We consider the threat model of unrestricted adversarial examples — adversarial examples that are beyond small perturbations.

We propose to effectively construct unrestricted adversarial examples with conditional generative models. We consider the threat model based adversarial examples are special cases of unrestricted adversarial examples.

We show that the defenses against perturbation-based adversarial examples, including provable defenses, are susceptible to unrestricted adversarial examples. Our attacks uniformly achieved over 84% success rates across all the datasets in our experiments and showed moderate degree of transferability.

Code: