CHAPTER 4

Physical adversarial attacks

Beyond digital space, adversarial examples can also be realized in physical world to evade the prediction of machine learning-based applications such as object detection, known as *physical adversarial examples*. In this chapter, we cover the attack formulation for physical adversarial examples and provide an overview of such attacks in the physical world.

4.1 Physical adversarial attack formulation

Physical adversarial examples are often realized as robust adversarial examples crafted from digital spaces (e.g., simulated physical environments) that remain the adversarial objective in the physical space. To bridge the gap between adversarial examples in digital and physical spaces, there are two major obstacles: (i) The digital adversarial examples cannot be precisely realized in the physical space. For example, some RGB color values in the digital space cannot be printed exactly. (ii) Digital adversarial examples may not well generalize to physical environments. For example, simple rotation or zooming in and out a printed adversarial image may make the adversarial effect in vain.

To address challenge (i), during the attack process, we can project the digital adversarial examples to the space of realizable actions in the physical space (e.g., the space of printable colors) or add an additional regularization loss in the attack loss to minimize inconsistency between the digital adversarial examples and the resulting physical versions.

To address challenge (ii), a common practice is to introduce multiple transformations during the generation process such that the resulting adversarial example will remain effective. The methodology can also be thought as learning a universal adversarial manipulation that is simultaneously effective to most of (if not all) considered transformations.

Expectation over transformation (EOT) (Athalye and Sutskever, 2018) introduces a set of data transformations denoted by $\mathcal{T}$ in generating a robust adversarial example $\boldsymbol{x}$ of a data sample $\boldsymbol{x}_0$ via solving the following optimization problem:

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmin}}\ \mathbb{E}_{t\sim\mathcal{T}}\, g(t(\boldsymbol{x})) + r(\boldsymbol{x}), \tag{4.1}$$

Adversarial Robustness for Machine Learning
https://doi.org/10.1016/B978-0-12-824020-5.00013-2

where $\mathcal{X}$ is the space of feasible physical adversarial examples, t is a transformation function drawn from $\mathcal{T}$, g is the attack objective loss function as introduced in Chapter 2, and r is the regularizer (e.g., a penalty function on unrealizable printing colors).

The optimization objective of EOT attack can be viewed as minimizing the average loss over all data transformations in $\mathcal{T}$. Instead of minimizing the average loss, Wang et al. (2021a) proposed a MinMax attack formulation that optimizes the worst-case transformation loss through

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{w}\in\mathcal{P}} \sum_{i=1}^{K} w_i \cdot g(t_i(\boldsymbol{x})) - \gamma \|\boldsymbol{w} - \mathbf{1}/K\|_2^2 + r(\boldsymbol{x}), \tag{4.2}$$

where K is the number of transformations in $\mathcal{T}$, $\boldsymbol{w}$ is a nonnegative K-dimensional vector in the probability simplex $\mathcal{P}$ satisfying $\sum_{i=1}^{K} w_i = 1$, t_i is the ith transformation with w_i being its weighting factor, $\mathbf{1}/K$ is a uniform vector, and $\gamma \geq 0$ is a regularization coefficient. If $\boldsymbol{w} = \mathbf{1}/K$, then problem (4.2) reduces to the existing EOT formulation. Using the alternating one-step projected gradient descent-ascent (APGDA) algorithm for solving (4.2), it is shown by Wang et al. (2021a) that MinMax attack leads to more robust adversarial examples against multiple data transformations than EOT. MinMax attack also offers some interpretation on the difficulty level of each transformation through the associated learned weighting factor $\{w_i\}_{i=1}^{K}$ while solving the min-max optimization problem. It is also used by Xu et al. (2020c) in the design of physical adversarial T-shirts that are robust to nonrigid deformations.

4.2 Examples of physical adversarial attacks

Here we list some studies into crafting physical adversarial examples in different scenarios. Some of the physical adversarial examples are shown in Fig. 4.1.

- *Adversarial photo* (Kurakin et al., 2016): Demonstration that when feeding adversarial images obtained from cell-phone camera to an ImageNet Inception classifier, a large fraction of adversarial examples are classified incorrectly even when perceived through the camera.
- *Adversarial eyeglasses* (Sharif et al., 2016): Realization through printing a pair of eyeglass frames such that when worn by the attacker whose image is supplied to a face-recognition algorithm, the eyeglasses allow her to evade being recognized or to impersonate another individual.

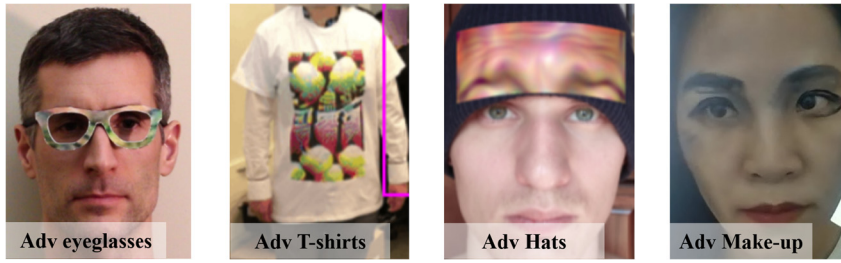

Figure 4.1 Examples of physical adversarial examples. Left to right: Adversarial eyeglasses (Sharif et al., 2016), adversarial T-shirt (Xu et al., 2020c), adversarial hat (Komkov and Petiushko, 2021), and adversarial make-up (Lin et al., 2021).

- *Adversarial patch* (Brown et al., 2017): A physical adversarial object such that its presence causes the classifiers to ignore the other items in the scene and report a chosen target class.
- *3D-printed adversarial object* (Athalye and Sutskever, 2018): A 3D-printed adversarial object that leads to incorrect predictions when taking its pictures from different views.
- *Adversarial stop sign* (Evtimov et al., 2017): An adversarial stop sign created by adding color patches to evade the detection of real-time object detectors.
- *Adversarial board* (Thys et al., 2019): An adversarial patch able to hide a person from a person detector when wearing it.
- *Adversarial sticker* (Li et al., 2019b): Demonstration of a carefully crafted and mainly translucent sticker over the lens of a camera that can create universal perturbations of the observed images that are inconspicuous, yet misclassify target objects as a different (targeted) class.
- *Adversarial T-shirt* (Xu et al., 2020c): A robust physical adversarial example for evading person detectors, even if it undergoes nonrigid deformation due to a moving person's pose changes.
- *Adversarial hat* (Komkov and Petiushko, 2021): Designing and printing a rectangular paper sticker on a common color printer and put it on the hat to attack Face ID systems.
- *Adversarial make-up* (Lin et al., 2021): Adversarial full-face makeup guided by generative adversarial networks to impersonate a target person and fool facial recognition systems.

4.3 Empirical comparison

Here we compare the performance of EOT attack (avg.) (Athalye and Sutskever, 2018) and MinMAX attack (Wang et al., 2021a). For each input sample (*ori*), we transform the image under a series of functions, e.g., flipping horizontally (*flh*) or vertically (*flv*), adjusting brightness (*bri*), performing gamma correction (*gam*) and cropping (*crop*), and group each image with its transformed variants. ASR_{all} is reported to measure the attack success rate (ASR) over groups of transformed images (each group is successfully attacked signifies successfully attacking an example under all transformers). In Table 4.1, compared to EOT, MinMax leads to 9.39% averaged lift in ASR_{all} over given models on CIFAR-10 by optimizing the weights for various transformations.

Table 4.1 Comparison of average and min-max optimization on robust attack over multiple data transformations on CIFAR-10. Acc (%) represents the test accuracy of classifiers on adversarial examples. Table adopted from Wang et al. (2021a).

Model	Opt.	Acc_{ori}	Acc_{flh}	Acc_{flv}	Acc_{bri}	Acc_{gam}	Acc_{crop}	ASR_{all}	Lift (↑)
A	*avg.*	10.80	21.93	14.75	11.52	10.66	20.03	55.88	–
	min max	12.14	18.05	13.61	13.52	11.99	16.78	**60.03**	**7.43%**
B	*avg.*	5.49	11.56	9.51	5.43	5.75	15.89	72.21	–
	min max	6.22	8.61	9.74	6.35	6.42	11.99	**77.43**	**7.23%**
C	*avg.*	7.66	21.88	15.50	8.15	7.87	15.36	56.51	–
	min max	8.51	14.75	13.88	9.16	8.58	13.35	**63.58**	**12.51%**
D	*avg.*	8.00	20.47	13.46	7.73	8.52	15.90	61.13	–
	min max	9.19	13.18	12.72	8.79	9.18	13.11	**67.49**	**10.40%**

4.4 Extending reading

- Unadversarial examples by realizing input perturbations to design robust objects that are explicitly optimized to be confidently detected or classified (Salman et al., 2020a).