vmhieu17@gmail.com
linkedin/vmhieu
Hanoi, Vietnam

Hieu Vu

AI/ML Research Engineer

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

Summary

- Researcher/Engineer in AI/ML with 6+ years of experience.
- Experienced in Information Extraction/Retrieval, Language Modelling and Document Image Understanding.
- Published papers at peer-reviewed conferences, including a Best Paper Award at DIL-ICDAR 2021.
- Bachelor of Computer Science (Honours Programme) - Highest Ranking Graduate (1st/600+).

Experience

AI Engineer **July 2024 — Present**
Castalk

- **Building virtual AI companions.**
  - Core chat bot with memory and tool use.
  - User's engagement detection with head-pose tracking and emotion detection.

AI Engineer/Researcher **Nov 2018 — May 2024**
Cinnamon AI *Hanoi, Vietnam*

- **Developing RAG-based applications.**
  - Co-creator of **kotaemon**: An open-source tool for local RAG application built for both end users and developers.
    - An easy-to-use local application for end users to chat with their documents.
    - A framework for developers to use RAG pipelines.
  - Built demos for LLM-powered applications targeting the Insurance domain.
- **Researched, developed, and implemented AI solutions for Document Image Understanding.**
  - Published papers at peer-reviewed conferences, including 1 Best Paper Award at DIL-ICDAR 2021.
  - Key Information Extraction on document images low-resource languages:
    - Implemented MVLM pre-training task for LayoutLM (and variants).
    - Cross-lingual adaptive pre-training for a low-resource language (Japanese).
    - Created new technical assets by introducing new Information Extraction models that became the new standard for client projects. Increased the f1-score by 2% - 7%.
  - Document Image Classification, over 85% accuracy achieved on a client data set with 20+ classes.
  - Other: Document Segmentation; Document Object Detection (logos, stamps, check marks, etc.); Data Synthesis/Augmentat (Image Processing based); Text Segmentation; printed/handwriting OCR.
- **Developed data-driven products and processes.**
  - Worked on Data Management CLI tool: Synchronization and local version control, used by AI Engineers and Researchers to query data from a central database and manage the local copy.
  - Roadmap planning for data-related objectives: lead discussions, identify issues, propose solutions, decide action items for data centralization, data management, data integrity, labeling UI improvement, etc.
  - Initiated data standardization: defined and implemented processes regarding data life cycle and organization, enabling datasets from different client to be re-useable collectively.
  - Related skills: Data Management, Label Schema Design, Data-driven Development, Software Engineering.
- Other tasks: Supporting cross-department alignment; Conducting/Facilitating technical sharing sessions; Training/Mentoring.

Undergraduate Research Assistant **Aug 2017 — Sep 2019**
IOT Lab, UET - Vietnam National University, Hanoi *Hanoi, Vietnam*

- **Developed the Machine Learning Toolkit of an online wellbore data interpretation and management platform.**
  - Built predictive models for geophysical data using machine learning and statistical methods.
  - Worked in conjunction with domain experts and FE/BE engineers to ensure requirements are met.
  - Problem worked on: Facies/rock type classification; Time-series Analysis; Permeability Regression; Integrated Prediction Error Filter Analysis (INPEFA) curve calculation; Cumulative and Federated Learning.

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*.

Education

Bachelor Degree, Computer Science (Honours Programme) **GPA: 3.83/4.00**
UET - Vietnam National University, Hanoi *Rank: 1st/600+*

- **Highest Ranking Graduate.**
- Merit for Excellent Graduation Certificate recipient.
- Excellent Thesis Defence Certificate recipient.
- Excellent Student Certificate recipient (3 times).
- Academic Encouragement Scholarship recipient (4 times).

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.

Skills

Techincal Fields	**Information Extraction, Document Understanding, LLMs**, NLP, Computer Vision, Image Processing, Time-series Analysis.
ML/AI Development	**Pytorch, Transformers, RAG**, LangChain, Haystack, LlamaIndex, Tensorflow/Keras, Scikit-learn, OpenCV.
Software Development	**Git, Github Action**, Docker, CircleCI, DVC.
Programming Languages	**Python**, C/C++, Java, Shell Script.
Industrial Domains	Insurance, Manufacturing, Virtual Companion.
Environments	GCP, AWS, Linux, Windows/WSL.
Natural Languages	Vietnamese (native), English (IELTS 7.5), Japanese (JLPT N4).
Other	Research (Google Scholar), Problem Solving (Leetcode).

Personal Projects

Kotaemon - An open-source tool for local RAG application. **Jan 2024 — May 2024**
Open-source project, Co-creator

- A local RAG-based tool for chatting with your documents. Built with both end users and developers in mind.
- For end users: A local Question Answering UI for RAG-based QA.
- For developers: A framework for building your own RAG-based QA pipeline.

A local application for chatting PDF documents **Jan 2024 — Feb 2024**
Personal project

- A Retrieval Augmented Generation (RAG) application for question answering on PDF documents.
- Complete pipeline from PDF parsing to indexing, retrieval and generation.
- Instead of continue working on this, I moved to build kotaemon.

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.
- Conducted experiments on improving data utility of the DECAF method using alternating graph during synthesis while still achieving similar level of fairness.
- Gave discussion and suggestions on the choice of data utility metrics.

Channel-invariant Deformable Convolution **Feb 2020 - May 2020**
A part of my Undergrade Thesis

- A modified version of Deformable Convolution where the convolution offsets stay the same for all channels.
- Sped up the Deformable Convolution operation by an order of magnitude while still achieving similar performance.

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

- Problem description: Classify the speaker's voice in a recording (typically under 3 seconds) by gender (male/female) and regional accent (northern/central/southern).
- **4th place** on the Private Leaderboard, achieved **79.208% accuracy** within 10 days as an individual participant.
- About the competition: 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.

Training and Mentoring

Cinnamon AI Bootcamp 2020, 2022, 2023
Teaching/Mentoring

- Mentored groups of 3-4 students.
- Designed syllabus, prepared entrance tests, interviewed candidates.
- Prepared materials and gave lectures on Language Modelling and Transformers.