Blacklight: Defending Black-Box Adversarial Attacks on Deep Neural Networks

Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y. Zhao
Summarised by Siddhartha Dutta

July 10, 2020

1 Summary

1.1 Introduction

Deep Neural Networks (DNNs) are vulnerable to adversarial attacks. Adversarial examples can easily be defined by adding small image perturbations to input images. These perturbations are hardly misleading to humans but can produce misclassifications in even well-performing DNN models. Two types of adverarial attacks have been defined in this regard - white box and black box attack. In a white box attack, the attacker has access to all internal and external model details like architecture, model weights, outputs etc., while in black box attack the attacker is only restricted to querying the model through queries and classification outputs. *Li et al.* in the paper *Blacklight: Defending Black-Box Adversarial Attacks on Deep Neural Networks* [1] proposes a novel light weight scalable system against (query based) black-box adversarial attacks in image based DNNs.

Blacklight relies on the fact that a query-based attacker generates a similar query image to a previously queried image to generate a perfect adversarial example, i.e., he or she must submit atleast two queries which have small differences in image space. Blacklight aims to detect any such event from happening. The core of Blacklight, which is also the central contribution of the paper, lies a probabilistic fingerprint algorithm that computes hash values for each input query. The idea is simple. A recent query whose probabilistic fingerprint matches any previously queried input by a certain threshold is flagged as a possible attack. However, the challenge lies in creating an efficient algorithm that is powerful enough to detect a sophisticated attack as well as few false positives.

1.2 Existing literature

A lot of literature is available concerning the vulnerability of DNNs regarding adversarial attacks. Even after continuous advancement in white box defense, a foolproof white box defense is still not available. In contrast, black box attacks are more realistic to approach. Majorly black box attacks are classified into two types - *Substitute model attacks* and *Query based attacks*. While there have been effective defenses against substitute model based attacks, a successful query based defense is yet to be seen. This paper by *Li et al.* concerns about adversarial defense, where the authors show *Blacklight's* capability in defending different query based adversarial attacks.

1.3 Threat model

The authors define a threat model and a defender model that they used for designing *Blacklight*. Given a DNN model, $\mathbb{F}$ which will be used to design adversarial attacks on, input x_0, target label L_t. An adversary repeatedly queries $\mathbb{F}$ with $(x_0, x_1, \ldots, x_n)$, where x_n is the last query. An attack is considered successful if:

$$\mathbb{F}(x_n) = L_t \text{ and } ||x_n - x_0||_p < \epsilon \tag{1}$$

where ϵ is the perturbation target and n is in order of 10^3 and 10^6. Further assumptions are made such as - the attacker does not know any model details, has sufficient computation resources, and has several user accounts and IP addresses. The defender, on the other hand, hosts the target model $\mathbb{F}$, has limited resources, and has access to all the queries submitted to it. The authors add a further constraint that the system periodically resets the past queries, *e.g.,* every 1 or 2 days, to prevent the system from overflowing.

1.4 Design

In lieu of the above threat and defender models, the authors propose the design of *Blacklight*. It consists of three important parts: (1) *Pixel quantization*, (2) *Fingerprint computation*, (3) *Comparing and matching fingerprints*. Pixel quantization converts continuous pixel value space to a finite set of discrete hash values and therefore nullifies any minor pixel modifications. To compute fingerprint, the quantized image is first flattend and represented as a vector. A sliding window of size *w* and a slide step of size *p* is applied to the vector to extract numerous vectors of size *w*.

For each vector a hash value using *SHA-3* combined with a random salt is computed. This creates a large set of hash values. Instead of using all hash values, the authors decides to take top S values sorted in numerical order and call this as probabilistic fingerprint. Finally, Blacklight compares the fingerprint with all fingerprint stored in the database. Any fingerprint in the database that has same or more than T common hash entries when compared to the input image is flagged as an attack image. The values of w, p, S and T are configured in the system. Typically values of S and T determine accurate detection and low false positive rate.

1.5 Experiments and Results

The authors apply *Blacklight* against five representative black box attacks and four different image classification tasks. The representative black box attacks were *NES Query Limited and Label Only*, *HopSkipJumpAttack*, *ECO*, and *Boundary Attack*. Whereas the image classification tasks were *digit recognition using MNIST*, *traffic sign recognition using GTSRB*, *object recognition using CIFAR10* and *ImageNet* datasets. The authors select random 100 images from each test dataset as test image and only 50 images from ImageNet due to its increased computation costs on high resolutions. Following the model in eq. 1, ϵ was set to 0.05 and the distance metric (p) was set to L_2. The only exception was for MNIST, where the authors claim L_∞ and $\epsilon = 0.1$ were necessary for an attack to succeed. Among the other config parameters chosen were $w = 20$, $p = 1$, $S = 50$, and $T = 25$ with the exception for MNIST, where $w = 50$ as it has mostly black background. The authors do not provide any specific reason for choosing these values. Each attack is run until either it is successful or it has completed 100K queries.

To evaluate *Blacklight* the authors use 6 different metrics - (1) *False positive rate*, (2) *Attack detection rate*, (3) *Detection coverage*, (4) *Average no. of queries to detect an attack*, (5) *Attack success rate with mitigation*, and (6) *Detection overhead*. The authors claim to achieve a 100% detection rate, 96% detection all attack queries with exception of *Boundary Attack* with performed the worst for MNIST at 61%. The system also detects most of attack in 2-6 queries with exception of *Boundary Attack* where it took 48 queries for ImageNet. Attack success rate for all of them was 0%. False positive rate was less than 0.1% for T=25. The authors also show *Blacklight*'s matching cost is near constant and independent of number of fingerprint data in the database. Finally the authors test their system against *Skilled adversary* (*e.g.,* using generic image augmentation) and *Oracle Attacks*. While the detection rate remained at 100% for *Skilled Adversary, Oracle Attacks* could not be successfully defended.

2 Discussion

In my opinion, the authors have managed to develop a thorough query-based black box defense system and have tested it in majority aspects. The strength of the system lies in the fact that the system was designed considering a reasonable threat model and a limited capability defender. There are, however, a few aspects where the system might fail, or would be interesting to see how it behaves.
(1) The authors only evaluate the systems on datasets that have a unique set of images. They do not mention if this approach will work on datasets with continuous frames, *e.g.*, *KITTI* [2] or *Human3.6m*[3] datasets. My understanding is the system might flag consequent frames as an attack since they are similar to the previous.
(2) Efficient image quantization is an important aspect of detecting similar images. The authors do not provide in-depth details of how this quantization has been done, *e.g.,* no mention of quantization thresholds.
(3) With regards to fingerprinting, consider a specific attack where the adversary keeps perturbing exactly T pixels by more than ϵ in subsequent attack images such that they always remain lowest after doing a sort. Fingerprint computation for each image will be unique and hence will be accepted by the system. Using the rest of the pixels, that attacker can create an adversarial image. T can be found in approximately $\log_2(T)$ attack trails by using a binary search approach, initially assuming a high T.
(4) In my opinion, the sample size 100 chosen by authors is very less. Moreover, since the sampling was random, it would be interesting to see how the system behaves when you have a specific problem, *e.g.,* when the adversary tries to fool a face identification network using General Adversarial Networks (GANs).
(5) For higher resolution images the response 400ms can affect realtime performance of the original application.

References

[1] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Blacklight: Defending black-box adversarial attacks on deep neural networks. *arXiv preprint arXiv:2006.14042*, 2020.

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.