컴퓨터 비전을 통한 물리적 환경에서의 적대적 공격에 관한 연구

수랸토 나우팔, 김용수*, 홍윤영, 김호원

부산대학교, *스마트엠투엠

naufalso@pusan.ac.kr, yongsu@smartm2m.co.kr, hyy0238@pusan.ac.kr,
howonkim@pusan.ac.kr

A Study on the Physical World Adversarial Attack in Computer Vision

Naufal Suryanto, Yongsu Kim*, Yoonyoung Hong, Howon Kim
Pusan National University, *SmartM2M

Abstract

Deep learning has become the heart of modern computer vision. The ability to solve complex problems with a high level of accuracy has made deep learning models used for various computer vision tasks, including security and safety-critical applications. Unfortunately, recent studies have shown that deep learning models are vulnerable to adversarial attacks that can manipulate the model prediction by adding designed perturbation to the input. The studies have shown that the attacks are applicable not only in the digital domain but also in the physical world. This paper investigates the recent adversarial attack techniques in computer vision that specifically work on the physical world. We outline essential factors, including key challenges and proposed solutions such as losses that make adversarial attacks work, robust, and applicable in the real world.

Ⅰ. Introduction

Deep learning has emerged as the driving force behind modern computer vision. *AlexNet* [1] is one of the earliest examples of deep learning raising with Deep Convolutional Neural Networks (CNNs), which took first place beyond traditional computer vision in the 2012 large-scale visual recognition challenge [2]. Deep learning remains the current state-of-the-art technique in various computer vision applications [3], including security and ‘safety-critical applications.

Despite the outstanding performance of the deep learning models on computer vision, *Szegedy et al.* [4] realized that they are vulnerable to adversarial attacks. This attack tries to change the network prediction by slightly changing the input with imperceptible perturbation. Early after this discovery, researchers later proposed ways to fool deep learning models more effectively [5] and higher success rates [6]. The research was initially conducted in the digital domain, where an adversary can directly change each image pixel to produce an adversarial example.

Research on adversarial attacks did not just stop at the digital domain. *Kurakin et al.* [7] tried to print the generated adversarial examples with several methods and found that some are still working to fool the tested model in the physical world. They hypothesized that the failed case was due to perturbation which was easier to destroy by real-world transformation. In the next section, we recap challenges and researchers' proposed solutions to make adversarial attacks work, robust, and applicable in the real world.

Ⅱ. Main subject

The adversary must solve several challenges so that adversarial attacks can work in the real world. By referring to *Eykholt et al.* [8], we summarize that the adversary must generate the adversarial example that can survive against physical world transformation, such as varying environmental conditions and viewpoints, overcoming spatial constraints, fabrication error, and physical limits on imperceptibility. Here we summarize how the researchers handle each of the challenges.

Robustness against physical world transformation. *Athalye et al.* [9] proposed a method called *Expectation Over Transformation (EOT)* to construct a robust physical adversarial example. The idea is to optimize the adversarial example x' over the chosen transformation distribution T instead of optimizing the single example such as:

$$\arg\max_{x'} E_{t\sim T}[\log P(y_t|t(x'))]$$

The transformations include varying viewpoints and lighting conditions for the physical world case. They demonstrate their method's effectiveness by making 3D printed models of a turtle and baseball—which remain adversarial over a wide distribution of viewpoints.

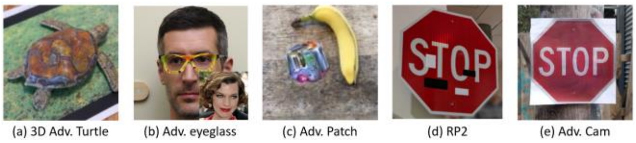

Figure 1. Examples of successful physical-world attacks fool the recognition system proposed by [9,11,10,8,12].

EOT method becomes a standard for generating robust physical adversarial examples in later studies.

Overcoming spatial constraints. In a physical adversarial attack, the adversary cannot perturb the whole part of the input image, especially the background. The typical proposals are to either perturb the target object [8,9,11,12] or create an adversarial patch [10]. The adversarial patch has some advantages, such as being universal and not tied with target object.

Minimizing fabrication error. To realize an adversarial example into the real world, the adversary must either print or paint it, which may result in fabrication errors.

Sharif et al. [11] proposed a *non-printability score (NPS)* to craft adversarial perturbations that can mostly be reproduced by the printer. They select a set of printable RGB colors P and penalize each pixel based on the distance of the closest color on the set. They define the *NPS* of pixel $\hat{p}$ as:

$$NPS(\hat{p}) = \prod_{p\in P} |\hat{p} - p|$$

$NPS(\hat{p})$ will be low if $\hat{p}$ belongs to P or closed to $p \in P$. Otherwise, $NPS(\hat{p})$ will be high.

Furthermore, the authors proposed *total variation (TV)* loss to keep the smoothness of their adversarial by penalizing the square root distances between each neighbor pixel. For perturbation *r* on each pixel, they define *TV(r)* as:

$$TV(r) = \sum_{i,j} \sqrt{\left(r_{i,j} - r_{i+1,j}\right)^2 + \left(r_{i,j} - r_{i,j+1}\right)^2}$$

The smoothness is essential because natural images captured by the camera are smooth and consistent, where colors change gradually. The adversary can minimize the error between digital and physical domains using both losses, enhancing physically realizable adversarial examples.

Handling Physical Limits on Imperceptibility. Tiny imperceptible perturbations are more likely to be destroyed when transferred to the real world. Instead of finding tiny perturbations, the physical attacks aim to produce natural and unsuspicious adversarial examples to hide the attack.

Eykholt et al. [8] proposed masked perturbation that resembles graffiti. *Duan et al.* [12] proposed camouflage loss consisting of style, content, and smoothness losses for generating their natural style adversarial examples. Both authors show unsuspected attacks using natural perturbations to physical traffic signs, which adds to the danger of this invisible attack.

Ⅲ. Conclusion

The adversarial attacks remain a threat to deep learning-based computer vision applications, including in the physical world. This paper outline how the adversary can craft robust adversarial examples that can withstand physical world transformation and solve all challenges beyond digital adversarial attack.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2020-0-01797) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

참고문헌

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, 2017.

[2] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int. J. Comput. Vis., 2015.

[3] "Computer vision", Papers With Code. [Online]. Available: https://paperswithcode.com/area/computer-vision. [Accessed: 14-Sep-2022].

[4] C. Szegedy et al., "Intriguing properties of neural networks," in 2nd International Conference on Learning Representations, ICLR, 2014.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in 3rd International Conference on Learning Representations, ICLR, 2015.

[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in 6th International Conference on Learning Representations, ICLR, 2018.

[7] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in 5th International Conference on Learning Representations, ICLR, 2017

[8]. K. Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.

[9] A. Athalye, L. Engstrom, A. Ilyas, and K. Kevin, "Synthesizing robust adversarial examples," in 35th International Conference on Machine Learning, 2018.

[10] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," CoRR, vol. abs/1712.09665, 2017.

[11] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in Proceedings of the ACM CCS, 2016.

[12] R.Duan, X.Ma, Y.Wang, J.Bailey, A.K.Qin, and Y.Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2020.