Syllabus for CS6787

Advanced Machine Learning Systems — Spring 2024

Term	Spring 2024	**Instructor**	Christopher De Sa
Room	Phillips Hall 101	**E-mail**	[email hidden]
Schedule	MW 7:30pm – 8:45pm	**Office hours**	W 2:30pm – 3:30pm
Forum	Ed Discussion	**Office**	Gates 426

So you've taken a machine learning class. You know the models people use to solve their problems. You know the algorithms they use for learning. You know how to evaluate the quality of their solutions.

But when we look at a large-scale machine learning application that is deployed in practice, it's not always exactly what you learned in class. Sure, the basic models, the basic algorithms are all there. But they're modified a bit, in a bunch of different ways, to run faster and more efficiently. And these modifications are really important—they often are what make the system tractable to run on the data it needs to process.

CS6787 is a graduate-level introduction to these system-focused aspects of machine learning, covering guiding principles and commonly used techniques for scaling up learning to large data sets. Informally, we will cover the techniques that lie between a standard machine learning course and an efficient systems implementation: both statistical/optimization techniques based on improving the convergence rate of learning algorithms and techniques that improve performance by leveraging the capabilities of the underlying hardware. Topics will include stochastic gradient descent, acceleration, variance reduction, methods for choosing hyperparameters, parallelization within a chip and across a cluster, popular ML frameworks, and innovations in hardware architectures. An open-ended project in which students apply these techniques is a major part of the course.

Prerequisites: Knowledge of machine learning at the level of CS4780. *If you are an undergraduate, you should have taken CS4780 or an equivalent course, since it is a prerequisite.* Knowledge of computer systems and hardware on the level of CS 3410 is recommended, but this is not a prerequisite.

Format: About half of the classes will involve traditionally formatted lectures. For the other half of the classes, we will read and discuss two seminal papers relevant to the course topic. These classes will involve presentations by groups of students of the paper contents (each student will sign up in a group to present one paper for 15-20 minutes) followed by breakout discussions about the material. Historically, the lectures have occurred on Mondays and the discussions have occurred on Wednesdays, but due to the non-standard timeline this semester, these course elements will be scheduled irregularly (see schedule below).

Grading: Students will be evaluated on the following basis.

20%	Paper presentation
10%	Discussion participation
20%	Paper reviews
10%	Programming assignments
40%	Final project

Paper review parameters: Paper reviews should be about one page (single-spaced) in length. The review guidelines should mirror what an actual conference review would look like (although you needn't assign scores or anything like that). In particular you should at least: (1) summarize the paper, (2) discuss the paper's strengths and weaknesses, and (3) discuss the paper's impact. For reference, you can read the ICML reviewer guidelines. Of course, your review will not be precisely like a real review, in large part because we already know the impact of these papers. You can submit any review up to two days late with no penalty. Students who presented a paper do not have to submit a review of that paper (although you can if you want).

Final project parameters (subject to change): The final project can be done in groups of up to three (although more work will be expected from groups with more people). The subject of the project is open-ended, but it must include:

- the **implementation of a machine learning system** for some task,
- exploring one or more of the **techniques discussed in the course** (or similar techniques subject to instructor approval),
- to **empirically evaluate the performance** and compare it with some baseline method, in two ways:
  - statistical performance (e.g. iterations to converge to some accuracy threshold), and
  - hardware performance (e.g. throughput or wall-clock time).

The project proposal should satisfy the following constraints:

- The main body should be about one page in length.
- It should describe the project you intend to do.
- It should contain at least one citation of a relevant paper that we did not cover in class (but preferably more).

- It should include some preliminary or exploratory work you've already done, that helps to support the idea that your project is feasible (this preliminary work can be very minimal, but should indicate that you've got started—or at least have a clear idea how to do so).
- In addition to the one-page text proposal, it should contain one short **experiment plan** per person, which should consist of:
    - a hypothesis
    - a proxy statement which describes what metric you are going to use to measure the variables you care about
    - a short protocol statement describing what you are going to do
    - the results you expect to get

  The experiment plan should not be longer than half a page, and may be much shorter.

The project will culminate in a project report of at least four pages, not including references. The project report should be formatted similarly to a workshop paper, and should use the [ICML 2019 style] or a similar style. The project proposal is due on **Monday, March 25, 2024**. A draft of the final abstract is due for presentation and discussion in class on **Monday, April 29, 2024**. Per the registrar, the final project report is due on **May 15, 2024 at 4:30 PM**.

Course Calendar

Monday, January 22
In Person
Jan 21 | Jan 22 | Jan 23 | Jan 24 | Jan 25 | Jan 26 | Jan 27

Lecture #1: Overview.
[Slides] [Demo Notebook]

- Overview
- Course outline and syllabus
- Learning with gradient descent
- Stochastic gradient descent: the workhorse of machine learning
- Theory of SGD for convex objectives: our first look at trade-offs

Wednesday, January 24
In Person
Jan 21 | Jan 22 | Jan 23 | Jan 24 | Jan 25 | Jan 26 | Jan 27

Lecture #2: Backpropagation & ML Frameworks.
[Slides] [Demo Notebook]

- Backpropagation and automatic differentiation
- Machine learning frameworks I: the user interface
- Overfitting
- Generalization error
- Early stopping

Optional extra reading. Some older papers on SGD and backpropagation!

- Hinton, Geoffrey E. Learning distributed representations of concepts. Proceedings of the eighth annual conference of the cognitive science society. Vol. 1. 1986.
- Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. Cognitive modeling 5.3 (1988): 1.
- Tong Zheng. Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

Monday, January 29
In Person
Jan 28 | Jan 29 | Jan 30 | Jan 31 | Feb 1 | Feb 2 | Feb 3

Lecture #3: Hyperparameters and Tradeoffs.
[Slides] [Demo Notebook]

- Our first hyperparameters: step size/learning rate, minibatch size
- Regularization
- Application-specific forms of regularization
- The condition number
- Momentum and acceleration
- Momentum for quadratic optimization
- Momentum for convex optimization

Released: Programming Assignment 1.

Wednesday, January 31
In Person
Jan 28 | Jan 29 | Jan 30 | Jan 31 | Feb 1 | Feb 2 | Feb 3

Paper Discussion 1a.
Attention is all you need
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin.
In Advances in neural information processing systems (NeurIPS), 2017.

Paper Discussion 1b.
Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.
Sergey Ioffe, Christian Szegedy.
Proceedings of the International Conference on Machine Learning (ICML), 2015.

Monday, February 5
In Person
Feb 4 | Feb 5 | Feb 6 | Feb 7 | Feb 8 | Feb 9 | Feb 10

Lecture #4: Kernels and Dimensionality Reduction.
[Slides] [Demo Notebook]

- The kernel trick
- Gram matrix versus feature extraction: systems tradeoffs
- Adaptive/data-dependent feature mappings
- Dimensionality reduction

Wednesday, February 7 **In Person**	**Paper Discussion 2a.** [Palm: Scaling language modeling with pathways.] Aakanksha Chowdhery et al. *Journal of Machine Learning Research (JMLR)*, 2023. **Paper Discussion 2b.** [Language models are few-shot learners.] Tom Brown et al. *In Advances in neural information processing systems (NeurIPS)*, 2020. **Due: Review of paper 1a or 1b.**
Monday, February 12 **In Person**	**Lecture #5: Adaptive Methods & Non-Convex Optimization.** [Slides] [Demo Notebook] • Adaptive methods • AdaGrad • Adam • Non-convex optimization **Due: Programming Assignment 1.**
Wednesday, February 14 **In Person**	**Paper Discussion 3a.** [Random features for large-scale kernel machines.] Ali Rahimi and Benjamin Recht. *In Advances in Neural Information Processing Systems (NeurIPS)*, 2007. **Paper Discussion 3b.** [Feature Hashing for Large Scale Multitask Learning.] Kilian Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford and Alex Smola. *Proceedings of the International Conference on Machine Learning (ICML)*, 2009. **Released: Programming Assignment 2.**
Monday, February 19 **Online Only**	**Lecture #6: Hyperparameter Optimization.** [Slides] [Demo Notebook] • Hyperparameter optimization • Assigning parameters from folklore • Random search over parameters
Wednesday, February 21 **In Person**	**Paper Discussion 4a.** [Random shuffling beats sgd after finite epochs.] Jeff Haochen and Suvrit Sra. *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. **Paper Discussion 4b.** [Adam: A method for stochastic optimization.] Diederik Kingma and Jimmy Ba. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. **Due: Review of paper 3a or 3b.**
Monday, February 26	**February Break: No classes.**
Wednesday, February 28 **In Person**	**Paper Discussion 5a.** [Random search for hyper-parameter optimization.] James Bergstra and Yoshua Bengio. *Journal of Machine Learning Research (JMLR)*, 2012. **Paper Discussion 5b.** [Scaling laws for neural language models.] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. *arXiv preprint arXiv:2001.08361*, 2020.
Monday, March 4 **In Person**	**Lecture #7: Parallelism.** [Slides] [Demo Notebook] • Hardware trends that lead to parallelism • Sources of parallelism in hardware • Data parallelism • Extracting parallelism at different places in the computation • Simple parallelism on multicore **Due: Programming Assignment 2.**

Wednesday, March 6
In Person
Mar 3 4 5 6 7 8 9

Paper Discussion 6a.
[Map-reduce for machine learning on multicore.]
Cheng-Tao Chu, Sang K Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Andrew Y. Ng, and Kunle Olukotun
In Advances in Neural Information Processing Systems (NeurIPS), 2007.

Paper Discussion 6b.
[Hogwild: A lock-free approach to parallelizing stochastic gradient descent.]
Feng Niu, Benjamin Recht, Christopher Re, and Stephen Wright.
In Advances in Neural Information Processing Systems (NeurIPS), 2011.

Monday, March 11
In Person
Mar 10 11 12 13 14 15 16

Lecture #8: Distributed Learning.
[Slides]

- Learning on multiple machines
- SGD with all-reduce
- The parameter server
- Asynchronous parallelism on multiple machines
- Decentralized and local SGD
- Model and pipeline parallelism

Due: Review of paper 5a or 5b.

Wednesday, March 13
In Person
Mar 10 11 12 13 14 15 16

Paper Discussion 7a.
[Flashattention: Fast and memory-efficient exact attention with io-awareness.]
Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré.
In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Paper Discussion 7b.
[A System for Massively Parallel Hyperparameter Tuning.]
Liam Li et al.
Proceedings of the 2nd Conference on Machine Learning and Systems (MLSys), 2020.

Monday, March 18
In Person
Mar 17 18 19 20 21 22 23

Lecture #9: Low-Precision Arithmetic.
[Slides]

- Memory
- Low-precision formats
- Floating-point machine epsilon
- Low-precision training
- Scan order

Due: Review of paper 6a or 6b.

In-class project feedback activity.

Wednesday, March 20
In Person
Mar 17 18 19 20 21 22 23

Paper Discussion 8a.
[Large scale distributed deep networks.]
Jeff Dean et al.
In Advances in Neural Information Processing Systems (NeurIPS), 2012.

Paper Discussion 8b.
[Towards federated learning at scale: System design.]
Keith Bonawitz et al.
In Proceedings of the 2nd MLSys Conference (MLSys), 2019.

Monday, March 25
In Person
Mar 24 25 26 27 28 29 30

Lecture #10: Inference and Compression.
[Demo Notebook]

- Efficient inference
- Metrics we care about when inferring
- Compression
- Fine-tuning
- Hardware for inference

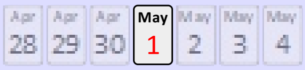
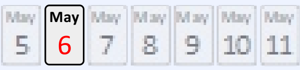
Due: Review of paper 7a or 7b.

Due: Final project proposals.

Wednesday, March 27
In Person
Mar 24 25 26 27 28 29 30

Paper Discussion 9a.
[Gpipe: Efficient training of giant neural networks using pipeline parallelism.]
Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, and Yonghui Wu.
In Advances in Neural Information Processing Systems (NeurIPS), 2019.

	Paper Discussion 9b. [Efficiently scaling transformer inference.] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. *In Proceedings of Machine Learning and Systems (MLSys)*, 2023.
Monday, April 1	**Spring Break: No classes.**
Wednesday, April 3	**Spring Break: No classes.**
Monday, April 8 **In Person**	**Lecture #11: Machine Learning Frameworks II.** • Large scale numerical linear algebra • Eager vs lazy • ML frameworks in Python **Due: Review of paper 8a or 8b.**
Wednesday, April 10 **In Person**	**Paper Discussion 10a.** [Deep learning with limited numerical precision.] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. **Paper Discussion 10b.** [LoRA: Low-Rank Adaptation of Large Language Models.] Edward J. Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
Monday, April 15 **In Person**	**Lecture #12: Hardware for Machine Learning.** • CPUs vs GPUs • What makes for good ML hardware? • How can hardware help with ML? • What does modern ML hardware look like? **Due: Review of paper 9a or 9b.**
Wednesday, April 17 **In Person**	**Paper Discussion 11a.** [Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding.] Song Han, Huizi Mao, and William J Dally. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. **Paper Discussion 11b.** [GPTQ: Accurate post-training quantization for generative pre-trained transformers.] Frantar, Elias, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
Monday, April 22 **In Person**	**Lecture #13: Modern Generative AI.** • Scaling for large language models • Challenges for LLM inference • What does the future of generative AI look like? • What are the policy and social implications of this technology? **Due: Review of paper 10a or 10b.**
Wednesday, April 24 **Online Only**	**Paper Discussion 12a.** [In-datacenter performance analysis of a tensor processing unit.] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. *In Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA)*, 2017. **Paper Discussion 12b.** [A Configurable Cloud-Scale DNN Processor for Real-Time AI.] Jeremy Fowers, Kalin Ovtcharov, Michael Papamichael, Todd Massengills, et al. *In Proceedings of the 45th Annual International Symposium on Computer Architecture (ISCA)*, 2018.
Monday, April 29 **In Person**	**Lecture #14: Large Scale ML on the Cloud.** [Slides] • Challenges of deployment • Distributed learning at datacenter scale

	Due: Review of paper 11a or 11b. **Due: Final project abstract draft. Can be submitted late until Wednesday afternooon; will discuss in class on Wednesday.**
Wednesday, May 1 **In Person** 	**Lecture #15: Final Project Disussion.**
Monday, May 6 **In Person** 	**Lecture #16: Final Project Disussion.**