Differences between adversarial examples in the digital and physical worlds

Ju Jinquan, Park Sujin, Lee Hoonjae

Dongseo Univ.

jujinquan4@gmail.com

디지털 세계와 물리적 세계의 적대적 사례에 대한 차이점 분석

쥐진취안, 박수진, 이훈재

동서대학교

Abstract

So far, most work has assumed attacks on a model in the digital world, where the attacker can feed data directly into the machine learning system. The situation is more complex for systems operating in the physical world, such as those that use signals from cameras and other sensors as input, which are also vulnerable to adversarial examples. We propose a method to compare adversarial examples' attack success rate, robustness, and performance impact on clean test data in different scenarios. Experimental evaluation shows that adversarial examples have good deception capabilities in different scenarios. At the same time, there are threats from defense techniques and human perception.

Ⅰ. Introduction

Suppose there is a machine learning system M and an input sample X, which we call a clean example. We assume that sample X is correctly classified by machine learning system, i.e: $M(X) = y_{true}$ Constructing an adversarial example X_{adv} that is perceptually indistinguishable from X but misclassified, i.e., $M(X_{adv}) \neq y_{true}$[4] is possible.

1.1 Challenges

Three attributes can be used to describe the differences in adversarial attacks: 1) Attack success rate represents a measure of the ability of adversarial attacks to damage neural networks. 2) Robustness represents the ability to deceive machine learning systems. 3) Performance impact on clean test data reflects whether the model's ability to classify regular inputs is reduced in defending against adversarial attacks.We quantify the performance of clean test data by comparing the stealth and flexibility of adversarial examples.

1.2 Contributions

We analyze the existence of robust adversarial examples and adversarial objects in the physical and digital worlds. Moreover, we propose a general way to compare their differences and prove that the method is robust. Specifically, our contributions are as follows:

- We fabricate two typical adversarial examples in both digital and physical environments and demonstrate their ability to attack machine learning systems, demonstrating our approach's end-to-end effectiveness and the existence of robust adversarial objects.

- Experiments in digital and physical world scenarios demonstrate that disguised adversarial examples are highly stealthy and flexible in test data while effectively deceiving state-of-the-art machine learning systems.

Ⅱ. Method

We randomly selected 150 clean images from 5 categories of the ImageNet ILSVRC2012[3] test set.We then applied two methods (PGD[1] and AdvPatch[2]) to craft a targeted adversarial example for each clean image. The threat model adopts a gray-box setting; the target network is the VGG-19 network. We consider the backdoor attack setting, where the attacker operates as follows. First, choose a target label $y_{adv} \in [K]$ and a function $T: X \rightarrow X_{adv}$ that applies adversarial examples to the input. Then, given

access to input label pairs from the data distribution D, they generate a finite number of arbitrary samples (x, y) and inject these samples into the algorithm training set.

2.1 Attack Success Rate

To evaluate the success rate of the backdoor attack, we are interested in the behavior of the model trained on the poisoned dataset when we apply the backdoor trigger to a previously unseen test sample D_{test}. The attack success rate is defined as:

$$\underset{(x,y)\sim D}{pr}\left[f_\theta(T(x)) = y_{adv} \middle| y \neq y_{adv}\right] \quad (1)$$

2.2 Digital World Analysis

For PGD[1], we attack the central object region obtained by manual selection, while for AdvPatch[2], we further select a circular attack region within the object region. For PGD, we use the maximum perturbation E = 16/255 (denoted as PGD-16). For a fair comparison, we filter out failed adversarial samples. Finally, we collect 132 and 122 adversarial samples for PGD and AdvPatch, respectively. Figure 1 shows the adversarial examples for these two methods.

(a) Original (b) PGD-16 (c) AdvPatch

Figure 1: Original image, PGD-16 and AdvPatch generated adversarial images

PGD usually has a higher attack success rate on models, especially those that have not been adversarially trained. Therefore, the robustness of the model under PGD attacks directly depends on whether it has undergone dedicated adversarial training and the anti-perturbation ability of the model structure itself.

A notable feature of AdvPatch is its physical feasibility. Even though the model is robust to common adversarial examples in the digital space, AdvPatch can generate adversarial patches in the real world to attack the model in the physical environment. The robustness analysis of such attacks is not limited to digital models; it also needs to consider the application scenarios in the physical world.

2.3 Physical World Analysis

We conducted a human perception study, asking human evaluators to choose whether a displayed image was “natural and realistic” or “unnatural or unrealistic”. To simulate real-world adversarial examples, we showed users two adversarial images presented in random order, individually rather than in pairs. Ultimately, AdvPatch was selected as “natural and realistic” with a probability of 19.0±1.68% and PGD was selected with a probability of 77.3±1.53%. We summarize these statistics as the stealth scores of the two methods, and Table 1 shows the difference between the two adversarial examples. This also confirms that the PGD method is much stealthier than the AdvPatch method in the digital world.

Table 1: Summary of existing attacks

Attack	Condition	Attack success rate	Robustness	Hiddenness
PGD	Digital	99.2%	☆☆	77.3±1.53%
AdvPatch	Physical	92.7%	☆☆☆	19.0±1.68%

Ⅲ. Conclusion

This paper explored the differences in creating adversarial examples for machine learning systems operating in the digital and physical worlds. In the digital world, we successfully avoided the search for machine and human perception by controlling the perturbations within reasonable range, demonstrating the excellent performance of adversarial examples. At the same time, the problem of adversarial examples in the physical world is more complicated due to viewpoint transformation, camera noise, and other natural transformations.In future work, we foresee strengthening physical-world attacks against different types of machine learning systems, such as reinforcement learning agents, with the goal of improving the performance of adversarial examples in the physical world.

ACKNOWLEDGMENT

Put sponsor acknowledgments.

REFERENCES

[1] Brown, Tom B., et al. "Adversarial patch." arxiv preprint arxiv:1712.09665 (2017).

[2] Madry, Aleksander. "Towards deep learning models resistant to adversarial attacks." arxiv preprint arxiv:1706.06083 (2017).

[3] Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 acm sigsac conference on computer and communications security. 2016.

[4] Gu, Tianyu, et al.. "Badnets: Identifying vulnerabilities in the machine learning model supply chain." arxiv preprint arxiv:1708.06733 (2017).