Caroline Zhu

(717) 552 6372 | wzhucaroline@gmail.com | linkedin.com/in/carzh/ | github.com/carzh

Experience

Software Engineer January 2023 – Present

Microsoft – ONNXRuntime *Redmond, WA*

- Facilitated scalable ML model training on limited-resource devices by expanding on-device training runtime to support ONNX models exceeding 2GB by redesigning the flatbuffers schema and modifying the C++ core.
- Designed, architected, and implemented a TypeScript API for training machine learning models in the browser.
- Presented a talk at the ONNX conference on the TypeScript ML training API.
- Finetuned Phi-3 for Q&A, developing a training script for a generative LLM and modifying an inference ONNX file for training.
- Implemented automated UI testing for Android & iOS apps for on-device testing purposes, by adding an Appium UI orchestration layer to a .NET MAUI app which was built into IPA and APK files and integrating automated testing steps into the Azure DevOps pipelines.
- Conducted performance testing for the CoreML ONNXRuntime backend on MacOS. Adjusted and wrote CoreML operators to cover the performance gaps.

Software Engineer Intern May 2022 - July 2022

Microsoft *Bellevue, WA*

- Refactored benchmarking repository for running NLP models locally and on Azure Machine Learning (AML) Compute Clusters, enhancing performance and scalability.
- Developed and maintained Dockerfiles, and contributed examples and bug fixes to the Hugging Face Optimum library, improving usability and functionality.

Software Engineering Intern July 2021 - Jan 2022

Best Buy Health *Remote*

- Led the development and launch of an Android app as the primary software engineer, ensuring high-quality performance and user experience.
- Implemented timing and tracking features for two research-focused Android apps using Vue.js, Node.js, and Java. Configured and maintained AWS servers for backend support for the apps.

Education

Northeastern University Boston, MA

Bachelor of Science in Computer Science, Minor in Interaction Design *September 2019 - May 2023*

GPA: 3.7

Relevant Coursework: Natural Language Processing, NLP & Robotics, Algorithms, Object-Oriented Design, Computer Systems, Artificial Intelligence, Database Design, Machine Learning

Projects

Synthetic Data Generation for Vision-Language-Navigation Agents | *Python* September - December 2022

- Finetuned a Pegasus model (trained on task of paraphrasing) on semantically-paired instructions dataset Room-to-Room, using the transformers library for the Pegasus model and tokenizer.
- Trained the Discrete-Continuous-VLN model on synthetically generated data then compared to baseline Discrete-Continuous-VLN model which were run in the Habitat sim.

Political Classifier | *Python* June 2021

- Classified tweets by political party based on American politics. Improved performancce metrics by roughly 15% using normalization and hyperparameter tweaking.
- Scraped and processed tweets from Twitter with a Python library. Applied numerous normalization techniques such as removing stopwords, normalizing case, and manipulating emojis and hashtags using Python.
- Employed the transformers library and multiple BERT models to process tweets into vectors. Used Tensorflow logistic regression trained on the labeled tweet vectors.

Technical Skills

Languages: Python, C++, Bash, Java, TypeScript, JavaScript, C, C#
Developer Tools: Git, Docker, Azure, AWS, BrowserStack
Libraries: transformers, pandas, NumPy, Matplotlib, protobuf, flatbuffers, ONNX, ONNXRuntime, PyTorch