Abstract

Adversarial examples represent a major security threat for emerging technologies that use machine learning models to make important decisions. Adversarial examples are typically crafted by performing gradient descent on the target model, but recently, researchers have started to look at methods to generate adversarial examples without requiring access to the target model at decision time. This thesis investigates the place that generative attacks have in the machine learning security threat landscape and improvements that can be made on existing attacks. By evaluating attacks and defenses on image classification problems, we found that generative attackers are most relevant in black box settings where query access is given to the attacker before test time. We evaluate generative methods in these settings as standalone attacks and we give examples of how generative attacks can reduce the number of queries or amount of time required to produce a successful adversarial example by narrowing the search space of existing optimization algorithms.