

Hieu (Ian) considers himself an individual who is intellectually curious, passion-driven and eager to better himself. He has been working in the industry for over 5 years, his work consists of both conducting research and developing solutions to suit the needs of clients. He's deeply interested in working with AI/ML and is open to any interesting topic/problem. He is looking for opportunities where he could stay current with the literature, develop and implement AI solutions that people enjoy using.

PUBLICATIONS

- 1 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.
- 2 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.
- 3 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, DIL - ICDAR 2021**. Springer, pp. 353–363.
- 4 **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) Sep 2016 — Aug 2020
University of Engineering and Technology - Vietnam National University, Hanoi GPA: 3.83/4.00

- **Highest Ranking Graduate.**

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.

Certificate of Highest Ranking Graduate Aug 2020
University of Engineering and Technology - Vietnam National University, Hanoi
Awarded to students graduate with the highest GPA amongst the graduating class.

Certificate of Merit for Excellent Graduation Aug 2020
Vietnam National University
Awarded by the President of Vietnam National University, Hanoi to students with excellent academic performance and level of conduct during a 4-year undergraduate programme.

Certificate of Excellent Thesis Defence Aug 2020
University of Engineering and Technology - Vietnam National University, Hanoi
Awarded to the best thesis of the Undergraduate Thesis Defence Committee.
Thesis title: A Layout-aware key-value relation predicting model for document images.

Certificate of Excellent Student 2016 - 2020
University of Engineering and Technology - Vietnam National University, Hanoi
Awarded annually to students with excellent academic performance and level of conduct.
3 times awarded.

Academic Encouragement Scholarship 2016 - 2020
University of Engineering and Technology - Vietnam National University, Hanoi
Awarded to top 8% student every semester on the basis of academic performance and extracurricular activities.
4 times awarded, 6 times nominated.

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.

EXPERIENCE

AI Research Engineer

Nov 2018 — Present

Cinnamon AI

Hanoi, Vietnam

- **Developing LLMs-powered applications for Insurance Domain.**

- Developing a platform for efficient 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.
- Building demo for LLM workflows targeting the Insurance domain.
- Building mock UI for pipeline visualization and execution.
- Working closely with Sale/BizDev teams to shape the product.
- 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.**

- Co-authored 3 peer-reviewed papers on Key Information Extraction.
- 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, L^AT_EX, 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, University of Engineering and Technology - 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.

PERSONAL PROJECTS

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 synthesis 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 hardware.

SKILLS

Programming	Python , C/C++, Java, Shell Script.
ML/AI Technologies	Pytorch , Transformers (Hugging Face) , Langchain , Langflow , Tensorflow/Keras, Scikit-learn, OpenCV.
Tools and Technologies	Git , Github Action , CircleCI, DVC, Docker.
AI Domains	Information Extraction , Large Language Models , Document Intelligence, 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 (https://leetcode.com/lone17 Leetcode profile), Attentive to detail, Presentation, Communication, Academic Research (Google Scholar profile).

OUTREACH

Cinnamon AI Bootcamp 2023

Mentoring

- Mentored a group of 4, which won the 3rd prize for their graduation project.
- 2 students received internship offers (out of 5 total offers).
- (These kids gave me serious peer-pressure).
- Design syllabus, prepared entrance tests.

Cinnamon AI Bootcamp 2022

Mentoring

- Mentored a group of 4, which won the 2nd prize for their graduation project.

- (These kids are amazing, far more gifted than I could ever be).
- Design syllabus, prepared entrance tests, interviewed candidates.

Cinnamon AI Bootcamp 2020

Teaching/Mentoring

- Mentored a group of 3.
- Gave lectures on Language Modelling and Transformers to a class of 15 students, most of them hold a Master degree or are Master students.
- Contributed to syllabus design, interviewed candidates.