Nguyen Hung Quang

quanghngnguyen@gmail.com
https://nguyenhungquang.github.io/

Research Interest

I am interested in understanding the behavior and robustness of deep neural networks. Specifically, my work analyzes the vulnerability of neural networks to adversarial attacks and data poisoning, and leverages mechanistic interpretability tools to reveal knowledge attribution of components in generative models. My research aims to establish a rigorous understanding of models' mechanisms, facilitate systematic evaluation, and enable the deployment of reliable AI systems.

Education

Vietnam National University, University of Engineering and Technology (VNU-UET) *2017 - 2021*
B.S. Computer Science *GPA: 3.37*
Thesis: Link Prediction on Knowledge Graph Using Graph Neural Network *Grade: 9.6/10*

Experience

VinUniversity *2023 - Now*
Research Assistant. Supervised by Prof. ***Khoa Doan****.*

- **Randomized Feature Defense.** Developed and analyzed a light-weight randomized defense to safeguard neural networks against black-box adversarial attacks with minimal accuracy trade-off *(accepted at ICLR 2024 & ACL 2024).*
  - We proved that injecting noise into the feature space is more effective against black-box attacks than into the input.
  - Our defense improves the accuracy of ResNet50 under Square attack from 0.1% to 53.4% in CIFAR10.
- **Selective Data Poisoning.** Proposed selective poisoning strategies that boost the attack success rate of clean-label backdoor attacks in practical scenarios *(accepted at ICLR 2025).*
  - We showed that hard samples are more important for clean-label backdoor attacks and improve the attack success rate of BadNet by 60% in CIFAR10.
- **Concept Attribution.** Developed a model attribution framework for diffusion models and proposed efficient model editing algorithms to enhance or suppress knowledge *(accepted at NeurIPS 2025).*
  - We proved the knowledge localization hypothesis and discovered the existence of negative components.
  - Our unlearning method decreases the probability of generating nudity content by more than 90% even under adversarial prompts, while remaining other knowledge.
- **LLM Routing.** Developed multi-arm bandit algorithms that balance LLM inference cost and performance by routing input prompts to suitable LLMs.

VinUniversity *2023 - Now*
Teaching Assistant

- Conducted labs and office hours, prepared and evaluated homework and examinations for Artificial Intelligence, Machine Learning, and Data Mining courses.

Sun* R&D Unit *10/2021 - 11/2022*
AI Engineer

- Developed algorithms, collected and pre-processed data to build high-quality text-to-speech models that automatically generate audiobooks, in collaboration with Voice of Vietnam.

Math and Science Summer Program *7/2021*
Mathematics Mentor

- Prepared materials and tutorials, delivered lectures about Error-correction code to Vietnamese high school students.

Knowledge Technology Laboratory - UET *2020/2021*
Undergraduate research student. Supervised by Prof. ***Phan Xuan-Hieu****.*

- Developed complex and quaternion vector transformations in graph neural networks for more efficient link prediction in knowledge graphs.

Publications

[1] **Quang H Nguyen**, Hoang Phan, Khoa D Doan. "Unveiling Concept Attribution in Diffusion Models." Neural Information Processing Systems (NeurIPS) 2025.

[2] **Quang H Nguyen**, Nguyen Ngoc-Hieu, The-Anh Ta, Thanh Nguyen-Tang, Kok-Seng Wong, Hoang Thanh-Tung, Khoa D Doan. "Wicked Oddities: Selectively Poisoning for Effective Clean-Label Backdoor Attacks." International Conference on Learning Representations (ICLR) 2025.

[3] Cao-Duy Hoang, **Quang H Nguyen**, Saurav Manchanda, Minlong Peng, Kok-Seng Wong, and Khoa D Doan. "Fooling the Textual Fooler via Randomizing Latent Representations." Findings of the Association for Computational Linguistics (ACL Findings) 2024.

[4] **Quang H Nguyen**, Yingjie Lao, Tung Pham, Kok-Seng Wong, and Khoa D Doan. "Understanding the Robustness of Randomized Feature Defense Against Query-Based Adversarial Attacks." International Conference on Learning Representations (ICLR) 2024.

Workshop Papers

[1] **Quang H Nguyen**, Ngoc-Hieu Nguyen, Thanh Nguyen-Tang, Hoang Thanh-Tung, Khoa D Doan. "Clean-label Backdoor Attacks by Selectively Poisoning with Limited Information from Target Class." NeurIPS Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly 2023.

[2] Nguyen Ngoc-Hieu, **Quang H Nguyen**, The-Anh Ta, Thanh Nguyen-Tang, Khoa D Doan, Hoang Thanh-Tung. "A Cosine Similarity-based Method for Out-of-Distribution Detection." ICML Workshop on Spurious Correlations, Invariance and Stability 2023.

Preprints

[1] **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." 2024.

[2] Sze Jue Yang, Chinh D La, **Quang H Nguyen**, Kok-Seng Wong, Anh Tuan Tran, Chee Seng Chan, Khoa D Doan. "Synthesizing Physical Backdoor Datasets: An Automated Framework Leveraging Deep Generative Models." 2023.

Technical Background

Programming skills: Python, SQL, Git, Numpy, Pandas, Pytorch, HuggingFace, vLLM, Pyspark.
Machine learning: Learning theory, diffusion models, large language models, mechanistic interpretability.
Mathematics: Probability theory, statistics, real analysis, linear algebra, optimization.

Community Services

Reviewer

- NeurIPS (Top Reviewer) *2024*
- ICLR, AISTATS, CVPR, ICML, TMLR, NeurIPS (Top Reviewer) *2025*
- ICLR, AISTATS *2026*

Mentoring

- Hoang Phan. Undergraduate student at VinUniversity.

Additional Activities

- Volunteered for the ACML 2024 conference.
- Contributed to the Vietnamese translation of chapter 5 of the book "Interpretable machine learning" by Christoph Molnar.

Grants

VinUniversity Research Grant ($4,500).

ICLR 2025 Travel Grant ($1,745).