Physically Realizable Adversarial Attacks and Defenses – A Review

Dhruv Behl[1], Dr. B Sathish Babu[1]

[1]*Department of Computer Science and Engineering, R V College of Engineering, Bengaluru, India*

Abstract - *Deep neural networks have become the approach of choice in a multitude of domains, especially computer vision related tasks like image classification, localization and segmentation. However, numerous demonstrations have shown that deep neural networks may be easily deceived by precisely perturbing pixels in an image, which is commonly referred to as an adversarial attack. As a result, a considerable amount of literature has evolved on defending deep neural networks against adversarial examples, with approaches for learning more robust neural network models or detecting malicious inputs being proposed. Oddly, while considerable attention has been devoted to defending against adversarial perturbation attacks in the digital space, there are no effective methods specifically to defend against such physically-realizable attacks. We study the problem of defending deep neural network approaches for image classification from physically realizable attacks. First, we demonstrate all the physically realizable attacks that have come up recently and tabulate their attack performance on different datasets. Then, we discuss the existing defenses against physical attacks, their robustness, and their shortcomings. Finally, we discuss the challenges faced by most of the current defenses and present future research perspectives needed to achieve true adversarial robustness.*

Key Words: Adversarial Robustness, Convolutional Neural Networks, Deep Learning, Image Classification, Safety, Security.

1.INTRODUCTION

Computer Vision is the study of how computers can extract high-level information from digital photographs or films. Various categorization challenges make up a large part of the field of Computer Vision. The task of providing a label to an image is known as image classification.

The approaches for handling classification problems have been widely researched in both academic and commercial businesses as a core challenge in computer vision and machine learning, and significant progress has been made. Convolutional neural networks (CNNs) are the most popular picture classification algorithms, with better-than-human performance on a variety of benchmark datasets, while their real-world performance across new institutions and curated collections is still unknown. Fig 1 shows a demonstration of a CNN being used for the image classification task.

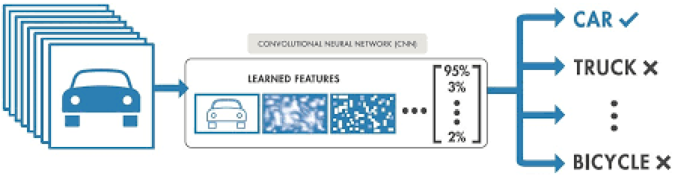

Fig 1. CNN being used for Image Classification.

State-of-the-art effectiveness of deep neural networks has made it the technique of choice in a variety of fields, including computer vision, natural language processing and speech recognition. However, there have been a myriad of demonstrations showing that deep neural networks can be easily fooled by carefully perturbing pixels in an image through what have become known as adversarial example attacks. In response, a large literature has emerged on defending deep neural networks against adversarial examples, typically either proposing techniques for learning more robust neural network models, or by detecting adversarial inputs.

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the (trained) model to make a mistake [1]. Adversarial images appear visually and semantically the same to us, but the model ends up predicting the wrong class with very high confidence, which is worrying. Fig 2. gives an example of an adversarial image created to fool the classifier [2].

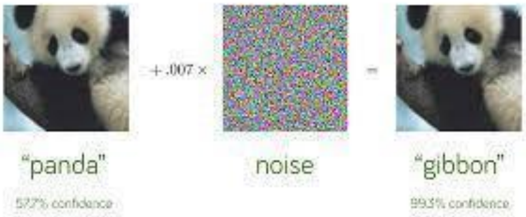

Fig 2. Adversarial image being used to fool the classifier.

The size of the perturbation is in the core of the adversarial attack, a small perturbation is the fundamental premise of such models. When designing an adversarial example, the attacker wants the perturbed input to be as close as possible to the original one, in the case of images, close enough that a human can not distinguish one image from the other.

- **Perturbation Scope**: The attacker can generate perturbations that are input specific, in which we call individual, or it can generate a single perturbation which will be effective to all inputs in the training dataset, which we call universal perturbation.

- **Perturbation Limitation**: Two options are possible, optimized perturbation and constraint perturbation. The optimized perturbation is the goal of the optimization problem, while the constraint perturbation is the set as the constraint to the optimization problem.
- **Perturbation Measurement**: Is the metric used to measure the magnitude of the perturbation. The most commonly used metric is the lp-norm, with many algorithms applying l0, l2, l∞ norms.

Particularly concerning, however, have been a number of demonstrations that implement adversarial perturbations directly in physical objects that are subsequently captured by a camera, and then fed through the deep neural network classifier [3].

Properties of Physically Realizable Attacks [12]:
1. Attacks can be implemented in physical space (eg: putting a sticker on a stop sign)
2. Attacks should have low suspiciousness
3. Attacks cause misclassification of SOTA neural network

Among the most significant of such physical attacks on deep neural networks are three that we specifically consider here:
1. the attack which fools face recognition by using adversarially designed eyeglass frames
2. the attack which fools stop sign classification by adding adversarially crafted stickers
3. the universal adversarial patch attack, which causes targeted misclassification of any object with the adversarially designed sticker (patch).

While much emphasis has been paid to guarding against adversarial disturbance attacks in the digital realm, no viable strategies for fighting against comparable physical attacks exist.

The goal of this review paper is to raise concerns and awareness about the dangers of physically-realizable attacks. We study the current physical attacks, their threat models and methodology and success rate. Then we evaluate performances of some defences against these attacks. We conclude by discussing the challenges faced by most of the current defenses and present future research perspectives needed to achieve true adversarial robustness.

2. Physical Adversarial Attacks

2.1 Fooling Face Recognition Systems

The goal of this attack [12] is to fool Face Recognition Systems (FRS). Here, the attacker is allowed to change only the physical objects, not individual pixels. The attacker can't control camera position, lightning, etc. The attack should be inconspicuous, i.e., the defender shouldn't notice the attack.

Threat model:
They assume an attacker who gains access to the pre-trained FRS to mount a dodging (untargeted) or impersonation (targeted) attack. The adversary cannot "poison" the FRS by altering training data, injecting mislabeled data, etc.

Adversaries can alter only the composition of inputs to be classified; attacks should be physically realizable. It assumes a white-box scenario: the attacker knows the internals (architecture, weights, feature space) of the system being attacked

Full attack approach:

- Train a model for face-recognition (eg: VGG, OpenFace)
- Pick attacker and target (for targeted attack)
- Generate eyeglasses using the discussed approach for attack
- Print the eyeglasses
- Collect image(s) of attacker wearing the eyeglasses
- Classify the collected images

Success metric: fraction of images misclassified as target.

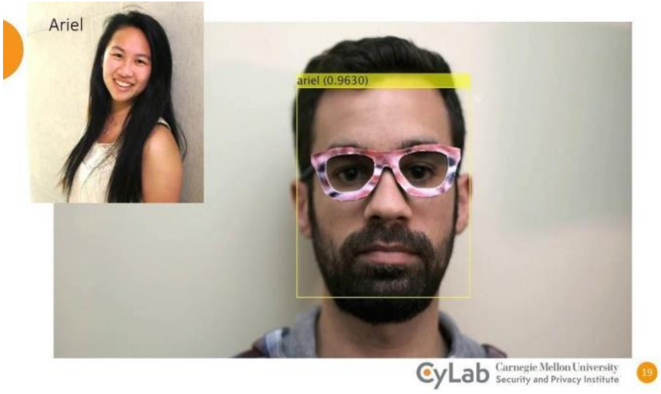
Ariel
ariel (0.9630)
CyLab Carnegie Mellon University Security and Privacy Institute
19

Fig 3. Fooling FRS.

2.2 Fooling Road Sign Classifier

In this attack, the attacker adds physical perturbation to road signs in order to cause misclassification in the vehicle's road sign classifier [5].

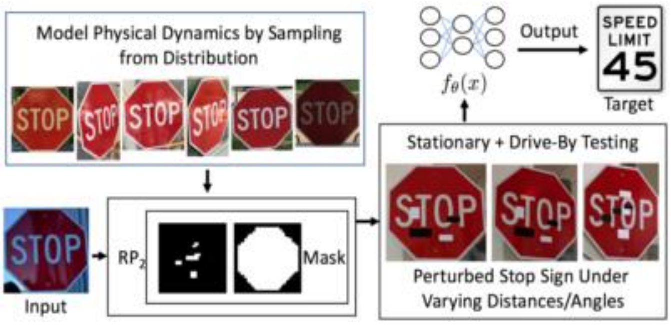
Model Physical Dynamics by Sampling from Distribution
$f_\theta(x)$
Output
SPEED LIMIT 45
Target
Stationary + Drive-By Testing
Input
RP$_2$
Mask
Perturbed Stop Sign Under Varying Distances/Angles

Fig 4: Fooling Road Sign Classifier.

Their threat model assumes a white-box scenario, where the adversary has access to the network architecture and weights.

Attacker does not have control over the vehicle’s systems, but is able to modify the objects in the physical world that a vehicle might depend on for classification.

It can be difficult to carry out a successful adversarial attack on real-world classifiers. It was discovered that when hostile images were treated to small changes, their effectiveness was reduced. When an adversarial item is put in a real-world setting, it may be seen from various perspectives and under various lighting conditions. All of those transformations, some of which can be malicious, must be survived for a successful attack. Because of the difficulty of the endeavour, several researchers concluded that physical adversarial attacks are not practical and should not be regarded a danger.To face the challenge of performing adversarial attacks in the physical world Athalye et al. have proposed the Expectation Over Transformations method, shown in Fig 4.

2.3 Adversarial Patch

This is a method to create adversarial image patches in the real world that are [6]:

- **Universal**: can be used to attack any scene
- **Robust**: work under a wide variety of transformations
- **Targeted**: can cause a classifier to output any target class

These adversarial patches may be printed, placed in any scene, photographed, and then presented to image classifiers; they cause the classifiers to disregard the rest of the scene and report a certain target class.

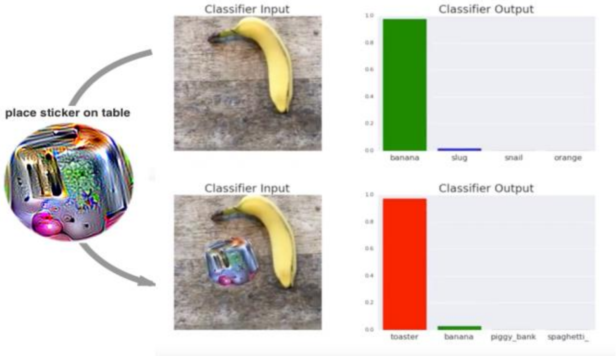

Fig 5: Adversarial Patch.

The attacker crafts perturbations that are not bounded by an ε value but are bounded to a small region or location in the image. This approach enables attackers to launch a physical-world attack without knowing the lighting circumstances, camera angle, classifier type being attacked, or even the other items in the scene.It successfully fools the classifier in both white-box and black-box settings (black-box setting requires a larger patch size for effective transferability).

The significance of this attack is that the attacker does not need to know what image they are attacking when generating it. Following the creation of an adversarial patch, it might be widely distributed via the Internet for other attackers to print and utilise. Additionally, because the attack uses a large perturbation, the existing defense techniques which focus on defending against small perturbations may not be robust to larger perturbations such as these. Indeed recent work has demonstrated that state-of-the art adversarially trained models on MNIST are still vulnerable to larger perturbations than those used in training either by searching for a nearby adversarial example using a different metric for distance, or by applying large perturbations in the background.

3. Physical Adversarial Defenses

3.1 Defense by Pre-processing

Watermark removal: An image has been corrupted through scratches or random noise and the task is to restore the image and remove such noise.

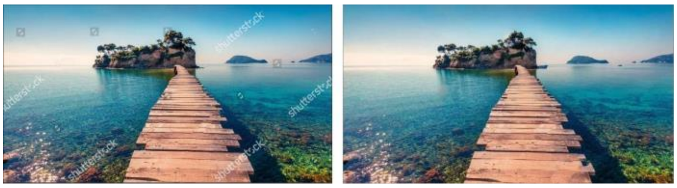

Fig 6: Watermark removal.

The problem of removing visible localized adversarial perturbations (as seen in adversarial patches) is similar to the problem of watermark removal - we have a corrupted copy of an image and wish to remove the noise and restore the image [11].

Steps:

1. Construct a saliency map of the image using the guided backpropagation method.
2. Use a combination of erosion and dilation to remove small “holes”.
3. Find the contour area of positive regions within the updated saliency map, and if the contour area is below a threshold, we zero out this area.
4. Finally, we use the remaining positive regions of the saliency map as locations to mask the adversarial image

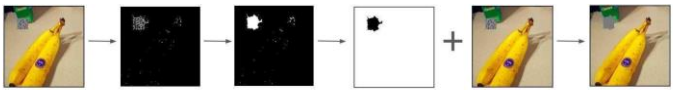

Fig 7: First find the saliency map of the image. The following two steps construct a mask that is applied to the adversarial image, blocking the adversarial perturbation.

3.2 DOA Defense

They demonstrated that traditional adversarially robust model training approaches on digital images, such as Robust Adversarial Training with PGD and Randomized Smoothing, perform poorly against physical attacks.

The traditional attack concept is far too incompatible with realistic physical attacks. The insertion of hostile occlusions to a portion of the input is a fundamental common factor in many physical attacks. The amount of the adversarial occlusion, but not its shape or position, is a frequent restriction of such attacks [12].

White-box scenario is considered.

Fig 8: Demonstration of ROA defense.

They propose the following simple abstract model of adversarial occlusions of input images. The attacker introduces a fixed-dimension rectangle. This rectangle can be placed by the adversary anywhere in the image, and the attacker can furthermore introduce l∞ noise inside the rectangle with an exogenously specified high bound (for example, = 255, which effectively allows addition of arbitrary adversarial noise). This model bears some similarity to l0 attacks, but the rectangle imposes a contiguity constraint, which reflects common physical limitations. The model is clearly abstract: in practice, for example, adversarial occlusions need not be rectangular or have fixed dimensions (for example, the eyeglass frame attack is clearly not rectangular), but at the same time cannot usually be arbitrarily superimposed on an image, as they are implemented in the physical environment. Nevertheless, the model reflects some of the most important aspects common to many physical attacks, such as stickers placed on an adversarially chosen portion of the object we wish to identify. They call this attack model a rectangular occlusion attack (ROA). An important feature of this attack is that it is untargeted: since the ultimate goal is to defend against physical attacks whatever their target, considering untargeted attacks obviates the need to have precise knowledge about the attacker’s goals.

3.3 Certified Defense (IBP)

Interval-Bound Propagation is the first certified defense against patch attacks [10]. It’s based on the fact that if we specify the patch location, one can represent the feasible set of images with a simple interval bound:

- For pixels within the patch, the upper and lower bound is equal to 1 and 0
- For pixels outside of the patch, the upper and lower bounds are both equal to the original pixel value

We can then use constraints to apply IBP for training a provably robust model, i.e., a model that has a certified lower-bound accuracy.

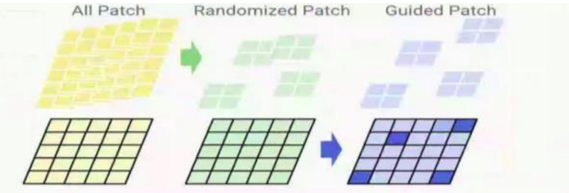

Fig 9: Demonstration of IBP Defense.

By extending interval bound propagation (IBP) protections, they provide the first certifiable defence against patch attacks. They also recommend changes to IBP training in the patch configuration to make it more efficient. They further investigate the generalisation of certified patch defences to patches of various shapes, finding that robustness is consistent across patch types. Preliminary results on verified defence against the tougher sparse attack model, in which a set number of potentially non-adjacent pixels can be freely manipulated, are also shown.

The caveats of this certified approach are that the IBP defense has relatively poor clean and provable robust accuracy. The IBP-based method is not likely to effectively scale up to ImageNet.

4. CONCLUSIONS

Adversarial attacks pose a huge threat to security, safety and trust in our ML models. Physically realizable attacks are dangerous and haven’t been explored as much as digital adversarial attacks, which greatly impedes progress to building robust deep learning models. There exist physical attacks that can dodge face recognition systems, fool road sign classifiers, and generate universal adversarial stickers/patches to cause misclassification in any scene. Some recent works introduce defenses to deal with these attacks on small-scale datasets. Further work is needed to evaluate progress on larger-scale datasets.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations (ICLR), 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR), 2015.

[3] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In International Conference on Learning Representations (ICLR), 2017.

[4] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In International Conference on Learning Representations (ICLR), 2018.

[5] Adith Boloor, Xin He, Christopher Gill, Yevgeniy Vorobeychik, and Xuan Zhang. Simple physical adversarial examples against end-to-end autonomous driving models. In IEEE International Conference on Embedded Software and Systems, 2019.

[6] Tom B. Brown, Dandelion Mane, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial patch. CoRR, abs/1712.09665, 2018.

[7] Edward Chou, Florian Tramer, and Giancarlo Pellegrino. Sentinet: Detecting physical attacks against deep learning systems, 2018. arxiv preprint.

[8] Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In Security and Privacy (SP), 2017 IEEE Symposium on, pp. 39–57. IEEE, 2017.

[9] Engstrom, L., Ilyas, A., and Athalye, A. Evaluating and understanding the robustness of adversarial logit pairing. arXiv preprint arXiv:1807.10272, 2018.

[10] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning, 2019.

[11] J. Hayes. On visible adversarial perturbations & digital watermarking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 1597–1604, 2018

[12] Tong Wu, Liang Tong, Yevgeniy Vorobeychik. Defending Against Physically Realizable Attacks on Image Classification. ICLR 2020.