
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) **2020**
UET - Vietnam National University, Hanoi *GPA: 3.83/4.00 (Rank: 1st/600+)*
• **Highest Ranking Graduate.**

HONOURS AND AWARDS

Best Paper Award **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 **2020**
UET - Vietnam National University
Awarded to students graduate with the highest GPA amongst the graduating class.

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

Certificate of Excellent Thesis Defence **2020**
UET - Vietnam National University
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**
UET - Vietnam National University
Awarded annually to students with excellent academic performance and level of conduct.
3 times awarded.

Academic Encouragement Scholarship **2016 - 2020**
UET - Vietnam National University
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 (Individual participant) **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.

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.

SKILLS

Technical 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).
Other	Research (Google Scholar), Problem Solving (Leetcode).

EXPERIENCE

AI Engineer/Researcher

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 (IR), CI/CD.
 - Related tools: LangChain, LlamaIndex, HuggingFace, ChromaDB, Github Action, Docker, FastAPI, Pydantic.
- **Research, develop, and implement solutions for Visually-rich Documents Understanding on low-resource languages.**
 - Information Extraction and 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%.
 - Publish [papers](#) on Information Extraction at peer-reviewed conferences, including 1 Best Paper Award at DIL-ICDAR 2021.
 - Related skills: Research, Deep Learning, Image Processing, Computer Vision, Natural Language Processing.
 - Related tools: Pytorch, Tensorflow/Keras, Transformers, OpenCV, Scikit-learn, \LaTeX , DVC, CircleCI, Docker.
- **Internal data management for Visually-rich documents.**
 - Implement synchronization and local version control for the internal data management system.
 - Worked on labelling schema standardization, data life-cycle.
 - Related skills: Data Management, Data-driven Development, Software Engineering.
- Other tasks: Support cross-department alignment; Conduct/Facilitate technical sharing sessions; Training/Mentoring.

Undergraduate Research Assistant

Aug 2017 — Sep 2019

IOT Lab, UET - Vietnam National University

- **Predictive models for wellbore 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.

PERSONAL PROJECTS

Question Answering on PDF documents

Jan 2024 — Present

Personal 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



- 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.

Electric Meter OCR

Oct 2019 — Nov 2019

University Coursework Project



- Extract 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**.