Jerry Chee

Department of Computer Science
Cornell University

JerryChee@cs.cornell.edu
Jerry-Chee.github.io

I am interested in developing machine learning methods to meet the needs of practitioners.

Education

Cornell University — Ithaca, NY
Ph.D. in Computer Science — 2019 - 2025 (expected)
Advisor: Chris De Sa

University of Chicago — Chicago, IL
B.S. in Computational and Applied Mathematics — 2013 - 2017
Advisor: Panos Toulis

Publications

J. Chee, Y. Cai, V. Kuleshov, C. De Sa. *QuIP: 2-Bit Quantization of Large Language Models with Theoretical Guarantees* In *NeurIPS 2023* **(Spotlight)**

J. Chee, H. Kim, P. Toulis. *"Plus/minus the learning rate": Easy and Scalable Statistical Inference with SGD.* In *AI and Statistics 2023*

J. Chee, M. Renz, A. Damle, C. De Sa. *Model Preserving Compression for Neural Networks.* In *NeurIPS 2022*

J. Chee, S. Braun, V. Gopal, R. Cutler. *Performance Optimizations on U-Net Speech Enhancement Models.* In *IEEE Multimedia Signal Processing 2022*

C. Yang, Z. Wu, **J. Chee**, C. De Sa, M. Udell. *How Low Can We Go: Trading Memory for Error in Low-Precision Training.* In *ICLR 2022*

J. Chee, P. Li. *Understanding and Detecting Convergence for Stochastic Gradient Descent.* In *IEEE Big Data 2020*

J. Chee, P. Toulis. *Convergence Diagnostics for Stochastic Gradient Descent.* In *AI and Statistics 2018* **(Oral)**

Talks

Statistical Properties of Stochastic Gradient Descent — Denver, CO
Joint Statistics Meeting, with Panos Toulis. — Jul 2019

Convergence Diagnostics for Stochastic Gradient Descent — Canary I.
AISTATS 2018, with Panos Toulis. — Apr 2018

Projects

Practical 2-Bit Quantization of Large Language Models (In preparation)
- with C. De Sa, V. Kuleshov
Making 2 bits the optimal use of memory, with an efficient implementation.

Harm-Mitigation in Recommender Systems (In submission)
– with S. Ernala, S. Kalayanaraman, S. Ioannidis, S. Dean, U. Weinsberg
Study recommender policies which mitigate user engagement with harmful content.

Predicting lincRNA functionality in short and long ORF (In preparation)
– with C. Railey, C. De Sa, A. Nelson
Predict protein coding ability of lincRNA using deep recurrent NN models.

Industry Experience

Meta, Core Data Science — Menlo Park, CA
Research Engineer Intern — Jun–Sept 2022

- Prototyped deep learning-based metric to estimate the likelihood a user would interact with borderline harmful content based on previous interaction history.
- Compiled requisite datasets using SQL, performed data analysis and visualization in notebooks, and trained distributed DNNs at scale.

Amazon, Supply Chain Optimization Technologies — Seattle, WA
Applied Scientist Intern — Dec 2021–May 2022

- Estimated 12× training speedup for a causal inference model used to estimate the value of in-stock items on Amazon.com.
- Saved and reused repeated computation via repeated linear regressions with common set of controls.

Microsoft, IC3-AI — Redmond, WA
Intern — Jun–Sept 2021

- 7× inference speedup of deep background noise suppression models used real-time in Teams.
- Identified and implemented model compression methods supported by the neural network inference engines ONNX Runtime, CoreML, and TFLite.

Baidu, Cognitive Computing Lab — Bellevue, WA
Research Intern — Mar–Jul 2019

- Developed statistical convergence tests for variants of stochastic gradient descent with momentum and gradient compression.
- Utilized multi-task learning to increase the available training data in order to improve the predictive performance of graph neural networks.

McKinsey & Company — Boston, MA
Senior Analytics Fellow — Oct 2017 - Feb 2019

- Implemented data science solutions at client organizations, working closely with business leaders and domain experts.
- Led several data science initiatives in predictive maintenance for the network technology division of a top telecommunications company.
  - Utilized a cost (of true positive, false positive, etc.) analysis for selecting the prediction target and implementation strategy which maximized business impact and modeling feasibility.
  - Built classification models for network and customer service use cases.

Teaching

TA, CS 4780/5780: Machine Learning for Intelligent Systems — Fall 2019
TA, CS 4787: Principles of Large-Scale Machine Learning — Spring 2020
TA, CS 6787: Advanced Machine Learning Systems — Fall 2020

Outreach

Skype A Scientist Volunteer — Apr 2020-May 2021
Video call with classrooms across the country to help educate students about research in computer science and career options as a quantitative scientist.

Other Information

Programming: Python (PyTorch), SQL, R (RCpp), C (MPI)
Languages: Chinese (Limited oral proficiency)