Adversarial Camouflage:A Review

1st Naman Goenka
Department of Computer Science and Information Systems
Birla Institute of Technology and Science, Pilani
Rajasthan, India
f20180398@pilani.bits-pilani.ac.in

***Abstract*—This review paper has been written in partial fulfilment of the selection process for project-based course under Dr. Pratik Narang, CSIS, BITS Pilani. We attempt to review and present our understanding of technical contribution, findings and future scope of one research paper related to the topic of Adversarial Attacks. It commences with an introduction the topic, then proceeds to summarize technical contribution of the paper and research gaps identified and finally concludes with scope of future work of the topic.**

***Index Terms*—Adversarial Attacks, Adversarial Strength, Camouflage Flexibility, Adversarial Stealthiness**

I. Introduction

Well-designed input samples refered to as adversial examples are proven to be vulnerable to Deep Neural Networks (DNNs) [2]. This is a major concern in deploying DNN based models to security-critical applications such as self-driving cars. Two different setups in which adversarial attacks can be employed is: 1) **Digital Setting**, whereby input images are directly given to DNN network, perturbations of less magnitude are enough in this setting ; 2) **Physical-world setting**, whereby DNN network can only be fed from a camera, perturbations of large or unrestricted magnitude are required in this setup because of the kind of perturbations observed in this setup.

Three main properties to evaluate a adversarial attack is: 1) **camouflage flexibility** refers to the amount of control in the process of generation of adversarial examples; 2) **adversarial flexibility** indicating the possibility of detection of adversarial examples by a human observer; and 3) **adversarial strength** means the extent to which an adversarial attack can make a DNN classifier to do a wrong prediction. Adversarial attacks can also be classified as **targeted and untargeted attacks**. Targeted attacks prompts network to misclassify adversarial example into a specific label not equal to true label of the sample. Whereas untargeted attacks do not focus on the incorrect label to which adversarial sample is classified.

Targeted attack is to fool the network to misclassify the adversarial example into the class that attacker expects, while untargeted attack is to fool the network to misclassify the adversarial example into any incorrect classes

Adversarial attacks helps to examine robustness of DNN's. **Neural Style Transfer** used by the authors is based on notion of difference between content and style associated with an image. Style information can refer to various aspects such as texture, pattern etc. in the image. Convolutional neural network (CNN) can be used to separate the two, and recombine content of given image with target style information provided to get a style-transformed image of a given image.

II. Related Work and Research Gaps identified

Existing approaches like Projected Gradient Decent(PGD) [3], Carlini and Wagner(CW) [4], generative adversarial networks (AdvGAN) [5] in digital setting focus on crafting adversarial examples from small and imperceptible distortions. While methods in physical-setting like adversarial patch (AdvPatch) [6], robust physical perturbations (RP2) [7] focus on high magnitude unrealistic observations. Most of the existing adversarial attack methods both in digital and physical-world setting fail to build a flexible approach for developing highly stealthy camouflaged examples with high degree of distortion.

First problem identified by the authors in existing approaches is the need to specify optimum perturbation size to get highly stealthy adversarial examples, but it is quite difficult to do so while balancing adversarial stealthiness and strength. The proposed approach *AdvCam* eliminates the need of specifying distortion size as rather than restricting perturbations manually, it allows to incorporate more natural based perturbations via neural style transfer.

Second problem addressed is the amount of camouflage flexibility with the attacker in the process of generation of adversarial examples. Hence, authors target to present an **flexible, strong and stealthy approach** for large perturbations especially adapted for **robustness in physical-word setting**.

Even the methods which are capable of generating large unrestricted perturbations in digital setting lack the ability to simulate real-word situations via complex adversarial patterns. On the other side, majority of methods in physical-world setting attempt to generate large distortions but unfortunately often these generated adversarial examples become unrealistic while modelling natural transformations like light shift, rotation, movement of camera, etc.

III. Summary

A. Technical contribution

- A mechanism *AdvCam* is introduced which allows creating and camouflaging adversarial attacks adapted for physical-world with large perturbations.
- Attacker is given the flexibility to specify neural style in which adversarial example should be camouflaged in along with region of attack in target image.

- To adapt generated adversarial examples to natural transformations, physical adaptation training is performed by a technique similar to Expectation Over Transformation (EOT) [8].
- In a positive sense, *AdvCam* can be to **ensure privacy** by camouflaging private information leading it to be secure from both DNN and human observers.
- Calibrating style image x^8 along with background image o in reference with original image helps to make adversarial samples with high stealthiness and adversarial strength.

B. Methodology

The problem statement of adversarial attack is : Provided with a m-dimensional test image x and a true label y, a DNN classifier which maps m-dimensional image input to one of the k labels and a target class y_{adv} different from true label y, search an adversarial example $x^{'}$ corresponding to x by realizing this strategy for optimization:

$$\begin{aligned} &\text{minimize } \mathcal{D}(x, x') + \lambda \cdot \mathcal{L}_{adv}(x') \\ &\text{such that } \quad x' \in [0, 255]^m \end{aligned} \tag{1}$$

where first and second term represent trade-off between adversarial stealthiness and adversarial strength. In the experiments, apart from adversarial strength parameter $lambda$ all other parameters are treated as constants. The net loss called as *adversarial camouflage loss* is given by :-

$$\mathcal{L} = (\mathcal{L}_s + \mathcal{L}_c + \mathcal{L}_m) + \lambda \cdot \mathcal{L}_{adv} \tag{2}$$

$$Loss = (styleloss + contentloss + smoothnessloss) + \lambda \cdot adversarialloss \tag{3}$$

The first term represent style distance between style image and adversarial example generated which defines stealthiness; second term represent loss of content between original image and $x^{'}$; third term represents the differences between adjacent pixels of the generated image;with all three terms defining camouflage in final loss.

C. Experimental Setup summary

Baseline development is done through comparison with one of the best **models for mass-use** in digital (PGD) and physical-world (AdvPatch) settings with targeted and untargeted attack success rate and visual effect as evaluation parameters. Further authors conduct an ablation study for gaining insights into the attack examples generation process by focusing on adversarial strength parameter,camouflage losses and region size and shape. **Human perception study** analysis concludes that AdvCam distortions aren't restricted. Further analysis shows that generated samples are capable of fooling both DNN and human observer and achieve high stealthiness.

IV. Improvements and Scope of Future Work

Adversarial camouflage (AdvCam) [1] is introduced, which deploys neural style transfer technique to flexibly make adversarial examples capable of fooling both DNN and human observers. But, the proposed model still has need to manually calibrate neural style and the adversarial attack region. Hence there is scope for automation of the same via various image segmentation techniques.

Developing defense mechanism against such high stealthy attacks remains a crucial research opportunity for the future. Since only adversarial strength parameter λ is variable in all the experiments conducted, varying other parameters in (1) can be explored in the future.

References

[1] Duan, Ranjie, Xingjun Ma, Yisen Wang, James Bailey, A. Kai Qin, and Yun Yang. "Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1000-1008. 2020.

[2] Yuan, X., He, P., Zhu, Q., Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems, 30(9), 2805-2824.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt,Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In ICLR, 2018. 1, 2, 5

[4] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In IJCAI, 2018. 1, 3

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In IEEE SP, 2017. 1, 2

[6] Tom B Brown, Dandelion Man´e, Aurko Roy, Mart´ın Abadi,and Justin Gilmer. Adversarial patch. In NIPS Workshop,2017. 2, 3, 5

[7] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. In CVPR, 2018. 1, 2, 3

[8] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In ICLR, 2017. 1, 3, 5