Zhenyu Lin

AI/ML Engineering
(415)-794-5746 | zhenyulin.cs@gmail.com | www.linkedin.com/in/zhenyu-lin/ | www.zhenyulincs.com

Technical Skills

Languages: Java, Python, C/C++, SQL (Postgres), JavaScript
Frameworks: Pytorch, Tensorflow, React, Node.js, Spring,LangChain,Lambda
Developer Tools: Git, Docker, Amazon Web Services, Google Cloud Platform, Linux,

Publication

Conference Paper May 2023

- Z. Lin, P. Liang, X. Zhang, and Z. Qin, "Toward robust high-density emg pattern recognition using generative adversarial network and convolutional neural network," in NER'23, IEEE.

Experience

Machine Learning Researcher Sep. 2021 – Present
Sony, MIC Lab *San Francisco, CA*

- Researched methods to reduce deep learning model parameters using Functionality-based Pruning
- Minimized deep learning model from 14 MB to 0.55 MB through functionality-based pruning, achieving 95% image recognition accuracy.
- Further compressed the model from 0.55 MB to 0.21 MB using 8-bit quantization, achieving 500 ms processing time for 18 KB image.

Projects

Real-Time Deep Learning for Mobile Devices Jun. 2023 – Aug. 2023

- Optimized a deep learning model by 85%, shrinking the baseline model size from 463kB to 73kB with less than 1.5% accuracy drop through 8-bit Quantization
- Implemented a CNN-based Bionic Arm control on a IoT device with 1.5MB sRAM, achieved 85% accuracy and 160ms clinical-grade control latency using C++
- Accelerated sampling rates of async muscle signal data streams by over 200% on an Android device by conducting rigorous, iterative runtime profiling and data structures optimization in Java.

Robust Muscle Movements Gesture Recognition Framework Jun. 2022 – Dec 2022

- Developed a deep learning model using Python and PyTorch to interpret muscle movements signals for gesture recognition.
- Developed framework to generate synthetic muscle signals by utilizing a Generative Adversarial Network (GAN) , achieving a 95% similarity.
- Boosted gesture recognition accuracy from 64% to 94.89% for deep learning models affected by noisy signals.
- Applied deep transfer learning to adapt a pre-trained deep learning model to over 100 users' muscle signals.

Personalized Learning Platform Using Large Language Models Jun. 2024 – Aug. 2024

- Designed a responsive user interface and cross-browser compatibility using React.js and Node.js, deployed using AWS and Docker.
- Applied Flask-Caching and Redis to cache frequently requested data, reducing Restful API response times from 3.3s to 0.15s.
- Fine-tuned Mistral 7B on over 200k instruction-answer pairs and implemented Retrieval-Augmented Generation (RAG), improving response accuracy by 5%.

Education

San Francisco State University San Francisco, CA
Bachelor of Science in Computer Science *Aug. 2019 – Jun. 2023*

San Francisco State University San Francisco, CA
Master of Science in Electrical and Computer Engineering *Aug. 2023 – Dec. 2025*