Wenbin Hu

Yr-4 Undergraduate Student
Department of Computer Science and Engineering
Hong Kong University of Science and Technology, HK

+86-15917019800
whuak@connect.ust.hk
GitHub Profile
LinkedIn Profile

Education

- **Hong Kong University of Science and Technology, HK** *2020-24*
  B.Eng in Computer Science (Awards: Dean's List)
  - **Basic Computer Science Courses**: Python Programming, C++, Object-Oriented Programming, Data Structures & Algorithms, Software Engineering, Computer Organization.
  - **Basic Mathematics Courses**: Multivariate Calculus, Linear Algebra, Probability, Discrete Math, Differential Equation, Mathematical Analysis, Convex Optimization, Abstract Algebra, Advanced Probability Theory
  - **ML/DL/CV/CG Courses**: Deep Learning in Computer Vision, Machine Learning for Graphical and Sequential Data, Deep Generative Models, Machine Learning Theory, Computer Graphics.

- **Technical University of Munich, Munich** *Summer 2023*
  Exchange Programme in TUM Informatics

- **Fudan University, Shanghai** *Summer 2022*
  Summer School Programme

Publications

Attacking by Aligning: Clean-Label Backdoor Attacks on Object Detection, Yize Cheng*, **Wenbin Hu***, Minhao Cheng, arXiv preprint arXiv:2307.10487 [arxiv]

- Inspired by an inherent property of deep neural network, we've uncovered that malicious attackers can effortlessly implant backdoors into object detection models by aligning their hazardous objectives with those of the detectors.
- Based on this insight, we have developed an simple, yet highly effective clean-label attack strategy against object detection in scenarios involving both Object Disappearance Attacks and Object Generation Attacks.
- Numerous experiments and thorough ablation studies were carried out to confirm the efficacy of the proposed attack method and its sustainability when subjected to fine-tuning.

Mitigating the Alignment Tax of RLHF, Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, **Wenbin Hu**, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, Tong Zhang, arXiv preprint arXiv:2309.06256 [arxiv]
(* represents equal contribution.)

Projects

- **Identification of Adversaries from Collusive Adversarial Examples** *On Going*
  Advisor: Dr. Minhao Cheng - Hong Kong University of Science and Technology, HK
  - Based on an existing research in forensic investigation of adversaries, we extended it to identification of collusive adversaries from a adversarial sample.
  - Drawing inspiration from the Logical-and property of the Balanced Incomplete Block Design (BIBD) code, we have successfully incorporated traditional anti-collusion codes into our framework, enabling us to effectively trace collusive adversaries.
  - Although our framework has limitations in the data-free scenario, we are committed to developing an improved anti-collusion code to address this issue.

- **Generative Graph Embedding Inversion Attack** *On Going*
  Advisor: Dr. Yangqiu Song - Hong Kong University of Science and Technology, HK
  - In this work, we proposed generative embedding inversion attack on inductive graphs. We utilize a generative language model (GPT-2) to recover the semantic information and neighbourhood from a node embedding.
  - We found that the re-encoded embedding from the recovered text is close, but not identical, to the victim embedding. To bridge this gap, we developed an iterative algorithm that gradually refines the generated content, bringing it closer to the desired target.

- **Clean-Label Backdoor Attacks on Object Detection** *Oct 2022 - Aug 2023*
  Advisor: Dr. Minhao Cheng - Hong Kong University of Science and Technology, HK
  - Inspired by an inherent property of deep neural network, we've uncovered that malicious attackers can effortlessly implant backdoors into object detection models by aligning their hazardous objectives with those of the detectors.

– Based on this insight, we have developed an simple, yet highly effective clean-label attack strategy against object detection in scenarios involving both Object Disappearance Attacks and Object Generation Attacks.
– Numerous experiments and thorough ablation studies were carried out to confirm the efficacy of the proposed attack method and its sustainability when subjected to fine-tuning.

• **Enhance Resilience against Backdoor Attacks on Commonsense Inference** *Oct 2022 - Jan 2023*

Advisor: Dr. Yangqiu Song - Hong Kong University of Science and Technology, HK

– We investigated the role of inductive graph reasoning on improving the resilience of Commonsense Knowledge Graph (CSKG) against backdoor attacks.
– The experiments indicated that there is a distribution shift in the prediction heads, and hence, inductive graph reasoning can enhance the model's resilience to some extent. Nevertheless, the impact is rather modest, as the attack success rate remains virtually unchanged.

• **UROP 2022 Summer: A Research Survey on Autonomous Driving** *June 2022 - Aug 2022*

Advisor: Dr. Qifeng Chen - Hong Kong University of Science and Technology, HK

– This is a research-based course provided by Undergraduate Research Opportunities Program (UROP) in HKUST.
– We conducted a survey on motion forecasting for moving cars.

Research Interests

My research interest is centered around the expansive field of **Trustworthy Machine Learning**. Through my research, I have unveiled a critical insight: machine learning models are inherently susceptible to malicious attacks. This revelation underscores the imperative for our community to scrutinize the dependability of these models. I have focused on: 1) uncovering potential strategies that adversaries could exploit to compromise machine learning algorithms, and developing resilient models that can withstand such attacks; 2) develop privacy-preserving methods for training models on sensitive data; 3) mitigate biases in machine learning and produce models that are fair, impartial, and truly representative; 4) reduce the uncertainty and enhance the interpretability in machine learning. My ultimate objective is to construct a **robust, trustworthy, and interpretable** framework for machine learning.

Skills

Programming Language: C/C++, Python

Library/Framework: Pytorch, TensorFlow/Keras, Scikit-Learn, Numpy, OpenCV, OpenGL.

Tools: Latex, Markdown, Git, Microsoft Office.

Language: Chinese (Native), English (Fluent), Cantonese (Good).

Soft Skill: Academic Writing, Communication, and Presentation; Strong Self-motivation and Self-learning Ability.

Personal Interests

Piano, Basketball, Running, Exploring New Things.