שם הקורס: נושאים מתקדמים בהתקפות עוינות על מודלי למידה עמוקה ומערכות אבטחה
Advanced topics in adversarial attacks on deep learning models and cyber security systems
מספר הקורס: 236207
סמסטר: חורף תשפ"ד

Lecturers:	**Prof. Avi Mendelson(Avi.mendelson@technion.ac.il) and Dr. Leonid Azriel(leonid.azriel@gmail.com)**
T.A	**Mr. Yaniv Nemcovsky(yanemcovsky@gmail.com)**
When	**Monday, lecture 15:30 – 17:30, Tutorial 17:30 – 18:30**
Where	**Taub 7 + zoom (*)**
Language	**The course will be taught in English**
Credit	**3 academic points**
Format	**Frontal presentations + students' presentations**
Grade	**Based on assignments and final project (see below)**
Prerequisites	**Deep learning course 236781 or 046211 or 097200, or a similar one Computer Architecture course 236267 or a similar one**
Attendance	**(*)Students are required to attend, however zoom attendance and recordings will be provided in special cases.**

Course Description:

This graduate-level course aims to allow students to experience the robustness of machine learning algorithms and their usage in the Cyber security world. The course discusses theoretical, practical, and research aspects, and aims to encourage students to pursue future research and projects in these directions.

The course will focus on **two** main topics: adversarial attacks on deep learning models, and the usage of such models for reverse engineering.

In the scope of adversarial attacks, we will discuss various settings of adversarial attacks and their success in harming the proper conduct of models and systems, as well as the defenses aiming to mitigate such effects. Among other topics, we will discuss white-box and black-box attacks, universal attacks, sparse attacks, patch-based attacks, attacks in real-world scenarios, adversarial training, empirical smoothing and unadversarial examples.

In the reverse Engineering section, we will cover the application of machine learning methods to hardware security. In particular, we will study the usage of GNNs in reverse engineering of integrated circuits, hardware trojans and side channel analysis.

Course assignments and Grade distribution:

- Lectures and presentation's feedback (**10% total**): Answering short open questions concerning each class topic. There will be 8-12 such assignments with binary grades.
- Homework assignments with code (**30% total**, 15% each): submitting 2 homework assignments which will be in the scopes of adversarial attacks on image classification models, and IC reverse engineering tools and frameworks.
- Final project (**60% total**): Each student will freely choose a topic for the final project based on a paper relevant to the course, and correspondingly fulfill the following assignments under the guidance of the course staff:
    - Presenting the original paper (*15%*)
    - Suggesting and implementing an extension to the paper, in code (*30%*)
    - Project presentation and final report *(15%*)

Full syllabus:

In the scope of adversarial attacks on deep learning models, we would review several topics and refer to relevant papers. We detail the list of class topics and relevant papers below:

- Introduction to adversarial attacks:
  - Explaining and Harnessing Adversarial Examples: https://arxiv.org/abs/1412.6572
  - Towards Evaluating the Robustness of Neural Networks: https://arxiv.org/abs/1608.04644
- Adversarial defenses:
  - Ensemble Adversarial Training: Attacks and Defenses: https://arxiv.org/abs/1705.07204
  - Adversarial robustness via noise injection in smoothed models: https://link.springer.com/article/10.1007/s10489-022-03423-5
- Adversarial attack settings:
  - Universal Adversarial Perturbations Against Semantic Image Segmentation: https://openaccess.thecvf.com/content_iccv_2017/html/Metzen_Universal_Adversarial_Perturbations_ICCV_2017_paper.html
  - Efficient and effective sparse adversarial attacks (recent paper by the course staff)
- Applied adversarial attacks:
  - Physical passive patch adversarial attacks on visual odometry systems: https://openaccess.thecvf.com/content/ACCV2022/html/Nemcovsky_Physical_Passive_Patch_Adversarial_Attacks_on_Visual_Odometry_Systems_ACCV_2022_paper.html
  - unadversarial Examples: Designing Objects for Robust Vision: https://proceedings.neurips.cc/paper/2021/hash/816a6db41f0e44644bc65808b6db5ca4-Abstract.html

In the scope of IC reverse engineering and hardware security, we will have several introductory lectures on hardware security and then focus on the machine learning methods in IC reverse engineering and side channel attacks. We detail the list of class topics and relevant papers below:

- Introduction to hardware security:
  - M. Tehranipoor, C. Wang, Introduction to Hardware Security and Trust
- Crypto and crypto implementation 101:
  - Paar, Petzl, Understanding Cryptography
- IC reverse engineering and GNN:

    - Fyrbiak et al, HAL- The Missing Piece of the Puzzle for Hardware Reverse Engineering, Trojan Detection and Insertion
    - Azriel et al, A survey of algorithmic methods in IC reverse engineering
    - Hong et al, ASIC Circuit Netlist Recognition Using Graph Neural Network
- Machine learning methods in side-channel attacks:
    - Mangard et al, Power Analysis attacks: Revealing the secrets of smart cards
    - Barkewitz, Leakage Prototype Learning for Profiled Differential Side-Channel Cryptanalysis
- Hardware Trojans or Security verification

Selected papers:

- Theoretical papers on the phenomenon of adversarial perturbations:
    - Explaining and Harnessing Adversarial Examples: https://arxiv.org/abs/1412.6572
    - Adversarial Spheres: https://arxiv.org/abs/1801.02774
    - The Dimpled Manifold Model of Adversarial Examples in Machine Learning: https://arxiv.org/abs/2106.10151
    - Intriguing properties of neural networks: https://arxiv.org/abs/1312.6199
- Implementations and applications of adversarial attacks:
    - Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks: https://arxiv.org/abs/2003.01690
    - Towards Evaluating the Robustness of Neural Networks: https://arxiv.org/abs/1608.04644
    - Universal Adversarial Perturbations Against Semantic Image Segmentation: https://openaccess.thecvf.com/content_iccv_2017/html/Metzen_Universal_Adversarial_Perturbations_ICCV_2017_paper.html
    - Adversarial Patch: https://arxiv.org/abs/1712.09665
    - Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors: https://link.springer.com/chapter/10.1007/978-3-030-58548-8_1
    - Evading real-time person detectors by adversarial t-shirt: https://www.researchgate.net/profile/Hongge-Chen/publication/336797050_Evading_Real-Time_Person_Detectors_by_Adversarial_T-shirt/links/5dcabe87458515143506854e/Evading-Real-Time-Person-Detectors-by-Adversarial-T-shirt.pdf
    - Over-the-Air Adversarial Flickering Attacks Against Video Recognition Networks: https://openaccess.thecvf.com/content/CVPR2021/html/Pony_Over-the-Air_Adversarial_Flickering_Attacks_Against_Video_Recognition_Networks_CVPR_2021_paper.html
    - Sparse and Imperceivable Adversarial Attacks: https://openaccess.thecvf.com/content_ICCV_2019/html/Croce_Sparse_and_Imperceivable_Adversarial_Attacks_ICCV_2019_paper.html
    - Adversarial Patch Attacks on Monocular Depth Estimation Networks: https://ieeexplore.ieee.org/abstract/document/9207958

    - Physical passive patch adversarial attacks on visual odometry systems: https://openaccess.thecvf.com/content/ACCV2022/html/Nemcovsky_Physical_Passive_Patch_Adversarial_Attacks_on_Visual_Odometry_Systems_ACCV_2022_paper.html

- Adversarial defenses:
    - Ensemble Adversarial Training: Attacks and Defenses: https://arxiv.org/abs/1705.07204
    - CAT: Customized Adversarial Training for Improved Robustness: https://arxiv.org/abs/2002.06789
    - Adversarial training for free!: https://proceedings.neurips.cc/paper/2019/file/7503cfacd12053d309b6bed5c89de212-Paper.pdf
    - Adversarial robustness via noise injection in smoothed models: https://link.springer.com/article/10.1007/s10489-022-03423-5

- Other utilization of adversarial attacks:
    - unadversarial Examples: Designing Objects for Robust Vision: https://proceedings.neurips.cc/paper/2021/hash/816a6db41f0e44644bc65808b6db5ca4-Abstract.html
- IC reverse engineering and hardware security:
    - Alrahis et al, "GNN-RE: Graph Neural Networks for Reverse Engineering of Gate-Level Netlists"
    - Yu et al, "HW2VEC: A Graph Learning Tool for Automating Hardware Security"
    - Chowdhury et al, "ReIGNN: State Register Identification Using Graph Neural Networks for Circuit Reverse Engineering"
    - Yasaei et al, "GNN4IP: Graph Neural Network for Hardware Intellectual Property Piracy Detection"
    - Hong et al, "ASIC Circuit Netlist Recognition Using Graph Neural Network"
    - Yasaei et al, "GNN4TJ: Graph Neural Networks for Hardware Trojan Detection at Register Transfer Level"
    - Alrahis et al, "Embracing graph neural networks for hardware security"
    - Li et al, "Characterize the ability of GNNs in attacking logic locking"