Jonny Li

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

Experience

Machine Learning Engineer II - SoundHound AI @ Toronto, Canada — Aug 2023 - Present

- Lead the research and development of an ASR error correction LLM that rewrites and corrects entity name mistranscriptions in an ASR model's output given a list of names in the prompt, achieving up to 90% accuracy gain in transcribing contact names, menu items, etc
- Researched and applied supervised fine tuning, distillation, reinforcement learning, and synthetic data generation techniques to fine tune open source LLMs for specific tasks
- Optimized and balanced tradeoffs between accuracy, latency, throughput, and cost of production LLMs with load testing, cost analysis, and techniques such as distillation, quantization, vector DBs, RAG
- Authored cross-team distributed LLM training and evaluation library with Ray, DeepSpeed, Accelerate, Transformers, TRL, PEFT, Kubernetes
- Researched, implemented, and trained multi-modal speech/text LLMs with neural speech adapters and speech tokenization methods
- Implemented data loading pipelines for streaming large amounts of audio and text data from S3 and applying async/lazy transformations on IO-bound datasets using Boto3, PyTorch, Torchaudio
- Developed in-house Realtime API service with support for custom ASR+LLM+TTS stack or multimodal LLM, leading to significant cost savings over OpenAI API and improved voice agent product experience
- Developed forks of LLM inference servers like vLLM and TGI for task-specific constrained generation in Python and Rust
- Implemented and trained state of the art streaming ASR models in PyTorch

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

- 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 large scale knowledge graphs and synthetic data generation pipelines with Python and ArangoDB
- Fine-tuned GPT-2 for ASR transcription rescoring/reranking with Tensorflow and PyTorch
- Implemented and trained large scale n-gram LMs for ASR in Python, Java, C++, and Bash

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

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

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 from scratch with numpy only
- Tutored basic ML theory and engineering to high school students

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