Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks

We consider adversarial examples for image classification in the black-box decision-based setting. Here, an attacker cannot access confidence scores, but only the final label. Most attacks for this scenario are either unreliable or inefficient. Focusing on the latter, we show that a specific class of attacks, Boundary Attacks, can be reinterpreted as a biased sampling framework that gains efficiency from domain knowledge. We identify three such biases, image frequency, regional masks and surrogate gradients, and evaluate their performance against an ImageNet classifier. We show that the combination of these biases outperforms the state of the art by a wide margin. We also showcase an efficient way to attack the Google Cloud Vision API, where we craft convincing perturbations with just a few hundred queries. Finally, the methods we propose have also been found to work very well against strong defenses: Our targeted attack won second place in the NeurIPS 2018 Adversarial Vision Challenge.