Jonny Li

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

ML engineer with strong software engineering background.

Work Experience

ML Engineer II, LLM R&D - SoundHound AI @ Toronto, Canada — Aug 2023 - Present

- Implemented and conducted experiments for audio+text multi-modality by integrating pre-trained LLMs with audio encoders
- Trained LLMs on instruction following and agentic function calling tasks for voice assistant NLU product
- Implemented LLM training and evaluation library from scratch with Ray, DeepSpeed, Accelerate, Transformers, TRL, PEFT, Kubernetes
- Developed LLM inference server (vLLM, TGI) forks for constrained/guided text generation in Python, Rust, gRPC
- Implemented LLM inference server load testing library and deployment tools with Python, Kubernetes, and Helm
- Conducted/implemented training experiments on transducer ASR models for domain adaptation with fine-tuning and LoRA

Software Engineer, ML Algorithms - SoundHound AI @ Toronto, Canada — Oct 2021 - Aug 2023

- Implemented knowledge graphs and synthetic data generation for ASR transcription reranking models with Python and ArangoDB
- Conducted fine-tuning experiments with GPT-2 for ASR transcription reranking with Tensorflow
- Implemented hyperparameter tuning algorithm for optimizing training dataset weights and realized 30% accuracy gain in production ASR model
- Created MLOps library for automating training workflows with Python, Docker, Kubernetes, Gitlab CI/CD, Jenkins, Airflow, and reduced time spent on manual tasks by 5x
- Built ETL pipelines with Apache Spark in Python/Java and achieved 2x speed gain over old MapReduce pipelines
- Implemented and trained large scale n-gram LMs for ASR in Python, Java, C++, and Bash

SDE Intern, Backend Search Engine - Amazon @ Vancouver, Canada — Jun 2020 - Sep 2020

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

Personal Projects

Vector Similarity Search Engine for RAG @ github.com/jonnyli1125/similarity-search — Jul 2024

- Implemented GPU-accelerated vector similarity search from scratch with C++, CUDA, Pybind
- Implemented embedding indexes and RAG from scratch with numpy, scikit-learn, and OpenAI

Neural Network Training 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 — Nov 2022

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

Japanese Grammatical Error Correction BERT @ github.com/jonnyli1125/gector-ja — Jun 2021

- Implemented BERT model from 2020 Grammarly paper with HF Transformers, Keras/Tensorflow, and trained with TPUs on Google Colab
- 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 on Japanese GEC evaluation dataset

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