

Yue Gao

☎ +1 (608) 733-8789 ✉ ygao234@gmail.com 🏠 pages.cs.wisc.edu/~gy 🌐 ygao234 📶 Lodour 📍 San Mateo, CA

RESEARCH INTERESTS

AI Security	ML System Security, Adversarial Robustness, Attacks and Defenses, Security and Privacy.
Machine Learning	Vision Recognition, LLMs, Multi-Modality, Diffusion Models, Natural Language Processing.
Cybersecurity	Multi-component ML Systems, Web-based Applications, Cryptography, Linux Kernel Memory.

EDUCATION

University of Wisconsin–Madison

Madison, WI

Ph.D. in Computer Sciences (advised by Prof. Kassem Fawaz)

Sep 2018 – Mar 2024

- Thesis: *Challenges and Advances in Adaptive Response Strategies for Machine Learning Security*
- Selected Courses: *Mathematical Foundations of Machine Learning, Advanced Algorithms, Intro to Information Security.*

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*
- Selected Courses: *Operating Systems, Computer Network, Assembly Language, Software Engineering.*

WORK EXPERIENCE

Product Security Engineer @ Snowflake Inc.

San Mateo, CA

Managed by Sharath Sarangpur

Apr 2024 – Present

- Red-teaming LLM-based production systems for **prompt injection**, **jailbreaking**, and privacy attacks.
- Developed and deployed **robust and scalable** vulnerability scanners using Python, Docker, and Kubernetes.

Research Assistant @ University of Wisconsin–Madison

Madison, WI

Advised by Prof. Kassem Fawaz

Nov 2018 – Mar 2024

- Researched the security of **multi-component ML Systems** deployed in real-world environments.
- Systematized the security analysis of ML-based and web-based systems in **black-box** settings.

ML System Security Research Intern @ Microsoft Research

Redmond, WA

Mentored by Dr. Jay Stokes and Dr. Emre Kiciman

Jun 2021 – Sep 2021

- Proposed a research project on defenses against imperceptible textual **backdoor attacks on language models**.
- Discovered blind spots in state-of-the-art attacks and defenses, and published stronger defenses at MILCOM.
- Successfully reduced the attack success rate from 100% to 12%, even at a challenging poisoning rate of 10%.

ML Research and Development Intern @ TuCodec (Startup)

Shanghai, China

Mentored by Dr. Chunlei Cai

Jan 2018 – Jul 2018

– *ML Research*

- Secured 1st place as a **primary contributor** in the CVPR 2018 Challenge on Learned Image Compression.
- Improved the average runtime of DNN-based image compression algorithms **from 1 min to 4 secs** per 4K-res image.

– *ML Engineering & Security*

- Independently developed ML-based desktop apps on Ubuntu, MacOS, and Windows using TensorFlow and C++.
- Independently developed ML-based image compression systems using TensorFlow, Python, Docker, and Kubernetes.
- Protected ML systems from **security intrusion** and **model stealing** with security measures and anomaly analysis.

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 Earlene Fernandes
31st USENIX Security Symposium (USENIX Security), 2022
Bug Bounty (\$1500)
- [4] **SEA: Shareable and Explainable Attribution for Query-based Black-box Attacks**
Yue Gao, Ilia Shumailov, and Kassem Fawaz
Under Submission, 2023
- [5] **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] **Human-Produced Adversarial Examples**
David Khachaturov, Yue Gao, Ilia Shumailov, Robert Mullins, Ross Anderson, 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

Real-world ML Systems Security: Threat Modeling, Defending, and Characterization Sep 2020 – Jun 2023

- Investigated the security of **multi-component ML systems** exposed to diverse security threats, e.g., dependencies.
- Revealed **threats amplified by 9x** and **broke state-of-the-art defenses** by jointly exploiting multiple vulnerabilities.
- **Formally proved the non-robustness** of randomization-based defenses beyond demonstrating empirical attacks.
- Characterized the attack's progression for **forensic purposes** and **human-explainable intelligence sharing**.

GARD: Guaranteeing AI Robustness Against Deception Jun 2019 – Feb 2024

– Leadership

- Led a 9-member cross-university team to **1st and 2nd place in grant competitions hosted by DARPA**.
- Developed initial code bases and mentored **team members with varying technical backgrounds**.
- Onboard and mentored new team members to maintain adversarial mindsets with designing defenses.

– Individual Contribution

- Performed **red teaming** and **broke over 10 internal defenses** proposed by team members prior to submission.
- Proposed CLIP-like and diffusion-based methods to **enforce robust features** across RGB and Depth modalities.
- Successfully reduced the disappearance rate from 62% to 9% even **under the red-team evaluation from MITRE**.
- Contributed plug-and-play modules to the official upstream evaluation team and received acknowledgment.

Security Analysis of Online Business Collaboration Platforms Mar 2021 – Dec 2021

- Automated the analysis of security principle violations for 3K+ third-party apps in Slack and Microsoft Teams.
- Reverse engineered OAuth designs to bypass access control, and received bug bounty for medium severity.
- Demonstrated POC attacks of eavesdropping on private chats, spoofing video calls, and unauthorized code merging.

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, ICML, and ICLR	2022 – 2024
External Reviewer: IEEE Symposium on Security and Privacy	2021 – 2023
External Reviewer: USENIX Security Symposium	2021 – 2022
External Reviewer: ACM Conference on Computer and Communications Security	2019
Team Leader: Collegiate ICPC Team at Shanghai University	2016 – 2017

TALKS

1. Forensics and Intelligence Sharing for ML Security <i>DARPA GARD PI Meeting, IBM Research</i>	Oct 2023
2. The Vulnerabilities of Preprocessing in Adversarial Machine Learning <i>ML Red Team, Google</i>	Oct 2023
3. The Vulnerabilities of Preprocessing in Adversarial Machine Learning <i>TrustML Young Scientist Seminar, RIKEN AIP</i>	Apr 2023
4. On the Limitations of Stochastic Pre-processing Defenses <i>DARPA GARD PI Meeting, University of Southern California (virtual)</i>	Oct 2022
5. The Interplay Between Vulnerabilities in Machine Learning Systems <i>DARPA GARD PI Meeting, University of Michigan</i>	Sep 2022
6. Experimental Security Analysis of the App Model in Business Collaboration Platforms <i>USENIX Security 2022</i>	Aug 2022
7. The Interplay Between Vulnerabilities in Machine Learning Systems <i>ICML 2022</i>	Jun 2022

TEACHING AND MENTORING

Project Mentor: DARPA GARD Project, University of Wisconsin–Madison	Fall 2023
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	Expertise in ML and Security Research ^(2018 – present) and Backend Development ^(2016 – 2018) .
PyTorch	Expertise in ML Research ^(2019 – present) and Distributed Training ^(2020 – present) .
Docker	Expertise in ML Research ^(2018 – present) with familiarity in Kubernetes Cluster ^(2017 – present) .
C / C++	Proficient in Security Research ^(2018 – 2019) , Software Development ^(2017 - 2019) , and ICPC ^(2014 – 2018) .
Golang	Proficient in Software Engineering ⁽²⁰²⁴⁾ .
TensorFlow	Familiarity in ML Research ^(2017 – 2020) and Service Deployment ⁽²⁰¹⁸⁾ .

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