Efficient adversarial robustness evaluation of AI models with limited access

January 31 2019, by Pin-Yu Chen, Sijia Liu

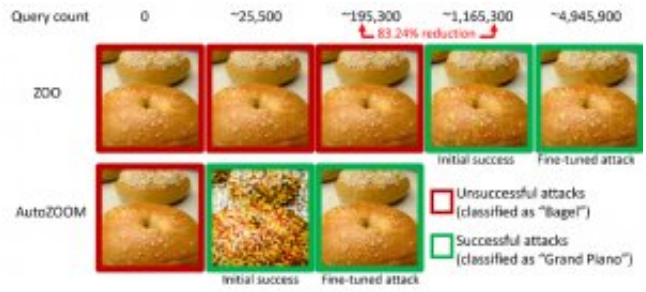

Figure 1: Performance comparison in turning a bagel image into an adversarial bagel image classified as a “grand piano” using ZOO and AutoZOOM attacks. Credit: IBM

Recent studies have identified the lack of robustness in current AI models against adversarial examples—intentionally manipulated prediction-evasive data inputs that are similar to normal data but will cause well-trained AI models to misbehave. For instance, visually imperceptible perturbations to a stop sign can be easily crafted and lead a high-precision AI model towards misclassification. In our previous paper published at the European Conference on Computer Vision (ECCV) in 2018, we validated that 18 different classification models trained on ImageNet, a large public object recognition dataset, are all vulnerable to adversarial perturbations.

Notably, adversarial examples are often generated in the "white-box"

setting, where the AI model is entirely transparent to an adversary. In the practical scenario, when deploying a self-trained AI model as a service, such as an online image classification API, one may falsely believe it is robust to adversarial examples due to limited access and knowledge about the underlying AI model (aka the "black-box" setting). However, our recent work published at AAAI 2019 shows the robustness due to limited model access is not grounded. We provide a general framework for generating adversarial examples from the targeted AI model using only the model's input-output responses and few model queries. Compared to the previous work (ZOO attack), our proposed framework, called AutoZOOM, reduces at least 93% model queries in average while achieving similar attacking performance, providing a query-efficient methodology for evaluating adversarial robustness of AI systems with limited access. An illustrating example is shown in Figure 1, where an adversarial bagel image generated from a black-box image classifier will be classified as the attack target "grand piano". This paper is selected for oral presentation (January 29, 11:30-12:30 pm @ coral 1) and poster presentation (January 29, 6:30-8:30 pm) at AAAI 2019.

In the white-box setting, adversarial examples are often crafted by leveraging the gradient of a designed attack objective relative to the data input for the guidance of adversarial perturbation, which requires knowing the model architecture as well as the model weights for inference. However, in the black-box setting acquiring the gradient is infeasible due to limited access to these model details. Instead, an adversary can only access to the input-output responses of the deployed AI model, just like regular users (e.g., upload an image and receive the prediction from an online image classification API). It was first shown in the ZOO attack that generating adversarial examples from models with limited access is possible by using gradient estimation techniques. However, it may take a huge amount of model queries to craft an adversarial example. For example, in Figure 1, ZOO attack takes more than 1 million model queries to find the adversarial bagel image. To

accelerate the query efficiency in finding adversarial examples in the black-box setting, our proposed AutoZOOM framework has two novel building blocks: (i) an adaptive random gradient estimation strategy to balance query counts and distortion, and (ii) an autoencoder that is either trained offline with unlabeled data or a bilinear resizing operation for acceleration. For (i), AutoZOOM features an optimized and query-efficient gradient estimator, which has an adaptive scheme that uses few queries to find the first successful adversarial perturbation and then uses more queries to fine-tune the distortion and make the adversarial example more realistic. For (ii), as shown in Figure 2, AutoZOOM implements a technique called "dimension reduction" to reduce the complexity of finding adversarial examples. The dimension reduction can be realized by an offline trained autoencoder to capture data characteristics or a simple bilinear image resizer which does not require any training.

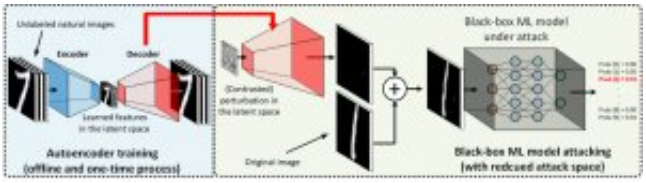

Figure 2: Illustration of the dimension reduction technique used in AutoZOOM for query redemptions. The decoder can be either an offline trained autoencoder or a bilinear resizing operation that does not require any training. Credit: IBM

With these two core techniques, our experiments on black-box deep neural network based image classifiers trained on MNIST, CIFAR-10 and ImageNet show that AutoZOOM attains a similar attacking performance while achieving a significant reduction (at least 93%) in the mean query counts when compared to the ZOO attack. On ImageNet, this drastic reduction means millions of fewer model queries, rendering

AutoZOOM an efficient and practical tool for evaluating adversarial robustness of AI models with limited access. Moreover, AutoZOOM is a general query redemption accelerator that can readily apply to different methods for generating adversarial examples in the practical black-box setting.

The AutoZOOM code is open-sourced and can be found here. Please also check out IBM's Adversarial Robustness Toolbox for more implementations on adversarial attacks and defenses.

More information: Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models. European Conference on Computer Vision (ECCV), 2018

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh, "ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models," ACM Workshop on AI-Security, 2017

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Pai-Shun Ting, Shiyu Chang, and Lisa Amini. Zeroth-Order Stochastic Variance Reduction for Nonconvex Optimization. *Advances in Neural Information Processing Systems* (*NeurIPS*), 2018

Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin Black-box Adversarial Attacks with Limited Queries and Information. International Conference on Maching Learning (ICML), 2018

Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. International Conference on Learning

Representations (ICLR), 2019

This story is republished courtesy of IBM Research. Read the original story here.

Provided by IBM

Citation: Efficient adversarial robustness evaluation of AI models with limited access (2019, January 31) retrieved 11 May 2024 from https://phys.org/news/2019-01-efficient-adversarial-robustness-ai-limited.html