Abstract

Deep learning-based computer vision models have achieved remarkable progress in recent years, leading to their deployment in a wide range of real-world applications, including safety-critical systems in domains like healthcare, surveillance, defense, and autonomous driving. The rapid transition from controlled lab environments to real-world applications has, however, resulted in a lack of critical scientific scrutiny, leaving these systems vulnerable to a range of security and privacy threats from malicious actors.

This thesis studies two distinct yet complementary threats to the trustworthiness of computer vision systems: *backdoor attacks*, which corrupt a model's behavior, and *model stealing attacks*, which compromise model confidentiality. Backdoor attacks implant hidden behavior during training, causing the model to make targeted predictions on inputs of the attacker's choosing, for example, misclassifying a bus stamped with a specific trigger as a bicycle. Model stealing attacks, on the other hand, replicate proprietary models by repeatedly querying them and using the responses to train a substitute model, enabling the recreation of commercial APIs without access to the original internals. While differing in goals and execution, both backdoor and model stealing attacks reveal systemic vulnerabilities in the machine learning pipeline and threaten the secure deployment of computer vision models.

To address threats to model integrity, we first focus on defending face recognition systems against physically realizable backdoor attacks, such as those triggered by face accessories like hats and sunglasses. Existing defenses primarily target diffuse or pixel-level triggers and do not generalize to structured patterns introduced by physical objects such as face accessories. We bridge this gap by proposing a backdoor detection framework that both flags compromised

systems and recovers the underlying physically-realizable trigger, enabling proactive mitigation. Expanding our investigation of backdoor attacks to object detection, we investigate the backdoor vulnerability of state-of-the-art open-vocabulary object detectors that can recognize arbitrary object categories via natural language prompts. While these models expand recognition capabilities, their security under backdoor attacks remains unexplored. We demonstrate that they are indeed susceptible, and introduce a novel multi-modal attack that exploits prompt tuning to implant backdoors in these networks.

The latter half of the thesis focuses on model stealing attacks. We first address medical imaging models deployed via restricted-access APIs, where both data scarcity and high query costs make stealing particularly challenging. We design a query-efficient attack tailored for such settings and validate its effectiveness by replicating models for gallbladder cancer and COVID-19 diagnosis, underscoring confidentiality risks in medical AI. Finally, we analyze how the shift from conventional models to large foundation models affects model stealing capabilities. Our analysis reveals that fine-tuned vision transformers and contrastive models are significantly more vulnerable to stealing attacks as compared to smaller convolutional networks, raising serious concerns for commercial API deployment.

Overall, the thesis contributes new attack and defense methodologies across conventional and foundation models, and highlights the urgent need for principled safeguards in trustworthy computer vision.