강의계획서

년도	2026	학기	1학기	학수번호	CSED490H	분반	01
개설학과	컴퓨터공학과					학점	3-0-3
과목명	특강: 인공지능보안					이수구분	전공선택
선수과목							
강좌유형	강의실 강좌						
강의시간	화 목 (14:00 ~ 15:15) -제2공학관 강의실 [109호]						

담당교수	박상돈	소속	인공지능대학원
E-Mail	SANGDON@POSTECH.AC.KR	Homepage	HTTPS://SANGDON.GITHUB.IO/
연구실		연락처	054-279-2396
Office Hours			

강의목표

As AI advances and is being-practical, its safety and security concerns dramatically emerge. In this class, we learn the art of attacking AI systems along with necessary concepts and tools in AI. In particular, we will learn two core concepts, victim models (e.g., LLMs, VLAs, and Agentic AI) and attack methods (e.g., adversarial examples and jailbreaking) along with core optimization tools (e.g., gradient descent, policy optimization, and prompt tuning with LoRA). At the end of this class, students will have a good understanding of trendy AI models, broad aspects of AI red teaming methods, and necessary AI tools. Note that this course is designed for undergraduates -- graduate students may audit.

강의선수/수강필수사항

- Artificial Intelligence

성적평가

중간고사	기말고사	출석	과제	프로젝트	발표/토론	실험/실습	퀴즈	기타	계

비고

- Assignment/Presentation 80% -- three HWs and one final project#
- Participation: 20%

강의교재

도서명	저자명	출판사	출판년도	ISBN

참고문헌 및 자료

Related references include the following:#

강의계획서

년도	2026	학기	1학기	학수번호	CSED490H	분반	01
개설학과	컴퓨터공학과					학점	3-0-3
과목명	특강: 인공지능보안					이수구분	전공선택
선수과목							
강좌유형	강의실 강좌						
강의시간	화 목 (14:00 ~ 15:15) -제2공학관 강의실 [109호]						

- Ian J. Goodfellow et al., “Explaining and Harnessing Adversarial Examples,” ICLR’15.#
- Ashish Vaswani et al., “Attention Is All You Need,” NIPS’17.#
- John Schulman et al., " Trust Region Policy Optimization," ICML’15.

강의진도계획

Week 1: #
- Introduction to AI Security#

Week 2: #
- Preliminary: Neural Networks / SGD #
- Inference-time Attacks: Adversarial Examples / Adversarial Patches / Transfer Attacks#

Week 3: #
- Preliminary: Transformers / LLMs / LCMs / LRMs#
- Preliminary: RAG#

Week 4:#
- Student Presentation and Discussion on HW 1#

Week 5:#
- Preliminary: Diffusion Models#
- Preliminary: Vision-Language-Action Models#

Week 6:#
- Student Presentation and Discussion on HW 2#

Week 7:#
- Preliminary: Optimization for Whitebox Victim Models -- Prompt tuning methods (e.g., LoRA)#
- Preliminary: Optimization for Blackbox Victim Models -- Zero-th Order Optimization#

Week 8:#
- Preliminary: Optimization for Blackbox Victim Models -- RL / Policy Optimization#
- Inference-time Attacks: Prompt Leaking, Prompt Injection, Jailbreaking#

Week 9:#
- Preliminary: Agentic AI / Tool-calling Agents#
- Inference-time Attacks: Current Trends on Red Teaming#

강의계획서

년도	2026	학기	1학기	학수번호	CSED490H	분반	01
개설학과	컴퓨터공학과					학점	3-0-3
과목명	특강: 인공지능보안					이수구분	전공선택
선수과목							
강좌유형	강의실 강좌						
강의시간	화 목 (14:00 ~ 15:15) -제2공학관 강의실 [109호]						

Week 10:#

- Student Presentation and Discussion on HW 3#

Week 11:#

- Training-set Attacks: membership inference attacks#
- Training-set Attacks: data poisoning attacks#

Week 12:#

- Model Attacks: model extraction attacks#

Week 13:#

- Final Remarks: Overview on defense methods#

Week 14:#

- Student Presentation and Discussion on Final Projects#

Week 15:#

- Student Presentation and Discussion on Final Projects

수업운영

학습법 소개 및 기타사항