Patching the Cracks: Detecting and Addressing Adversarial Examples in Real-World Applications

1st Niklas Bunzel [April 2025]
Fraunhofer SIT / ATHENE / TU-Darmstadt
Darmstadt, Germany
bunzel@sit.fraunhofer.de 0000-0002-8921-1562

***Abstract*—Neural networks, essential for high-security tasks such as autonomous vehicles and facial recognition, are vulnerable to attacks that alter model predictions through small input perturbations. This paper outlines current and future research on detecting real-world adversarial attacks. We present a framework for detecting transferred black-box attacks and a novel method for identifying adversarial patches without prior training, focusing on high entropy regions. In addition, we investigate the effectiveness and resilience of 3D adversarial attacks to environmental factors.**

***Index Terms*—Adversarial Attacks, Detection, Adversarial 3D Objects and Patches, Image Classification, Object Detection**

I. Introduction

Neural networks are increasingly being used in production systems and, in part, autonomously. They are also often used in security-related applications such as autonomous vehicles, facial verification or identification, enabling video identification systems, or mobile payment authorization. However, neural networks are prone to adversarial attacks that allow to control the outcome of their predictions, by adding small adversarial perturbations. The attacks can be broadly categorized as white-box, black-box, or a combination of the two, i.e., gray-box attacks. Most of the presented attacks can be considered as white-box attacks, which calculate an adversarial noise for a neural network [1], [2], [3], [4], [5], [6]. In a real world scenario, these attacks would most likely not be applicable as the attacker often does not have access to the neural network model. In turn, black-box attacks do not require any additional information about the target and thus, are more suitable in practice. Some examples of black-box attacks are gradient estimation [7] or gradient-free attacks [8], [9]. A major challenge of these attacks is, however, the efficiency in regards to the amount of queries needed for a successful attack. Therefore, transferred black-box attacks [10] can be used. Another attack approach are adversarial patches, where the adversarial perturbation is limited to a small patch that can be placed anywhere in an image [11]. Adversarial patches can be applied digitally to an image [12] or can be printed and placed anywhere and attached to objects in the real world [13], [14], [15], [16]. Adversarial attacks in the real world can also be used to protect the privacy of citizens by challenging systems designed to exploit facial recognition and other forms of invasive surveillance, turning the technology's vulnerabilities into a tool for enhancing personal privacy and security [17]. Defense techniques consist of adversarial training [18], [2], where the system is trained with adversarial attacks in addition to benign training data, thus learning to correctly classify objects even under adverse conditions. Adversarial training is computational and resource expensive, especially for systems already in production, and requires a trade-off between benign accuracy and attack robustness. Another defense technique is to detect adversarial attacks [19], [20]; this task is as difficult as correctly classifying the attacks [21] and also requires a trade-off. The advantage is that detection methods are typically cheaper, especially for production systems. However, in some use cases, such as autonomous driving, it is not enough to simply detect attacks - you also need to mitigate them in order to ensure the safe operation of the system.

Research Goal: The overarching goal of this research is to develop comprehensive and effective methodologies for detecting and mitigating adversarial attacks, including transferred black box attacks and adversarial patches, in both digital and real-world domains. The research aims to leverage explainable features and innovative techniques to enhance the security and robustness of systems against these types of attacks, thereby ensuring the integrity, reliability and accuracy of computer vision systems in various contexts.

II. Multiclass Black Box Detection

In cybersecurity, a variety of attacks are carried out by automatically exploiting known vulnerabilities, SQL injections, large-scale credential stuffing, or ransomware attacks. Off-the-shelf tools make it easy for attackers to identify vulnerabilities or execute attacks. It is expected that these types of automated attacks will eventually be applied to neural networks, with an increasing number of easy-to-use frameworks such as the adversarial robustness toolbox[1] or the foolbox[2]. Therefore, we focus on the detection of adversarial attacks with default parameters in [22]. The detection of these off-the-shelf black box attacks is even more relevant as we have shown in [23] that Projected Gradient Descent (PGD) attacks [2], initially designed for CNN image classification [2], can be partially transferred to state-of-the-art Large vision language models as GPT-4 [24]. We propose a multi-class detector framework

[1] https://adversarial-robustness-toolbox.readthedocs.io/en/latest/
[2] https://foolbox.readthedocs.io/en/stable/

2833-292X/24/$31.00 ©2024 IEEE
DOI 10.1109/DSN-S60304.2024.00020

Fig. 1: Left: Adversarial patch image [14], right: Detected and mitigated patch.

based on image statistics to detect transferred off-the-shelf adversarial attacks. We implemented a detection scheme for Attack on Attention (AoA) [25] and Fast Gradient Sign Method (FGSM) [18], based on explainable features. The detection rates achieved were 70% and 75% respectively, with a False Positive Rate (FPR) of less than 5%. The multi-class detector identified 77% of attacks as adversarial while correctly classifying 90% of the benign images, demonstrating its capability to detect out-of-the-box attacks.

III. Adversarial Patch Detection & Mitigation

Deep neural networks have significantly advanced image recognition, but are vulnerable to several security threats, especially when used in high-security applications such as facial recognition and autonomous driving. One notable type of attack is the adversarial patch, which alters pixels in a constrained area, causing object detectors to misidentify or miss objects entirely. While there have been defense proposals against these attacks, in some use cases, such as autonomous driving, it is not enough to just detect attacks; one must mitigate them to have safe system operation. In [26], we introduced a defense method against arbitrary patch shapes, different attack types, real patches, and digital patches. Our method is based on the observation that adversarial patches often appear extremely noisy with frequent pixel intensity changes. The proposed defense scheme uses edge detection to detect the presence of adversarial patches in an image without training. The framework operates as follows:

1) Initial Object Detection: The image is first processed by an object detector to identify initial object labels. For this, any state-of-the-art detector can be used. In this case, Faster R-CNN was employed.
2) Preparing Filter Masks: The image undergoes edge detection to identify potential patches. Each suspected patch results in a mask.
3) Patch Verification and Mitigation: To counteract patch attacks, each adversarial patch in the image is masked and inpainted one by one as shown in Figure 1. After each inpainting, the image is reprocessed by the object detector, and the labels are compared to identify adversarial patches.

Experimental results show that this framework can detect adversarial patches in the real-world with a 72% true positive rate. Additionally, the framework can determine if an image is adversarial and provide a corrected output post-mitigation. An approach using Error Level Analysis (ELA) for digitally applied adversarial patches and GLCM-based texture analysis for real-world adversarial patches show, that we can reach over 80% accuracy in detecting and mitigating adversarial patches[27].

IV. Adversarial Attacks from 3D to Real-World

Recent research has demonstrated the effectiveness of adversarial attacks in the digital domain. The translation of these attacks to real-world scenarios typically involves the use of adversarial patches. However, the effectiveness of such patches in real-world situations remains an area of active research. Evaluation in real-world contexts is relying on limited visual evidence provided by researchers. This raises the important question of the actual effectiveness and robustness of these attacks when exposed to real-world environmental conditions. We have implemented 3D adversarial objects to bridge the gap between the digital domain and real-world applicability. Our focus is to rigorously evaluate the robustness of these adversarial objects against varying environmental conditions such as rain, snow, fog, changing lighting conditions, varying camera angles and distances, and varying degrees of object overlap on the adversarial patches.
To this end, we developed a comprehensive 3D simulation environment capable of emulating these physical conditions.This environment is designed to facilitate the evaluation of both existing adversarial patches and our 3D adversarial objects. Subsequently, we intend to validate our simulated results with real-world experiments, highlighting the potential of our 3D adversarial objects and simulation environment to generate and evaluate attacks tailored for real-world effectiveness.

V. Conclusion & Future Work

This research aims to develop strategies for detecting and neutralizing adversarial attacks, with a particular focus on transferred black-box attacks and adversarial patches. We have shown the feasibility of using explainable features to detect black-box attacks, despite some trade-offs with benign accuracy. Our findings on identifying adversarial patches through regions of high entropy lead to a novel detection method. Future efforts will explore other signal-based, such as ELA, GLCM, and Wavelet, and neural network-based methods for improved defense, and delve into the practical use of adversarial 3D objects under varying conditions, anticipating further significant contributions to improving system robustness.

ACKNOWLEDGMENT

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[2] 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, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[3] A. Ghiasi, A. Shafahi, and T. Goldstein, "Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[4] E. Wong, F. R. Schmidt, and J. Z. Kolter, "Wasserstein adversarial examples via projected sinkhorn iterations," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 6808–6817.

[5] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[6] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 86–94.

[7] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C. Hsieh, "ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, B. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, and A. Sinha, Eds. ACM, 2017, pp. 15–26. [Online]. Available: https://doi.org/10.1145/3128572.3140448

[8] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1111–1119. [Online]. Available: https://doi.org/10.1145/3321707.3321749

[9] N. Bunzel and L. Graner, "A concise analysis of pasting attacks and their impact on image classification," in *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2023, pp. 136–140.

[10] N. Papernot, P. D. McDaniel, and I. J. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *CoRR*, vol. abs/1605.07277, 2016. [Online]. Available: http://arxiv.org/abs/1605.07277

[11] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017. [Online]. Available: https://arxiv.org/abs/1712.09665

[12] X. Liu, H. Yang, Z. Liu, L. Song, Y. Chen, and H. Li, "DPATCH: an adversarial patch attack on object detectors," in *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*, ser. CEUR Workshop Proceedings, H. Espinoza, S. Ó. hÉigeartaigh, X. Huang, J. Hernández-Orallo, and M. Castillo-Effen, Eds., vol. 2301. CEUR-WS.org, 2019.

[13] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on machine learning models," *CoRR*, vol. abs/1707.08945, 2017.

[14] A. Braunegg, A. Chakraborty, M. Krumdick, N. Lape, S. Leary, K. Manville, E. M. Merkhofer, L. Strickhart, and M. Walmer, "APRICOT: A dataset of physical adversarial attacks on object detection," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12366. Springer, 2020, pp. 35–50.

[15] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive GAN for generating adversarial patches," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 1028–1035.

[16] T. Sato, J. Shen, N. Wang, Y. Jia, X. Lin, and Q. A. Chen, "Dirty road can attack: Security of deep learning based automated lane centering under physical-world attack," in *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, M. Bailey and R. Greenstadt, Eds. USENIX Association, 2021, pp. 3309–3326.

[17] N. Bunzel, N. Zander, R. Landwirth, and A.-K. Riedel, "Adversarial examples zum selbstdatenschutz? der fall biometrischer gesichtserkennung im öffentlichen raum," *INFORMATIK 2021*, 2021.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[19] D. Hendrycks and K. Gimpel, "Early methods for detecting adversarial images," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[20] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *ArXiv*, vol. abs/1704.01155, 2018.

[21] F. Tramèr, "Detecting adversarial examples is (nearly) as hard as classifying them," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 21 692–21 702.

[22] N. Bunzel and D. Böringer, "Multi-class detection for off the shelf transfer-based black box attacks," in *Proceedings of the 2023 Secure and Trustworthy Deep Learning Systems Workshop*, ser. SecTL '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3591197.3591305

[23] N. Bunzel, "Transferrability of adversarial attacks from convolutional neural networks to chatgpt4," Nov 2023. [Online]. Available: https://publica.fraunhofer.de/handle/publica/456854

[24] OpenAI, "Gpt-4 technical report," 2023.

[25] S. Chen, Z. He, C. Sun, J. Yang, and X. Huang, "Universal adversarial attack on attention and the resulting dataset damagenet," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2188–2197, 2022.

[26] N. Bunzel, A. Siwakoti, and G. Klause, "Adversarial patch detection and mitigation by detecting high entropy regions," in *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2023, pp. 124–128.

[27] N. Bunzel, R. A. Frick, G. Klause, A. Schwarte, and J. Honermann, "Signals are all you need: Detecting and mitigating digital and real-world adversarial patches using signal-based features," in *Proceedings of the 2024 Secure and Trustworthy Deep Learning Systems Workshop*, ser. SecTL '24. New York, NY, USA: Association for Computing Machinery, 2024.