Harry Dong

harryd@andrew.cmu.edu | www.andrew.cmu.edu/user/harryd/ | www.linkedin.com/in/dongharry

Education

Carnegie Mellon University *Pittsburgh, PA*

Electrical & Computer Engineering PhD Candidate *2021 - present*

- Advisor: Prof. Yuejie Chi
- GPA: 4.00
- **Research interests**: Deep Learning Efficiency, Hardware-aware Algorithms, Sparsity in Large Language Models, Generation

UC Berkeley *Berkeley, CA*

Statistics BA & Computer Science BA *2017 - 2021*

- GPA: 3.96 (High Distinction)

Relevant Coursework

- **Statistics/Math**: Theoretical Statistics, Linear Algebra, Stochastic Processes, Time Series, Discrete Math, Real Analysis
- **Electrical Engineering/Computer Science**: Deep Learning, Algorithms & Intractable Problems, Convex Optimization, Data Structures, Database Systems, Linear Systems, Adaptive Control
- **Economics**: Econometrics, Microeconomics

Publications

Preprints

Prompt-prompted Mixture of Experts for Efficient LLM Generation, **Harry Dong**, Beidi Chen, Yuejie Chi, 2024.

Get More with LESS: Synthesizing Recurrence with KV Cache Compression for Efficient LLM Inference, **Harry Dong**, Xinyu Yang, Zhenyu Zhang, Zhangyang Wang, Yuejie Chi, Beidi Chen, 2024.

Journals

A Lightweight Transformer for Faster and Robust EBSD Data Collection, **Harry Dong**, Sean Donegan, Megna Shah, Yuejie Chi, *Scientific Reports*, 2023.

- Oral presentation at the *Machine Learning for Scientific Imaging Conference at Electronic Imaging*, 2024.
- Poster presentation at the *Joint Workshop at the Intersection of Materials Science and Machine Learning*, 2023.

Fast and Provable Tensor Robust Principal Component Analysis via Scaled Gradient Descent, **Harry Dong**, Tian Tong, Cong Ma, Yuejie Chi, *Information and Inference*, 2023.

- Contributed talk at *SIAM MDS22*, 2022.

Conferences

Deep Unfolded Tensor Robust PCA with Self-supervised Learning, **Harry Dong**, Megna Shah, Sean Donegan, Yuejie Chi, *ICASSP*, 2023.

- Also presented at the *Third Workshop on Seeking Low-Dimensionality in Deep Neural Networks*, 2023.

Learning Optimal Traffic Routing Behaviors Using Markovian Framework in Microscopic Simulation, Theophile Cabannes, Jiayi Li, Fangyu Wu, **Harry Dong**, Alexandre Bayen, *TRB 99th Annual Meeting*, 2020.

Workshops

Towards Structured Sparsity in Transformers for Efficient Inference, **Harry Dong**, Beidi Chen, Yuejie Chi, *ICML Workshop on Efficient Systems for Foundation Models*, 2023.

Honors & Awards

Wei Shen and Xuehong Zhang Presidential Fellowship, 2024
Liang Ji-Dian Graduate Fellowship, 2023
Michel and Kathy Doreau Graduate Fellowship in Electrical and Computer Engineering, 2023
NSF GRFP Honorable Mention, 2023
UC Berkeley High Distinction, 2021

Research Experience

Yuejie Chi Group *Pittsburgh, PA*
Advisor: Prof. Yuejie Chi *Sep 2021 - present*

- Developed a fast, learnable, and provable tensor robust principal component analysis algorithm
- Designing hardware-aware algorithms for efficient transformer models by leveraging/inducing structured sparse patterns
- Reducing computational footprint (e.g. KV cache) for long context tasks using principles from recurrent networks

Mobile Sensing Lab *Berkeley, CA*
Advisor: Prof. Alexandre Bayen; Mentor: Theophile Cabannes *May 2019 - May 2021*

- Constructed a model to optimize multi-agent network games with applications in traffic routing
- Explored stochastic controller designs for efficient flow through networks

Lawrence Berkeley National Laboratory & UCSF *Berkeley, CA*
Mentors: Roy Ben-Shalom, Jan Balewski *Jun 2019 - May 2021*

- Improved robustness and model interpretability for prediction of neuron ion conductance properties from voltage responses to stimuli

Professional Experience

Air Force Research Lab *Wright-Patterson AFB, OH*
Research Intern *May 2022 - Aug 2022*

- Designed a method for efficient high-dimensional materials data recovery using transformer models
- Continuing collaboration to build multimodal microscopy diffusion models
- Mentored by Megna Shah & Sean Donegan

Amazon Web Services *Seattle, WA (Remote)*
Software Development Engineer Intern *Jun 2021 - Aug 2021*

- Full stack development of internal service for cloud operations cost modeling
- Received but declined full-time offer to pursue PhD

Amazon Web Services *Seattle, WA (Remote)*
Software Development Engineer Intern *May 2020 - Aug 2020*

- Full stack development of internal services that facilitate server testing for hardware engineers
- Received return offer

Teaching Experience

CMU 18-786 (Introduction to Deep Learning) *Pittsbrugh, PA*
Teaching Assistant *Jan 2024 - May 2024*

- Teaching recitations, maintaining the course website, and hosting office hours for a graduate deep learning class

CMU 18-202 (Mathematical Foundations of Electrical Engineering) *Pittsbrugh, PA*
Teaching Assistant *Jan 2022 - May 2022*

- Taught recitations, hosted office hours, and created material (homework and exams) for an undergraduate class

UC Berkeley Student Association of Applied Statistics *Berkeley, CA*
Education Director *Jun 2020 - Dec 2020*

- Led a team of lecturers to teach data science concepts and skills to undergraduates of all levels of expertise

Outreach / Other Experience

Faculty Hiring Student Council *Pittsburgh, PA*

Council Member *Jan 2023 - Apr 2023*

- Evaluated CMU ECE faculty candidates' interpersonal relationships between colleagues and students

Cal Ballroom *Berkeley, CA*

Competition Coordinator *May 2019 - May 2020*

- Organized all competition-related events with the Cal Ballroom team
- Publicized events, hired judges, negotiated with other organizations, and hosted competitions with hundreds of participants