**Ensemble Adversarial Training: Attacks and Defenses**

Adversarial examples are perturbed inputs designed to fool machine learning models. Adversarial training injects such examples into training data to increase robustness. To scale this technique to large datasets, perturbations are crafted using fast single-step methods that maximize a linear approximation of the model's loss. We show that this form of adversarial training converges to a degenerate global minimum, wherein small curvature artifacts near the data points obfuscate a linear approximation of the loss. The model thus learns to generate weak perturbations, rather than defend against strong ones. As a result, we find that adversarial training remains vulnerable to black-box attacks, where we transfer perturbations computed on undefended models, as well as to a powerful novel single-step attack that escapes the non-smooth vicinity of the input data via a small random step. We further introduce Ensemble Adversarial Training, a technique that augments training data with perturbations transferred from other models. On ImageNet, Ensemble Adversarial Training yields models with strong robustness to black-box attacks. In particular, our most robust model won the first round of the NIPS 2017 competition on Defenses against Adversarial Attacks. However, subsequent work found that more elaborate black-box attacks could significantly enhance transferability and reduce the accuracy of our models.