Simple Black-box Adversarial Attacks

Seungmin Lee (profile2697@gmail.com; 2013-11420), Dept. of Computer Science and Engineering, Seoul National University

1. Motivation

Even though white-box attack methods have shown their effectiveness, these methods are impractical in that they require the architectures of target models. On the other hand, black-box attack methods are more applicable because they do not assume the knowledge of target models and only utilize an output distribution. However, the black-box attack methods also should minimize the number of queries needed to create a successful adversarial image. Therefore, finding an effective black-box attack method is still an open problem. Previous works try to tackle the black-box attack with some complex procedures such as training proxy models. This paper, on the other hand, suggests a simple and effective algorithm inspired by the fact that the vulnerability is inevitable in high-dimensional spaces with some reasonable assumptions [1, 2].

2. Notations

The *untargeted* black-box attack can be formulated as the following optimization problem:

$$\arg\min_{\delta} \mathcal{S}_y(\mathbf{x}+\delta) \text{ subject to: } \|\delta\|_2 < \rho, \text{ quires} \leq B \quad (1)$$

where $\mathcal{S}_y(\cdot)$ is a confidence of a model h classifiying a input as y, δ is a adversarial pertubation that the norm of it is bounded by ρ, and B is a query limit. Similarly, we can define the *targeted* black-box attack as $\arg\max_{\delta} \mathcal{S}_{y'}(\mathbf{x}+\delta)$ where $y \neq y'$. For the sake of simplicity, I will focus on the untargeted black-box attack.

3. Proposed Method

The proposed method is based on the previous studies that the vulnerability of a model is in high-dimensional spaces [1, 2]. These prior studies mean merely adding a set of orthogonal vectors to an image can make an adversarial example. Therefore, for each iteration, the proposed method samples a vector from a basis and add it to an input image by multiplying a small step size ϵ. Then, the method queries the modified image to the target model to get a confidence change. If the confidence becomes lower, the modified image is maintained. Otherwise, the method rollbacks the perturbation and tries another perturbation by sampling another vector from the basis. The proposed method repeats the iteration until the output label is changed or until it runs out of the budget.

4. Results

The proposed method tests its effectiveness using a cartesian basis and discrete cosine basis. The cartesian basis is simply sampling one pixel of an image and perturbing it. The proposed method shows better results than the previous techniques, even though it is simple.

5. Personal Memo

I think the relation between the proposed method and the white-box methods is similar to the relationship between black-box optimization methods and policy gradient methods in policy-based methods of reinforcement learning. Therefore, I think using other effective black-box optimization methods such as evolutionary strategy or cross-entropy method would be better than the proposed method. I think this is definitely worth a try.

References

[1] A. Fawzi, et al. Adversarial vulnerability for any classifier. In *NeurIPS*, 2018. 1

[2] A. Shafahi, et al. Are adversarial examples inevitable? In *CoRR, abs/1809.02104*, 2018. 1