Andrew Z. **Luo**

Machine Learning · Systems · Software Engineering

425-241-9772 | andrew.zhao.luo@gmail.com | andrew-zhao-luo

Summary

A software engineer with expertise in machine learning, systems, and performance. Highly interested in making systems run fast, cheaply, and reliably. Prior experience in deep learning includes training, profiling, and specialized compilers like TVM. **Proficient in Python, C/C++, Java, and CUDA**. Enjoys solving problems, quickly iterating, and taking ownership in ambiguous environments.

Experience

Staff Engineer — *Seattle, WA*

Octoml, Applied Compiler Engineering - Performance — *Mar. 2021 - Now*

- **Official committer to Apache TVM (9.2k stars on GitHub)**, an open source autotuning compiler written in C++ and Python.
- Contributing quantization, mixed precision, ONNX support in TVM. Making general performance improvements to compiler.
- Leading latency initiatives of high-impact models on CPU and GPU, **bettering model latency by up to 50% over old baselines**.
- Improving collaboration with core compiler and product teams to help set future roadmap.
- Implementing features in SaaS product such as compiler cache, adding new tuning algorithms, and lowering resource usage.
- Improving experimentation tooling in SaaS, reducing iteration times from hours to minutes while gaining more insights.

Machine Learning Engineer — *Seattle, WA*

Apple, AI/ML Machine Intelligence Neural Design — *Jan. 2020 - Mar. 2021*

- Using quantization, sparsity, and hardware-specific knowledge to train models for Siri, Homepod, and future products
- Developing in-house solutions for training vision models and deploying/benchmarking on FPGA and ASIC environments
- Developing demos show-casing technologies to internal stake-holders.

XNOR.ai, Machine Learning Team (Acquired by Apple) — *Aug. 2019 - Jan 2020*

- Training performant computer vision models that can run on bespoke and edge hardware. Part time until Jan 2019.
- Creating demos showcasing XNOR's technologies to key executives at major tech companies.

Education

University of Washington — *Seattle, WA*

Double Major in Computer Engineering and Bioengineering — *Sep. 2015 - Jun. 2019*

- **Coursework**: Machine Learning, Probability and Statistics, Real Analysis, Operating Systems, Compilers, Embedded Systems
- **GPA**: 3.95, *Summa Cum Laude*

Projects and Potpourri

Past Internships — *Summers 2015-2018*

- **Sift**: I rewrote HBase snapshot system, **saving over $1.5 million in S3 costs a year** and integrated data store with BigQuery.
- **Facebook**: I implemented statistical models to predict ad reach demographics for customers with multi-million yearly spend
- **The Institute For Systems Biology**: I helped scientists analyze large datasets of gene expression data.

Project: FPGA Image Convolution Photobooth — *2019*

- Implemented streaming algorithm to run kernel convolutions on streamed images, implemented in FPGA on Altera Cyclone V
- Integrated with camera and VGA, creating a variety of filters like Sobel edge detector, Gaussian blur, and image sharpening

Publication — *2017*

Automatic Characterization of User Errors in Spirometry. **Andrew Luo**, Eric Whitmire, James Stout, Drew Martenson, Shwetak Patel. *IEEE EMBC 2017 (Oral Presentation + Paper)*

Patent Application — *2020*

Compressed Neural Network Models. US Patent App. 16/788261. James Gabriel et al. and **Andrew Luo**. Filed 13 August 2020.