Nguyen Hung Quang

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

Education

- **Vietnam National University - University of Engineering and Technology**
  B.S. Computer Science *2017-2021*
  Thesis: Link Prediction on Knowledge Graph Using Graph Neural Network

Research Interest

My current interest mainly focuses on the robustness and trustworthiness of deep models, ranging from adversarial attacks and backdoor attacks to generalization. I aim to understand what makes the model vulnerable to such problems and how to make the model more robust and resilient to security threats. I am also interested in interpreting the behavior of diffusion models and language models.

Experience

- **Knowledge Technology Laboratory - VNU** *2020/2021*
  Undergraduate research student
- **Math and Science Summer Program** *7/2021*
  Mathematics Mentor
  Topic: Error-correction code
- **Sun* R&D Unit** *10/2021 - 11/2022*
  AI Engineer
  Worked with Voice of Vietnam to build a text-to-speech model to generate high-quality audiobooks.
- **MAIL Research - VinUni** *2023 - Now*
  Research Assistant & Teaching Assistant
  Conducted research on adversarial attacks, backdoor attacks, generative models, and interpretability.

Publications

- 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 2023 Workshop on Spurious Correlations, Invariance and Stability (2023).
- **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 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly. (2023).
- **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 (2023).
- 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 (2024).

Preprints

- Sze Jue Yang, **Quang H Nguyen**, Chee Seng Chan, Khoa D. Doan. "Everyone Can Attack: Repurpose Lossy Compression as a Natural Backdoor Attack." (2023).

- 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).

- **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.” (2024).

- **Quang H Nguyen**, 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).

- **Quang H Nguyen**, Hoang Phan, Khoa D Doan. “Unveiling Concept Attribution in Diffusion Models.” (2024).

Professional Services

- Reviewer at NeurIPS 2024 (Top Reviewer), ICLR 2025, AISTATS 2025, CVPR 2025.

Technical Background

- **Programming languages**: Python. Experience working with Pytorch, Numpy, HuggingFace, Pyspark.
- **Machine learning**: Security in machine learning, diffusion models and large language models.
- **Mathematics**: Probability theory, statistics, analysis, linear algebra.

Additional Activities

- Teaching Assistant of Artificial Intelligence, Machine Learning, and Data Mining in VinUniversity.
- Contributed to the Vietnamese translation of the book ”Interpretable machine learning”.