Khoi Nguyen

nguyentienkhoi01@gmail.com | 438-493-3478 | Montreal, Canada
khoitiennguyen.github.io | KhoiTienNguyen | khoi-tien-nguyen

Education

McGill University — Montreal, Canada
B.Sc. COMPUTER SCIENCE, GPA: 3.51/4.0 — *Sep 2019 - May 2023 (expected)*

- Awards: Major Entrance Scholarship
- Courses: Intro to Data Science, Applied Machine Learning, Artificial Intelligence, Intelligent Software Systems, Natural Language Understanding with Deep Learning, Reinforcement Learning, Algorithms and Data Structures

Work Experience

McGill University, Distributed Digital Music Archives & Libraries Lab — Montreal, Canada
SOFTWARE DEVELOPER — *May 2022 - Present*

- Build workflow allowing users to label image layers and train optical recognition models to separate them.
- Develop essential image processing functionalities using **Python** to prepare training data for the models.
- Work closely with researchers to apply iterative changes to the modelling process and perform experiments.
- Design and implement a dynamic data loading scheduler to improve image loading efficiency and speed up model training by three times.

Accreon Inc. — Fredericton, Canada
CO-OP STUDENT — *Sep 2018 - Jan 2019*

- Implemented new functionalities in the ETL pipeline using **Java** for more flexible data processing and retrieval.
- Extracted client information from a central database and processed it as compressed JSON files to improve data storage and data sharing.

Skills

Programming Python · Java · OCaml · C · Bash · JavaScript · HTML · CSS · Assembly · SQL
Technologies TensorFlow · PyTorch · HuggingFace · Scikit-learn · NumPy · Pandas · OpenCV · Git · Docker · AWS

Projects

For more projects, visit https://khoitiennguyen.github.io

Automatic Music Transcription

- Converted instrumental audio to Mel spectrograms and tokenized music symbols for model training using **Python**.
- Experimented with various custom **CRNN** model architectures in **Keras** using CTC loss for automatic transcription.
- Trained model across seven instruments, achieving less than 10% character error rate on four instruments.

Movie Recommendation System

- Developed a pipeline to continuously collect data from a **Kafka** stream, pre-process the data, and retrain models.
- Trained an **XGBoost** model with hand-crafted features and **SVD** collaborative filtering, achieving 0.15 RMSE.
- Trained a **K-Means clustering** model with user profiles using **scikit-learn** to mitigate cold start problem for new users.
- Implemented CI/CD with unit-testing, offline and online evaluations using **GitHub Actions**. Performed load-balancing with canary releases using **Docker** and **Nginx**. Monitored the system using **Prometheus** with **Grafana**.

Data Size Probing

- Fine-tuned **BERT** on GLUE and SuperGLUE tasks with specific data size intervals using **HuggingFace** and **PyTorch**.
- Evaluated each model and probed them with tasks from SentEval benchmark to inspect the semantic and syntactic information encoded in the models using **TensorFlow**.
- Iteratively fine-tuned models on different tasks to transfer knowledge and assess linguistic recoverability.