1 Summary

The advancements and overwhelming success of Machine Learning has profoundly affected the future of computer architecture. Not only is performing learning on big-data the leading application driver for future architectures, but also machine learning techniques can be used to improve hardware efficiency for a wide variety of application domains.

This course will explore, from a computer architecture perspective, the principles of hardware/software codesign for machine learning. One thrust of the course will delve into accelerator, CPU, and GPU enhancements for ML algorithms, including parallelization techniques. The other thrust of the course will focus on how machine learning can be used to optimize conventional architectures by dynamically learning and adapting to program behavior.

Assignments using parallel-programming frameworks (eg. CUDA). Open-ended team-based project involving some aspect of ML/architecture codesign.

2 Student Responsibilities

Participation (30%)

- **Paper Reading** Read the assigned papers in detail for that class day.

- **Online Discussion via Piazza** Students will participate in online discussion before each class by posting either a review or responding with thoughtful questions/comments on papers.

- **In-class Discussion** You should aim to participate in every in-class discussion. I will do my best to track who is participating. Remember, this class will be fun and interesting if you guys make it so!

Mini-Project (15%) There will be two mini projects which can also be performed in groups:

- Parallelizing a machine learning kernel using CUDA on our V100 GPU.

- Build a ML-accelerator simulator, which is correct and produces accurate performance estimates.

Leading Class Discussion (15%) Each student (or group) will lead 1 lecture.

Project (40%) Group based research/implementation project with 2-4 students. Please see the project handout, and feel free to use Piazza to help form groups. You will need to propose a project by the beginning of the 5th week of class, so please start thinking early.

3 Course Schedule and Specific Topics

The course schedule from last year can be found here: [https://docs.google.com/document/d/1cDOiy-sW2crkD10tC7b7ekisetygA-bMRHQuZlPABhs/edit](https://docs.google.com/document/d/1cDOiy-sW2crkD10tC7b7ekisetygA-bMRHQuZlPABhs/edit)

Last year we focused mainly on accelerators for neural network computations, and that will likely be true this year. I will adjust the schedule during the first week or two of based on the following criteria:

- Foundational or simple-to-understand topics/papers.

- New topics in hardware for ML that have just been published.

- Topics/papers that you have a special interest in. Your input and curiosity matters to me!
4 Piazza Discussion Directions

The abstract of the papers we are reading on Piazza in advance, and each student should contribute >10 meaningful sentences worth of discussion across that day’s papers. (I will check these at 9am on the day of the presentation, so please be done by then. Also, participating earlier will help the presenter more!!!). The goal of these pre-class discussions is to bring up meaningful questions and insights for the presenter, and to prime the class to be thinking critically about the work. Some suggestions for types of meaningful posts are on piazza.

Piazza signup link: piazza.com/ucla/spring2019/cs259

5 Class Discussion Directions

Some advice when leading a class discussion:

First, read the paper thoroughly, and read/skim several additional papers for further context on the paper. This includes papers cited as background/prior work, as well as newer work that builds upon the existing work (if such papers exist). Try searching Google Scholar or ACM/IEEE libraries for finding such links. Discuss with your group members about what you feel are the important parts of the paper.

Prepare material for a 40 minute presentation (think about one slide per minute), and have discussion questions ready. You may want to look at the Piazza page to check if other students have raised interesting questions. You should cover the key ideas in the paper, including the problem the authors were trying to solve, the proposed solution, main intellectual contributions, relationship to previous work, basic methodology, key results, and discussion questions.

It is okay to borrow slides (feel free to contact the author if the slides are not available - this is common practice!), but please cite the sources. Please email me the slides, with [cs259] in the email, by noon the day of. I will post to the course website.

This goes without saying, but please rehearse your presentation. Most importantly: Don’t stress too much or worry about saying something wrong, we are a friendly audience! The point is simply to provoke some good discussions.

6 Academic Honesty

Please make sure that you give proper attribution for any works you use in this class. Also, it is acceptable to share this class’s project with other research you are currently doing independently – however, please clear it with me first. You may not reuse prior work for these projects.