Kuan Zhou

linkedin.com/in/kuanzhou/
github.com/kzhoulatte

Santa Clara, CA, 95054
zhoukuan1@gmail.com
669-356-2086

Skills

- **Proficiency**: Python, Pytorch, Golang, C/C++, Typescript, React, Docker, Kubernetes, gRPC, Mermaid
- **Familiarity**: TensorFlow, JAX, MLIR, LLVM, Java, Rust, SQL, Mathematica, Spark, ORTools, Numba, Julia

Experience

- **Principal Software Engineer** April 2020 - Present
  SambaNova Systems *Palo Alto, CA*
  - Lead a team in integrating foundation models into the Kubernetes platform, focusing on service performance optimizations
  - Contributed to the design and development of core features in the SambaNova AI framework
  - Co-designed and co-developed a distributed learning infrastructure for extremely large models
  - Implemented various deep learning models leveraging dataflow architecture and advanced software platforms
- **Software Engineer, Machine Learning** February 2019 - March 2020
  Petuum *Sunnyvale, CA*
  - Leveraged OCR engines and deep learning models to process logistic bills automatically with 0.87 accuracy
  - Collaborated in implementation of various anomaly detection models for equipment health prediction
  - Contributed in machine learning pipeline refactoring and model improvement based on various use cases
- **Artificial Intelligence Fellow** June 2018 - September 2018
  Insight Data Science(Bootcamp) *Palo Alto, CA*
  - Architected SketchTML that takes in several hand drawn sketches and produces an interactive HTML website
  - Leveraged the framework of pix2code to build a more robust image captioning model with different styles
  - Improved BLEU score up to 0.88 through inventive data augmentation methods and weighted loss functions

Relative Projects

- **Competition Expert (top 1%)** July 2017 - December 2017
  Kaggle
  - **Santa Gift Matching Challenge**:
    - Optimized a integer programming problem with a cubic objective for a toy matching algorithm using ORTools
    - Conducted literature study and implemented a relaxation approach to handle triplets and twins requirement
    - Reduced memory usage from more than 200G to less than 35G with non trivial arcs formation
  - **TensorFlow Speech Recognition**:
    - Implemented various convolutional neural networks (VGGNet, ResNet, etc.) on spectrogram and mel-frequency cepstrum coefficients of spoken commands to understand speech
    - Ensembled different networks and filters with bagging to improve accuracy up to 88.1%
- **Independent Project** September 2017 - October 2017
  Coursera
  - **Movie Recommender System with Hadoop**:
    - Built a movie recommender system based on item collaborative filtering using Hadoop in Java
    - Worked on preprocessing raw data and building co-occurrence matrix and rating matrix
    - Implemented MapReduce jobs including cooccurrence matrix normalization and matrix multiplication

Education

- **PhD in Computational Physics** December 2018
  University of California, Riverside *Riverside, CA, USA*
- **BSc in Physics, Zhongyao Zhao Applied Physics Elite Class** June 2013
  University of Science and Technology of China *Hefei, Anhui, China*