Weixin CHEN

+1 217-778-6107 | weixinc2@illinois.edu | chenweixin107.github.io

EDUCATION

University of Illinois at Urbana-Champaign - *The Grainger College of Engineering* *Aug. 2023 - Present*

- Ph.D. in Computer Science
- Advisor: Prof. Han Zhao
- Research interests: Trustworthy AI, Trustworthy LLMs, Adversarial Machine Learning

Tsinghua University - *Tsinghua Shenzhen International Graduate School* *Aug. 2020 - Jun. 2023*

- M.E. in Electronic and Information Engineering (Artificial Intelligence)
- Advisor: Prof. Haoqian Wang
- GPA: 4.0 / 4.0 Rank: 1 / 1067
- Main courses: Convex Optimization, Stochastic Processes, Artificial Neural Network

Sun Yat-sen University - *School of Mathematics (Zhuhai)* *Aug. 2016 - Jun. 2020*

- B.S. in Information and Computing Science
- Advisor: Prof. Zhiwei Wu
- GPA: 4.0 / 4.0 Rank: 1 / 36
- Main courses: Mathematical Analysis, Numerical Analysis, Geometry and Algebra, Numerical Algebra, Probability Theory, Mathematical Statistics, Foundation of Information Theory, Data Structure and Algorithms

PUBLICATIONS

DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models [Code] *2023*

Boxin Wang, **Weixin Chen***, Hengzhi Pei*, Chulin Xie*, Mintong Kang*, Chenhui Zhang*, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, Bo Li*

Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023, Oral, Outstanding Paper Award)

TrojDiff: Trojan Attacks on Diffusion Models with Diverse Targets [Code] *2023*

***Weixin Chen**, Dawn Song, Bo Li*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2023)

Effective Backdoor Defense by Exploiting Sensitivity of Poisoned Samples [Code] *2022*

***Weixin Chen**, Baoyuan Wu, Haoqian Wang*

Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022, Spotlight)

PROFESSIONAL EXPERIENCES

Research Intern - *Secure Learning Lab, University of Illinois at Urbana-Champaign* *Jul. 2023 - Dec. 2023*

Advisor: Prof. Bo Li

- Proposed **an effective offline reinforcement learning paradigm for large language models** to enhance the truthfulness.
- Experiments on TruthfulQA's MC1 and MC2 datasets showed the SOTA performance of the learned 7B model, compared with a variety of open and proprietary models scaling from 7B to 70B.

Research Intern - *Secure Learning Lab, University of Illinois at Urbana-Champaign* *Jul. 2022 - Jun. 2023*

Advisor: Prof. Bo Li

- Proposed **the first Trojan attack on diffusion models**, TrojDiff, with diverse targets and triggers.
- We proposed (1) Trojan diffusion process with novel transitions to diffuse adversarial targets into a biased Gaussian distribution, (2) Trojan generative process based on a new parameterization that leads to a simple training objective for attack.
- Experiments on 2 benchmark datasets showed the superior attack performance of TrojDiff against 2 diffusion models, considering 3 types of adversarial targets and 2 types of triggers, in terms of 6 evaluation metrics.

Research Intern - *SCLBD, The Chinese University of Hong Kong, Shenzhen* *Jun. 2021 - May. 2022*

Advisor: Prof. Baoyuan Wu

- Proposed **two effective backdoor defenses**, D-ST and D-BR, by exploiting sensitivity of poisoned samples to transformations.
- We proposed (1) a secure training module with semi-supervised contrastive learning to train a secure model from scratch, (2) a backdoor removal module based on unlearning and relearning to remove backdoor from a backdoored model.
- Experiments on 3 benchmark datasets showed the superior defense performance of D-ST and D-BR against 8 widely used backdoor attacks, to 6 state-of-the-art backdoor defenses with different defense paradigms.

ACADEMIC SERVICES

Journal Reviewer: TPAMI (IEEE Transactions on Pattern Analysis and Machine Intelligence), TIFS (IEEE Transactions on Information Forensics & Security)
Conference Reviewer: AISTATS 2024

SELECTED HONORS

Outstanding Paper Award, NeurIPS	*2023*
Wing Kai Cheng Fellowship, UIUC	*2023*
First Prize Scholarship (top 3%), Tsinghua University	*2021*
First Prize Scholarship (top 5%), Sun Yat-sen University	*2017, 2018, 2019*
National Scholarship (top 2%) / Giordano Donation Scholarship (top 3%), Sun Yat-sen University	*2018, 2019 / 2017*
First Prize (top 1%), Chinese Mathematics Competitions (CMC)	*2018*

SKILLS

Programming: Python, PyTorch, LaTeX
Languages: English (fluent), Mandarin (native), Cantonese (native)