Mahmood Sharif TAU

Assessing and enhancing ML systems’ adversarial robustness in the real world

Abstract:

Adversarial examples have emerged as a profound challenge for machine learning (ML), sparking interest in developing adversarially robust ML models and reliable techniques for assessing robustness. This talk will present our recent efforts on these fronts, offering attacks and defenses under practical constraints. First, I will introduce our work on black-box attacks composing and integrating data augmentations into evasion techniques to promote transferability to unseen models. Second, I will describe our work fooling face-recognition systems via physically realized adversarial accessories attackers can wear to dodge recognition or impersonate others. We evaluated face-recognition systems relying on different sensor types, including visible-light and near-infrared cameras, against evasion and found all were highly vulnerable, even when defended by state-of-the-art techniques. Third, I will present practical attacks against ML-based malware detection. Our attacks interweave binary-diversification techniques and optimization frameworks to mislead malware detection while preserving binaries’ functionality. Unlike prior work, ours manipulate instructions that are a functional part of the binary, rendering them particularly challenging to defend against. I will conclude with our attempts to enhance the robustness of ML-based malware detection via adversarial training.