vmhieu17@gmail.com
linkedin/vmhieu
Hanoi, Vietnam

Hieu (Ian) Vu

AI Research Engineer / Applied Scientist

gscholar/Hieu M. Vu
github/lone17
leetcode/lone17

Summary

- Researcher/Engineer in AI/ML.
- Experienced in Document Image Understanding (Layout Analysis, OCR, Information Retrieval, Information Extraction, Language Modelling).
- Published papers at peer-reviewed conferences, including 1 Best Paper Award at DIL-ICDAR 2021.
- Bachelor of Computer Science (Honours Programme) - Highest Ranking Graduate (1st/600+).

Experience

AI Research Engineer **Nov 2018 — Present**

Cinnamon AI

- **Develop RAG-based Question Answering solutions for Insurance Domain.**
  - Related skills: Large Language Models (LLMs), Retrieval Augmented Generation (RAG), Prompt Engineering, Information Retrieval, CI/CD.
  - Related tools: LangChain, LlamaIndex, HuggingFace, ChromaDB, Github Action, Docker, FastAPI, Pydantic.
- **Research, develop, and implement solutions for Visually-rich Documents** Understanding of low-resource languages.
  - Publish papers on Information Extraction at peer-reviewed conferences, including 1 Best Paper Award at DIL-ICDAR 2021.
  - Related skills: Research, Deep Learning, Training/Pre-training/Fine-tuning, Image Processing, Computer Vision, Natural Language Processing.
  - Related tools: Pytorch, Tensorflow/Keras, transformers, OpenCV, Scikit-learn, LaTeX, DVC, CircleCI, Docker.
- **Internal CLI tool for data management.**
  - Focused on data and labels for Visually-rich documents.
  - Labelling schema standardization, data life-cycle, upstream-local data synchronization,
  - Related skills: Data Management, Data-driven Development, Software Engineering.
- Other responsibilities: Supporting cross-department alignment; Conducting/Facilitating technical sharing sessions; Training/Mentoring.

Undergraduate Research Assistant **Aug 2017 — Sep 2019**

IOT Lab, UET - Vietnam National University

- **Predictive models for geophysical data using machine learning and statistical methods.**
  - Facies/rock type classification; Time-series Analysis; Permeability Regression; Integrated Prediction Error Filter Analysis (INPEFA) curve calculation; Cumulative and Federated Learning.
  - Related skills: Data Science, Machine Learning, Time-series Analysis.
  - Related tools: Python, Keras/Tensorflow, OpenCV, Scikit-learn, XGBoost, Javascript.

Education

Bachelor Degree, Computer Science (Honours Programme) **GPA: 3.83/4.00**

UET - Vietnam National University, Hanoi *Rank: 1st/600+*

- **Highest Ranking Graduate.**

Skills

Techincal Fields	**Information Extraction, Document Understanding**, Data Science, Natural Language Processing, Computer Vision.
ML/AI Development	**Pytorch, Transformers, RAG**, LangChain, LlamaIndex, Tensorflow/Keras, Scikit-learn, OpenCV.
Software Development	**Git, Github Action**, CircleCI, DVC, Docker.
Programming Languages	**Python**, C/C++, Java, Shell Script.
Industrial Domains	Insurance, Manufacturing.
Environments	GCP, AWS, Linux, Windows.
Natural Languages	Vietnamese (native), English (fluent), Japanese (JLPT N4).
Misc	Problem Solving (Leetcode), Attentive to detail, Presentation, Communication, Research (Google Scholar).

Publications

1 Nguyen, Dat, **Hieu M Vu**, Cong-Thanh Le, Bach Le, David Lo, and Corina Pasareanu (2024). "Inferring Properties of Graph Neural Networks". In: *arXiv preprint arXiv:2401.03790.*

2 Nguyen, Bao-Sinh, Dung Tien Le, **Hieu M Vu**, Tuan-Anh D Nguyen, Minh-Tien Nguyen, and Hung Le (2022). "Improving Document Image Understanding with Reinforcement Finetuning". In: *International Conference on Neural Information Processing.* **Oral presentation**. Springer, pp. 51–63.

3 Son, Nguyen Hong, **Hieu M Vu**, Tuan-Anh D Nguyen, and Minh-Tien Nguyen (2022). "Jointly Learning Span Extraction and Sequence Labeling for Information Extraction from Business Documents". In: *2022 International Joint Conference on Neural Networks (IJCNN).* **Oral presentation**. IEEE, pp. 1–8.

4 Nguyen, Tuan-Anh D, **Hieu M Vu**, Nguyen Hong Son, and Minh-Tien Nguyen (2021). "A Span Extraction Approach for Information Extraction on Visually-Rich Documents". In: *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16.* **Best Paper Award**. Springer, pp. 353–363.

5 **Vu, Hieu M** and Diep Thi-Ngoc Nguyen (2020). "Revising FUNSD dataset for key-value detection in document images". In: *arXiv preprint arXiv:2010.05322.*

Personal Projects

Question Answering on PDF documents **Jan 2024 — Present**

Open-source project

- A Retrieval Augmented Generation (RAG) application for question answering on PDF documents.
- Including a complete pipeline from PDF parsing to indexing, retrieval and answer generation; a FastAPI backend and a chat interface.

Data Utility Improvement Experiment for DECAF **Oct 2022 — Nov 2022**

Personal research

- A personal research on Causal Inference, Algorithmic Fairness and specifically the paper DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative networks.
- Improved data utility of the DECAF method using alternating graph during synthesis while still achieving similar level of fairness.

Gender/Accent Classification for Vietnamese short voice recordings **Aug 2018 — Sep 2018**

Zalo AI Challenge 2018

- Classify the speaker's voice in a recording (typically under 3 seconds) by gender and regional accent.
- **4th place** on the Private Leaderboard, achieved **79.208% accuracy** in 10 days as an individual participant.

Electric Meter OCR **Oct 2019 — Nov 2019**

University Coursework Project

- Develop a solution for extracting the value on the dial from images of electric meters. The solution is meant to be used in embedded hardware.
- Achieved **0.08 on edit distance** with total code size **under 10MB** and processing time **under 0.3s/image**.

Honours and Awards

Best Paper Award **Sep 2021**

ICDAR 2021, Workshop on Document Images and Language

Paper title: A Span Extraction Approach for Information Extraction on Visually-Rich Documents

Accepted for oral presentation and awarded the Best Paper Award at Workshop on Document Images and Language, ICDAR 2021.

Top 4 Zalo AI Challenge 2018 - Voice Track (Individual participant) **Sep 2018**

Zalo, VNG Corporation

Finished at 4th place on the Private Leaderboard of the Voice Gender/Accent Classification challenge.

Zalo AI Challenge is an annual Kaggle-like competition hosted by Zalo - one of the biggest tech companies in Vietnam. In 2018, the competition attracted over 700 teams competed in 3 challenges.

Certificate of Highest Ranking Graduate

UET - Vietnam National University

Awarded to students who graduated with the highest GPA amongst the graduating class.