Jonny Li

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

Machine Learning Engineer. Focus in ASR (automatic speech recognition), language modeling, distributed computation, MLOps. Strong background in software engineering, algorithms & DS, full stack web development.

Employment

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

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

- Implemented distributed cluster ETL pipelines with Apache Spark from scratch. Technical debt cleanup of 10+ year old legacy Hadoop MapReduce library to optimize compute times for large scale datasets. On-prem Hadoop and Spark cluster setup, maintenance, DevOps.
- Implemented model training and pipelining Python library from scratch for building large scale statistical n-gram LM using Hydra, DVC, and open source + internal n-gram LM libraries. Applied and combined modern ML techniques (hyperparameter tuning, ensemble methods, etc) with legacy n-gram LM research for significant accuracy improvement in WFST-based classical ASR systems.
- Implemented automation for model training pipelines using Airflow, Docker/k8s, Jenkins, Gitlab CI/CD. Introduced standard techniques in software engineering to internal ML codebases, like readable/user-friendly API design, code quality control, unit/integration testing, and CI/CD.
- Implemented knowledge graph DB + LLM-based text generation pipelines for low-resource domain synthetic data generation. Fine tuned smaller, domain-specific GPT models used for ASR rescoring.

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

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

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

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

Projects

LLM Docs Assistant @ github.com/jonnyli1125/docs-assistant — May 2023

- Created a Slack chatbot assistant to search Confluence documentation wikis with vector DB semantic search + LLM embeddings using OpenAI API, LangChain, llama.cpp, ChromaDB, and Slack API.

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

- Created statistical + neural hybrid machine translation model for translating webspeak to English using Python and Keras.
- Implemented a similarity learning-based neural model for recognizing similarly spelled words, memory efficient n-gram language model optimized for low latency, and beam search decoder, all from scratch.
- Implemented synthetic data generation methods for neural model training on low resource webspeak domain.
- 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 Google Cloud + Colab.
- Created interactive web app demo using Python/Flask and HTML/CSS/JavaScript.

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 Webmaster

Skills

Programming Languages: Python, Java, Bash, JavaScript

Technologies/Frameworks: Apache Spark, Hadoop MapReduce, Docker, Kubernetes, Airflow, Jenkins, Gitlab CI/CD, Keras/Tensorflow, PyTorch, scikit-learn

General: MLOps, DevOps, NLP, Machine Learning, Distributed Computing