Yue Gao

+1 (608) 733-8789 gy@cs.wisc.edu pages.cs.wisc.edu/~gy ygao234 Lodour Madison, WI

Research Interests

Trustworthy Machine Learning (adversarial robustness, black-box evasion attacks and defenses)
System Security (machine learning systems, web-based applications and services)

Education

University of Wisconsin–Madison Madison, WI
Ph.D. Candidate in Computer Science (advised by Prof. Kassem Fawaz) *Sep 2018 – May 2024 (expected)*
- Thesis: *Characterizing the Limitations of Defenses in Adversarial Machine Learning*

Shanghai University Shanghai, China
B.S. in Computer Science and Technology (GPA 3.99/4.00, Ranked 1/292) *Sep 2014 – Jul 2018*
- Thesis: *A Deep Neural Network based Image Compression Method*

Work Experience

ML Security Research Intern @ Microsoft Research Redmond, WA
Mentored by Dr. Jay Stokes and Dr. Emre Kiciman *Jun 2021 – Sep 2021*
- Initiated a research project on defenses against imperceptible textual backdoor attacks on language models.
- Proposed a defense strategy leveraging the limitation of imperceptible backdoors (published at MILCOM).
- Achieved a reduction in attack success rate from 100% to 12% at a challenging poisoning rate of 10%.

ML Research and Development Intern @ TuCodec Shanghai, China
Mentored by Dr. Chunlei Cai *Jan 2018 – Jul 2018*
- Achieved 1st place in the CVPR 2018 Challenge on Learned Image Compression as a primary contributor.
- Improved the average runtime efficiency of DNN-based compression from 1 min to 4 secs per 4K-res image.
- Independently developed DNN-based apps on Ubuntu, MacOS, Windows, and self-hosted cloud services.

Publications

Conference

[1] On the Limitations of Stochastic Pre-processing Defenses
Yue Gao, Ilia Shumailov, Kassem Fawaz, and Nicolas Papernot
Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), 2022

[2] Rethinking Image-Scaling Attacks: The Interplay Between Vulnerabilities in Machine Learning Systems
Yue Gao, Ilia Shumailov, and Kassem Fawaz
Proceedings of the 39th International Conference on Machine Learning (ICML), 2022
Oral Presentation (Top 2%)

[3] Experimental Security Analysis of the App Model in Business Collaboration Platforms
Yunang Chen*, **Yue Gao***, Nick Ceccio, Rahul Chatterjee, Kassem Fawaz, and Earlence Fernandes
31st USENIX Security Symposium (USENIX Security), 2022
Bug Bounty ($1500)

[4] I Know Your Triggers: Defending Against Textual Backdoor Attacks With Benign Backdoor Augmentation
Yue Gao, Jack W. Stokes, Manoj Prasad, Andrew Marshall, Kassem Fawaz, and Emre Kiciman
IEEE Military Communications Conference (MILCOM), 2022

Workshop

[1] Variational Autoencoder for Low Bit-rate Image Compression
Lei Zhou*, Chunlei Cai*, **Yue Gao**, Sanbao Su, and Junmin Wu
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018
Winner of the first Challenge on Learned Image Compression

Preprints

[1] SEA: Shareable and Explainable Attribution for Query-based Black-box Attacks
Yue Gao, Ilia Shumailov, and Kassem Fawaz
arXiv, 2023

[2] Analyzing Accuracy Loss in Randomized Smoothing Defenses
Yue Gao*, Harrison Rosenberg*, Kassem Fawaz, Somesh Jha, and Justin Hsu
arXiv, 2020

Selected Projects

Shareable and Explainable Attribution for Black-box Attacks on ML systems *Jan 2023 – Aug 2023*
- Characterized the attack's progression for forensic purposes and human-explainable intelligence sharing.
- Fingerprinted and attributed zero-day attacks on their first and second occurrence, respectively.
- Discovered specific minor implementation bugs in popular ML attack toolkits.

The Role of Randomization in Adversarial Robustness *Feb 2022 – May 2022*
- Characterized the limitations of using randomization to defend ML models.
- Established theoretical and empirical results for the non-robustness of randomization-based defenses.

Security Analysis of Online Collaboration Platforms *Mar 2021 – Dec 2021*
- Analyzed the permission model of third-party apps in Slack and Microsoft Teams with closed-source access.
- Exploited OAuth designs to bypass access control and user privacy, received bug bounty for medium severity.
- Demonstrated POC attacks that eavesdrop chatting, launch fake video calls, and merge codes.

Trustworthy Machine Learning in Real-World Systems *Sep 2020 – Jan 2021*
- Investigated the robustness of real-world ML pipelines exposed to diverse security threats.
- Revealed 9x amplified threats and broke state-of-the-art defenses by exploiting multiple vulnerabilities jointly.

Defending against Evasion Attacks in Multimodal Scenarios (Collaborative) *Since 2019 (semiannual)*
- Led a 9-member team to 1st and 2nd place in competitions for robust multimodal object detection.
- Proposed a defense strategy using diffusion-based modality reconstruction (from RGB to Depth).
- Achieved a reduction of disappearance rate from 62% to 9% under strong adaptive attacks.
- Contributed plug-and-play modules to the official upstream evaluation team and received acknowledgment.
- Developed initial code bases and eval pipelines for team members from varying technical backgrounds.

Selected Honors & Awards

Slack Bug Bounty: Medium Severity, $1500	2022
Top 10% Reviewers Award: NeurIPS	2022
CVPR Competition Winner: Challenge on Learned Image Compression	2018
National Scholarship: China	2017
Top 100 Elite Collegiate Award: China Computer Federation	2017
Scholarship for Exceptional Leadership: Shanghai University	2017
City Scholarship: Shanghai	2016
Outstanding Student Award: Shanghai University	2016
Outstanding Volunteer Award: ACM ICPC Asia Regional Contest	2016
Scholarship for Exceptional Innovation: Shanghai University	2016
Scholarship for Exceptional Academic Achievements: Shanghai University	2015 – 2018
Bronze Prize for Programming Contest: ACM ICPC Asia East-Continent Final Contest	2015
Bronze Prize for Programming Contest: ACM ICPC Asia Shanghai Regional Contest	2015

Professional Activities

Reviewer: NeurIPS and ICML	2022 – 2024
External Reviewer: USENIX Security Symposium	2021 – 2022
External Reviewer: IEEE Symposium on Security and Privacy	2021 – 2022
External Reviewer: ACM Conference on Computer and Communications Security	2019
Team Leader: Collegiate ICPC Team at Shanghai University	2016 – 2017

Talks

1. **The Vulnerabilities of Preprocessing in Adversarial Machine Learning** *Apr 2023*
   TrustML Young Scientist Seminar, RIKEN AIP
2. **On the Limitations of Stochastic Pre-processing Defenses** *Oct 2022*
   University of Southern California (virtual)
3. **The Interplay Between Vulnerabilities in Machine Learning Systems** *Sep 2022*
   University of Michigan
4. **Experimental Security Analysis of the App Model in Business Collaboration Platforms** *Aug 2022*
   USENIX Security 2022
5. **The Interplay Between Vulnerabilities in Machine Learning Systems** *Jun 2022*
   ICML 2022

Teaching and Mentoring

Teaching Assistant: CS 368 (C++ for Java Programmers), University of Wisconsin–Madison	Fall 2018
Guest Lecturer: Advanced Algorithms & Data Structures, Shanghai University	2015 – 2017
Problem Designer: Undergraduate Programming Contests, Shanghai University	2015 – 2017
Student Mentor: Undergraduate Computer Science Coursework, Shanghai University	2015 – 2017

Technical Skills

Python	Research (2018 – present), System Optimization (2018), Backend Development (2016 – 2017).
PyTorch	Research (2019 – present), Distributed Training (2020 – 2022).
Docker	Research (2018 – present), Computing Cluster (2017 – 2018).
C / C++	Linux Kernel (2019), ML System (2018), Programming Contest (2014 – 2018).
TensorFlow	Service Deployment (2018).
Java EE	Backend Development (2016).

Articles and Media Coverage

CleverHans. Can stochastic pre-processing defenses protect your models?	2022
USENIX login. Experimental Security Analysis of the App Model in Business Collaboration Platforms	2022
Wired. Slack's and Teams' Lax App Security Raises Alarms	2022