentation (10%): The students are required to give a presentation, in week 15, that includes: problem definition, project timeline, key analysis and findings, and final conclusion. Each team will have a 20-minute allocated time including a 15-minute presentation and a 5-minute Q&A session. The group will decide how to break up the workload, so long as there is a fair contribution among all members.

2. Team project final report (20%): A 6-10 pages team project technical report should be included as a substitute for the final exam. The formatting should be according to the NeurIPS conference papers without the appendices (Latext files are https://www.overleaf.com/ latex/ templates/ neurips-2024/). The report should be written together with all team members and cover important deliverables required by the team project. It should follow a similar structure of the presentation and include problem definition, project timeline, important figures and plots with analysis, and a summary of key insights from the project. Students will also need to discuss lessons learned during the team project and what additional work could be done in the report.

Assignments:
There will be a total of 5 assignments, at the end of Week 2, 5, 7, 11, and 12.

Exams:
There will be one midterm exam. The team project will replace the final exam.

Assignment Submission Policy
Assignments have to be submitted by Friday 5pm of the following week.

Additional Policies
Late assignment submissions will receive a 15% penalty in grading. No assignment will be accepted 2 weeks after the deadline. No late submission of the team project will be accepted. Students are expected to attend all classes.

Course Schedule: A Weekly Breakdown

Week 1 & 2: Advanced Computing in Data Science

This section provides an overview of High-Performance Computing system architecture, with a focus on USC-CARC resources, and introduces the main tools to interact with this system including terminals, job scheduling system, and data transfer platforms. This is followed by a gentle introduction to parallel computing and its role in machine learning. The students will also learn how to setup their deep learning environment on modern HPC systems.

- The components of the HPC cluster: compute nodes, interconnects, storage, scheduler, etc.
- Parallel computing concepts
- A brief history of Deep Learning, Why HPC+DL is important
- Connecting to the cluster (VPN, USC Secure Wireless) via OnDemand
- Setting up the Deep Learning environment on an HPC cluster: Conda and Pytorch
- Start working on JupyterLab

Week 3 -5: Introduction to Deep Learning & Neural Networks

This section teaches the fundamental concepts of deep learning and neural networks. Students will use popular deep learning software packages such as Pytorch/Keras to build simple neural networks and solve common machine learning problems. An example of deep learning image classification problem will be introduced and sample codes will be taught. Students will then practice how to run such deep learning algorithms on the CARC cluster and compare training results with different types of GPU architectures.

- Learn basics of neural networks
- How to use software packages like Pytorch to build deep neural networks
- Applying Deep Learning to Image Classification tasks
- Use transfer learning or fine tuning to train deep learning tasks
- Use tensorboard to visualize datasets and training results

Week 6 & 7: Advanced Computing with GPU

Graphic Processing Units have been proven to outperform the CPUs when it comes to many Machine Learning algorithms. This section begins with introducing the GPU architecture and why/how it is able to deliver superior performance in data-intensive applications. We then overview the main steps to develop and run a simple program on a GPU in a low-level language (like C). In hands-on practice, we use CuPy to write this simple program in Python and compare the results with the CPU.

In the next part, we introduce the Nvidia RAPIDS, focusing on CuDF and CuML, and use these tools for GPU-accelerated data preparation (similar to Panda) and classical Machine-Learning techniques and compare the performance against running the code on the CPUs.

- GPU Architecture
- Steps to develop and run a simple program on a GPU in a low-level language
- Hands-on practice to write a simple program in Python via CuPy
- Introducing Nvidia RAPIDS, focusing on CuDF and CuML
- Data preparation and Classical Machine Learning with GPU and comparing the performance against CPU

Week 8: Midterm Exam (10/15) and Team Project Introduction (10/17) and Advanced Computing with GPU continued (10/17)

Week 9-10: Training Deep Learning Models with Multiple GPUs
This section discusses both data and model parallelism techniques to enhance the efficiency of deep learning tasks. We explore how data parallelism involves distributing the dataset across multiple devices or processors, allowing for concurrent processing of subsets of data. Additionally, we introduce model parallelism, which partitions the model across devices, enabling the computation of different model segments simultaneously.

- **Distributed Data Parallelism**
  - Introduce PyTorch distributed data-parallel algorithms and how it can be applied to multi gpu training with hands-on exercises
- **Model Parallelism**
  - Introduce PyTorch model-parallel algorithms and how it can be applied to multi gpu training with hands-on exercises

Week 11: Transformer and Large Language Model
This week, the focus will be on the basic understanding the transformer architecture and the attention mechanism. We will then delve into the computational costs associated with training and fine-tuning Large Language Models.

- Attention and Transformer architecture
- Introduction to Large Language Models
- Training and fine-tuning costs of Large Language Models

Week 12: Data Science on the Cloud
In this section, we overview the basic concepts in Cloud Computing such as regions, zones, VPC, and subnets. Different machine types, as well as storage options, are briefly described and the possible use cases for each type are discussed. Special attention is paid to the Machine Learning (and Deep Learning) Accelerators available on AWS. In hands-on practice, we use the AWS SageMaker to prepare the data and train a deep learning model on the cloud.

- Basics of Cloud Computing (VMs, Storage, Accelerators)
- Hands-on practice using AWS SageMaker to prepare the data and train a Deep Learning Model on Cloud

Week 13-15: Team Project & Invited Lecture
- Last week of the class includes a guest lecture on the “large-scale AI computations”.
- Final presentation is due on the last class
- The final report for the team project is due on the date of the final exam.

Week	Topics	Readings	Assignments
1-2	Advanced computing in Data Science	Sterling, T., Brodowicz, M., & Anderson, M. (2017). *High performance computing: modern systems and practices*. Morgan Kaufmann. Chapter 1 (sections: 1.1, 1.2, 1.3), Chapter 2 (sections: 2.1, 2.2, 2.3, 2.7, 2.8, and 2.9) Chapter 3 https://www.carc.usc.edu/user-information/user-guides	Example code runs and Slurm job submission
3-5	Introduction to Deep Learning and Neural networks	Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, Inc. Chapters 10 and 14	Building neural networks with Keras/Pytorch and solve an image classification problem
6-7	GPU Computing	Bandyopadhyay, A. (2019). *Hands-On GPU Computing with Python: Explore the capabilities of GPUs for solving high performance computational problems*. Packt Publishing Ltd. Chapters 4, 6, and 8	Using CUDA for data processing and comparing the performance against CPU
8	**Midterm Exam (Tuesday 10/15)** **Team Project Introduction (Thur)**		Project introduction, team formation
9-10	Training Deep Learning Models with Multiple GPUs	PyTorch distributed training tutorials. https://pytorch.org/tutorials/distributed/home.html	
11	Transformer and Large Language Model	Vaswani et al. (2017). *Attention is all you need.* https://proceedings.neurips.cc/p	Conceptual problems regarding Transformer architecture.

		aper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf Brown et al. (2020). *Language Models are Few-Shot Learners.* https://arxiv.org/pdf/2005.14165.pdf	
12	Data Science on the Cloud	Dabravolski, V. (2022). *Accelerate Deep Learning Workloads with Amazon SageMaker.* Packt Publishing Ltd. Chapter 1	Train and test a Deep Learning model on AWS SageMaker
13	Team project work, office hours	Teams discuss the progress of their projects and instructors will be available for assistance.	
14	Team project Thanksgiving (11/28)	Project development status report and discussion	
15	Project Presentation	Students will be required to do about a 20-minute presentation, which is due in place of the final exam. The class anticipates about 15–20 students in the course. Each group will consist of 3–4 students, so there will likely be about four groups, and each will have about 20 minutes to present in the final week. (If enrollment is larger, the final exam period will also be used for group presentations to ensure sufficient time for each group to present its work.)	
FINAL	Report due	The project report is due on the same date as the final exam for this class. The students should consult the USC Schedule of Classes at classes.usc.edu/.	

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct

Sharing of course materials outside of the learning environment

SCampus Section 11.12(B)

Distribution or use of notes or recordings based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study is a violation of the USC Student Conduct Code. This includes, but is not limited to, providing materials for distribution by services publishing class notes. 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. (See Section C.1 Class Notes Policy).

Support Systems:

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

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED) - (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
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
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility and Services - (213) 740-0776
osas.usc.edu

Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Campus Support and Intervention - (213) 821-4710
campussupport.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu
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
dps.usc.edu, emergency.usc.edu
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) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.