vmhieu17@gmail.com
linkedin/vmhieu
Hanoi, Vietnam

Hieu (Ian) Vu

AI Research Engineer / Applied Scientist

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

Summary

- Working in AI/ML since 2018 with focus on Information Extraction and Document Understanding.
- 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 — *Hanoi, Vietnam*

- **Developing LLMs-powered applications for Insurance Domain.**
  - Developing an internal framework for Large Language Model (LLM) pipeline building: allowing users to build workflows that consists of LLM, Prompts, Vector Stores, Indexers and Retrievers, Agents, Output Parsers, etc.
  - Built demos for LLM workflows targeting the Insurance domain.
  - Related skills: Large Language Models, Prompt Engineering, Information Retrieval, Software Engineering, Business-oriented Development.
  - Technologies used: OpenAI, Cohere, LangChain, LangFlow, Github Actions, HuggingFace, LlamaIndex, Haystack.
- **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).
    - Adapted the English pre-trained weights to a low-resource language (Japanese).
    - Pre-trained LayoutLM-based models for the Japanese language and performed fine-tuning on several client data sets, 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/Augmentation (Image Processing based); Text Segmentation; printed/handwriting OCR.
  - Related skills: Research, Training/fine-tuning, Language Model Pre-training, Image Processing, Computer Vision, Natural Language Processing.
  - Technologies used: Python, Pytorch, Tensorflow/Keras, Transformers (Hugging Face), OpenCV, Scikit-learn, LaTeX, DVC, Neptune, CircleCI, Docker.
- **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.
- Related skills: Data Science, Machine Learning, Time-series Analysis.
- Technologies used: Python, Keras/Tensorflow, OpenCV, Scikit-learn, XGBoost, Javascript.

Skills

Programming	**Python**, C/C++, Java, Shell Script.
ML/AI Technologies	**Pytorch, Transformers (Hugging Face), RAG**, LangChain, Tensorflow/Keras, Scikit-learn, OpenCV.
Tools and Technologies	**Git, Github Action**, CircleCI, DVC, Docker.
AI Domains	**Information Extraction, Large Language Models, Document Understanding**, Data Science, Natural Language Processing, Computer Vision.
Environments	GCP, AWS, Linux, Windows.
Languages	Vietnamese (native), English (fluent), Japanese (JLPT N4).
Misc	Problem Solving (my Leetcode profile), Attentive to detail, Presentation, Communication, Academic Research (my Google Scholar profile).

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 **Sep 2018**
Zalo, VNG Corporation (Individual participant)
Finished at 4th place on the Private Leaderboard of the Voice Gender/Accent Classification challenge.
Zalo AI Challenge 2018 is a Kaggle-like competition hosted by Zalo - one of the biggest tech companies in Vietnam. The competition attracted over 700 teams competed in 3 challenges.

Certificate of Highest Ranking Graduate
University of Engineering and Technology - Vietnam National University, Hanoi
Awarded to students graduate with the highest GPA amongst the graduating class.

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.
- Complete pipeline from PDF parsing to indexing, retrieval and generation.

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

- Studied about 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 synthesization while still achieving similar level of fairness.
- Gave discussion and suggestions on the choice of data utility metrics.

Voice Gender/Accent Classification **Aug 2018 — Sep 2018**
Zalo AI Challenge 2018

- **4th place** on the Private Leaderboard, achieved **79.208% accuracy**.
- Individual participant, participated only during the last 10 days/1 month+ of the competition.
- Problem description: Classify the speaker's voice in a recording (typically under 3 seconds) by gender(male/female) and regional accent (northern/central/southern).
- About the competition: Zalo AI Challenge 2018 is a Kaggle-like competition hosted by Zalo - one of the biggest tech companies in Vietnam. The competition attracted over 700 teams competed in 3 challenges.

Electric Meter OCR **Oct 2019 — Nov 2019**
University Coursework Project

- Achieved **0.08 on edit distance** while having the size of just **under 10MB** and processing time of under **0.3 seconds** per image on a normal laptop.
- Problem description: Extract the value on the dial from images of electric meters. The solution is meant to be used in embedded hardwares.

Outreach

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.