Summary

University of California, Berkeley — August 2015 - May 2019

B.S. Electrical Engineering and Computer Science — GPA: 3.7/4.0

Core Programming Languages: Python, Java, C, Tensorflow, Pytorch

Natural Language Processing (CS 231n) Relevant Coursework:

Machine Learning (CS189) Convex Optimization (CS 127) Operating Systems (CS162) Database Management (CS 186) Graduate Algorithms (CS 270) Artificial Intelligence (CS 188) Random Processes (EE 126) Compilers (CS 164)

Graphics (CS 184)

Professional Experience

Forethought.ai | San Francisco, CA

July 2019 - Present

Software Engineer — ML & Backend | Sami Ghoche, Prof. Christopher Manning

- · Accelerated neural network models for ticket answering in customer support using hardware-optimized kernels
- Trained fp16 models on BERT/XLNet models, built TensorRT engines with specialized C++ kernels for efficient GPU global memory access, and deployed TensorRT server images on AWS Kubernetes clusters using Spinnaker
- Decreased model latency from 100ms to 15ms (6.6x) on 80,000 API requests/day

Berkeley AI Lab | Berkeley, CA

February 2018 - August 2018

Research Assistant — Adeptlab | Bichen Wu, Prof. Kurt Keutzer

- Optimized model inference using sharing and pruning methods on residual networks for super resolution
- Benchmarked baselines for NVIDIA Taylor pruning using Pytorch forward hooks, analyzed FLOPs and memory usage of neural network architectures, and designed an auxilary Pytorch module for dynamic pruning
- · Increased theoretical model inference by 4x and memory footprint by 17x using dynamic pruning

Facebook, Inc. | Menlo Park, CA

May 2017 - August 2017

Software Developer Intern — Recruiting Tools | Zef Rosnbrick

- · Improved candidate ranking system of 4M candidates for the recruiting tool used by over 800 Facebook recruiters
- Normalized Hive data with DaggerDS, tested various models with the in-house FBLearner tool, and integrated a data pipeline with Dataswarm to add additional structured features into Sourcing Tool's scoring model
- Increased click-through rates by 320 clicks/day (28%) and claim rates by 3 candidates/day (5%)

UC Berkeley | Berkeley, CA

August 2016 - December 2016

Undergraduate Student Instructor — *Probability* | *Prof. Satish Rao*

· Dedicated 8 hours a week to holding discussions, leading homework parties, and grading exams for 700 students

Technical Skills

- · Languages: Python, Java, C, HTML, CSS, Javascript
- Frameworks: Pytorch, Tensorflow, Kubernetes, MongoDB, ElasticSearch
- Software: Spinnaker, Tensorflow Serving, AWS, GCP, Jupyter

Honors & Awards

- · First Place, Berkeley Innovation Collider (2017): Awarded for most innovative use case and MVP for AI chatbots
- Best Hack, Pebble and Here API, CalHacks 2.0 (2016): Awarded for BearWalk's best use of the Pebble smartwatch
- · HKN Initiation (2016): Invites top 25% of EECS students with senior standing to the EECS honors society
- Edward Frank Kraft Scholarship (2016): Finished the first year with a 13 unit course load and a GPA of 4.0.

Skills & Hobbies

· Dances hip-hop; Sings in a cappella; Plays guitar, flute and the piano