Thinh Dao

+84 343 952 570 | 21thinh.dd@vinuni.edu.vn | LinkedIn | GitHub | Google Scholar | Personal website

Education

Vin University — Ha Noi, Viet Nam

B.Sc. in Computer Science; ***CGPA: 3.68/4.00 (Core GPA: 3.8/4.00)*** — *Sep 2021 – Jun 2024*

Relevant coursework: Machine Learning, Artificial Intelligence, Data Mining, Linear Algebra, Calculus, Advanced Probabilities and Statistics, Algorithm Design, Object-oriented Programming.

Academic & Work Experience

BIDV Data Center — Ha Noi, Viet Nam

AI Engineer Apprentice (on-contract) — *May 2025 – Aug 2025, Full-time*

The technology hub for BIDV (Bank for Investment and Development of Vietnam), one of the largest banking institutions in Vietnam with over 25 million customers.

- Developed an advanced Mixed Raster Compression (MRC) compression system for PDF documents, achieving **85–90%** size reduction in internal files and surpassing a leading commercial baseline ABBYY, with projected annual cost savings of **$700,000** at BIDV.
- Fine-tuned Tesseract OCR models on a custom-built dataset of private bank documents, **reducing word error rates** on Vietnamese text and handwriting by **20%** on a private benchmark.
- Architected and built an end-to-end document processing pipeline using FastAPI, integrating the custom MRC and OCR models to convert scanned images into searchable, PDF/A-compliant documents for A/B testing.
- Benchmarked state-of-the-art OCR technologies (including Surya and VLMs like Qwen2-VL), leading to the development of a superior GPU-based pipeline that outperformed the Tesseract baseline on CPU-only device.

FPT.AI (a division of FPT Corporation) — Ha Noi, Viet Nam

AI Engineer Intern — *Jan 2025 – May 2025, Full-time*

FPT.AI is the largest enterprise AI solutions provider in Vietnam, developing platforms for major international clients like Millipore Sigma.

- Developed a content-based image retrieval system to serve a key pharmaceutical client (Millipore Sigma), achieving 98% accuracy in matching new product images against a database of 8,200+ CAD diagrams.
- Engineered the core of the solution by using CLIP-based embedding model to generate highly discriminative, fine-grained similarity scores for complex industrial parts.
- Built and deployed a full-stack proof-of-concept application using FastAPI for the backend API, Streamlit for the interactive UI, and integrated with Azure S3 for scalable image storage and retrieval.

Security and Artificial Intelligence Lab @ VinUniversity — Ha Noi, Viet Nam

Research Assistant (Volunteer/Part-time/Contractual Full-time) — *Dec 2021 – Present, Part-time*

- Authored two research papers on Clean-Label Backdoor Attacks and Backdoor Attacks in Federated Learning (FL), currently under review at NeurIPS'25 and AAAI'25.
- Deploy FL models on edge devices such as Jetson Nano and Android mobile phones with Flower framework, benchmarking the performance of FL training on MNIST dataset.
- Mentored 3 junior undergraduate researchers on experimental design and scientific writing, leading to the successful completion of their undergraduate research projects.
- Presented notable papers in weekly seminars and research findings at lab meetings to Prof. Kok-Seng Wong and Prof. Khoa D. Doan, contributing to brainstorming and refining project strategy.

COMP2050 – Artificial Intelligence course in VinUni — Ha Noi, Viet Nam

Teaching Assistant — *Jan 2024 – May 2024*

- Designed 2 challenging projects: Tic Tac Toe and Gomoku AI (**Github**). Built upon the mechanics of these two classic board games, the projects require students to creatively devise different AI algorithms that could beat the programmed bots. In Gomoku, the programmed bots have 4 levels of intelligence, from beginner (a simple reflex agent) to master (based on the Threat-search space algorithm). Custom test cases were also created to assess the decision-making capabilities of the AI agents.

Unifix – Peer-to-peer academic mentorship platform Ha Noi, Viet Nam

Founder, Product Manager *Jul 2022 – Dec 2022*

- Unifix is an online platform that offers peer-mentorship services for college-level students on academic subjects. Our startup successfully secured an initial funding of 3,000 USD from the program TechYouth Incubator and Entrepreneurship Lab at VinUni.
- Led a cross-functional team of 7 people (3 in business, 4 developers) to build the website using Agile methodology.
- Conducted user interviews, surveys, and market research to understand user needs, pain points, and preferences.
- Gathered user inputs (needs, requirements, and feedback) to define key features and the product roadmap.

Projects

Startup Visualization System – *Software engineer* *Sep 2024 – Dec 2024*

- Developed the Startup Visualization System for the VinUni Entrepreneurship Lab (Elab) with Next JS front-end and Django back-end. This product is designed to help Elab efficiently track and manage startup information with a streamlined interface for administrators and an intuitive experience for viewers.
- Worked as Product Manager, acting as the primary point of contact with clients with clients and coordinating task assignments across the team.
- Led the front-end team, overseeing the development of the user interface and ensuring a seamless user experience.
- Develop the login page with JWT authentication, the dashboard page with real-time filters, visualization page with interactive graphics for Elab to draw key insights about VinUni startups.

Student-management Discord bot *March 2024 – April 2024*

- This bot was developed for the fintech startup Finful as my freelance project. It automates several tasks on Discord, including (1) assigning students to the appropriate study rooms based on subscriptions, (2) responding to student queries for educational resources (slides, lecture notes, etc), and (3) monitoring for inappropriate language. The bot was built using the Discord API and MongoDB as the backend.

Pacman AI (Github, Report) – *Reinforcement Learning* *Sep 2022 – Feb 2023*

- Pacman AI is a term project for the Artificial Intelligence COMP2080 course, where we implemented intelligent Pacman agents using Minimax, Monte Carlo Tree Search, and Reinforcement Learning algorithms (Tabular, Approximate, and Double Deep Q-Learning). We benchmark these algorithms in challenging maps with advanced ghosts, such as Dot-Eatable Ghosts and Smart Ghosts that move intelligently based on A* algorithm. Our project was the only one in the class to achieve a perfect score (10/10).

VinUni Easybooking (Github) – *Software engineer* *Aug 2022 – April 2023*

- This is a term project for the Object-Oriented Programming course for the first year. VinUni Easy Booking is a multi-functional Android app that combines booking services including Office Hour, Discussion Room, and Check Calendar/Create Event for VinUni students and professors. The report can be found here.
- Designed the UI and core functionalities for the app.
- Integrated Firebase backend to store data of booking appointments

Skills

Programming Languages: Python, Java, JavaScript, Typescript, C/C++, SQL, LaTeX, HTML, TailwindCSS.

Technologies: Git, Docker, AWS, PostgreSQL, Firebase, MongoDB, Bash.

Frameworks and Libraries: PyTorch, TensorFlow, NumPy, Pandas, Next JS, React, OpenCV, Django.

Soft Skills: Leadership, Problem-Solving, Presentation, Teamwork, Paper Writing/Reading.

Awards & Achievements

Scholarship: 90% Scholarship, *College of Engineering and Computer Science.* (Sep 2021 – Present)

Dean's List: Freshman year, Vin University (Jul 2022)

Vietnam Science and Engineering Fair (ViSEF): Third Prize, National. (Sep 2018)

Standardized Scores

IELTS: 7.5/9 (August, 2020)

SAT: 1570/1600 (March, 2021)

SAT Maths: 800/800 (November, 2020)

Preprints

Thinh Dao, Khoa Doan, Kok-Seng Wong. *Clean-Label Physical Backdoor Attacks with Data Distillation.*

- This paper introduced Clean-Label Physical Backdoor Attack (CLPBA), a physical backdoor paradigm that backdoors models by imperceptibly perturbing a small set of target-class images—without label tampering or trigger injection during training. Framed as dataset distillation (parameter/gradient/feature matching), CLPBA generalizes to real scenes and outperforms dirty-label baselines under both linear-probe and full fine-tuning on face and animal datasets.

Thinh Dao, Thuy Dung Nguyen, Khoa Doan, Kok-Seng Wong. *BackFed: An Efficient & Standardized Benchmark Suite for Backdoor Attacks in Federated Learning.*

- This paper introduced **BackFed**, a standardized and reproducible benchmark suite for federated learning (FL) backdoor attacks and defenses that reflects real-world constraints; it includes a high-throughput **multiprocessing engine** for fast experiments, a modular **plug-and-play** API for adding new methods, and a unified evaluation spanning CV and NLP across diverse models and settings. Using BackFed, we conducted large-scale empirical studies that surfaced previously unreported limitations and failure modes of backdoor attacks and defenses under practical conditions, providing actionable guidance for more secure FL system design and enabling fair, reliable comparisons in future research.

Quang H. Nguyen, **Thinh Dao**, Duy C. Hoang, Juliette Decugis, Saurav Manchanda, Nitesh V. Chawla, Khoa D. Doan. *MetaLLM: A High-performant and Cost-efficient Dynamic Framework for Wrapping LLMs.*

- This paper proposed **MetaLLM**, a query-level routing framework that selects the optimal LLM among multiple providers/models for classification and MCQA, improving both accuracy and cost-efficiency. Frames model selection as a multi-armed bandit to balance performance and price under uncertainty; validated on OpenAI, Together AI, and open-source LLMs.