Abstract

Recent studies have highlighted the vulnerability of deep neural networks (DNNs) to adversarial examples - inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed inputs result in the model outputting incorrect answers with high confidence. Gradient-based optimization is used to find said adversarial examples by jointly minimizing the perturbation while maximizing the probability that the generated example causes the target model to misclassify. This approach can be readily applied in the white-box case, where the attacker has complete access to the target model and thus can compute the gradients via backpropagation. We extend such approaches to the black-box case, where the attacker is only given query access and therefore incapable of directly computing the gradients. We introduce ZOO, which uses the finite difference method to estimate the gradients for optimization from the output scores. Furthermore, we also improve the state-of-the-art in the no-box case, where the attacker is not even capable of querying the target model. We introduce EAD, which incorporates L_1 minimization in order to encourage sparsity in the perturbation, hence generating more robust adversarial examples in the white-box case which can transfer to unseen models. Through experimental results attacking state-of-the-art models trained on the MNIST, CIFAR-10, and ImageNet datasets, we validate the effectiveness of the proposed attacks. In addition, we demonstrate that the proposed attacks can successfully attack recently proposed defenses in these limited access settings. We show that ZOO can succeed against the state-of-the-art ImageNet defense, Ensemble Adversarial Training, while EAD can succeed against the state-of-the-art MNIST defense, the Madry Defense Model, and input transformation defenses, such as Feature Squeezing.