CS 562: Advanced Topics in Security, Privacy, and Machine Learning

Jan 12, 2024

Session-VC: Machine Learning for Sys, Networks, and Security

Home | Campuswire | Paper Signup Sheet | Project

Instructor: Varun Chandrasekaran (varunc@illinois.edu)

TA: Qilong Wu (qilong3@illinois.edu)

Time/Location: Wednesday 03:00 - 06:00 PM. Siebel Center for Comp Sci Room 0216

Office Hour: By Appointment

Announcement

1/17/2024: [First week of class] Enrolled students will be added/invited to CS 562 Campuswire before the first week of the class. If you registered during/after the first week and did not get the Campuswire invitation, please email the instructor (varunc@illinois.edu) for the invitation code.

Class Description

Advanced topics in security and privacy problems in machine learning systems, selected from areas of current research such as: This section will primarily focus on using machine learning for system, networking, and security applications. Example topics include using ML to build novel security defenses (e.g., detecting network intrusions, cybercrime, and disinformation, and performing user authentication and vulnerability analysis), launch novel attacks (e.g., privacy attacks, password guessing, deepfake-based social engineering), and support system optimizations. We will explore new research directions and seek to understand the limitations and potential risks of ML-based approaches. Students will read, present, and discuss research papers, and work on an original research project. The goal of the project is to extend machine learning techniques to new problems and produce publishable results.

Expected Work

- **Reading**: students will be reading and reviewing all the required papers, and participating in paper discussions during the class and over the online discussion board.
- **Participation**: students are required to attend all the lectures. Please inform the instructor via email if you cannot make it to the class due to travel or sickness.
- **Team Project**: 2-3 students will form a team to work on a single research project throughout the semester. The project should aim to solve a real problem in the intersection area of machine learning and security/system. Each team will write a project proposal, perform literature surveys, give a short talk in the midterm, and give a final presentation at the end of the semester. Each team is also expected to write up a final project report.
- **Paper Presentation**: students will present papers during the class to lead the discussion.

All deadlines are 11:59 PM (CT) of the specific date (not including paper reviews).

Class Schedule

Date	Area	Week	Link	Notes
Jan 17	Introductions + Guest Lectures	1		Select your project team
Jan 24	Evasion	2	- Towards Evaluating the Robustness of Neural Networks - Towards Deep Learning Models Resistant to Adversarial Attacks - On the Robustness of Domain Constraints	
Jan 31	Poisoning	3	- Trojaning Attacks on Neural Networks - Certified Defenses for Data Poisoning Attacks - Poisoning Web-scale Datasets is Practical	
Feb 7	Membership Inference	4	- Membership Inference Attacks From First Principles - When is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning? - White-box vs Black-box: Bayes Optimal Strategies for Membership Inference	
Feb 14	Model Extraction	5	- High Accuracy and High Fidelity Extraction of Neural Networks - Exploring Connections Between Active Learning and Model Extraction - On the Difficulty of Defending Self-Supervised Learning against Model	Project Proposal: Share <5 slides with prescribed format

			Extraction	
Feb 21	Explanations	6	- "Why Should I Trust You?" Explaining the Predictions of Any Classifier - Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers - Interpretable Deep Learning under Fire	
Feb 28	Privacy Introductions	7	- Deep Learning with Differential Privacy - Bolt-on Differential Privacy for Scalable Stochastic Gradient Descent-based Analytics - Certified Robustness to Adversarial Examples with Differential Privacy	
Mar 6	More Privacy	8	- Privacy Auditing with One (1) Training Run - Learning Differentially Private Recurrent Language Models - Manipulation Attacks in Local Differential Privacy	
Mar 13	Spring Break	9		Break; no class
Mar 20	Mid-Term Presentation	10		Mid-term check-in: share project report
Mar 27	Foundation Models	11	- Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions - Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP) - A Watermark for Large Language Models	
Apr 3	Foundation Models: Attacks	12	- Universal and Transferable Adversarial Attacks on Aligned Language Models - Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection - Poisoning Language Models During Instruction Tuning	
Apr 10	Copyright in ML	13	- Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models - Unlearnable examples: Making personal data unexploitable - On Provable Copyright Protection for Generative Models	
Apr 17	Unlearning	14	- Machine Unlearning - Unlearn What You Want to Forget: Efficient Unlearning for LLMs - Graph Unlearning	Three-quarter check-in: share project report
Apr 24	Unlearning: Potpourri	15	- Adaptive Machine Unlearning - Experimenting with Zero-Knowledge Proofs of Training - When Machine Unlearning Jeopardizes Privacy	
May 1	Crypto + ML	16	- SecureML: A System for Scalable Privacy-Preserving Machine Learning - Cerebro: A Platform for Multi-Party Cryptographic Collaborative Learning - GAZELLE: A Low Latency Framework for Secure Neural Network Inference	
May 8	Final Presentation	17		Final presentation

Grading

- Class attendance and participation: 10%
- In class quiz: 5%
- Paper reviews: 25%
- Paper presentation in class: 10%
- Project: proposal: 10%
- Project: midterm presentation: 10%
- Project: final presentation: 10%
- Project: midterm report + progress update slides: 10%
- Project: final report: 10%

To calculate final grades, I simply sum up the points obtained by each student (the points will sum up to some number x out of 100) and then use the following scale to determine the letter grade: [0-60] F, [60-62] D-, [63-66] D, [67-69] D+, [70-72] C-, [73-76] C, [77-79] C+, [80-82] B-, [83-86] B, [87-89] B+, [90-92] A-, [93-100] A.

Paper Review

We read 3 papers before each class meeting. Before each class, students are expected to read both papers and submit a short review via Campuswire. The deadline for the review is 11:59 PM (CT) on Tuesday before class.

The review should contain sufficient content (about 200-500 words; it can be longer if needed). The review can focus on the key contributions of the paper, the strengths and weaknesses, or potential issues with the experiment methodologies and results. You can also discuss the practical implications of the paper and suggest new ideas. The review should reflect your own thoughts. All the students will post the reviews under the given paper's Campuswire thread. If you are the first to review the paper, you get to summarize the paper and comment on the key contributions. Other students who come later should avoid repeating the same arguments/comments that the previous reviews have already covered. Each review needs to have some original comments that are different from others.

Policies

Late Policy

All the deadlines are hard deadlines. Any late submissions will be subject to point reduction. For paper reviews, and project-related assignments: submitting within 3 days (72 hours) after the deadline = 60% of the points. This policy does not apply to the final project report, for which a late submission is not allowed.

Academic Integrity

Students must follow the university's guidelines on academic conduct (quick link). This course will have a zero-tolerance policy regarding plagiarism. You (or your team) should complete all the assignments and project tasks on your own. When you use the code or tools developed by other people, please acknowledge the source. If an idea or a concept used in your project has been proposed by others, please make the proper citations. All electronic work submitted for this course will be archived and subjected to automatic plagiarism detection. Whenever in doubt, please seek clarifications from the instructor. Students who violate Academic Integrity policies will be immediately reported to the department and the college.

When presenting research papers in the class, you may NOT use the authors' slides directly. Please make your own slides.

Special Accommodations

If you need special accommodations because of a disability, please contact the instructor in the first week of classes.

Diminished Mental Health

Diminished mental health, including significant stress, mood changes, excessive worry, substance/alcohol abuse, or problems with eating and/or sleeping can interfere with optimal academic performance, social development, and emotional wellbeing. The University of Illinois offers a variety of confidential services including individual and group counseling, crisis intervention, psychiatric services, and specialized screenings at no additional cost. If you or someone you know experiences any of the above mental health concerns, it is strongly encouraged to contact or visit any of the University's resources provided below. Getting help is a smart and courageous thing to do – for yourself and for those who care about you.

- **Counseling Center**: 217-333-3704, 610 East John Street Champaign, IL 61820
- **McKinley Health Center**: 217-333-2700, 1109 South Lincoln Avenue, Urbana, Illinois 61801

Varun Chandrasekaran

Assistant Professor