Jonny Li

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

ML engineer with strong software engineering background. Focus in NLP, language models, handling large datasets with distributed systems, model serving/inference optimization, and MLOps/DevOps.

Employment

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

- Wrote LLM inference server load testing library from scratch with Python, Kubernetes, and Docker to measure and compare latency/throughput of various open source libraries.
- Implemented custom forks of open source LLM inference servers, such as huggingface/text-generation-inference (Rust, PyTorch, HF/transformers) and vLLM (Python, HF/transformers).
- Implemented novel second pass ASR hypothesis rescoring mechanism with instruction fine tuned LLMs such as T5 and Llama 2.

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

- Performed domain adaptation of GPT-2 ASR rescoring model by fine tuning with synthetic data generated from knowledge graph + LLM pipeline, using ArangoDB, HF/transformers, and Tensorflow.
- Implemented Apache Spark library from scratch for processing terabyte scale datasets and billion parameter scale n-gram LMs on distributed clusters, and reduced compute times by 2x over original codebase.
- Improved accuracy by relative 30% in n-gram LM-based ASR model with ML techniques such as hyperparameter tuning with Bayesian optimization, bagging, boosting, etc. using open source libraries such as scikit-learn, Hydra, nevergrad.
- Implemented perplexity/entropy-based n-gram pruning technique in Apache Spark to reduce parameter count by 100x while maintaining negligible loss in accuracy.
- Implemented Python library from scratch for invoking training workflows of large scale n-gram LMs, integration with Docker/Kubernetes, automated unit/integration testing, CI/CD pipelines, and automated scheduling with Airflow.

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

RNN-Transducer in 100 lines @ gist.github.com/jonnyli1125/e5bab12ed6f36711c57807b7f1528f3a — Oct 2023

- Implemented the RNN Sequence Transducer (Graves 2012) in 100 lines of numpy, including a beam search decoder.

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

- Implemented and trained character-level CNN with contrastive learning in Keras for recognizing similarly spelled words.
- Implemented beam search decoder with n-gram language model to create statistical/neural hybrid translation model.
- Implemented synthetic data generation pipelines based on 2019 grammatical error correction paper to overcome lack of training data.
- Implemented Discord chat bot interface using Discord API, and deployed to cloud with Docker and Heroku.

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

- Implemented state of the art BERT-based grammatical error correction model with 10% gain over previous best model.
- Reproduced 2020 Grammarly research paper in Tensorflow/Keras and Huggingface transformer modules.
- Implemented scalable, memory-efficient, synthetic data generation and training pipelines for large datasets that don't fit in memory, with Tensorflow and GCP + Colab.
- Created interactive web app demo using Python/Flask and HTML/CSS/JavaScript.

MCTS + Transformer-based Chess Engine @ github.com/jonnyli1125/chess-bert-mcts — Jun 2021

- Implemented an AlphaZero-inspired, experimental chess engine with multi-threaded Monte Carlo Tree Search and BERT for policy/value networks, using HF/transformers, PyTorch, and PyTorch Lightning for model implementation/training.

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