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. Enjoys solving problems, quickly iterating, and taking ownership in ambiguous environments.

Experience

Staff Engineer — *Seattle, WA*

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

- Committer to TVM, an open source autotuning compiler written in C++ and Python.
- Contribute to quantization, mixed precision, and framework support. Support key features in SaaS product.
- Improve internal experimentation workflow, significantly reducing iteration time on tasks from hours to minutes.
- Improve performance of high-impact models on CPU and GPU by up to 50% over old baselines.
- Languages: *C++, Python*. Technologies: *TVM, CUDA*

Machine Learning Engineer — *Seattle, WA*

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

- Use 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
- Languages: *Python, C*. Technologies: *PyTorch, Tensorflow*

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.
- Created face identification demo 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

Sift Science - Engineering Intern — *Jun. 2018 - Sep. 2018*

- Rewrote HBase snapshot system, saving over $1.5 million in S3 costs a year. Added BigQuery integration with HBase.

Facebook, Ads Core - Engineering Intern — *Jun. 2017 - Sep. 2017*

- Implemented back-end statistical models to predict demographics of ad reach for customers with multi-million yearly spend

Project: FPGA Image Convolution Photobooth — *2019*

- Created 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

Publications

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.