RUOYU (Ryan) **HUANG**

huangruoyu666@gmail.com
https://www.linkedin.com/in/ruoyu/

SUMMARY

Experienced Software Engineer specializing in LLM and MLOps infrastructure. Proficient in building scalable distributed systems, optimizing ML workflows and model performance. Skilled in PyTorch, Python, Kubernetes, Ray, Triton, Go, and AWS. Demonstrated success in delivering distributed training and inference systems in production environments.

EXPERIENCE

2022 - Present, Tech Lead, AI Platform - Omniva, Seattle, WA

- Build distributed LLM training and inference systems using Ray and Kubernetes.
- Optimize model training with faster dataloading, model parallelism, custom Triton kernel fusion.
- Optimize model inference with quantization, tensor parallelism, and continuous batching.
- Build scalable Retrieval-Augmented Generation (RAG) applications with LLMs and vector databases.
- Develop tools for LLM performance profiling and capacity planning.

2020 - 2022, Sr. Software Engineer, Machine Learning Platform – Uber, Seattle, WA

- Engineered a Kubernetes-based control plane for internal ML platforms (Built on k8s CRDs and operator).
- Designed and implemented a workflow system to streamline ML model lifecycles (training, evaluation, monitoring).
- Enhanced ML developer experience through custom CLI tools for infrastructure interaction.
- Built an orchestration framework to improve reliability of training pipelines on Spark and Ray.

2017 - 2020, Sr. Software Engineer, Amazon Transcribe – AWS, Seattle, WA

- Led the development of Amazon Transcribe (ASR) from scratch, creating multi-tenant cloud services
- Build real-time streaming service speech recognition, bidirectional streaming using HTTP/2 and GRPC.
- Optimize system efficiency across the stack (model runtime, job scheduling, fleet auto scaling)
- Led a team to develop applications for special domains, e.g., medical and call center analytics.
- Built MLOps training, evaluation pipelines using AWS Lambda and ECS.

2017 - 2017, Software Engineer – Thomson Reuters, Sunnyvale, CA

- Built an A/B testing platform leveraging Spark, Kafka, Cassandra, and Redis for data processing and analytics.

2014-2017, CPU Physical Design Engineer – Broadcom, Santa Clara, CA

- Physical design of ARM v16 CPU core, timing closure, place & route, and CAD tools, etc.

EDUCATION

2014 - Master of Science (Electrical & Computer Engineering), University of Florida - FL, United States

2013 - Bachelor of Science (Micro-electronics), UESTC - Chengdu, China

PATENT

No. US10777186B1: Streaming Real-time Automatic Speech Recognition Service.