Tran Duc Trung

Email: trung1803lucky@gmail.com
Phone: +(84) 325647395
Github | Linkedin

1. Goals and objectives

I love coding and exploring something new, big data and AI are my interests domain. Over the next three years, my ambition is to launch my career as a data engineer while delving deeper into the field of big data and AI. My goal is to master the complexities of data engineering, while also expanding my knowledge and skills in AI and ML. With dedication and ongoing learning, I expect to transition into roles like MLE or MLOps, where I can leverage my expertise to build and deploy models in real-life pipelines.

2. Projects

Data Engineering:

ETL Data Pipeline For Trip Record Oct – Dec. 2023

- Motivation: Build an ELT data pipeline with the TLC Trip Record Data. Helping taxi businesses acquire sufficiently good and clean data before conducting statistical analysis or building models for prediction, thus deriving insights to enhance taxi services based on observed patterns within the data.
- Github: NYC-TripRecord with #Demo
- Tasks:
  - Ingest data into database.
  - Develop end-to-end data platform following Lambda architecture by build ETL pipeline, batch processing with Apache Spark.
  - Transfer data to data warehouse.
  - Make some analysts and demo data with a web app written by Streamlit.
- Technologies: Docker, Dagster, Apache Spark, MySQL, MinIO, PostgreSQL, Streamlit.

AI / Machine Learning:

Visual Question Answering in Medical VQA-RAD data June – July. 2024

- Motivation: Developing a medical VQA system aims to enhance diagnostic accuracy and efficiency by providing instant, precise answers to questions based on medical images, which supports healthcare
- Github: Project-NLP-VQA
- Task:
  - Applied MUMC (SOTA in 2023) for VQA-Rad. Focus on understanding the relationships between different approaches (image-text, image-image, text-text)
  - In MUMC, there are 2 phases, pre-training phase and fune-tuning phase
  - Train both pre-training phase (as Image Captioning) and training phase (as VQA)
  - In this model will has image encoder leverages a 12-layer Vision Transformer (ViT), text encoder leverages the first 6 layers of pre-trained BERT, multimodal encoder is the last 6 layers of BERT and answering decoder is a 6-layer transformer-based decoder
- Technologies: Pytorch, NLP knowledge

House Price Prediction May – June. 2023

- Github: House-Price-Prediction
- Tasks:
  - Using beautifulSoup and regex to crawl data on web
  - EDA, cleaning and preprocessing data (using MICE, one-hot encoding, scaling, remove outlier)
  - Build and compare to choose best model for predict house price (Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, Random Forest, CatBoost, XGBoost, Stacking model)
- Technologies: beautifulSoup, regex, scikit-learn, matplotlib, streamlit

3. Education

VNU-HCM University of Science
2021 – 2025 (Expected Graduation)

- Bachelor of Data Science
- GPA: 3.505/4.0

4. Skills & Coursework

- **Programming Languages:** Python, C/C++, R, Matlab.
- **My tech stack:** Basic Python libraries (like Numpy, Pandas, Matplotlib, Seaborn, Streamlit ...), Scikit-learn, Pytorch, Spark, Dagster, dbt, MinIO, Docker, Linux (Ubuntu).
- **Database**: SQL (Microsoft SQL Server, PostgreSQL, MySQL), NoSQL (MongoDB).
- **Relevant School Coursework:** DSA, OOP, Probability and Statistics, Discrete Math, Databases, Database Management Systems, Intro to AI, Data Mining, Machine Learning, Pattern Recognition, NLP, BigData.
- **Languages:** English (intermediate).

5. Activities & Certifications

- Fundamental Data Engineering — at #AIDE Institute
- VIASM The Summer School In Data Science 2023 — at #VNU-HCMUS
- Google Cloud Skills Boost — at #QuanQuanGCP
- HackerRank SQL (Basic to Advanced) Skills Certifications — #HackerRank