Lai Wei

Phone: 412-953-9889 | Email: royweilai@gmail.com | Github: roywei | Blogs: Medium

I am passionate about AI, with expertise in distributed systems, large-scale model training, ML frameworks and systems.

Work History

2018-01 - Present

Senior Software Development Engineer/Tech Lead - AWS AI

- Established a specialized team to develop, test, and deploy AWS-optimized PyTorch with out-of-the-box support for AWS EFA and enhanced NCCL. This solution was adopted by AWS's largest clients for training large language models (LLMs).
- Redesigned the AWS SageMaker Distributed Data Parallel Library to seamlessly integrate with PyTorch as a custom distributed backend, enhancing collective communications for large model training. Achieved a 40% reduction in training time for GPT NeoX, Dall-E, and guided diffusion models.
- Spearheaded loss function optimization efforts for the AWS MLPerf Benchmark in 2020, enabling model convergence with extremely large batch sizes. Successfully trained MaskRCNN using 512 GPUs under 7 minutes and T5-3B using 2048 GPUs in 4.68 days.
- Developed the Deep Java Library for framework-agnostic model inference in Java, designing and implementing TensorFlow and TensorFlow Lite inference engines.
- Served as the primary contributor and maintainer of Keras-MXNet, delivering the fastest multi-GPU training capabilities for the MXNet backend of the Keras API.
- Acted as a committer for Apache MXNet, contributing to the development of training APIs and CUDA operator enhancements while overseeing open-source MXNet releases

2016-06 - 2017-10

Software Engineer - Cheetah Mobile America AI Lab

- Designed and implemented the User2Vec feature pipeline, enhancing Ad Click-Through Rate (CTR) prediction accuracy.
- Constructed data pipelines to generate 5 million comprehensive user profiles, bolstering personalized news recommendation capabilities.

Education

2015-01 - 2016-05

Master of Science: Electrical and Computer Engineering

Carnegie Mellon University - Pittsburgh, PA

2010-08 - 2014-05

Bachelor of Engineering: Electrical and Electronic Engineering

Nanyang Technological University - Singapore

Publications

- Transfer Learning for Personalized Content and Ad Recommendation, *Industry Talk on RecSys 2017*
- Machine learning approach for shaft crack detection through acoustical emission signals, *2015 IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA)*