Jui-Tse Hung

howard88tw@gmail.com | ruizehung.github.io | ruizehung | ruize-hung | 650-391-5309

Education

Georgia Institute of Technology *Atlanta, GA, U.S.*

Computer Science B.S./M.S. - GPA 4.00/4.00 (Faculty Honors) *Aug 2020 - Dec 2024*

Selected Coursework: Design & Analysis of Algorithms, Systems and Networks, Computer Networking, Machine Learning, Deep Learning

Technical Skills

Programming Languages: Python, C/C++, Golang, Java, Typescript/JavaScript, SQL, HTML, CSS + I can learn anything

Frameworks & Tools: AWS (Certified Developer Associate), Docker, Kubernetes, Pytorch, Tensorflow, Express, FastAPI, React, Next.js, PostgreSQL

Experience

Scale AI *San Francisco, CA, U.S.*

Software Engineer Intern (Machine Learning Infrastructure Team) *May 2023 - Aug 2023*

- Reduced large language models (LLM) endpoints cold start time by 6x, reducing LLM serving costs. Link to the engineering blog post.
- Enhanced code organization and security by migrating model-serving endpoints into multi-container architecture with Kubernetes and Python.
- Built an internal model serving endpoints management dashboard using React, Tailwind CSS, and Express.
- Enabled external users to self-host LLM Engine, an engine for fine-tuning and serving large language models (LLM) from Scale AI written in Python.

Numbers Protocol *(Remote) Taipei, Taiwan*

Decentralized Data System Developer Intern *Feb 2022 - Mar 2022*

- Worked on multiple features and enhancement for Capture App, a web3 camera app, using Typescript, Angular, RxJS, Ionic.
- Developed a network application that stores media asset in the InterPlanetary File System (IPFS)/Filecoin network through Web3.Storage.

Facebook (Meta) *(Remote) Menlo Park, CA, U.S.*

Software Engineer Intern (Knowledge Infrastructure Team) *Sep 2021- Dec 2021*

- Optimized knowledge graph pre-build process speed by 7x. (Some component up to 36x.) Saved about 10 hours of developer time each week.
- Rewrote a 1400-line shell script that starts knowledge graph build into an efficient and user-friendly command line interface using Golang.
- Refactor 2000+ lines of Golang code in our knowledge graph core infrastructure code base.

Amazon *(Remote) Seattle, WA, U.S.*

Software Development Engineer Intern (Tax Engine Team) *May 2021- Aug 2021*

- Devised a standard operating procedure (SOP) for an issue accounting for 20% of our team's tickets with a business impact of € 1 million/month.
- Built an internal web app using Typescript, React, Java, and AWS that reduces the cycle time of tickets from a few weeks to a couple of days or less.
- Conducted detailed customer research and root cause analysis with multiple business and engineering teams across US, Europe, India, and Japan.

De Anza College Computer Science Department *Cupertino, California*

Teaching Assistant - Intermediate C++ Programming *Sep 2019 - Jun 2020*

- Created lecture notes, which include extra learning resources, and shared them with students. Link to my C++ notes.
- Resolved students' lecture and homework questions. Reviewed students' homework.

Technical Project

Zlind is an online forum leveraging zero-knowledge (ZK) proof to enable people to share and connect fearlessly and anonymously. People can sign-up using their work or school email, post or comment anonymously while proving that they belong to a specific company or school, and optionally reveal themselves later on if they want. I used Next.js, Tailwind CSS, tRPC, Prisma, Supabase, and Semaphore ZK protocol.

Royal Demons is a Dungeon Crawler game written in Java that won the Best Project Competition out of a total of 114 teams in Gatech Objects and Design course. Implemented maps, procedural generation, doors, spawning enemies, dropping items, NPC, and some UI. Demo video here.

Research

Pareto-Secure Machine Learning: Fingerprinting and Securing Inference Serving Systems *Georgia Institute of Technology*

Systems for Artificial Intelligence Lab - Prof. Alexey Tumanov *Jan 2023 - June 2023*

- Devised a query-efficient fingerprinting algorithm capable of victimizing and consistently triggering the same model within a model zoo served by a model-less ML model inference serving system.
- Built a shim layer in C++ over Clockwork, a state-of-the-art ML model inference serving system, enabling a model-less inference API.
- Paper link: https://arxiv.org/abs/2307.01292