Poster: Adversaries Don't Care About Averages: Batch Attacks on Black-Box Classifiers

Fnu Suya
University of Virginia
fs5xz@virginia.edu

David Evans
University of Virginia
evans@virginia.edu

Yuan Tian
University of Virginia
yt2e@virginia.edu

***Abstract*—We study black-box attacks on deep learning models where the adversary's goal is to acquire a batch of adversarial examples while minimizing the total number of queries. Our basic hypotheses are that (1) there is high variance on the number of queries across different seed images and (2) there exist efficient strategies to identify images which require fewer queries. Hence, the cost of generating each adversarial example in a batch attack can be much less than the average attack cost by focusing resources on the easiest seeds. Our preliminary results on CNN models for CIFAR-10 dataset show that both hypotheses hold and that a simple greedy strategy can provide close to optimal performance, reducing the total cost to find batch of adversarial examples to less than $1/25^{\text{th}}$ of the cost of a random search strategy when the attacker can select target seeds from a large pool of possible seeds.**

1. Introduction

Machine learning models are often vulnerable to small but carefully-crafted adversarial perturbations [1, 4, 7]. An important branch of research in adversarial machine learning focuses on attacking machine learning classifiers in black-box settings, where only query access to the underlying model is assumed. Black-box attacks can be coarsely categorized as either transferability-based attacks [6] or query-based gradient estimation methods [2]. The first category focuses on training a local model which mimics the decision boundary of the target model. With a trained local model, the adversary generates adversarial samples by attacking the local model and then transfers these samples to the target model. The second category is identical to existing white-box attack strategies while the gradient information is numerically approximated (in contrast to back-propagation in white-box scenarios). Hence, the former approach needs a one-time batch of queries to train the local model (and can then produce new likely-adversarial example in future without any additional queries). However, the number of queries needed to produce an adequate local model before finding the first adversarial example may be large, and transferability-based method suffer from transfer loss as not all adversarial examples from the local model successfully transfer to the target model. In contrast, the latter approach seems to require a large number of queries for each instance.

We consider an attacker whose goal is to produce a batch of different adversarial examples with the fewest total queries. Such an attacker goal is motivated by many potential uses of adversarial examples including medical image insurance fraud [3] where each adversarial example found can be exploited for some value by the attacker, but each query to the target model poses some risk of detection. A key hypothesis underlying this work, which we confirm experimentally, is that there is a large variance in the difficulty of finding adversarial examples across different seed images. When this is true, the next question is whether it is possible to identify easy-to-attack images with low cost, and reduce the total effort required by focusing an attacker's resources on attacking those seeds.

2. Seed Variability

Figure 1 shows the variation in the number of queries needed to find adversarial examples for different seeds for CIFAR10 dataset [5]. Each selected image is attacked individually in black-box scenario by ZOO attack proposed in [2] and AutoZoom (https://github.com/chunchentu/AutoZOOM, a query efficient version of ZOO). We use a standard definition of adversarial examples: the attacker's goal is given a seed $\mathbf{x}$, find an adversarial example $\mathbf{x}'$ such that the model's output on $\mathbf{x}'$ is in the target class and the distance between $\mathbf{x}$ and $\mathbf{x}'$ is below some perturbation magnitude limit D (for the distance measure, we use L_2 distance and $D = 3$ for the results shown).

The figure shows a high variance for the number of queries across the seeds, for both ZOO and AutoZoom method. Median number is less than the average and small fraction of images contribute significantly to the average number of queries. Hence, a smart attacker would avoid querying those hard-to-attack images and devote most of its resources on the promising images with potentially lower query numbers. Given the same maximum number of optimization iterations and perturbation magnitude limit, ZOO reliably finds adversarial samples for all of the images, while AutoZoom fails on some fraction of the images (for images > 320, all number of queries are identical because maximum query limit of 10,000 is reached and the query process is stopped).

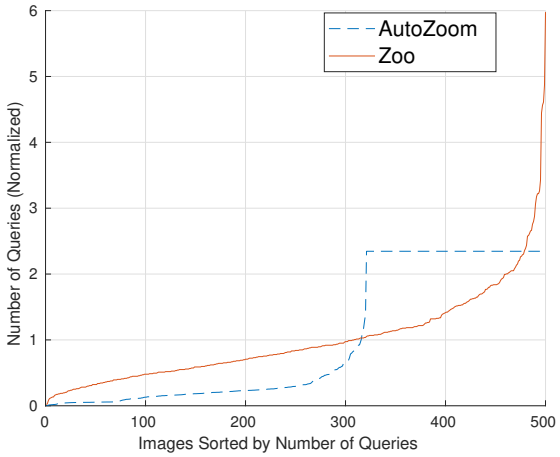

Figure 1: Distribution of Number of Queries. The images are sorted by number of queries for each attack, and the plotted values are normalized by the average number of queries for that attack over the full set of 500 seeds. For ZOO, the average is 32,590 queries; for AutoZoom, 4,263. We set coefficient $c = 1$, which balances the terms in mis-classification and perturbation magnitude [1, 2]. Maximum number of optimization iterations is set to 5000 for both ZOO and AutoZoom. For ZOO, one optimization iteration costs 256 queries to the model while AutoZoom costs 2 queries per optimization iteration. The results shown are for a targeted attack from seeds in class 3 (cat) to target class: 8 (ship). Our experiments with other seed and target classes produced similar results.

3. Batch Attacks

Next, we provide results on identifying the "promising" easy-to-attack images. We hypothesize that easy-to-attack images typically have higher target classification probability. Hence, the simple heuristic solution is to select the images with highest target class probability. In Figure 2 we compare this greedy strategy a random search strategy where the attacker selects images without any consideration to difficult. We also show the retroactive-optimal cost, for an attacker with oracle knowledge of the actual number of queries needed for each sample before starting the attack. The x-axis is the target number of adversarial samples attacker needs, and the y-axis gives the average number of queries to find each of those examples. To get rid of the influence of some extremely high values, we normalize all the query numbers of ZOO and AutoZoom by their retroactive optimal values and present them in log 10 scale. (For the random search strategy, we report the average over 100 executions.)

The results show that when the adversary has access to many more seeds than the target number of adversarial examples, the cost of the attack can be greatly reduced, even using the simple greedy heuristic. For AutoZoom, when the attacker is interested in attacking 50 images out of 500 images, this simple heuristic strategy can return images which take only 3.7% of that returned by random search strategy and average number of queries of the greedy strategy are only 1.5 times compared to the retroactive-optimal value.

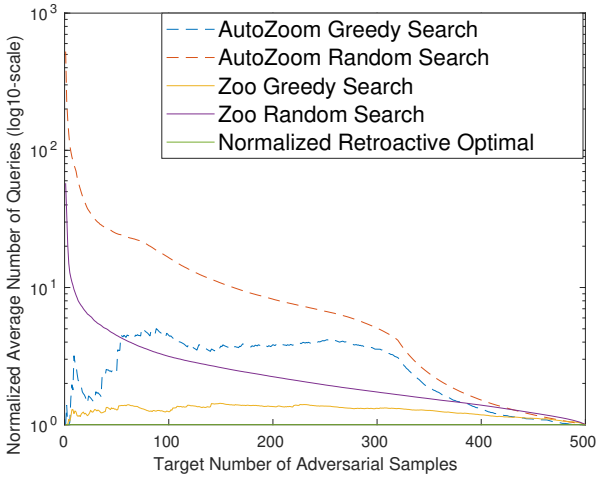

Figure 2: Performance of Greedy Search Strategy compared to Random Search and Retroactive Optimal Strategy

4. Conclusion

We study the problem of query efficient batch attacks to black-box classifiers. We verify that there exists high variance in the number of queries across different images and also propose a simple heuristic image search strategy based on target class probability. Experiments demonstrated the effectiveness of the greedy strategy in finding subset of images with fewest number of queries. At some level, these results are unsurprising—it is easier to find adversarial examples when the starting seed it closer to the target than when it is further away. This suggest that perhaps more consideration is needed about how adversaries will actually exploit the adversarial examples they find and how we should evaluate attacks and defenses.

References

[1] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017.

[2] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *10th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2017.

[3] S. G. Finlayson, I. S. Kohane, and A. L. Beam, "Adversarial Attacks Against Medical Deep Learning Systems," *arXiv preprint arXiv 1804.05296*, Apr. 2018.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[5] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[6] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*, 2017.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.