Robustness Benchmarking of Image Classifiers for Physical Adversarial Attack Detection

Ojaswee and Akshay Agarwal
IISER Bhopal, India
{ojaswee19, akagarwal}@iiserb.ac.in

Abstract

Deep neural network-based computer vision systems have become indispensable modules in modern days. Its rising popularity has caught the attention of fooling such networks. Recent works in this field have well-demonstrated the threat of adversarial attacks to real-world computer vision systems. While the majority of adversarial attacks are effective in the digital world, the adversarial patch attack is robust to be deployed in the physical world. Interestingly, existing literature on adversarial defense is heavily biased towards minute adversarial attacks, and little attention has been given to physical attack detection. To overcome this limitation, in this research, we have developed a novel adversarial patch attack dataset to benchmark the defense study in this critical direction. Using the collected dataset, we have conducted several experiments both under seen and unseen patch settings. On top of that, the generalization experimental setting, i.e., unseen dataset evaluation, shows that adversarial patch attacks are hard to defend. We assert that such a study along with the dataset and complexity in defending the patch attacks can inspire future defense works to add these attacks as well.

1. Introduction

The adversarial vulnerability of deep neural networks (DNNs) has received a lot of attention. The adversarial attacks can be broadly divided into three broad categories: (i) minute adversarial perturbations [12, 20], (ii) universal perturbations [16], and (iii) physical adversarial patches [8, 14]. While novel adversarial attacks ensure that the DNNs are secure from any possible vulnerability, the defense algorithms tackling them independently might be a severe concern [1, 6, 7]. It is seen that the transferability and applicability of minute and universal adversarial perturbations are limited in the physical world compared to adversarial patch attacks. Still, the majority of the adversarial defense algorithms target minute adversarial perturbation; whereas, little work tackles universal adversarial perturbations and physical patch attacks. Figure 1 shows the adversarial noises formed under three broad categories mentioned above. The minute and universal perturbations do not occlude any region of the images; however, the adversarial patches can occlude a small or significant portion of an image, depending on its size. These drastic natures in not only crafting but also applying an adversarial perturbation lead to a significant distribution shift among the adversarial images. Therefore, existing defenses tackling minute perturbations are not effective for physical patch detection. Hence, we assert that ignoring the impact of adversarial patch attacks can be dangerous, especially when aiming to deploy these state-of-the-art DNNs in the unconstrained physical world. Further, we want to mention that while several benchmark studies are also proposed in the literature to tackle the issue of adversarial defense, no work has included adversarial patch attacks. For example, Hendrycks and Dietterich [13] showcase the impact of several common corruptions, such as Gaussian noise, blur, and fog on DNNs, but no defense has been pro-

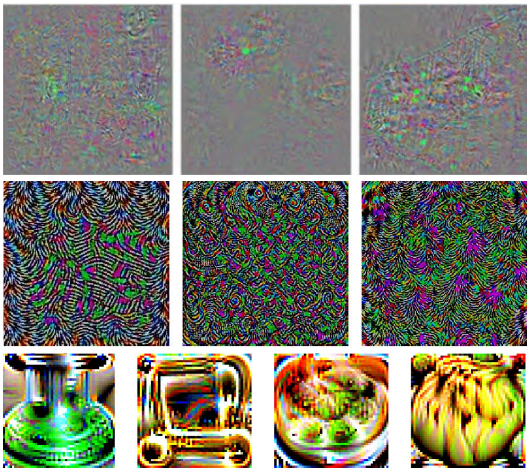

Figure 1. Distribution shift among different adversarial attacks. The first row is minute adversarial perturbation noise, the second is the universal perturbation vectors, and the third is the physical adversarial attacks.

posed in this study. Dong et al. [11] proposed a benchmark study to tackle only the minute adversarial perturbations. Several benchmark studies recently proposed to increase the defense umbrella by combing adversarial perturbations and common corruptions [2, 4]. Several defense works also exist that can effectively detect adversarial attacks in several generalized settings such as unseen datasets, unseen perturbation, and unseen threat model [1, 5, 6]. As discussed, no studies have benchmarked the defense against adversarial patch attacks; therefore, in this research, after creating patch attack datasets, we have performed the defense study using several deep image classification networks, including the network architecture search (NAS) method [22]. The prime reason for conducting a benchmark study on adversarial patches can also be understood from the distribution shift among the attacks and out-of-distribution handling limitations of the DNNs. We assert that the presence of the dataset and benchmark evaluation can help advance the research in this direction and make comparisons with new novel algorithms. In brief, the contributions of this research are:

- A novel adversarial patch attack dataset has been developed. The dataset contains images of multiple variations of patches. The presence of different style patches will ensure that the defense algorithms are not biased;

- A benchmark evaluation has also been conducted. For that, several real-world evaluations and protocols are developed to handle seen patches, unseen patches, and unseen datasets. A defined protocol can help make fair comparisons in future works, which is often missing in minute adversarial detection literature.

2. Adversarial Patched Dataset

Considering the lack of research in the field of detection of adversarial patch attacks. In this research, we have developed two unique adversarial patch datasets utilizing the images from the validation sets of ImageNet [10] and COCO [15] datasets. We have randomly selected 2000 images from each dataset and treated them as a real subset of the proposed dataset. On this real subset, we have applied different adversarial patches having significantly different styles from each other. To perform the adversarial patch attack, we have used the patches from ImageNet-Patch [17]. These adversarial patches are effective in the real world compared to the minute (additive) adversarial perturbations due to the transformation applied while learning the patches. These are targeted patches, each having a different target class; hence these patches not only have variations in texture and style but can also misclassify the images into different categories such as soap dispenser, cornet, plate, banana, cup, typewriter, electric guitar, hair spray, sock, a cellular telephone. Using these 10 patches, we have generated a large-scale adversarial patch dataset containing $20,000$ patched images of the validation set of the ImageNet dataset. We did the same with randomly selected 2000 images of the validation set of the COCO dataset, giving us 20000 more patched images. However, the images of COCO datasets are only used for evaluation to ensure the detection algorithms' unseen dataset generalizability. In total, the proposed dataset contains $40,000$ adversarial patch images and $4,000$ real images. Figure 2 shows some of the samples from the proposed dataset reflecting the challenge in detecting the adversarial attack not only due to significant style change of the patches but also their blending nature with the complex image regions.

Figure 2. Samples of the different adversarial style patches used in preparing the dataset. It can be seen that these patches can be blended with the image content and hence increase the complexity of its detection, especially when the detection network has not seen them in the training.

3. Benchmarking Adversarial Patch Detection Results and Analysis

Architecture: As mentioned, this research aims to overcome the limitation of the existing adversarial defense literature and benchmark the robustness evaluation of state-of-the-art (SOTA) image classifiers. Henceforth, we have used several CNN architectures to extensively study the robustness of classifiers and avoid any classifier bias to perform adversarial patch detection. The used architectures varied in terms of the number of layers, the connection between layers, and their formation and are as follows: XceptionNet [9], MobileNetv2 [18], NASMobileNet [22], and VGG16 [19]. These networks are finetuned by adding a few dense layers to extract the image features along with the classification layer. The reason for using these archi-

Table 1. Adversarial patch detection accuracy of the different architectures on ImageNet subset. The results are reported in terms of mean and standard deviation (SD), where the trained on one patch is tested on all the patches.

Models	Metric	Patch 0	Patch 1	Patch 2	Patch 3	Patch 4	Patch 5	Patch 6	Patch 7	Patch 8	Patch 9
Xception	Mean	81.51	75.23	82.39	76.43	84.13	70.90	72.99	78.54	81.85	78.34
	SD	07.80	07.72	06.82	09.68	05.89	10.20	09.72	09.30	08.57	10.24
VGG16	Mean	80.36	71.87	80.52	80.57	73.20	67.41	69.53	72.33	76.96	79.73
	SD	12.63	16.95	16.33	20.28	10.12	17.28	18.43	19.54	18.75	14.47
MobileNet	Mean	67.24	70.86	68.18	66.04	75.40	65.45	67.71	71.97	80.57	74.46
	SD	14.10	14.15	13.95	14.98	11.04	17.80	16.42	16.00	16.35	14.76
NASMobileNet	Mean	71.10	70.00	72.60	69.00	72.00	68.10	68.40	71.80	71.90	71.50
	SD	02.47	02.31	02.12	02.40	02.62	02.96	03.10	03.49	03.25	04.65

Table 2. Adversarial patch detection accuracy of the different architectures on COCO subset. The results are reported in terms of mean and standard deviation (SD), where the trained on one patch is tested on all the patches. In contrast to the COCO results, here, the results are generated using the detection models trained on the ImageNet subset and tested on the COCO subset.

Models	Metric	Patch 0	Patch 1	Patch 2	Patch 3	Patch 4	Patch 5	Patch 6	Patch 7	Patch 8	Patch 9
Xception	Mean	74.90	68.59	75.75	69.80	77.49	64.28	66.36	71.91	75.25	71.73
	SD	09.94	05.77	08.37	09.50	08.19	07.66	07.20	10.30	10.77	11.36
VGG16	Mean	78.84	70.36	79.00	79.05	71.70	65.89	68.02	70.82	75.43	78.21
	SD	12.89	16.19	16.43	20.05	10.06	16.53	17.62	19.61	18.97	14.27
MobileNet	Mean	66.63	70.21	67.56	65.41	74.79	64.83	67.09	71.36	79.93	73.82
	SD	14.21	13.85	14.07	15.08	11.11	17.58	16.20	16.03	16.41	14.74

tectures is that they are heavily popular for image classification; therefore, understanding their adversarial patch detection effectiveness can raise the alarm to pay undue attention to this attack as well. Further, NAS-style architectures have never been explored for adversarial patch detection, and NAS is a promising direction to find an effective architecture for image classification. Therefore, benchmarking its robustness can pave the way to incorporate the adversarial nature of the images while crafting network architecture.

Results and Analysis: *We have performed extensive experiments in seen patches, unseen patches, and unseen patched dataset settings to effectively evaluate the performance and robustness of the image classifiers for adversarial patch detection.* This setting reflects the real-world nature, as in the real world, novel attack patches or out-of-distribution datasets can occur, which can easily break the developed adversarial defense. Due to the space limitation, we have reported the results in terms of the average classification performance of each network along with the standard deviation (SD) observed in their detection performance. Table 1 shows the average performance of different image classifiers when trained and tested on the same and different patches. Whereas Table 2 shows the '*dual generability*' of the classifiers, i.e., in this setting, the classifiers are trained on ImageNet and tested on the COCO subset. Deep networks for image classification are found vulnerable against out-of-distribution samples [21]. The same can be seen where the patch detection performance is significantly low when tested on an unseen dataset compared to the seen dataset. *Regarding image classifiers, VGG is found to be the most effective, whereas, MobileNet is found least effective in detecting adversarial patches under unseen patch detection settings.*

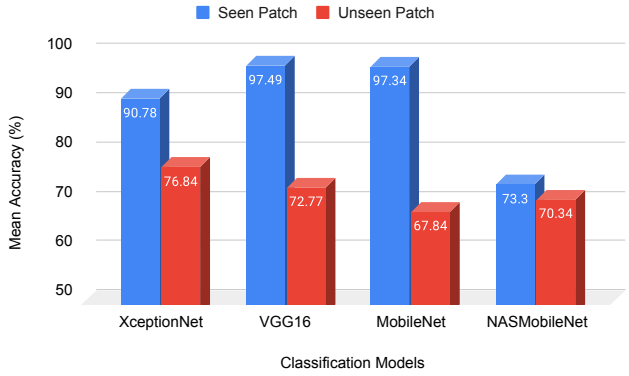

Figure 3. Average adversarial patch attack detection performance on the ImageNet subset under seen and unseen patches evaluation setting. The results reflect that when unseen patches come for classification, the performance of the networks drops drastically.

Figure 3 shows the robustness of each classifier when trained and tested on seen and unseen attack patches. It is clear that when the networks try to classify the adversarial patches not seen during training, they suffer significant drops in performance. The prime reasons as mentioned are, that the style texture distribution shift across patches and their blended nature with the complex image regions. Interestingly, NASMobilenet shows the highest level of generalizability; however, the network's capacity is found least. The network's capacity is defined as its performance in seen patches training testing conditions. For example, the perfor-

mance of NASMobileNet is $\mathbf{3}\%$ lower in the unseen patches setting than seen patches setting; however, its performance is at least $\mathbf{17.4}\%$ lower than other architectures in seen evaluation setting. However, the architecture shows the possibility of developing a robust adversarial patch detection classifier when intelligent adversarial patch information can be incorporated while crafting the architecture search.

4. Conclusion

Adversarial vulnerability of 'any' and 'every' style of convolutional neural networks, including vision transformers [3] raises a severe concern about their deployment in the real world. One complex adversarial attack is known as an adversarial patch attack; can be deployed in the physical world; however, the defense against this attack has yet to receive attention. To make an impact in this direction and advance the research in handling this real-world attack, we have developed a large-scale dataset containing $\mathbf{44,000}$ real and adversarial patched images. We have utilized these images to perform an extensive adversarial patch detection benchmark study using several SOTA deep image classifiers. The experimental analysis reveals that detecting adversarial patch attacks is challenging due to their varying texture style, especially when the patches are not known at the time of training a detection network or coming from out-of-distribution images. In the future, we aim to extend the dataset further and build a unified and robust patch attack detection architecture.

References

[1] Akshay Agarwal, Gaurav Goswami, Mayank Vatsa, Richa Singh, and Nalini K. Ratha. Damad: Database, attack, and model agnostic adversarial perturbation detector. *IEEE TNNLS*, 33(8):3277–3289, 2022. 1, 2

[2] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Benchmarking robustness beyond l_p norm adversaries. In *ECCVW*, 2022. 2

[3] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Crafting adversarial perturbations via transformed image component swapping. *IEEE TIP*, 31:7338–7349, 2022. 4

[4] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Exploring robustness connection between artificial and natural adversarial examples. In *IEEE CVPRW*, pages 179–186, 2022. 2

[5] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In *IEEE BTAS*, pages 1–7, 2018. 2

[6] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Image transformation-based defense against adversarial perturbation on deep learning models. *IEEE TDSC*, 18(5):2106–2121, 2021. 1, 2

[7] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Cognitive data augmentation for adversarial defense via pixel masking. *PRL*, 146:244–251, 2021. 1

[8] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *ECCV*, pages 514–532. Springer, 2022. 1

[9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE CVPR*, pages 1251–1258, 2017. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009. 2

[11] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *IEEE CVPR*, pages 321–331, 2020. 2

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

[14] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *IEEE CVPR*, pages 7848–7857, 2021. 1

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE CVPR*, pages 1765–1773, 2017. 1

[17] Maura Pintor, Daniele Angioni, Angelo Sotgiu, Luca Demetrio, Ambra Demontis, Battista Biggio, and Fabio Roli. Imagenet-patch: A dataset for benchmarking machine learning robustness against adversarial patches. *PR*, 134:109064, 2023. 2

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE CVPR*, pages 4510–4520, 2018. 2

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[20] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[21] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 3

[22] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *IEEE CVPR*, pages 8697–8710, 2018. 2