Zhihan Lu

+1 713-301-9397 | zhihanlu.1@gmail.com | www.linkedin.com/in/zhihan-lu

Education

Carnegie Mellon University Pittsburgh, US
Master of Science in Machine Learning | GPA: 4.11/4.33 December 2022

Rice University Houston, US
Bachelor of Arts in Computer Science and Mathematics | GPA: 3.9/4.0 May 2021

Hwa Chong International School Singapore, SG
International Baccalaureate Diploma | IB score: 43/45 December 2016

Skills

Programming Languages: Python, C++, Golang, Java, SQL, C, JavaScript, Scala, C#, HTML, CSS
Frameworks: Pytorch, TensorFlow, React, Spark, Node.js, Angular, Flask, Hadoop, GraphQL, .NET
Software: Git, NumPy, Sklearn, OpenCV, AWS, MongoDB, Linux, Bash, Redis, Jira

Work Experience

Waymo (Google's self-driving company) | MLE Intern | Mountain View, US May 2022 – August 2022

- Productionized a gradient-based algorithm to find influential training examples for an eval example.
- Constructed the backend using Python, C++, Apache Beam, Tensorflow, and the frontend using Angular.
- Enabled fast example retrieval by creating a table-valued function server using nearest neighbor search.

Amazon | SDE Intern | Seattle, US May 2021 – August 2021

- Achieved a 2x higher invalidity detection rate in Amazon's catalog by creating a rule-based classifier.
- Designed a scalable REST API to keep human in the loop with Java and deployed it on AWS services.
- Accelerated data pre-processing by 10x with Java multithreading and database performance tuning.

LinkedIn | Software Engineering Intern | Sunnyvale, US May 2020 – August 2020

- Improved LinkedIn's people-match model's precision by 5% through enhanced feature engineering.
- Reduced the Spark pipeline's run-time by 50% through benchmarking with Scala and stage removals.

Lutron Electronics | Software Engineering Co-op | Coopersburg, US June 2019 – August 2019

- Developed an API using C# ORM to power an enterprise software for telemetry data discovery.
- Reduced data query time of >10 million records to less than 250ms with caching and query optimization.

SEA | Software Engineering Intern | Singapore, SG June 2018 – August 2018

- Constructed frontend logic for a workplace collaboration web client (SeaTalk) using JavaScript (React).

Research Experience

Sailing Lab (Professor Eric Xing), Carnegie Mellon University October 2021 – present

- Developed a CI/CD pipeline for an open-source distributed training compiler for large neural networks.
- Researching strategies for tensor re-sharding across hardware within the cluster during model training.

Efficient and Intelligent Computing Lab, Rice University December 2019 – May 2021

Early-Bird GCNs: Graph-Network Co-Optimization Towards Efficient GCN Training and Inference

- Proposed a novel method to jointly sparsify graph and weights in Graph Convolutional Network.
- Enabled >277 times inference FLOPs reduction; published the paper on *AAAI 2022* as the second author.

SACoD: Sensor Algorithm Co-Design Towards Efficient CNN-powered Intelligent PhlatCam

- Extended Neural Architectural Search (NAS) to enable co-design of IoT sensors and neural networks.
- Achieved 70% energy saving for IoT vision tasks; published the paper on *ICCV 2021* as a co-author.

Leadership & Community Experience

Rice Apps | Team Lead | Houston, US June 2020 – May 2021

- Led a 9-person agile team and created an SMS text bot that assisted discharged patients from hospitals.
- Used React, Node.js, and Flask to develop the text bot and its usage analytics dashboard.

HackTX at UT Austin | First-Place Winner | Austin, US October 2018

- Won First Place against 297 participants with a facial-expression-controlled speed-reading webapp.