Physical Adversarial Examples for Image Classifiers and Object Detectors

Submitted by akarns on Thu, 09/27/2018 - 9:35am. Contributor:
Ivan Evtimov

BIO

Ivan Evtimov is a second-year Ph.D. student in the Computer Security and Privacy Lab at the Paul G. Allen School of Computer Science at the University of Washington where he is advised by Tadayoshi Kohno and works closely with Earlence Fernandes. While he is fascinated by a variety of security and privacy problems, Ivan focuses on the intersection of security and machine learning. In particular, he seeks to understand how and whether adversarial examples pose a real threat to deployed systems. His research so far has shown that popular computer vision deep learning models are vulnerable to physical attacks with stickers that do not require digital access to the system. Ivan is also a member of the Tech Policy Lab where he collaborates with scholars from a variety of disciplines and regularly engages in discussions on topical issues in law and policy around technology.

ABSTRACT

Recent studies show that the state-of-the-art deep neural networks (DNNs) are vulnerable to adversarial examples, resulting from small-magnitude perturbations added to the input. Given that that emerging physical systems are using DNNs in safety-critical situations, adversarial examples could mislead these systems and cause dangerous situations. Therefore, understanding adversarial examples in the physical world is an important step towards developing resilient learning algorithms. This poster presents two works that made progress towards that goal. First, we introduce the robust physical perturbations (RP2) algorithm. It can generate visual adversarial perturbations that are robust under different physical conditions. Using the real-world case of road sign classification, we show that adversarial examples generated using RP2 achieve high targeted misclassification rates against standard-architecture road sign classifiers in the physical world under various environmental conditions, including viewpoints. Second, we extend physical attacks to more challenging object detection models, a broader class of deep learning algorithms widely used to detect and label multiple objects within a scene. Improving upon RP2, we create perturbed physical objects that are either ignored or mislabeled by object detection models.

Ivan Evtimov

Other available formats:

Physical Adversarial Examples for Image Classifiers and Object Detectors
Switch to normal viewerSwitch to experimental viewer

Adversarial Machine Learning object detection road signs classifiers adversarial examples Poster Posters