A NOTE ON HYPERPARAMETERS IN BLACK-BOX ADVERSARIAL EXAMPLES

Jamie Hayes
University College London
j.hayes@cs.ucl.ac.uk

ABSTRACT

Since Biggio et al. (2013) and Szegedy et al. (2013) first drew attention to adversarial examples, there has been a flood of research into defending and attacking machine learning models. However, almost all proposed attacks assume white-box access to a model. In other words, the attacker is assumed to have perfect knowledge of the models weights and architecture. With this insider knowledge, a white-box attack can leverage gradient information to craft adversarial examples. Black-box attacks assume no knowledge of the model weights or architecture. These attacks craft adversarial examples using information only contained in the logits or hard classification label. Here, we assume the attacker can use the logits in order to find an adversarial example. Empirically, we show that 2-sided stochastic gradient estimation techniques are not sensitive to scaling parameters, and can be used to mount powerful black-box attacks requiring relatively few model queries.

1 GRADIENT ESTIMATION METHODS

Black-box attacks usually rely on either the *transferability* property of adversarial examples (Papernot et al. (2016)) or gradient estimation techniques. Gradient estimation techniques such as finite differences come at the expense of the number of model queries. To estimate the gradient of a d dimensional vector requires $2d$ model queries, since the gradient of each dimension is measured independently. For this reason, recent work has explored stochastic methods for approximating the true gradient (Ilyas et al. (2018); Uesato et al. (2018)). Given a model f [1] and an input $x \in \mathbb{R}^d$, we estimate the gradient by one of the following methods:

$$\frac{1}{n}\sum_{i=i}^{n}\left[\frac{f(x+\delta\Delta_i)-f(x)}{\delta}\right]\cdot\xi_i \tag{1}$$

$$\frac{1}{n}\sum_{i=i}^{n}\left[\frac{f(x+\delta\Delta_i)-f(x-\delta\Delta_i)}{2\delta}\right]\cdot\xi_i \tag{2}$$

where $\delta \in \mathbb{R}$ is a small constant and $\xi_i, \Delta_i \in \mathbb{R}^d$ are random vectors sampled from some distribution P. We observed that the choice of P is largely unimportant; black-box attack success for $P \sim \mathcal{N}(0, I)$ (NES (Ilyas et al. (2018))), $P \sim \text{Bernoulli}_{\pm 1}$ (SPSA (Uesato et al. (2018))), $P \sim \mathcal{U}(-1, 1)$ (RDSA) is approximately equivalent. Gradient estimation given by (1) is referred to as 1-sided, and gradient estimation given by (2) is referred to as 2-sided. For all attacks, we take $\xi_i = \Delta_i^{-1}$, and the estimation can be viewed as finite differences on a random basis.

2 EXPERIMENTS

We run the PGD attack (Madry et al. (2017)) with gradient estimation, for $\epsilon = 0.05$ on the NIPS 2017 adversarial vision competition dataset (Kurakin et al. (2018)). This consists of 1000 Imagenet-

[1] For notational convenience, we represent both the model evaluation and loss function evaluation of an input with respect to a target label by f.

Table 1: Attack results for different estimation methods.

Attack	δ	**Success Rate (%)**	**Median Queries**
NES (2-sided)	1e-2	100	14815
	1e-3	100	17723
	1e-4	100	20222
RDSA (2-sided)	1e-2	100	13719
	1e-3	100	18564
	1e-4	100	19635
SPSA (2-sided)	1e-2	100	14586
	1e-3	100	17442
	1e-4	100	19916
SPSA (1-sided)	1e-2	9.88	50206
	1e-3	91.59	23816
	1e-4	99.89	**10244**

like images of size $299 \times 299 \times 3$. For each image, we select the least likely class as the adversarial example target class. We consider the attack successful if the predicted class is the target class and $\|x - x_{adv}\|_\infty < \epsilon$, and the attack requires fewer than one million queries to the model. Table 1 shows the results for different choices of random directions. For 2-sided attacks, all achieve perfect success rates in crafting adversarial examples, while exhibiting little sensitivity to the choice of δ. For 1-sided SPSA, the attack is extremely sensitive to the choice of δ; for a δ of 0.01, the attack is successful fewer than one times in ten, while a δ of 0.0001 the attack has near perfect success rate and also requires on average only 10244 queries, 3475 fewer than the best 2-sided attack.

In conclusion, we found that the choice of random direction is largely unimportant in practical attacks. An attacker choosing a 1-sided perturbation may require fewer queries to the model, however this is heavily dependent on the choice of δ. Reproducible code can be found at `https://github.com/jhayes14/black-box-attacks`.

REFERENCES

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598*, 2018.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. *arXiv preprint arXiv:1804.00097*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.