Kyung Min (Brian) Ko

[U.S. Permanent Resident]
https://ko120.github.io | ko120@purdue.edu | LinkedIn | Github |

Education

Purdue University, West Lafayette Aug 2019 - May 2024
Bachelor of Science in Electrical Engineering, graduated with distinction **GPA: 3.94/4.00**

- TA: ECE 570 Artificial Intelligence **(graduate)**, ECE 20875 Python for Data Science
- Coursework: Artificial Intelligence **(graduate)**, Statistical Machine Learning **(graduate)**, Natural Language Processing **(graduate)**, Probabilistic Method

Research interest

Truthyworthy Machine learning, Reinforcement Learning, LLM

Publication

Toward Trustworthy Machine Learning via Distribution Matching To be submitted to ICML, 2025
Kyung Min Ko, Jim Lim, Ziyu Gong, David Inouye. ***[Paper]***

Jailbreak via Reward Poisoning RLHF To be submitted to ICML, 2025
Kyung Min Ko, Han Wang, Arman Zharmagambetov, Haun Zhang. ***[Paper]***

Backward Curriculum Reinforcement Learning IEEE RO-MAN **(Oral)**, 2023
Kyung Min Ko. ***[Paper] [Code]***

V-advCSE: Virtual Adversarial Contrastive Learning for Sentence Embeddings Pre-print, 2023
Kyung Min Ko. ***[Paper] [Code]***

Exploiting Code Language Models and Contrastive Learning in Binary Code Authorship Pre-print, 2023
Kyung Min Ko, Nan Jiang, Lin Tan. ***[Paper] [Code]***

Experience

Research Assistant Aug 2024 - Present
UIUC (remote), Champaign, IL. ***Advised by Prof. Huan Zhang*** **To be submitted to ICML 25**

- Introduced a novel RLHF-based jailbreak method for the automated generation of adversarial suffixes.
- Enhanced controllability by designing and implementing a sophisticated reward function.
- Leveraged generated adversarial suffixes to improve safety-alignment methods for LLMs.

Research Assistant May 2024 - Present
Purdue University, West Lafayette, IN. ***Advised by Prof. David I. Inouye*** **To be submitted to ICML 25**

- Conducted research focusing on critical aspects of trustworthy machine learning (ML), including calibration, domain adaptation, and fairness.
- Developed a unified framework for trustworthy distribution matching (DM), incorporating methods such as Sinkhorn, MMD, and adversarial learning to address calibration, domain adaptation, and fairness tasks.
- Demonstrated the effectiveness of various DM methods for calibration, domain adaptation, and fairness, providing practical insights into selecting appropriate DM methods.

NSF Summer Undergraduate Research Intern [Paper & Code] May 2023 - Jan 2024
Purdue University, West Lafayette, IN. ***Advised by Prof.Lin Tan***

- Discovered the application of code language models for malware author classification
- Engineered a novel approach for function-level learning, transitioning from traditional file-level input
- Incorporated contrastive learning methodologies to address code authorship tasks, eliminating the need for labels

Human Resource Manager Nov 2021 - May 2023
Republic of Korea Army, South Korea

- Optimized the boundary protection schedule system by automating processes with programming
- Facilitated proper troop assignments by documenting the transferring process, considering current unit status
- Recognized for developing an AI object tracking system used in the guardroom, awarded by the chief of the general staff of the army

NSF Summer Undergraduate Research Intern [Code] Jun 2021 - Jan 2022
Georgia Tech, Georgia, Atlanta. ***Advised by Prof.Siva Theja Maguluri*** **IEEE ROMAN (Oral) 23**

- Implemented REINFORCE, A2C, and PPO algorithms applicable to both continuous and discontinuous action spaces
- Proposed a novel backward curriculum learning, enhancing sample efficiency via reverse order training
- Evaluated performance on different architecture settings to provide insight on choosing proper architecture

Projects

Guardroom Object Tracking System [Code] Jun 2022

Awarded commandment by the chief of the general staff of the army

- Developed a multi-object tracking system using Yolo-v4 and Deep Sort for automated CCTV surveillance in guardrooms
- Enhanced unit security by tracking objects entering selected regions and calculating real-time moving average distances to display object trajectories

SKILLS

Programming Languages: Python, C++, Java

Software: Pytorch Lightening, Hydra (for ML experiment), TensorFlow

Honors

Dean's List & Semester Honors All semester

NSF Summer Research Fellowship 2021,2023