

# 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, compilers, 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. Contribute to quantization, mixed precision, and framework support.
- Contributor to several SaaS features including autotune job-caching, automatic mixed precision.
- 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. Technologies: PyTorch, Tensorflow, Apple Neural Engine

### Machine Learning Engineer

Seattle, WA

XNOR.AI, MACHINE LEARNING TEAM

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
- Languages: Python, C, C++. Technologies: PyTorch, Bazel. **Acquired by Apple Jan. 2020**

## 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.