

Charlie Ruan

(626) 589-0364 | cfruan@cs.cmu.edu

EDUCATION

Carnegie Mellon University, Computer Science Department

M.S., Computer Science

GPA: 4.0/4.0

Pittsburgh, PA | Aug 2023 - May 2025

Cornell University, College of Engineering

B.S., Computer Science and Operations Research

GPA: 4.0/4.3; *Summa Cum Laude*

Ithaca, NY | Aug 2019 - May 2023

Relevant Courses: Deep Learning Systems, Reinforcement Learning, Parallel Architecture and Programming, Operating Systems, Distributed Computing Principles, Computer Networks, Functional Programming, Stochastic Processes

RESEARCH EXPERIENCE

Distributed LLM Serving *Research Assistant* (Distributed Systems, LLM Serving)

Pittsburgh, PA | Mar 2024 – Present

PI: Prof. Tianqi Chen, Prof. Zhihao Jia

- Working on distributed LLM serving systems, more details to come

MLC LLM & Web LLM *Research Assistant* (Machine Learning Compilation)

Pittsburgh, PA | Jun 2023 – Present

PI: Prof. Tianqi Chen

- Working on the open-source project MLC LLM, enabling universal native deployment of LLMs through machine learning compilation techniques including TVM, building an LLM serving system on top of it
- Leading the Web LLM project, bringing LLMs to run locally in client-side browser with WebGPU acceleration
- Github links available: <https://github.com/mlc-ai/mlc-llm>, <https://github.com/mlc-ai/web-llm>

RelaxML Lab *Research Assistant* (Distributed Machine Learning)

Ithaca, NY | Sep 2022 – May 2023

PI: Prof. Christopher De Sa

- Investigated finding provably better data permutations at large scale (i.e. distributed learning with decentralized data), using recently proposed example-ordering algorithm Gradient Balancing (GraB)
- Built a distributed machine learning system and implemented experiments (e.g. sequence classification with BERT, image classification with ResNet) in a decentralized fashion to analyze bounds on convergence rate and consensus error
- Paper accepted by *NeurIPS 2023*; manuscript available: <https://arxiv.org/abs/2302.00845>

Variance Reduction for Reinforcement Learning *Research Assistant* (RL)

Ithaca, NY | Dec 2021 – Sep 2022

PI: Prof. Jim Dai

- Used reinforcement learning (RL) to optimize the algorithm of matching drivers and customers on ride-hailing systems like Uber
- Formulated the application of variance-reduction method approximating martingale-process (AMP) in proximal policy optimization (PPO) when state space is large and state transitions are uncertain; experimented on ride-hailing and multiclass queueing networks
- Manuscript available: <https://arxiv.org/abs/2211.15886>

McMahon Lab *Research Assistant* (Physical Neural Network, Transformer)

Ithaca, NY | Oct 2021 – May 2022

PI: Prof. Peter McMahon

- Investigated a better digital twin for a physical neural network training algorithm (<https://www.nature.com/articles/s41586-021-04223-6>)
- Built multiple transformer-based neural networks to model chaotic physical systems (e.g. a coupled spring-pendulum system), including using Koopman Theory to create a physics-informed autoencoder

INDUSTRY EXPERIENCE

Google Core ML *Software Engineer Intern* (TensorFlow, Python)

Sunnyvale, CA | Jun 2023 – Aug 2023

- Worked under Core ML's Distributed Runtime team, optimizing TensorFlow (TF) runtime
- Worked on enabling TF asynchronous checkpoint in Keras, offloading model checkpointing to an asynchronous thread, hence reducing wasted TPU cycles, a critical mission within Alphabet
- Designed and implemented checkpoint APIs for all TF variables (~20 classes), committed 1700+ LOC to <https://github.com/tensorflow>
- Received a return offer

Google Cloud *Software Engineer Intern* (OpenBMC, Linux, C++)

Sunnyvale, CA | Aug 2022 – Oct 2022

- Worked on Google Cloud's Technical Infrastructure Platform team, deploying accelerators including GPUs in Google data centers
- Implemented a Linux daemon that interacts with D-Bus and I2C to monitor the health of data centers' GPUs using OpenBMC; built an API on D-Bus that provides out-of-band firmware updates
- Worked on pre-production hardware with limited debugging support; was responsible for communicating with external GPU vendor
- Received a spot bonus and a return offer as recognition for the solid delivery of the project

Amazon Robotics *Software Engineer Intern* (Full-Stack, Kotlin, Java)

Greater Boston, MA | May 2022 – Jul 2022

- Worked on the Human-Computer Interaction (HCI) team; implemented a full-stack configuration portal on Amazon Robotics's HCI software, allowing warehouse workers to personalize their interaction with the autonomous warehouse robots
- Improved workers' experience; alleviated burden of the HCI team as changing settings no longer requires submitting tickets
- Familiarized with the Spring Framework and various design patterns (e.g. Delegate, Singleton)
- Received return offer for a full-time position

XPeng Motors *Software Engineer Intern* (Sensor Fusion, Python, C++)

Shanghai, China | Jun 2021 – Aug 2021

- Processed and fused various sensor data (e.g. radars, cameras) of XPeng's self-driving cars to provide a reliable perception result
- Designed and implemented Key Performance Indicators that evaluate the Sensor Fusion algorithm's precision and detect issues (e.g. phantom objects, abnormal decelerations); optimized the Sensor Fusion algorithm (e.g. object tracking, measurement association)
- Integrated the KPIs into release management test, deciding whether new versions of autopilot code can be deployed on XPeng's cars

Morgina Information Technology *Software Engineer Intern* (C/C++, Embedded)

Shanghai, China | Jun 2020 – Jul 2020

- Optimized the multi-object tracking algorithm of millimeter-wave radars installed in intersections that monitor traffic information
- Evaluated radars' perception result and troubleshoot the specific parts of the algorithm that caused misalignments with ground truth

STUDENT ACTIVITIES

Cornell Electric Vehicles *Software Engineer* (Python, ROS, Linux)

Ithaca, NY | Aug 2019 – Mar 2022

- Designed the ROS (Robot Operating System) for the vehicle's autonomy system; engineered a platform for communications between sensors (e.g. LIDAR, IMU) and algorithms, as well as among algorithms (e.g. vision, localization)

TEACHING EXPERIENCE

Intro to Engineering Stochastic Processes *Teaching Assistant*

Ithaca, NY | Jan 2023 – May 2023

- Topics include: discrete-time/continuous-time Markov chain, Poisson process, queueing theory, Markov decision process
- Was the sole TA responsible for the design, direction, office hours, and grading for a coding project that compares traditional Monte Carlo simulation with neural networks

Intro to Machine Learning *Teaching Assistant*

Ithaca, NY | Aug 2021 – Dec 2021

- Topics include: decision trees, support vector machine, kernels, neural networks, statistical learning theory, online learning, boosting
- Participated in the design of homework and coding projects

PUBLICATIONS & MANUSCRIPTS

- A. Feder Cooper*, Wentao Guo*, Khiem Pham*, Tiancheng Yuan, **Charlie F. Ruan**, Yucheng Lu, Christopher De Sa. "CD-GraB: Coordinating Distributed Example Orders for Provably Accelerated Training." *NeurIPS 2023*. <https://arxiv.org/abs/2302.00845>
- **Charlie Ruan**. "Approximating Martingale Process for Variance Reduction in Deep Reinforcement Learning with Large State Space." *On arXiv November 2022*. <https://arxiv.org/abs/2211.15886>

AWARDS & HONORS

Cornell Engineering Dean's Honor List (for all semesters)

2019 – 2023

Omega Rho Honor Society for Operations Research

May 2023

Undergraduate Summer Research Funding, School of Operations Research (five students selected in total)

May 2022

Tau Beta Pi Engineering Honor Society

Mar 2022