ECE 491F: Computer Software - AI security Syllabus

In this 3-credit course, we will explore the hottest topic in the security community – AI security. You will build on your first experience of hacking AI models, in a way of crafting adversarial examples, and poisoning datasets to get backdoors of AI systems, and explore the up-to-date research papers on designing different safeguard mechanisms.

Aim: for students to gain the knowledge and insights to read and reproduce AI security papers.

Instructor:

Prof. Hanqing Guo
guohanqi@hawaii.edu

Course website:

- The course webpage (TBA) contains links to notes, recordings, and additional materials

Resources:

Introduction to AI Stanford CS231n
Introduction to AI safety Stanford CS120
Awesome Adversarial attacks: Adversarial papers
Awesome Backdoor attacks: Backdoor papers

Pre-requisites:

Solid coding background (Python, PyTorch, bash, git)

Course Schedule

	Topics covered	Reading	Notes
Introduction			
Week 1	Introduction. Course overview. Deep Learning basics.		
Week 2-3	Deep learning training basics. Primer on Pytorch, Colab;	BuildCNN_Pytorch	
Adversarial Attack and Defenses			

Week 4	White-box Adversarial Attack	Intriguing properties of neural networks	Reading report 1
Week 5	Black-box Adversarial Attack	Practical Black-Box Attacks against Machine Learning **Sign-opt**: A query-efficient hard-label adversarial attack	
Week 6	Adversarial Attack defenses	Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity	Reading report 2
Week 7-8	Paper reading and Code Reproduction	FGSM_Pytorch	Project 2
Backdoor Attack and Defenses			
Week 9	Backdoor Attack Basics	Badnets: Evaluating backdooring attacks on deep neural networks	
Week 10-11	Dirty label and clean label Backdoor Attacks	**Clean-label backdoor** attacks	Reading report 3
Week 12	Backdoor Attack Defenses	**Neural cleanse**: Identifying and mitigating backdoor attacks in **neural** networks	
Week 13	Paper reading and Code Reproduction	BackdoorBox	Project2
AI security applications			
Week 14	Watermarking – Protect model stealing	Black-box dataset ownership verification via backdoor **watermarking**	
Week 15	Watermarking – Protect IP leakage	An undetectable **watermark** for generative image models	
Week 16	Project presentation	Final project	Reading report 3

Logistics

There is no paper final exam. Instead, you will need to present your final project.
All deadlines TBD.

Grading

Reading Report	30%
Project	30%
Attendance	10%
Final Report/Presentation	30%

Assignments Guidelines:

- *Unless otherwise specified, all assignments and projects are individual work.*
- *Assignments and Late Penalty*: Assignments and projects will be posted at the class web site. Assignments & projects are due before the beginning of the class on the due day. See Topics and Notes for the due dates. Points will be deducted from late assignments: 50% for the first 24 hours after the due time, 100% after that. No extension will be granted except for documented emergency.
- Starting to work on the assignments as early as possible.
- Identification page: All assignments must have your name, and course number at the top of the first page.
- Please staple all the pages together at the top-left corner.
- Example **reading/code reports**(template):
- Final report example:
- https://arxiv.org/pdf/2303.10137

Policies:

- Attendance Policy: I will check the attendance for every class. If you miss a class without reasonable excuse, your missing will be recorded.
- If you think you have lost some points due to grading errors, make sure you approach the instructor within a week after the assignment, project, or test is returned to you.
- To get the most out of this class, you need to read the reading materials and spend time using computers regularly. Be prepared for a class by preview the material to be covered in that class and participate in discussions and problem-solving exercises, if applicable, in the class.
- *Academic dishonesty will not be tolerated in any form.* The integrity of our program depends on the integrity of the work done by each student. The University expects a student to maintain a high standard of individual honor in his/her scholastic work. Please refer to UH Student Conduct Code at http://www.studentaffairs.manoa.hawaii.edu/policies/conduct_code/ for Academic Honesty, Cheating, Plagiarism, Disciplinary Action, etc.