# An Assessment of Robustness for Adversarial Attacks and Physical Distortions on Image Classification using Explainable AI

K.T.Y. Mahima[1], Mohamed Ayoob[1] and Guhanathan Poravi[1]

[1]Department of Software Engineering, Informatics Institute of Technology, Colombo, Sri Lanka

## Abstract

Introducing defence mechanisms to overcome the vulnerability of adversarial attacks is a highly focused research area. However recent research highlights that introducing defence approaches for man-made adversarial attacks is not sufficient, because the deep learning models are vulnerable to the perturbations outside the scope of the training set and the physical world itself acts as an adversarial sample generator. Given this caveat, there is a necessity to introduce general defence approaches for both man-made and physical world adversarial samples. Prior to that, a brief explanation of how the model's decision-making process happens in the inference phase under the various adversarial perturbations is required. However, the deep learning models act as black boxes in the inference phase where the decision-making is not interpretable. As a result, research on model interpretability and explainability has been carried out in the domain which is collectively known as Explainable AI. Using a set of Explainable AI techniques, this study is investigating the deep learning networks' robustness; i.e., the decision-making process in neural networks and important pixel attributes for the predictions that are captured when the deep learning model inference gets adversarial inputs. These adversarial inputs are perturbed by adversarial attacks or the physical world adversaries using the deep learning network trained on the CIFAR10 dataset. The study reveals, that when the inference gets adversarial samples, the necessary pixel attributes for the prediction captured by the network vary everywhere in the image. However, when the network is re-trained using adversarial training or data transformation-based augmentation, it will be able to capture pixel attributes within the particular object or reduce the capture of negative pixel attributes. Based on the deductions gained from the findings, this paper states some potential research approaches to introduce a general adversarial defence method.

## Keywords

Computer Vision, Adversarial Robustness, General Robustness, Explainable AI, Model Interpretability

## 1. Introduction

Usage of Machine Learning (ML) has grown with the revolution of deep learning (DL) because of its superior generalization and applicability to perform advanced tasks in computer vision and across other domains which used traditional ML technologies [1]. The ability of the DL/ML to deliver its promises is, conditional on successfully mitigating the ethical and practical issues that occur such as algorithm bias and model interpretability. Moreover recently there is a huge discussion on Altruism AI which emphasizes how people can get a positive impact from the AI

*AI-CyberSec 2021: Workshop on Artificial Intelligence and Cyber Security, December 14, 2021, Cambridge, UK*

yasasmahima99@gmail.com (K.T.Y. Mahima); nazeem.2016343@iit.ac.lk (M. Ayoob); guhanathan.p@iit.ac.lk (G. Poravi)

without any biases and vulnerabilities [2]. In domains such as healthcare, trustworthiness and the transparency of the predicted outputs is of paramount importance and recent research have given their full potential to uncover the questions of trustworthiness and the fairness of the AI [3].

Recent research identified a significant drawback of these DL networks, the models' performance is degraded when the inference gets the samples out of distribution from the training set [4, 5]. Given priority to this phenomenon, a novel security threat for DL networks was introduced namely adversarial attacks where the DL model's predictions are able to be completely altered using the deliberately synthesized adversarial attacks which are visually similar to the clean inputs [6]. Moreover, recent research shows real-world computer vision applications deployed in robotic systems like Unmanned Vehicles (UV) and mobile devices also show considerable performance degradation under unintended physical world distortions also known as common corruptions such as noises and lightning level changes [5, 7, 8]. These physical adversarial conditions could appear as singular or mix-up instances. Further, there could be multiple degradations that appear at the same time. Potentially these corruptions are not stronger than digitally altered adversarial attacks, but strong enough to make security impacts.

To overcome the vulnerability of these two security threats in DL networks researchers introduced several robustness improvement approaches separately for each category. However, we identified that the integration of separate resilient approaches for adversarial attacks and physical distortions could increase the computational power of the system in the inference. In UV systems this would be a problem because those have various high resource-consuming tasks. Thus affording a greater proportion of computational resources to adversarial resilience is questionable and it could be allowed only if the defence method gives superior performance against attacks [9, 10]. Researchers tend to introduce general resilience approaches which enhance the robustness of existing DL networks and networks which are implemented from scratch against both human synthesized and naturally occurring adversarial inputs when these kinds of requirements are prioritized [11, 12].

Prior to introducing a general resilient approach to both the aforementioned types of adversarial perturbations (i.e., digital and physical adversaries), it is able to understand that, a comprehensive analysis of DL networks performance against both human synthesized and physical adversarial perturbations is required to identify the DL model's behavior under the influence of adversaries and to verify the relationships between each perturbation. However, the DL models act as black boxes in the inference phase where the decision-making process of the model is not accessible or interpretable [13, 14]. Explainable AI (XAI) as a field exists to pique interest in transparent decision-making of DL networks to better understand it. This has attracted increasing attention within the research community towards various XAI model interpretability algorithms and visualization approaches such as Saliency Maps, Integrated Gradients, Layer-wise Relevance Propagation (LRP), GradCam…etc. [15].

This is, to the best of our knowledge, the first paper to assess the general adversarial robustness and identify a relationship between adversarial attacks and physical world distortions using XAI. Consequently, we will assess previously presented model interpretation algorithms, using a set of chosen model interpretability approaches. Thereafter, we will conduct a comprehensive analysis of networks behavior and mark our observations on how it captures the pixel attributes

used to make the predictions under the human crafted digital adversaries and naturally occurring physical adversaries using the DL network implemented on CIFAR10 [16] dataset. Based on the main objectives of this study, the following research questions can be implied.
**RQ1** - How to assess general adversarial robustness using XAI algorithms?
**RQ2** - How to identify a relationship between adversarial attacks and physical distortions based on positive and negative pixel attribution captured by the network?

## 2. Related Works

### 2.1. General Adversarial Robustness

Data augmentation or transformation during the training phase is a wide research topic in the present since it will help to increase performance, avoid generalization issues, and improve the resistance against physical corruption [17]. In contrast, adversarial training is one of the most effective adversarial attack defence approaches where the network is re-trained using the adversarial samples synthesized on a particular attack [18]. Laugros et.al examined whether data augmentation is an appropriate way to improve the resilience against adversarial attacks and vice versa, adversarial training is capable of only improving the resilience against physical corruptions. However, the empirical results showed those two approaches are not mutually exclusive from each other [11]. Moreover, Daniel et al. also proved that model re-training individually with adversarially improved physical distortions like fog, snow…etc. , or with $l_\infty$, $l_2$ adversarial samples only improve the robustness for particular corruption [19].

Zhang et al. proposed a model robustification approach where the primary model is trained with the use of several supplementary classifiers which learn from the physically corrupted samples. Moreover, they showed when the primary classifier is trained using adversarial attack samples, the proposed approach could be used to improve the general robustness. However, the training time computational cost would be high in this approach [20]. DeepMind by Google proposed an adversarial data augmentation approach optimized by projected gradient ascent [21]. The results showed this approach is able to effectively increase the robustness against common physical corruptions. In particular, combined with other data augmentation techniques like AugMix [22] and DeepAugment [23] it improved the resilience against $l_\infty$ and $l_2$ attacks. Moreover, the proposed data augmentation approach is much robust than the $l_\infty$ and $l_2$ aversarial training for common corruptions.

Laugros et.al introduced a unified general defence approach using targeted labeling adversarial training with Fast Gradient Sign Method (FGSM) based adversarial assaults and image on image mixup [24] data augmentation approach. The empirical results show that, while their approach improves the general resilience, there is a slight performance degradation for some adversarial attacks while integrating the adversarial training with the image mix-up method. Further, exclusively using adversarial training alone, givesthe model more resilience to adversarial attacksthan the integrated general approach [12]. Park et.al showed, though data augmentation improves the clean accuracy, it will reduce the performance against adversarial attack samples. This emphasizes there is an internal corruption between adversarial attack perturbations and data augmentation [25].

The above-discussed works show general robustness on existing networks or DL networks built from the scratch required more research. If there is a solution that improves the resiliency of the network naturally without any modifications or no usage of any supporting tool, it is able to save computation power in the inference and be re-deployed without any additional dependencies. To achieve this, identifying the DL networks' behavior under each adversary is essential.

## 2.2. Model Interpretability

As discussed earlier DL model inference is a complete black-box where the models' decision-making is not transparent. Thus XAI technology was introduced. XAI is a reductive method to improve the transparency of the network which allows to examine and visualize how the model infers a particular prediction [15]. In this section, we will briefly discuss the chosen model interpretability algorithms for this study.

Recent research demonstrates model interpretability methods are able to use as an adversarial attack detector by analyzing the abnormal behavior of the pixel features captured by the network [26, 27]. So this mainly motivates us to conduct this assessment using XAI methodologies. For our study, we used the saliency maps-based XAI approaches which are post-hoc analysis methods; they, produce heatmaps to represent the contribution of the input features to the output of the network. By using the heatmaps it is possible to arrive at a fairly accurate picture of the regions of interest in an image where the DL model has high gradient flows to make the prediction [28]. Below we summarize the chosen XAI algorithms for this study.

### 2.2.1. Saliency

The Saliency [29] approach allows computing input features and output gradients concerning the input features in a heatmap. The generated heatmap from this method is at a fine-grained pixel level as we have gradients for each pixel [30]. However, this algorithm ignores the relevance of a singular pixel in an image to its surrounding pixels. This would be particularly problematic for any examination of fine-gradient classification models [31]. This would not be exactly relevant to our research but rather a scaffold to future works in accessing adversarial attacks using the XAI domain. In this scope of research, we focus on non-fine-gradient image classification.

### 2.2.2. DeepLift

Using the activation of each neuron, the proposed algorithm assigns contribution scores which are deliberated by comparing the output of the given input sample and a reference point backpropagated to each neuron. In particular, DeepLift gives a separate reflection of the positive and negative contributions. The state-of-the-art results exhibit DeepLifts superiority over gradient-based methods using models trained on images and genomic samples [32].

### 2.2.3. Integrated Gradients

Integrated Gradients is an axiomatic attribution approach that requires no modification to the network. The explanation was generated by attributing the predictions of the network to the input features. By using a few standard gradient calls, the variation of the gradients of the predicted output with respect to the input features was calculated [33]. The proposed approach satisfied a fundamental axiom, the desired property of the generated explanations which is known as completeness [15].

#### 2.2.4. GradientShap

Shapley Additive explanations (SHAP) enhance the model's interpretability by enumerating the essential values/SHAP values for each feature for a single prediction. GradientShap approximately estimates the SHAP values by computing the projection of gradients by randomly sampling from the distribution of the baselines. The authors of SHAP proposed SHAP values as a unified evaluation for calculating the feature importance [34].

## 3. Methodology

### 3.1. Standard Classification Model and Dataset

For the experiments of the study, we have chosen the CIFAR10 dataset [16] which is a balanced dataset, consisting 32x32 6000 RGB images classified into 10 classes. The standard classification model is a self-composed network with 1 dropout layer, 3 max-pooling 2d layers, 3 conv2d layers, and 2 linear layers with relu activation function. While training the initial model we have not used any color transformation or noises and the model was trained until the testing accuracy is unchanged throughout several epochs with model generalization.

### 3.2. Adversarial Attacks

For a fair assessment, we inspect both $l_\infty$ and $l_2$ norm bounded adversarial attacks. As the $l_\infty$ norm attacks we choose Fast Gradient Sign Method (FGSM) [6], Basic Iterative Method (BIM) [35], and Projected Gradient Descent (PGD) [36] attacks. Moreover as the $l_2$ norm attack we chose PGD $l_2$ norm attack. Equation 1-3 demonstrate the hypothesis functions of the FGSM, PGD and BIM attacks respectively. Here $x_{adv}$ is the adversarial input, $h$ is the DL network, $x$ input image, $y$ ground truth, $\epsilon$ epsilon, $\delta$ the adversarial perturbation, $\nabla_x$ the gradient of the model, $\Pi$ the project to the ball of interest (Clipping values between values [-$\epsilon$,$\epsilon$], $\alpha$ the step size and $l(h(x), y)$ the loss function of the network. For the evaluations of the $l_\infty$ attacks we kept $\alpha = 0.005$, number of iterations $T = 10$. Moreover we use worst case $\epsilon$value where $\epsilon = 0.01$. For $l_2$ attack $\epsilon = 5.0$, $\alpha = 0.5$.

$$x_{adv} = x + \epsilon . \sin(\nabla_x l(h(x), y)) \tag{1}$$

$$x^t_{adv} = x + \Pi_\epsilon(x^{t-1} + \alpha . \sin(\nabla_x l(h(x^{t-1}), y))) \tag{2}$$

$$x^{'}_{t+1} = clip_{(x,\epsilon)}\{x^{'}_t + \alpha . \sin(\nabla_x l(h(x^{'}_t), y))\} \tag{3}$$

**Table 1**
Selected Image Transformations as the Physical World Corruptions

| Color and Lightning Transformations | Geometric Transformations | Noise & other Transformations |
|---|---|---|
| Contrast | Swirl | Noise (Gaussian, speckle, pepper, salt, salt & pepper noise) |
| Exposure | Shearing | Scaling |
| Sharping | - | Blur |

### 3.3. Physical World Corruptions

Past research on DL network robustification used synthesized image transformations for the evaluations for the physical corruptions [37]. Thus as physical world corruptions, we have chosen 12 transformations classified into 3 categories. Table 1 summarizes the selected adversarial image transformations we have chosen as physical corruptions.

## 4. Experiments and Results

### 4.1. Experimental Setup

The PyTorch library was mostly utilized to build the DL network. The SkImage library was used to create physical corruptions via image transformations. To perform the above discussed XAI algorithms for the inspection we have used the Captum library by FaceBook AI [38]. Captum consists of various previously introduced model interpretability algorithms and appropriate visualization methods as well. It also contains a set of evaluation metrics for evaluating these XAI algorithms. This research used an image from CIFAR10's "Plane" class to display the assessment results. A particular picture of the “plane” class was chosen due to the high separation between the class level pixels and non-class pixels. The plane object is also of uniform linear shapes which helps us to qualitatively assess the XAI performance.

### 4.2. Assessment Results for Adversarial Attacks

Initially, the performance of the composed CIFAR10 classification model under the chosen adversarial attacks was assessed. Table 2 summarizes the Robustness Score ($R_N^\phi$ : Equation 4) [11] of the model under different attack scenarios.

$$R_N^\phi = \frac{A_\phi}{A_{clean}} \tag{4}$$

*Where $A_\phi$ is the accuracy of the adversarially corrupted (Dataset with adversarial examples) data set, $A_{clean}$ is the accuracy of the clean data set and N is the neural network.*

Next, we analyze the feature maps of each XAI algorithm under these attacks scenarios. Initially, the network's pixel feature's inferring accuracy of a clean non-adverse sample was examined. Based on the feature attributions generated by the XAI algorithms and displayed

**Table 2**
Robustness Scores of the Model under Different Attacks

| Adversarial Setting | Clean | FGSM | BIM | PGD $l_\infty$ | PGD $l_2$ |
|---|---|---|---|---|---|
| Robustness Score | 1.0 | 0.570 | 0.77 | 0.476 | 0.786 |

in the saliency maps in Table 3, we elaborate that while input is perturbed by adversarial attacks the important features captured are dispersed everywhere in the image. In addition, the evaluated XAI algorithms (Except the Saliency method) showed the adversarial attacks limit the capturing pixel feature attributions on top of the object in a particular image. According to the samples displayed in Table 3, the essential features captured on top of the "Plane" object while the input is a clean sample, are shifted and dispersed out of the "Plane" object while the adversarial attacks are performed. Thus we could arrive with reasonable premises for the following argument that this phenomenon would be the reason for misclassification of the predictions while inference gets human synthesized adversarial assaults. In contrast, the feature dispersing strength is relatively low under $l_2$ norm adversarial settings.

## 4.3. Assessment Results for Physical World Corruptions

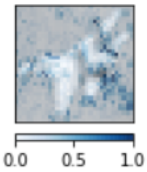
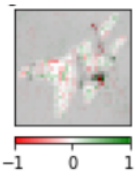
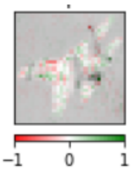
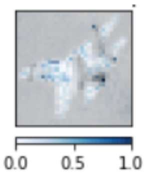
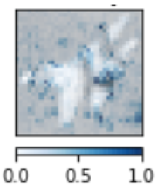
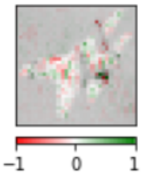
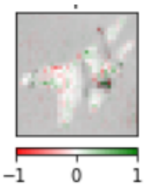
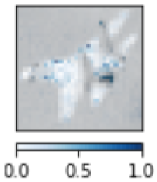
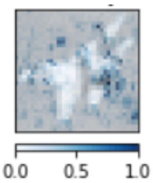
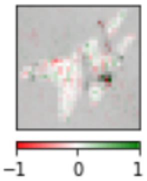
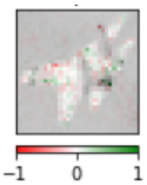
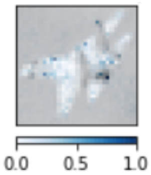
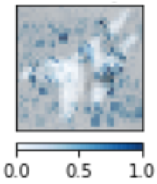
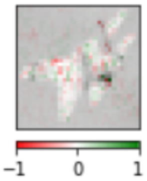
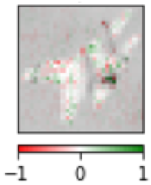
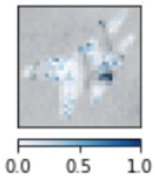
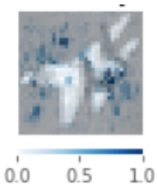
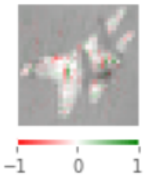
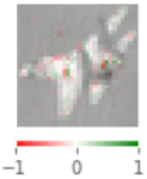
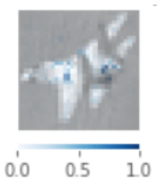
For evaluating the physical world adversarial settings we used the same procedure as the adversarial attacks and as mentioned earlier transformations presented in Table 1 were considered as the physical distortions. These physical world adversarial samples are not stronger than the adversarial attacks since they do use any knowledge about the model such as gradients. There is an average of 10% accuracy degradation for the evaluated physical corruptions. Sample Saliency maps of capturing feature attributes under the physical adverse conditions are demonstrated in Table 4.

Here we understood most of the physical world corruptions, act as the human synthesized adversarial assaults. In particular, the empirical results showed the network has the same behavior under contrast and scaling where it captures the pixel features dispersed out of the object when compared to the other distortions. In contrast exposure limits the capturing of the positive feature attributes withinside the object when analyzing the saliency approach. However other XAI algorithms that demonstrate exposure also result in dispersing the features. Moreover, under shearing, the captured feature attributes have shifted towards the angle between the original and image corrupted by shearing. A complete explanation of this phenomenon is displayed in Fig 1. In some scenarios, sharpening has increased the capturing feature attributes of the network far-better than the clean samples.

Among the assessed physical corruptions, the noise array was evaluated subjectively. There we understood all the noises result in dispersing the feature attributes everywhere in the image. Moreover, the behavior of the adversarial attack noises and these physical noise corruptions is almost similar. Among the investigated noise corruptions, Speckle and Gaussian noise transformations shifted and dispersed feature attributes away from the "Plane" object. In addition, when Salt, Pepper, and Salt & Pepper noise transformations are performed the feature attributes tend to superimpose on top of the noise points.

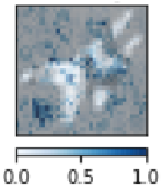
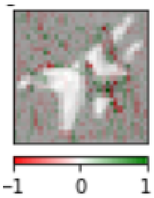
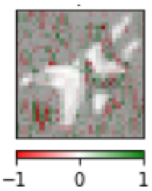
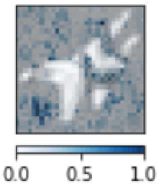
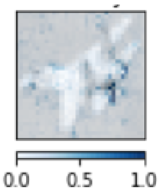
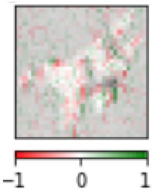
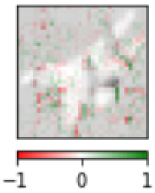
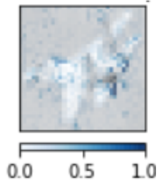
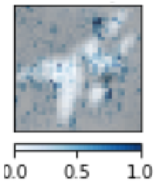
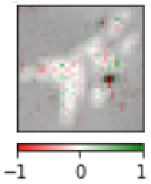
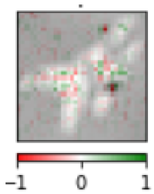
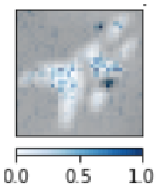
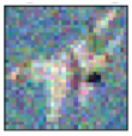
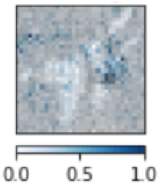
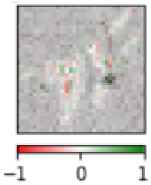
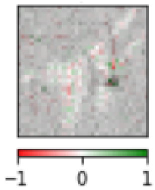
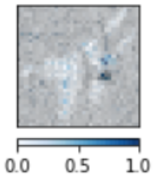
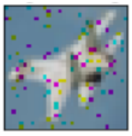
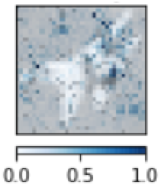
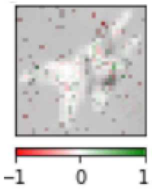
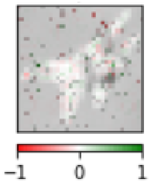
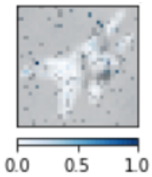
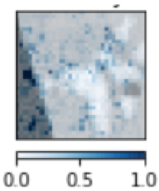
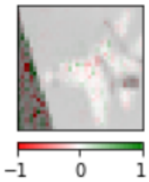
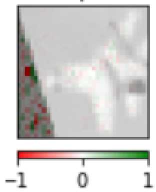
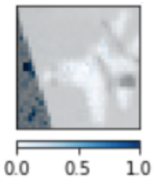
Overall, it can be deduced from the findings that the network's feature selection will be

**Table 3**
Feature attributions generated by the XAI algorithms for Adversarial Attacks (Pred:Prediction, Conf:Confidence)

| Adversarial Setting | Original Image | Saliency | Integrated Gradients | DeepLift | GradientShap |
|---|---|---|---|---|---|
| Clean | <br>Pred:Plane<br>Conf:0.99 |  |  |  |  |
| FGSM | <br>Pred:Deer<br>Conf:0.99 |  |  |  |  |
| BIM | <br>Pred:Deer<br>Conf:0.99 |  |  |  |  |
| PGD $l_\infty$ | <br>Pred:Deer<br>Conf:0.99 |  |  |  |  |
| PGD $l_2$ | <br>Pred:Deer<br>Conf:1.0 |  |  |  |  |

dispersed and shift away from the particular object, or that it would be limited while receiving adversarial samples. In particular, the network's feature attributes selected via backpropagation have similar behavioral patterns under physical and adversarial attack noise settings. Thus to overcome the adversarial noises and physical noises a mathematically optimized solution would be ideal. A complete summary of the XAI assessment details under the physical adversarial samples could be viewed at https://gitlab.com/a4855/xai-assessment.

**Table 4**
Feature attributions generated by the XAI algorithms for Physical Adversaries (Pred:Prediction, Conf:Confidence)

| Adversarial Setting | Original Image | Saliency | Integrated Gradients | DeepLift | GradientShap |
|---|---|---|---|---|---|
| Contrast | <br>Pred:Deer<br>Conf:0.99 |  |  |  |  |
| Exposure | <br>Pred:Plane<br>Conf:0.97 |  |  |  |  |
| Scaling | <br>Pred:Cat<br>Conf:0.99 |  |  |  |  |
| Gaussian Noise | <br>Pred:Frog<br>Conf:0.99 |  |  |  |  |
| Pepper Noise | <br>Pred:Bird<br>Conf:0.99 |  |  |  |  |
| Shear | <br>Pred:Plane<br>Conf:0.993 |  |  |  |  |

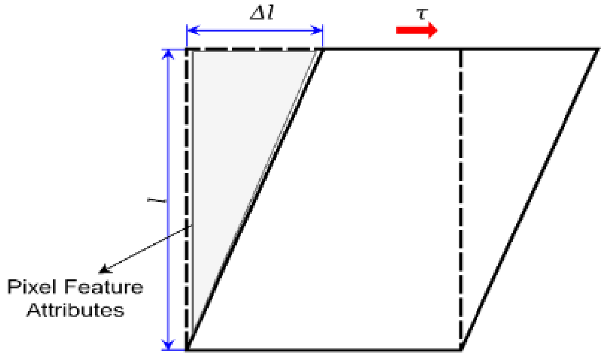


**Figure 1:** Feature shifting phenomenon under shearing transformation.

# 5. Discussion

This assessment was motivated by analyzing how the DL network captures and makes the decisions by capturing the pixel feature attributes under both human synthesized and naturally corrupted adversarial inputs prior to introducing a general defence approach. In summary, we understand that the performance degradation happens under these adversaries due to the networks capturing of the significant features dispersed everywhere in the image or limiting the capturing of the features.

Moreover, we conclude that the model has similar behaviors under the adversarial attack noises and physical noises. In addition, blur, scaling, and contrast transformation have quite similar behaviors to each other. Therefore we conclude, a correctly classified set of adversarial perturbations with an array of improved noise perturbations could be used to enhance the general resilience. In particular, training the given network along with Generative Adversarial Network (GAN) would be a promising research idea. To elaborate further on how the adversarial attack defence approaches and model robustification approaches improve resilience we analyze the performance of the adversarial training and image augmentation for adversarial attacks and physical distortions respectively.

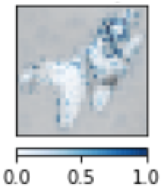
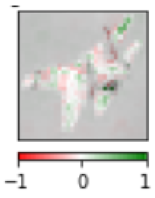
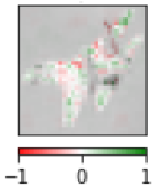
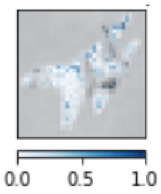
## 5.1. Impact of Adversarial Resilience Approaches

### 5.1.1. Adversarial Training

To perform the adversarial training we used PGD $l_\infty$ attack [36]. First, we have run several epochs with the clean samples and then gradually increased the epsilon $\epsilon$ value until $\epsilon = 0.01$ . Here the $\epsilon$ values are placed in an exponential series and all the other parameters are kept as constants. As the result, we got an average of 0.85 $R_N^\phi$ score for PGD $l_\infty$, FGSM and BIM attacks. Surprisingly we noticed adversarial training is able to improve the network's ability to capture the pixel features withinside the "Plane" object. Table 5 summarizes the pixel feature capturing of the adversarially trained model under the PGD $l_\infty$ adversarial samples and the same behavior was noticed under the other adversarial attack settings as well.

**Table 5**

Pixel feature capturing of the adversarially trained model under the PGD $l_\infty$ adversarial samples

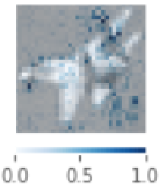
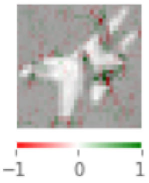
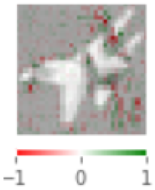
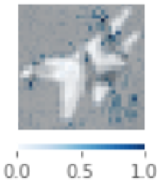
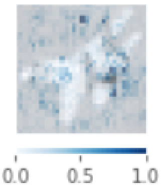
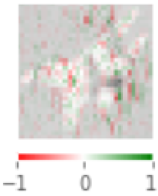
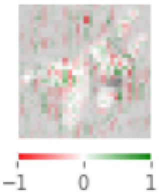
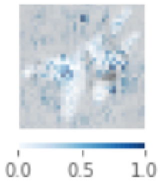
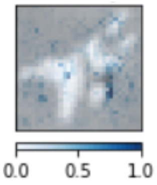
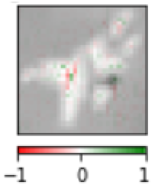
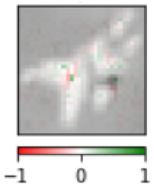
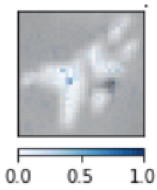
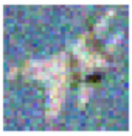
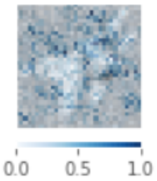
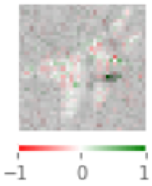
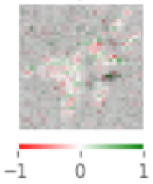
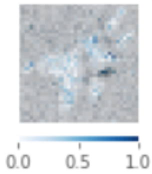
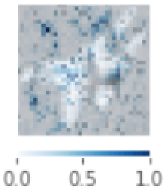
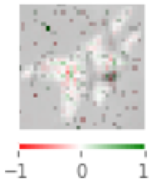
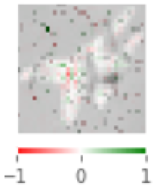
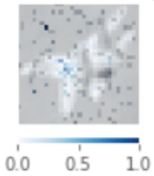
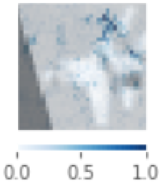
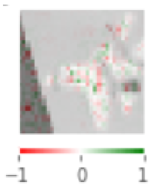
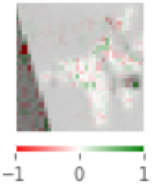
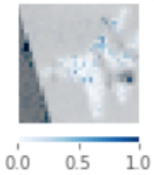
| Saliency | Integrated Gradients | DeepLift | GradientShap |
|---|---|---|---|
|  |  |  |  |

### 5.1.2. Input Data Transformations & Augmentation

To perform the input augmentation-based model re-training using the selected image transformation methods in section 3.3 a separate training set was augmented. Given priority to that as did by Laugros et al. [11] an array of severity levels of each transformation (except-swirl) was used while augmenting the dataset. Thereafter, we trained the initial classification model using the new training set. Moreover, to avoid generalization issues, clean samples to the training set were included randomly. As shown in Table 6, we noticed there is a robustness increment for most of the physical corruptions. Even in the shear condition, the model is now able to capture the pixel features within the "Plane" object and in other corruptions instances also there is a relatively low feature dispersing of the network. In contrast under the noise and exposure perturbations, the feature dispersing phenomena is still active, but now the features captured withinside the "plane" object are increased, and due to that models' robustness is increased. Though the data augmentation approach improves the robustness for physical corruptions, we noticed that it will increase or keep the same vulnerability for adversarial attacks as shown by the literature [11, 19]. For PGD $l_\infty$ attack $R_N^\phi$ is decreased by approximately 2% ($R_N^\phi$ = 0.458).

Both model resilience approaches which were assessed, are able to decrease the feature dispersing phenomena or it could improve the number of features captured within the object. These results gained by the assessment clearly emphasize the relationships and important insights of human synthesized adversarial attacks and naturally corrupted physical adversaries. Based on the insights obtained from the performance analysis of the adversarial training and data augmentation/transformation based robustification approaches, it is possible to further affirm the notion that, rather than mixing images perturbed by adversarial attacks and physical world distortions, a properly optimized and well-classified set of both adversarial attack and natural perturbation based training after analyzing the aforementioned model's behavior could improve the models' resilience naturally. We hope this will help to avoid using any auxiliary classifiers or tools in the inference which causes increasing cost and computation power in the inference. Another phenomenon observed was that these adversarial training acts as a sort of semantic segmentation of the objects (Plane in the above example). This could alternatively be due to the fact that this study choose to run the experiments on the relatively low-resolution CIFAR-10 dataset (due to practical, temporal, and cost constraints of running adversarial training on a high-resolution image dataset), however theoretically when scaling the networks should be able to extrapolate the results.

**Table 6**
Pixel feature capturing of the re-trained model using image transformation based augmentation (Pred:Prediction, Conf:Confidence)

| Adversarial Setting | Original Image | Saliency | Integrated Gradients | DeepLift | GradientShap |
|---|---|---|---|---|---|
| Contrast | <br>Pred:Plane<br>Conf:0.99 |  |  |  |  |
| Exposure | <br>Pred:Plane<br>Conf:0.99 |  |  |  |  |
| Scaling | <br>Pred:Plane<br>Conf:0.99 |  |  |  |  |
| Gaussian Noise | <br>Pred:Ship<br>Conf:0.98 |  |  |  |  |
| Pepper Noise | <br>Pred:Plane<br>Conf:0.99 |  |  |  |  |
| Shear | <br>Pred:Plane<br>Conf:0.99 |  |  |  |  |

While conducting this assessment we have noticed these XAI approaches (Especially Saliency Method) are meticulously fragile to adversarial attacks and physical noise corruption when compared to other adversarial perturbations. Even for attacks with small $\epsilon$ values, the interpretability is given as the dispersed feature attributes. According to [39, 40] the high fragility of the XAI algorithms to adversarial attacks is a common and an open research problem in feature attribution-based model interpretability approaches. This issue has to be addressed broadly in future research as this has raised a question about the model interpretability-based adversarial detection approaches. In addition, as a promising future research direction, an adversarially perturbed image detection with a model restoration method or a neural network model adaptation method could try out for both adversarial types as a general adversarial defence mechanism.

## 6. Conclusion

This research study empirically showed the connection between adversarial attacks and physical world corruptions using four feature attribution-based XAI algorithms. In particular, we demonstrated how each XAI approach visualizes the captured pixel features by the network under adverse conditions, and based on those heatmaps we further elaborated on several useful insights about the patterns of the model's feature capturing ability. In conclusion, this study demonstrates potential future research directions on model adaptation and restoration methodologies, or an optimization form of noise perturbations to introduce a general adversarial defence method for human-crafted adversarial attacks and physical adversarial corruptions.

## Acknowledgments

The first author would like to thank Ms. Suresha Perera (Lecturer at the Open University Sri Lanka) for proofreading and guiding the revisions to this paper.

## References

[1] N. Sharma, R. Sharma, N. Jindal, Machine learning and deep learning applications-a vision, Global Transitions Proceedings 2 (2021) 24–28. URL: https://www.sciencedirect.com/science/article/pii/S2666285X21000042. doi:https://doi.org/10.1016/j.gltp.2021.01.004, 1st International Conference on Advances in Information, Computing and Trends in Data Engineering (AICDE - 2020).

[2] H. Jeremie, Effective altruism, ai safety, and learning human preferences from the world's state, Towards Data Science (2020). URL: https://is.gd/zPYOFZ.

[3] A. Kerasidou, Ethics of artificial intelligence in global health: Explainability, algorithmic bias and trust, Journal of Oral Biology and Craniofacial Research 11 (2021) 612–614. URL: https://www.sciencedirect.com/science/article/pii/S2212426821000920. doi:https://doi.org/10.1016/j.jobcr.2021.09.004.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, CoRR abs/1312.6199 (2014).

[5] D. Hendrycks, T. G. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, ArXiv abs/1903.12261 (2019).

[6] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, CoRR abs/1412.6572 (2015).

[7] M. Shu, Y. Shen, M. C. Lin, T. Goldstein, Adversarial differentiable data augmentation for autonomous systems, 2021 IEEE International Conference on Robotics and Automation (ICRA) (2021) 14069–14075.

[8] D. Liu, B. Cheng, Z. Wang, H. Zhang, T. S. Huang, Enhance visual recognition under adverse conditions via deep networks, IEEE Transactions on Image Processing 28 (2019) 4401–4412.

[9] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, M. Kim, An analysis of adversarial attacks and defenses on autonomous driving models, 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom) (2020) 1–10.

[10] Y. Deng, T. Zhang, G. Lou, X. Zheng, J. Jin, Q.-L. Han, Deep learning-based autonomous driving systems: A survey of attacks and defenses, IEEE Transactions on Industrial Informatics 17 (2021) 7897–7912.

[11] A. Laugros, A. Caplier, M. Ospici, Are adversarial robustness and common perturbation robustness independent attributes ?, 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) (2019) 1045–1054.

[12] A. Laugros, A. Caplier, M. Ospici, Addressing neural network robustness with mixup and targeted labeling adversarial training, in: ECCV Workshops, 2020.

[13] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, ArXiv abs/2006.00093 (2020).

[14] S. Chakraborty, R. J. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. B. Srivastava, A. D. Preece, S. J. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, P. K. Gurram, Interpretability of deep learning models: A survey of results, 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (2017) 1–6.

[15] P. Linardatos, V. Papastefanopoulos, S. B. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, Entropy 23 (2021).

[16] A. Krizhevsky, Learning multiple layers of features from tiny images, 2009.

[17] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (2019) 1–48.

[18] T. Bai, J. Luo, J. Zhao, B. Wen, Q. Wang, Recent advances in adversarial training for adversarial robustness, in: IJCAI, 2021.

[19] D. Kang, Y. Sun, D. Hendrycks, T. B. Brown, J. Steinhardt, Testing robustness against unforeseen adversaries, ArXiv abs/1908.08016 (2019).

[20] L. Zhang, M. Yu, T. Chen, Z. Shi, C. Bao, K. Ma, Auxiliary training: Towards accurate and robust models, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 369–378.

[21] D. A. Calian, F. Stimberg, O. Wiles, S.-A. Rebuffi, A. Gyorgy, T. A. Mann, S. Gowal, Defending against image corruptions through adversarial augmentations, ArXiv abs/2104.01086 (2021).

[22] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, Aug-

mix: A simple data processing method to improve robustness and uncertainty, ArXiv abs/1912.02781 (2020).

[23] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. L. Zhu, S. Parajuli, M. Guo, D. X. Song, J. Steinhardt, J. Gilmer, The many faces of robustness: A critical analysis of out-of-distribution generalization, ArXiv abs/2006.16241 (2020).

[24] H. Zhang, M. Cissé, Y. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, ArXiv abs/1710.09412 (2018).

[25] S. Park, J. So, On the effectiveness of adversarial training in defending against adversarial example attacks for image classification, Applied Sciences 10 (2020) 8079.

[26] C. Zhang, Z. Yang, Z. Ye, Detecting adversarial perturbations with saliency, 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP) (2018) 271–275.

[27] S. Wang, Y. Gong, Adversarial example detection based on saliency map features, Applied Intelligence (2021).

[28] E. Tjoa, C. Guan, Quantifying explainability of saliency methods in deep neural networks, ArXiv abs/2009.02899 (2020).

[29] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, CoRR abs/1312.6034 (2014).

[30] S. Z. S. Samuel, V. Kamakshi, N. Lodhi, N. C. Krishnan, Evaluation of saliency-based explainability method, ArXiv abs/2106.12773 (2021).

[31] K. K. Nakka, M. Salzmann, Towards robust fine-grained recognition by maximal separation of discriminative features, in: ACCV, 2020.

[32] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: ICML, 2017.

[33] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, ArXiv abs/1703.01365 (2017).

[34] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, ArXiv abs/1705.07874 (2017).

[35] A. Kurakin, I. J. Goodfellow, S. Bengio, Adversarial examples in the physical world, ArXiv abs/1607.02533 (2017).

[36] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, ArXiv abs/1706.06083 (2018).

[37] D. Temel, T. A. Alshawi, M.-H. Chen, G. Al-Regib, Challenging environments for traffic sign detection: Reliability assessment under inclement conditions, ArXiv abs/1902.06857 (2019).

[38] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: A unified and generic model interpretability library for pytorch, ArXiv abs/2009.07896 (2020).

[39] A. Ghorbani, A. Abid, J. Y. Zou, Interpretation of neural networks is fragile, in: AAAI, 2019.

[40] H. Rasaee, H. Rivaz, Explainable ai and susceptibility to adversarial attacks: a case study in classification of breast ultrasound images, 2021 IEEE International Ultrasonics Symposium (IUS) (2021) 1–4.