RenChu Wang

+1-404-451-6225 rentruewang@gatech.edu /rentruewang /rentruewang /rentruewang.github.io

Education

MS CSE | Georgia Tech — Aug 2022 - May 2024

- Courses: Probability for Sci./Eng., Knowledge-based AI, Ubiquitous Computing, High Perform Computing, Modelling & Simulation, Data Science for Social, Computational Data Analysis, Data Visual Analysis

BS EE | National Taiwan University — Sep 2017 - Jun 2021

- Courses: Algorithms, Convex Optimization, Machine Learning, Linear Algebra, Digital Speech Processing, Data Structures, Signal and Systems, Computer Architecture, Integrated Circuit Design, Cloud Computing And Cyber Security

Technical Skills

Programming	Python, Go, Modern C++, Java, JavaScript, Dart
Machine Learning	NumPy, PyTorch, Pandas, TF, Keras, Spark, BoTorch, GPyTorch, Ax
Platforms	Azure, AWS, Google Cloud, DataBricks, Docker
Tools	Linux, SQL, Flask, Django, D3.js, Git

Selected Projects & Achievements

BoCoEL - 10 times faster LLM evaluation with Bayesian optimization and NLP — Jan 2024

- Creator and lead developer on a project with 200 stars on GitHub so far.
- Designed an algorithm based on Bayesian optimization that is able to speed up evaluation (benchmarking) of LLMs by more than 10 times with NLP techniques like dense retrieval and embedding search.

Tula - stock allocation advisor powered by explainable AI LLM agents — Oct 2022

- Led a team of 4 to win the 1^{st} place out of 1500 participants from the event's biggest sponsor, BlackRock.
- Designed a fintech ML model to make transparent decisions with NLP to aid financial professionals.
- Responsible for backend that utilizes volatility index and news aggregation for portfolio allocation.

Koila - ML library for solving CUDA (GPU) issues in 1 line — Nov 2021

- Creator and lead developer on the project that amassed 1800 stars and 50 forks on GitHub so far.
- Utilizes the idea of lazy evaluation to solve the notorious out of memory error for the most popular machine learning library for researchers, PyTorch, with a very minimalist API.
- Works with any PyTorch operation, such as CNN, GNN, RNN, Linear layers.

Positions of Responsibility

Software Designer — May 2023 – Jul 2023

Ponder (Acquired by SnowFlake)

- Ponder, a subsidiary of SnowFlake, focuses on connecting data warehouses with pandas API (SQL to Pandas), to enable data scientists to quickly design their pipeline without going into database infrastructure.
- Speed up the integrated vector database by 10 fold with SQL and ML techniques, integrated ML pipeline functionality, and implemented interactive debugging capabilities that fixed several high importance, day 0 SQL bugs in the deployment.

Data Scientist — Dec 2021 - Jun 2022

MediaTek Research

- Design and implementation of a reinforcement learning (RL) based AI / EDA system that speeds up the initial chip development process of the world's biggest chip designer by more than 3 times.
- Outlined and implemented a RAG system to reduce error rate by 40%.
- Sped up the existing distributed AI trainer by 30 times with careful profiling and reducing the critical path.
- Proposed and examined experiments to verify a reinforcement learning (AI/RL) based chip design model to justify solution quality.

Research Assistant — Sep 2019 - Aug 2022

National Taiwan University

- Engineered and developed a scalable quantum compiler that scales up to 20,000 qubits (previously only 128), a 200 times improvement.
- Created an alternative way of masking for transformers that speeds up the pre-training process for NLP pre-trained models by 10 times.