Jonny Li

https://jonny.li jonny.li@mail.utoronto.ca @jonnyli1125 jonnylii

ML research engineer with strong software engineering background.

I train LLMs and ASR models to build AI voice assistants, combining my expertise in machine learning with my experience in distributed computing. I also like to keep up with the latest ML research papers when I'm not coding.

Work Experience

ML Engineer II, Research & Development - SoundHound AI Aug 2023 - Present

- Trained LLMs on instruction following and audio multi-modality for voice assistant NLU research experiments
- Led the development of LLM training & evaluation library with Ray, DeepSpeed, Accelerate, Transformers, TRL, PEFT, Kubernetes
- Developed LLM inference server (vLLM, TGI) forks in Python and Rust and deployment tools with Helm
- Trained transducer-based ASR models for domain adaptation research experiments
- Implemented ASR training/hyperparameter tuning pipelines with PyTorch, Docker, Kubernetes

Software Engineer, MLOps - SoundHound AI Oct 2021 - Aug 2023

- Created MLOps library with Python, Docker, Kubernetes, Gitlab CI/CD, Jenkins, Airflow and reduced time spent on manual tasks by 5x
- Implemented hyperparameter tuning for n-gram ASR models and realized 30% accuracy gain in production
- Fine-tuned GPT-2 with knowledge graphs and synthetic data, improving domain adaptation accuracy by 10%
- Built ETL pipelines with Apache Spark and achieved 2x speed gain over old Hadoop MapReduce pipelines
- Setup, deployed and managed on-prem HDFS and Spark cluster

SDE Intern, Backend Search Engine - Amazon Jun 2020 - Sep 2020

- Optimized backend search engine efficiency and reduced tech debt by implementing a dependency graph analyzer in Python to safely remove redundant configuration objects

Software Intern, Full Stack Web - Mitsucari Sep 2018 - Aug 2019

- Full stack web development with Ruby on Rails, PostgreSQL, jQuery, Bootstrap, SASS, and Heroku

Personal Projects

Neural Network in Numpy @ gist.github.com/jonnyli1125/1ad95073ff218d00cc4faee133f05dcc Feb 2024

- Implemented neural network, SGD, backpropagation, and training pipeline with numpy only
- Tutored ML theory and engineering to high school students with notebook as example

RNN-Transducer in Numpy @ gist.github.com/jonnyli1125/e5bab12ed6f36711c57807b7f1528f3a Oct 2023

- Implemented RNN Sequence Transducer (Graves 2012) in 100 lines of numpy with a beam search decoder

Webspeak to English Translator @ github.com/jonnyli1125/piemanese-translator Feb 2022 - Nov 2022

- Trained contrastive loss CNNs in Keras/Tensorflow for spelling correction task
- Implemented evaluation and testing framework for translation models in Python
- Implemented beam search decoder and n-gram LM biasing module from scratch

Japanese Grammatical Error Corrector @ github.com/jonnyli1125/gector-ja Apr 2021 - Jun 2021

- Implemented and trained BERT model from 2020 Grammarly research paper in Keras/Tensorflow
- Fine-tuned with synthetic data generation pipeline to improve performance on class imbalanced dataset
- Achieved 10% accuracy gain over previous state of the art model

Education

University of Toronto Graduated Jun 2021

Honours Bachelor of Science, Computer Science and Linguistics

Courses: Machine Learning, AI, Computational Linguistics, NLP, Data Structures, Algorithms, Operating Systems

Activities: UTokyo Foreign Exchange, UofT Neurotech Workshops Lead, UofT Japan Assocation Staff