Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Doctoral Consortium Track

Adversarial Examples in Physical World

Jiakai Wang
Beihang University
jk_buaa_scse@buaa.edu.cn

Abstract

Although deep neural networks (DNNs) have already made fairly high achievements and a very wide range of impact, their vulnerability attracts lots of interest of researchers towards related studies about artificial intelligence (AI) safety and robustness this year. A series of works reveals that the current DNNs are always misled by elaborately designed adversarial examples. And unfortunately, this peculiarity also affects real-world AI applications and places them at potential risk. we are more interested in physical attacks due to their implementability in the real world. The study of physical attacks can effectively promote the application of AI techniques, which is of great significance to the security development of AI.

1 Background

Although deep neural networks (DNNs) have already made fairly high achievements and very wide range of impact, its vulnerability attracts lots of interests of researchers towards related studies about artificial intelligence (AI) safety and robustness this years. A series of works reveal that the current DNNs are always misled by elaborately designed **adversarial examples**. In general, there are several different ways to categorize adversarial attack methods, *e.g.*, targeted or untargeted attacks, white-box or black-box attacks, etc. Based on the domain in which the adversarial examples come into effect, adversarial attacks can be divided into digital attacks and physical attacks.

Digital attacks generate adversarial perturbations for input data in the digital pixel domain. Szegedy *et al.*first introduced adversarial examples and used the L-BFGS method to generate them. By leveraging the gradients of target models, Goodfellow *et al.*proposed the Fast Gradient Sign Method (FGSM) which could generate adversarial examples quickly. Moreover, Madry *et al.*proposed Projected Gradient Decent (PGD), which is currently the strongest first-order attack. Although these attacks achieve substantial results in the digital world, their attacking abilities degenerate significantly when introduced into the physical world.

On the other hand, physical attacks aim to generate adversarial perturbations by modifying the visual characteristics of the real object in the physical world. To achieve the goal, several works first generate adversarial perturbations in the digital world, then perform physical attacks by painting the adversarial camouflage on the real object or directly create the perturbed objects. By constructing a rendering function, Athalye *et al.*generated 3D adversarial objects in the physical world to attack classifiers. Eykholt *et al.*introduced NPS into the loss function which considers the fabrication error so that they can generate strong adversarial attacks for traffic sign recognition. Recently, Huang *et al.*proposed the Universal Physical Camouflage Attack (UPC), which crafts camouflage by jointly fooling the region proposal network and the classifier.

To sum up, we are more interested in physical attacks due to its implementability in the real-world. The study of physical attacks can effectively promote the application of AI techniques, which is of great significance to the security development of the AI.

2 Physical Adversarial Examples

Physical adversarial examples (PAE) aim to conduct attacks (i.e., make DNNs fail to predict the target objects accurately) in the real-world, which means that the generated adversarial examples should cross the digital-physical gap so as to affect the physical devices. As a rule, these physical adversarial examples are always able to be produced in some realistic way (*i.e.*, printer, 3D printer, drawing, and so on).

As for the definition, the basic formulation of PAE is similar with digital adversarial examples. Given a deep neural network and an input clean image I with the ground truth label y, a digital adversarial example I_{adv} in the **digital world** can make the model conduct wrong predictions as follows:

$$\mathcal{F}_\theta(I_{adv}) \neq y \quad s.t. \quad \|I - I_{adv}\| < \epsilon, \tag{1}$$

where $\|\cdot\|$ is a distance metric to quantify the distance between the two inputs I, I_{adv} sufficiently small, and θ is the parameters of the deep model $\mathcal{F}$.

When it comes to the physical world, the digital adversarial examples always suffer from the digital-physical transformation (*i.e.*, illumination, rigid transformations, and non-rigid transformations, and chromatic aberration, etc), which makes the digital adversarial examples lose their attacking ability.

Considering these difficulties, we redefine the physical adversarial examples as follows:

$$\mathcal{F}_\theta(\mathcal{T}(I_{adv})) \neq y \quad s.t. \quad \mathcal{S}(\mathcal{T}(I_{adv})) < \epsilon, \tag{2}$$

where the $\mathcal{T}$ means a implicit transformation which depicts the digital-physical gap, the $\mathcal{S}$ denotes a stealthiness function in order to evaluate the stealthiness of the physical adversarial examples. Note that there is always no fixed standard for $\mathcal{T}$.

3 Our Studies

We summarize challenges of physical adversarial examples into two aspects:
(1) **The transferablity of the generated adversarial examples**. Because of the differences between the generation and application domain, the transferability of physical adversarial examples has serious impacts on their performance, including transferability of environments (*i.e.*, digital and physical world), models (*i.e.*, white-box and black-box), and classes (class-specific and class-agnostic). (2) **The stealthiness of the generated adversarial examples**. Adversarial examples in the digital world are always restricted by $\|\cdot\|$, which can be a distance metric (*i.e.*, l-norm, latent layer feature variance). However, this metric is difficult to implement in the physical world. Thus, it should be replaced by a more flexible metric stealthiness function.

3.1 Bias-based Universal Adversarial Patch Generation

Recently, adversarial patch, with noise confined to a small and localized patch, has emerged for its easy feasibility in real-world scenarios. However, existing strategies failed to generate adversarial patches with strong generalization ability.

To address the problem, this paper proposes a bias-based framework to generate class-agnostic universal adversarial patches with strong generalization ability, which exploits both the perceptual and semantic bias of models. Regarding the perceptual bias, since DNNs are strongly biased towards textures, we exploit the hard examples which convey strong model uncertainties and extract a textural patch prior from them by adopting the style similarities. The patch prior is more close to decision boundaries and would promote attacks. To further alleviate the heavy dependency on large amounts of data in training universal attacks, we further exploit the semantic bias. As the class-wise preference, prototypes are introduced and pursued by maximizing the multi-class margin to help universal training. Taking Automatic Check-out (ACO) as the typical scenario, extensive experiments including white-box/black-box settings in both digital world (RPC, the largest ACO related dataset) and physical world scenario (Taobao and JD, the world's largest online shopping platforms) are conducted. Experimental results demonstrate that our proposed framework outperforms state-of-the-art adversarial patch attack methods [Liu *et al.*, 2020].

3.2 Attention-based Transferable Adversarial Camouflage Generation

As a more threatening type for practical deep learning systems, physical adversarial examples have received extensive research attention in recent years. However, without exploiting the intrinsic characteristics such as model-agnostic and human-specific patterns, existing works generate weak adversarial perturbations in the physical world.

Motivated by the viewpoint that attention reflects the intrinsic characteristics of the recognition process, this paper proposes the Dual Attention Suppression (DAS) attack to generate visually-natural physical adversarial camouflages with strong transferability by suppressing both model and human attention. As for attacking, we generate transferable adversarial camouflages by distracting the model-shared similar attention patterns from the target to non-target regions. Meanwhile, based on the fact that human visual attention always focuses on salient items (*e.g.*, suspicious distortions), we evade the human-specific bottom-up attention to generate visually-natural camouflages which are correlated to the scenario context. We conduct extensive experiments in both the digital and physical world for classification and detection tasks on up-to-date models (*e.g.*, Yolo-V5) and demonstrate that our method outperforms state-of-the-art methods [Wang *et al.*, 2021].

4 Discussion

Our long-term research interest is to investigate much more strategies to generate physical adversarial examples. We believe that these works can provide inhibit lines for real world AI applications and promote to design stronger DNN models. Moreover, we are also interested in studying the evaluation and defense approaches of AI applications. Now we are at-tempting to generate physical adversarial example to attack more scenarios in a multimodal condition. Aiming to per¬form more powerful attacks, we also wage studies on model comprehension.

Like the relationship between sword and shield, performing attacks and defenses in the physical world is becoming more and more important. By making an intensive study of physical adversarial examples, we can not only evaluate the security of deployed devices but also deepen our understanding of the DNNs, which may further help to improve the model performance and strengthen the model robustness. Prior studies have proved the exists of adversarial examples in both digital and physical world, revealing the threatens of adversarial examples, we support the opinions in full agreement and pay more attention to the physical scenario due to its great relevance. And we call on more researchers to devote themselves to the research in this field and make results.

References

[Liu *et al.*, 2020] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *Computer Vision – ECCV 2020*, pages 395–410, Cham, 2020. Springer International Publishing.

[Wang *et al.*, 2021] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. *CVPR*, abs/2103.01050, 2021.