Jonny Li

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

Full-stack ML engineer with strong software engineering background. Focus in language models and ASR.

Employment

ML Engineer II, Language Modeling - SoundHound AI @ Toronto, Canada — Aug 2023 - Present

- Implemented and conducted R&D experiments on supervised fine-tuning and multi-modality of LLMs for use with voice-based virtual assistants
- Lead development of distributed LLM training and evaluation infrastructure with Ray, Kubernetes, DeepSpeed, Accelerate, Transformers, and PEFT libraries
- Developed LLM inference server deployment and load testing library, along with forks of open source LLM inference servers (vLLM, TGI) in Python and Rust
- Conducted domain adaptation experiments for transducer-based streaming ASR models and implemented training/hyperparameter tuning infrastructure with PyTorch, Docker, Kubernetes

Software Engineer, Language Modeling - SoundHound AI @ Toronto, Canada — Oct 2021 - Aug 2023

- Authored from scratch large scale ETL pipelines with Apache Spark, achieved 2x speed gain over Hadoop MapReduce through Spark optimizations
- Developed from scratch Python MLOps library with unit/integration/end-to-end testing, Docker, Kubernetes, Gitlab CI/CD, Jenkins, Airflow, and reduced time spent on manual supervision/intervention tasks by 5x
- Implemented hyperparameter tuning infrastructure for n-gram based ASR models, realized 30% relative accuracy gain in production model
- Ran GPT fine-tuning experiments for ASR rescoring and implemented synthetic data generation pipelines with LLMs and knowledge graphs, achieved 10% in-domain accuracy gain

SDE Intern, Core Search Engine - Amazon @ Vancouver, Canada (Remote) — Jun 2020 - Sep 2020

Software Intern, Full Stack Web - Mitsucari @ Tokyo, Japan — Sep 2018 - Aug 2019

Personal Projects

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

- Implemented from scratch a feed-forward multi-layer neural network, SGD, backpropagation, and training pipeline with numpy only
- Tutored high school students basics of ML theory/engineering using notebook as example material

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

- Ran training experiments for contrastive learning CNN models in Keras/Tensorflow, and implemented an evaluation/testing framework for translation models
- Implemented from scratch beam search decoding with n-gram LM fusion and memory-efficient n-gram counting/inference

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

- Implemented 2020 Grammarly research paper in Keras/Tensorflow and ran BERT fine-tuning experiments, achieved 10% accuracy gain over previous state of the art model
- Implemented parallelized synthetic data generation and distributed training pipelines for larger-than-memory dataset with significant class imbalance

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