COS598D: Systems and Machine Learning
(Spring 2024)

Summary

In this graduate seminar, we plan to read and discuss recent research papers at the intersection between systems and machine learning (ML). The first category includes papers about building efficient hardware and software for ML such as parallel, distributed and privacy-preserving systems. The second category includes papers about how to apply ML in designing hardware and software systems, including new data structures and optimization methods. The class will have numerous guest speakers to give lectures.

This course is open to graduate students. For seniors who are interested in taking this course, it requires permission from the instructors. This course has no P/D/F option. All students are required to present papers and participate in discussions, and complete a programming assignment and a final project. This course satisfies the systems programming requirement for graduate students.

Administrative Information

- Classes Times: Friday 1:30 – 4:20 pm
- Classroom: Friends center 008
- Instructor: Professor Kai Li (li@cs.princeton.edu)
- Teaching assistants: Rui Pan (ruipan@princeton.edu) and Dongsheng Yang (dy5@princeton.edu)
- Office hours: appointments by email
- Online discussions and Q&A: Ed Stem

Requirements and Grading

Students taking this graduate seminar will be graded as follows:

- Presentation (15%)
  Students are required to work in a small team to give a presentation of a paper / topic (see syllabus below), and lead the discussion of the questions. Each presentation should take about 40 minutes, leaving 20 minutes for discussions. The presenters should meet the instructor one week before the presentation to finalize on the discussion questions.
- Reading and participation (15%)
  Students are required to read the main paper for each topic, submit a paragraph to answer the questions posted (not graded for correctness), and participate in the discussions in class.
- Programming Assignment (20%)
  We will have one small programming assignment with three options:
  - Network pruning,
  - Adaptive learned Bloom filter, and
  - Distributed training of a language model.
- Final projects (50%)
  There are two options: (1) extending one of the programming assignments (**work alone**) or (2) a new project (**work alone or in a two-student team**). For example, you can extend the explorations or evaluations of a recently published paper. For the two-student team, we expect your project to be significant such that each student can play different roles in the project.
  Students are required to
  - Submit a brief project proposal (what you plan to do) in one slide.
  - Submit a brief progress report in one slide.
  - Submit a final project report in the format of a conference paper, to concisely describe what you have done and your results. For a two-student team (option 2), the final report should state clearly who did what.
  - ~~Give a 10-min presentation for option 1 or a 15-min presentation for option 2.~~ The final reports are due at 11:59pm of the Dean's day, which is the deadline of the university. ~~The final presentations will be scheduled after the Dean's day.~~

Warmup

The goal of this warmup is to get familiar with ML frameworks with computing resources. Since some students may have done programming in Pytorch, this assignment is not required and will not be graded. We **strongly encourage** all students without experience to complete the MNIST classification warmup exercise. A related paper is Gradient-Based Learning Applied to Document Recognition. 1998.

MNIST dataset will be automatically downloaded if using MNIST classification. We have also set up the same on Google Colab:
https://colab.research.google.com/drive/1wUWfzIY81x7R6Xe8WZUkgzFoKmTws7co?usp=sharing

Computing Resources

Adroit Cluster: Princeton Research Computing Cluster for development
Google Colab: Free compute resources
Amazon AWS: Some free student credits
Microsoft Azure: Some free student credits

Tentative Syllabus

Dates	Presenters	Topics & Main Papers	Related Papers	Events
2/2	Kai Li	Introduction to COS 598D (slides)	A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution.	
	Dr. Jeff Dean & Dr. Amin Vahdat (Google)	Exciting directions for ML models and implications for computing hardware (Keynote at Hotchip 2023, Slides)		Start warmup
2/9	Prof. Jia Deng (Princeton)	Introduction to deep learning	Deep Learning. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, Nature 521, May 2015.	Start programming assignment
	Rui Pan (TA)	Introduction to ML frameworks and programming assignments	PyTorch: An Imperative Style, High-Performance Deep Learning Library. TensorFlow: A System for Large-Scale Machine Learning. OSDI 2016	
2/16	Kai Li	Intro to parallel/distributed training	TensorFlow: A System for Large-Scale Machine Learning. OSDI 2016 Large Scale Distributed Deep Networks. NIPS 2012. PipeDream: Generalized Pipeline Parallelism for DNN Training, SOSP 2019.	
	Dr. Bill Dally (Nvidia)	Hardware for Deep Learning (Keynote at Hotchip 2023, slides)	Domain specific hardware accelerators. CACM 2020. Vs-Quant: Per-Vector Scaled Quantization For Accurate Low-Precision Neural Network Inference. MLSys 2021. A 95.6-TOPS/W Deep Learning Inference Accelerator With Per-Vector Scaled 4-bit Quantization in 5 nm. IEEE JSSC, 2023.	
2/23	Team 1 Boyi Wei Minhao Jin Zixuan Wang	Network pruning The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks, ICLR 2019. Structured Pruning Learns Compact and Accurate Models, ACL 2022.	Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, ICLR 2015. Learning structured sparsity in deep neural networks, NeurIPS 2016.	Discussion questions
	Team 2 Yihan Wang Chenyu Wang Yihao Liang Wenzhe Li	Distillation **Knowledge distillation**: A good teacher is **patient** and consistent, CVPR 2022. LLM-Pruner: On the Structural Pruning of Large Language Models, NeurIPS 2023	Distilling the knowledge in a neural network, 2015	
2/25				Programming Assignment Due, 11:55pm
3/1	Prof. Danqi Chen, Tianyu Gao, Alexander Wettig (slides)	Transformers: Attention is all you need BERT: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding GPT-3: Language Models are Few-Shot Learners InstructGPT (instruction-tuning, alignment): Training language	T5: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer Scaling law: Scaling Laws for Neural Language Models Llama and CodeLlama (open-source models): LLaMA: Open and Efficient Foundation Language Models, Code Llama: Open Foundation Models for Code	Signup to discuss final projects by 11:59 PM on Sunday (Mar 3)

Dates	Presenters	Topics & Main Papers	Related Papers	Events
		models to follow instructions with human feedback Long context model: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context	DPO (alignment): Direct Preference Optimization: Your Language Model is Secretly a Reward Model	
3/8	~~Dr. Safeen Huda (Google)~~ (confirmed)	Google's TPU A Machine Learning Supercomputer With An Optically Reconfigurable Interconnect and Embeddings Support. Talk at Hotchips 2023. Tentative A full-stack search technique for domain optimized deep learning accelerators, ASPLOS 2022. Learning Semantic Representations to Verify Hardware Designs, NeurIPS, 2021.	TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings. ISCA 2023.	Submit final project proposals by the beginning of the class
3/22	Prof.. Jonathan Frankle, MosaicML and Harvard.	TBD		
3/29	Team 1 Ted Zadouri Amey Pasarkar Meenal Parakh David Yan Team 2 Andrew Sheinberg Tyler Zhu Jipeng Sun Haoran Wan	Large-scale training Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, 2022. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models, SC 2020. ========== Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism Pathways: Asynchronous Distributed Dataflow for ML. MLSys, 2022. PaLM: Scaling Language Modeling with Pathways, 2022.		Discussion questions Submit progress report by the beginning of the class
4/5	Prof. Song Han (MIT) **Emma Farkash** **Anika Maskara** **Austin Li** **Aditya Bhaskar**	Efficient ML SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. ICML 2023. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. 2023. Efficient Streaming Language Models with Attention Sinks. ICLR 2024. ========== EfficientViT: Multi-Scale Linear Attention for High-Resolution Dense Prediction. 2022. PockEngine: Sparse and efficient fine-tuning in a pocket, enable learning at the edge. IEEE MICRO, 2023.		Discussion questions
4/12	~~Dr. Yida Wang (AWS)~~	TBD		

Dates	Presenters	Topics & Main Papers	Related Papers	Events
	Stamatis Alexandropoulos Erich Liang Constantine Doumanidis Fengchen Gong Justin Wang **[1 student needs to move]**	GPT4 Technical Report, 2023. Gemini: A Family of Highly Capable Multimodal Models. 2023.	Towards Expert-Level Medical Question Answering with Large Language Models, 2023. LLaMA: Open and Efficient Foundation Language Models, 2023.	Discussion questions
4/19	Dr. Michael Cox (Nvidia) Confirmed Jiayi Geng Shuyao Zhou Jiahao Qiu Niusha Moshrefi Gudni Nathan	AI Safety Dataset Distillation/Condensation Dataset Distillation by Matching Training Trajectories. CVPR 2022. Dataset Condensation with Distribution Matching. CVPR 2023. You Only Condense Once: Two Rules for Pruning Condensed Datasets, NeurIPS 2023.	Dataset Distillation. 2018. Dataset Condensation With Gradient Matching. ICLR. 2021.	Discussion questions
4/26	Dr. Yida Wang (AWS) (rescheduled) Team 1: Jack Monas Yulai Zhao Olga Solodova Devan Shah Team 2: Katie Lee Mingqian Xue Xinyuan Li Dan Hu	Learned Data Structures ALEX: an updatable adaptive learned index, SIGMOD 2020. Can Learned Models Replace Hash Functions?, VLDB 2022. ML for Database systems Sagedb: A learned database system CIDR. 2019. SageDB: An Instance-Optimized Data Analytics System, VLDB 2022.	The case for learned index structures, SIGMOD 2018.	Discussion questions
~~5/3~~ 5/6	Dr. Safeen Huda (Google) 1:30-3 PM			
5/6	Posters Session 3-4:30 PM			
5/7 Dean's Day				Submit final project Reports Due 11:59pm

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023.

A Watermark for Large Language Models. ICML 2023.