Charlie Ruan

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

EDUCATION

Carnegie Mellon University, Computer Science Department **Pittsburgh, PA | Aug 2023 - Dec 2024**
M.S., Computer Science

Cornell University, College of Engineering **Ithaca, NY | Aug 2019 - May 2023**
B.S., Computer Science and *Operations Research*
GPA: 4.0/4.3; *Summa Cum Laude*

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

SKILLS

Programming Languages: Java, Python, C++, OCaml, SQL
Tools & Frameworks: Linux, git, PyTorch, TensorFlow, Ray, torch.distributed, Hugging Face, ROS, Linear Programming

RESEARCH EXPERIENCE

Catalyst Group *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 for large language models through machine learning compilation techniques like Apache TVM Unity, building an LLM chatting system and a serving system (distributed)
- Implemented various state-of-the-art LLM inference techniques including KV cache and sliding window attention
- Github link available: https://github.com/mlc-ai/mlc-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
- Manuscript available: https://drive.google.com/file/d/15mlQ_hOpJMqVrULqko4jrGvbkbB5TkAN/view?usp=sharing

PROFESSIONAL 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
- Project documentation: go/async-checkpoint-rollout

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 object tracking algorithm of millimeter-wave radars installed in traffic lights that monitor vehicles' velocity and status
- Evaluated radars' perception result and troubleshot 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, coding projects, and exams; was specifically responsible for the programming project on empirical risk minimization, which asks the students to build an email spam filter

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." *Under submission, 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
- **Charlie Ruan**. "Transformer-Based Models for Building a Precise Digital Twin for Physical Neural Network." *Manuscript updated May 2022*. https://drive.google.com/file/d/15mlQ_hOpJMqVrULqko4jrGvbkbB5TkAN/view?usp=sharing
- Wentao Guo*, **Charlie Ruan***, Claire Zhou*. "Photo Rater: Photographs Auto-Selector with Deep Learning." *On arXiv November 2022*. https://arxiv.org/abs/2211.14420

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 (top 12.5% of junior class invited)	**Mar 2022**