Exploring Robustness Connection between Artificial and Natural Adversarial Examples

Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh
University at Buffalo, USA and IIT Jodhpur, India
{aa298, nratha}@buffalo.edu, {mvatsa, richa}@iitj.ac.in

Abstract

Although recent deep neural network algorithm has shown tremendous success in several computer vision tasks, their vulnerability against minute adversarial perturbations has raised a serious concern. In the early days of crafting these adversarial examples, artificial noises are optimized through the network and added in the images to decrease the confidence of the classifiers against the true class. However, recent efforts are showcasing the presence of natural adversarial examples which can also be effectively used to fool the deep neural networks with high confidence. In this paper, for the first time, we have raised the question that whether there is any robustness connection between artificial and natural adversarial examples. The possible robustness connection between natural and artificial adversarial examples is studied in the form that whether an adversarial example detector trained on artificial examples can detect the natural adversarial examples. We have analyzed several deep neural networks for the possible detection of artificial and natural adversarial examples in seen and unseen settings to set up a robust connection. The extensive experimental results reveal several interesting insights to defend the deep classifiers whether vulnerable against natural or artificially perturbed examples. We believe these findings can pave a way for the development of unified resiliency because defense against one attack is not sufficient for real-world use cases.

1. Introduction

Deep learning has helped significantly in today's time for solving several computer vision tasks ranging from object recognition, object detection to person identification to reinforcement learning inspired autonomous tasks [23, 35, 50, 54]. However, the inception of minutely crafted adversarial perturbations and added in the images showcase that the deep neural networks are highly vulnerable [2, 24, 36, 38]. The interesting property of these perturbations is that they are imperceptible to the human examiners and hence can not be easily caught but can fool the deep neural networks with high confidence. The perturbed images are referred to as adversarial examples in the research community. Figure 1 shows one example of generating the adversarial example. In this case, an artificial perturbation randomly initialized is optimized using the classifier and added in the clean image. The final image is adversarial because it was mislabelled by the classifier into the wrong category. This randomly initialized perturbation can be optimized over each image or a single perturbation can be learned to fool the classifier on multiple images [39].

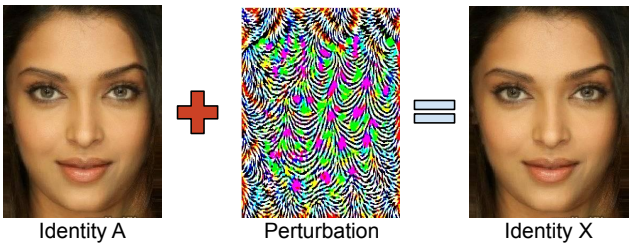

Figure 1. Process of generating adversarial examples through artificial perturbation. The left image is the clean image correctly classified into identity 'A'. When the artificial perturbation (middle) optimized using the classifier added in the clean image, we got the adversarial image (right). In this case, it got misclassified into the identity 'X'.

However, it is argued that these artificial perturbations are hard to find in the real world; therefore, the finding of natural adversarial examples have recently gained attention [25, 28, 37]. The natural adversary can be defined in the form of the presence of artifacts present in the face images such as random lines that might be added in the unconstrained environments [26] or the motion blur is a concern while capturing the images [28]. Similarly, changing the semantic information of an image while keeping the original information of an image intact can also help in developing the adversarial examples [30]. Interestingly, these type of adversarial examples shows that there is no need of adding any random noise optimized or not in an image to make it

an adversarial image. Agarwal et al. [2] have conducted an interesting study regarding the role of sign and direction in which these perturbations are optimized which make them perceptible or imperceptible.

As soon the vulnerability of deep neural networks is identified against the adversarial examples, research efforts are started to improve the robustness of the classifiers. The defense algorithms are broadly divided into three categories: (i) detection-based, (ii) mitigation or image transformation-based, and (iii) adversarial training or data augmentation-based. Out of these, adversarial examples detection and adversarial training received significant attention and have shown the potential of reducing the impact of adversarial perturbations. While some of the recent algorithms have shown tremendous success in defending the adversarial attacks; the majority of defenses are also proven ineffective [10, 12, 19]. Therefore, careful consideration is required while developing an effective defense algorithm. Another significant limitation of the existing defenses is that they are tackling only the adversarial examples containing artificial perturbations; leaving the natural adversaries free to attack. Therefore, in this research, for the first time, we have performed an experimental study to defend the deep neural network against both natural and artificial adversaries. Through the detection algorithm, we have tried to set up a possible connection between both eras of adversaries and analyze what will happen if the detection is trained on one era of adversaries and tested on another. For that purpose, we have to build a binary classification architecture utilizing several popular CNN architectures as a backbone. The results reveal a possible connection due to which decent detection performance is observed even when the different adversaries are unseen at the time of training. In brief, the contributions of this research are:

- A novel natural and artificial adversaries dataset is prepared to study a robustness connection between these two different eras of adversarial examples;
- Extensive experimental studies are conducted in seen and unseen adversarial examples algorithms to analyze how easy or difficult is the detection of a particular adversary if it is not seen at the time of training but came for evaluation.

2. Related Work

In this section, a brief overview of the existing adversarial examples generation algorithms along with the defense algorithms developed to counter them are provided. In 2014, Goodfellow et al. [24] have proposed the simplest adversarial examples generation algorithm by perturbing the image gradient into the image itself. The perturbation is applied once in the image and aims to increase the loss of a classifier. Later Kurakin et al. [36] have proposed the multi-step variant of the algorithm to increase the strength of the attack and control the perceptibility of the perturbation. Later, several research efforts are started to improve the strength of the adversarial attacks, and hence, newer attacks came into the picture. To name a few complex and effective attacks are: (i) PGD [38], (ii) DeepFool [40], (iii) Universal perturbations [11], and frequency-based perturbations [9, 17]. Interestingly, the majority of the attacks are performed in the white-box setting which requires complete access to the target model and hence witnesses poor transferability against the unseen classification models. To address those limitations, recently several research efforts are started to develop the transferable adversarial perturbations [33, 52]. Based on the assertion that the decision of deep classifiers is based on a few important regions of an image, Gao et al. [18] have proposed a push and pull technique to push the informative regions of true class close to incorrect class and pull the features of wrong class close to true class. A similar understanding is also used by Wang et al. [51] and perturb the important image features to develop transferable adversarial examples. Contrary to these artificial perturbation-based adversaries, recent efforts are also focused on finding the natural adversarial examples [2, 42, 53]. Hendrycks et al. [29] have selected the images of 200 classes that are misclassified by the ResNet50 classifier and termed them as natural adversarial examples. However, the authors claim that these adversarial examples are transferable across multiple classifiers. Later, Li et al. [37] have improved the adversarial examples proposed in [29] to strengthen the naturalness of the adversarial examples.

A similar advancement in the research efforts is also seen towards developing the defense algorithms to counter these adversarial examples. However, so far the defense algorithms are tackling the artificial adversarial examples. The possible reason might be that natural adversarial examples are recently highlighted in the research community and do not have a benchmark dataset covering both natural and artificial adversarial examples. One of the popular and effective defenses against the adversarial examples is to segregate them from the clean examples by detecting them. Towards that several binary classification algorithms ranging from simple binary classifier [3] to sophisticated deep learning-based algorithms are presented [1, 4]. These sophisticated algorithms have shown tremendous success in terms of generalizability by detecting the unseen adversarial examples [20–22, 27, 41]. Another popular and effective defense against is the training of classifiers either using the adversarial images itself or through the augmentation of images of different variations [5, 6, 8, 13, 45, 46]. While the adversarial training found the effective defense, its computational complexity and blind spots against unseen attacks are a major challenge [15, 43, 49, 55]. We refer the readers to the following survey papers to gain comprehensive insights

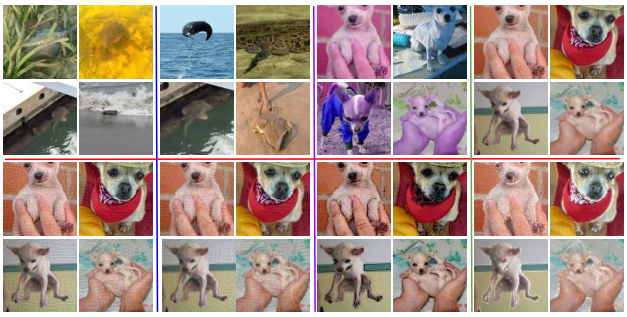

Figure 2. Samples of the different eras of adversarial examples that are considered in this research. Visually the adversarial examples look close to each other however perturb different features of the images. Each 2×2 image belongs to one adversarial examples algorithm. From the left of first/second rows: IN-A, IN-A+, DeepFool, and FGSM adversarial examples. From the left of third/fourth rows: PGD, NI+, NI-, and HSV adversarial examples.

about the adversarial examples research [7, 34, 44, 48].

3. Natural and Artificial Adversarial Dataset

In this research, we have presented a unique dataset containing both natural and artificial adversarial examples. The adversarial examples developed as part of this research are broadly divided into three groups: (i) artificial adversarial perturbations, (ii) natural adversarial examples, and (iii) semantic adversaries.

To develop the adversarial examples containing artificial adversarial perturbations, three state-of-the-art and challenging algorithms namely FGSM [24], PGD [38], and DeepFool [40]. FGSM works on the manipulation of images using the gradient information computed over an image. Mathematically, it can be described as follows:

$$X^* = X + \eta \cdot sign(\nabla_X J(X, Y_{true}))$$

where, X and X^* represent the clean and FGSM adversarial image, respectively. η controls the strength of added perturbation optimized through the loss function J computed over image X and its associate true label Y_{true}. ∇_X is the gradient concerning X and $sign$ represents the sign function. The attack is applied once per image and leaves the perceptible modification visible to the naked eye. Madry et al. [38] have proposed one of the strongest first-order universal adversaries. The optimization used in PGD iteratively searches for a perturbation vector that minimizes a l_p norm ball around the clean image. DeepFool (DF) is an optimization of attack which assumes that the deep networks behave linearly and tries to project the data onto the separating decision hyperplane to make sure it gets misclassified by the classifier. It iteratively perturbs the input image until it jumps the decision boundary and gets misclassified. For both FGSM and PGD attacks we have minimized the l_∞ norm of the perturbation whereas DeepFool works on the minimization of the l_2 norm.

In the case of natural adversarial examples, the images proposed by Hendrycks et al. [29] in the dataset namely ImageNet-A (IN-A) and Li et al. [37] in their dataset namely Imagenet-A-Plus (IN-A+) are selected. Hendrycks et al. [29] have downloaded the natural images of 200 classes from multiple images hosting websites that can fool the ResNet-50 classifier. In total, the dataset contains $7,500$ natural adversarial images. Li et al. [37] have improved these adversarial images by reducing the cluttered background in the images and cropping the image portion so that the foreground region takes a sufficient part in the images.

The final category referred to semantic adversaries contains the adversarial examples generated using the algorithms proposed by Hosseini and Poovendran [30] and Agarwal et al. [2]. The authors in [30] perturb the color components of images while keeping the value component untouched. For that purpose, the RGB images are converted into the HSV color space (we referred to the attack as HSV attack) and later, H and S components are modified iteratively using a scalar value until the network misclassified the images. While the attack algorithm is close to the human visual system but it utilizes the network information while perturbing the images. Whereas, the attack algorithm proposed by Agarwal et al. [2] utilizes the noise information naturally inherited (NI) in the images during their time of acquisition. The authors have extracted those noise patterns using several image filtering algorithms such as Gaussian, Laplacian, and Integral. The filtered image is subtracted from the clean image to get the noise pattern. Later this noise pattern is added (NI+) or subtracted (NI-) from the images to generate the adversarial examples. It leads to two completely different styles of adversarial images.

To generate the artificial perturbation-based adversarial images and semantic adversarial examples, we have first randomly selected the 3000 images from the validation set of the ImageNet dataset [16]. Later, an equal number of adversarial examples are generated from each algorithm belonging to an artificial adversary and semantic adversary. For artificial and HSV attacks, we have used the VGG architecture; however, similar detection analysis is observed on the images generated using MobileNet and DenseNet. The 3000 natural adversarial examples both from ImageNet-A and ImageNet-A-Plus datasets are directly taken without doing any modification. In total, the dataset contains $24,000$ adversarial images belonging to 8 different adversarial algorithms and $3,000$ clean images. We will release the dataset at a later stage after proper formatting of the directory structure to enhance the research in this direction. Figure 2 shows the adversarial images corresponding to different attack techniques used in this research.

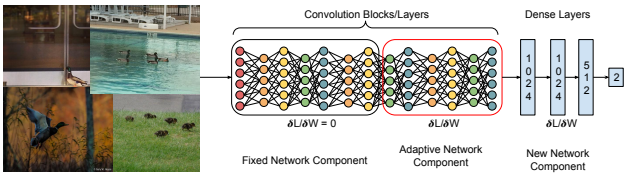

Figure 3. A broad view of the proposed adversarial example detection network. A binary classification decision boundary is learned using the images of both clean and adversarial classes.

4. Adversarial Detection Network for Robustness Connection Study

We want to highlight that the aim of this research is not to solve the adversarial examples detection; while the aim is to find whether there is any possible connection between different eras of adversaries. Henceforth, in this research, we have used the simple convolutional neural network (CNN) network for binary classification of the images into a clean and adversarial class. The schematic diagram of the proposed architecture is shown in Figure 3. The proposed architecture consists of three parts: (i) input images, (ii) convolutional blocks, and (iii) dense layers. To learn the parameters of the network, images of both clean and adversarial classes are fed into the network and the decision hyperplane learns to segregate the clean images from the perturbed images. The convolutional blocks are divided into two parts where one part is kept fixed, i.e., the gradient in these layers is assigned to 0. The second part of the network is adaptively updated using the batches of input images. The weights of the convolutional blocks are pre-trained on the ImageNet dataset. In the end, a few dense layers are added which are randomly initialized to learn the compact feature representation of both classes. The first two layers contain 1024 neurons and the following layer contains 512 neurons. The binary classification network is trained using binary class cross-entropy loss and the parameters are optimized using the 'Adam' optimization algorithm. The network is trained for 30 epochs using the batch size of images equal to 32. The initial learning rate is set to $1e^{-4}$ which is adaptively updated. to extensively study the connection and to avoid any classifier bias, we have used several CNN architectures to perform the adversarial examples detection. The used architectures varied in terms of number of layers, connection between layers, and their formation and are as follows: (i) VGG16 [47], (ii) DenseNet121 [32], (iii) MobileNet [31], (iv) InceptionV3 [37], and (v) Xception [14]. In different networks, a different number of layers are kept fixed and are optimized in a sense to keep the computational complexity low while maintaining high detection accuracy. For instance, in VGG16, the first 10 layers are kept fixed; whereas, in DenseNet121, wights of 110 layers are fixed.

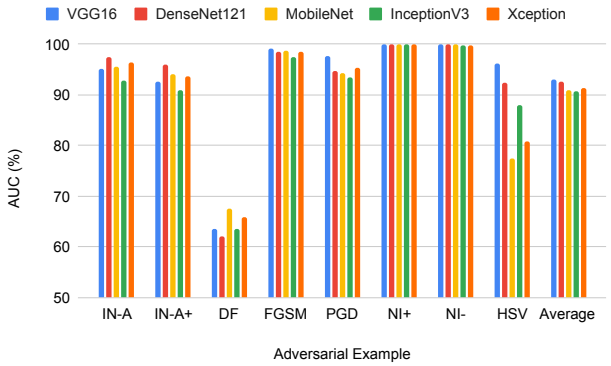

Figure 4. Adversarial examples detection (AUC %) where the adversarial examples algorithm used in training is also used for evaluation. NI adversarial examples are found approximately perfectly detectable in seen training-testing scenarios. Average represents the average performance of each network across the 8 different adversaries.

5. Experimental Results and Analysis

In this experimental results are described using the proposed adversarial examples detection algorithm developed using each CNN architecture. Images of each class are divided into two parts where the first half is used for training and the remaining half is used for testing. For example, the real class contains 3000 images, where the first 1500 images are used for training and the remaining 1500 images are used for evaluation. The experimental analysis can be described into two scenarios: (i) seen adversarial generation training-testing and (ii) unseen adversarial examples whether coming from the same broad category such as IN-A vs. IN-A+ of a broad natural adversary or different broad category such as Natural vs. Semantic adversarial examples. The experimental results are reported in terms of area under the ROC curve (AUC%) or otherwise specified.

5.1. Seen Adversary Detection

The results of seen adversarial examples training-testing experiment are reported in Figure 4. The analysis can be broken based on the category of the adversarial examples it belongs to. Among the natural adversarial examples, across each network used, the ImageNet-A examples yield better detection performance as compared to ImageNet-A+. The reason can be understood from the motivation of the development of the dataset by the authors who reduce the unnecessary background and object region features. The difference in accuracy lies in the range of 1.5% to 2.5%. Among all the adversarial examples category including the artificial perturbation category, DeepFool (DF) attack is found most challenging to detect. Whereas, the FGSM perturbation is found the easiest adversary to be defended among the artificial perturbations. In other words, the gradient-

Table 1. Unseen adversarial type detection performance in terms of AUC (%). It is observed that the detector architecture yields high performance when the training and testing adversarial examples belong to the same broad category. However, the performance drops are observed when the different category of adversarial examples comes for testing. – represents the seen training-testing scenario and hence removed to avoid any confusion. The seen scenario results are reported in Figure 4.

CNN	Type	Train ↓	Natural		Artificial			Semantic			Average
		Test →	IN-A	IN-A+	DF	FGSM	PGD	NI+	NI-	HSV	
VGG16	Natural	IN-A	–	93.66	55.24	57.14	52.17	70.74	61.5	59.63	64.30
		IN-A+	95.51	–	54.66	54.75	51.92	67.22	61.78	55.87	63.10
	Artificial	DF	62.66	58.98	–	95.16	83.97	79.19	73.34	52.09	72.20
		FGSM	59.44	52.25	60.33	–	88.1	70.98	70.51	50.94	64.65
		PGD	55.48	45.25	61.24	98.87	–	79.31	79.87	66.97	69.57
	Semantic	NI+	70.14	62.4	56.7	74.97	68.37	–	70.44	62.84	66.55
		NI-	72.24	64.38	55.64	69.14	66.22	62.63	–	60.71	64.42
		HSV	71.33	69.87	53.53	58.48	52.58	56.52	49.53	–	58.83
DenseNet121	Natural	IN-A	–	96.88	54.75	53.13	53.04	73.02	64.38	64.99	65.74
		IN-A+	97.71	–	53.7	53.52	52.71	73.05	62.23	57.9	64.40
	Artificial	DF	71.18	68.6	–	95.05	84.06	74.15	51.69	68.14	73.27
		FGSM	66.45	57.06	60.57	–	92.59	79.59	53.06	78.21	69.65
		PGD	69.65	62.5	59.68	97.26	–	83.33	72.08	84.78	75.61
	Semantic	NI+	72.65	66.03	53.35	61.8	62.33	–	91.56	67.24	67.85
		NI-	78.42	74.19	50.72	51.17	57.24	94.01	–	68.02	67.68
		HSV	51.15	55.61	52.19	50.53	48.72	47.01	57.45	–	51.81
MobileNet	Natural	IN-A	–	93.99	59.9	63.84	56.1	65.59	72.84	47.36	65.66
		IN-A+	96.09	–	56.8	64.53	55.79	71.4	72.59	46.9	66.30
	Artificial	DF	55.31	53.75	–	97.27	92.09	47.5	43.12	45.45	62.07
		FGSM	73.04	65.17	60.49	–	91.22	82.18	76.94	57.67	72.39
		PGD	71.64	71.21	90.46	97.83	–	79.12	76.97	64.42	78.81
	Semantic	NI+	88.77	81.99	54.66	68.19	60.12	–	87.08	55.1	70.84
		NI-	82.29	76.95	53.21	60.21	60.06	86.16	–	63.79	68.95
		HSV	31.66	31.68	52.17	70.12	60.86	59.38	58.76	–	52.09
InceptionV3	Natural	IN-A	–	90.78	55.87	61.54	57.62	69.67	68.62	58.04	66.02
		IN-A+	93.01	–	56.19	63.07	58.57	73.89	72	59.66	68.06
	Artificial	DF	57.7	54.53	–	95.65	88	68.91	55.25	48.34	66.91
		FGSM	73.78	65.71	59.07	–	85.78	79.14	71.17	66.48	71.59
		PGD	71.95	63	60.02	96.83	–	81.7	75.79	70.57	74.27
	Semantic	NI+	67.58	58.71	55.16	71.95	63.87	–	62.56	48.81	61.23
		NI-	76.87	66.75	55.71	72.73	68.01	82.62	–	58.14	68.69
		HSV	62.22	58.48	54.12	63.35	55.73	65.24	50.66	–	58.54
Xception	Natural	IN-A	–	96.03	55.26	58.91	51.88	69.58	63.77	49.83	63.61
		IN-A+	95.56	–	55.18	59.72	55.85	71.8	70.66	49.01	65.40
	Artificial	DF	64.73	61.9	–	98.14	92.23	59.75	58.36	39.91	67.86
		FGSM	75.94	68.66	59.93	–	88.05	83.8	77.46	62.55	73.77
		PGD	75.73	67.39	60.08	98.6	–	86.64	80.95	68.74	76.88
	Semantic	NI+	74.21	69.33	55.5	70.5	62.04	–	46.81	61.72	62.87
		NI-	71.05	68.72	54.85	69.72	64.84	40.2	–	52.23	60.23
		HSV	42.09	44.44	50.58	55.19	49.47	47.39	45.81	–	47.85

based attacks (FGSM and PGD) yield higher detection performance as compared to the decision boundary-based optimization attack (DF). In terms of semantic adversarial examples, naturally inherited adversarial examples are found approximately perfectly identifiable; whereas, the HSV attack poses a slight challenge in its detection. The reason might be that the HSV attack does not touch the image feature component and only modifies the color components; whereas, the NI attacks perturb the frequency features of an image and significantly alter image distribution. The VGG backbone CNN architecture yields more than 96% detection performance of the HSV attack adversarial examples.

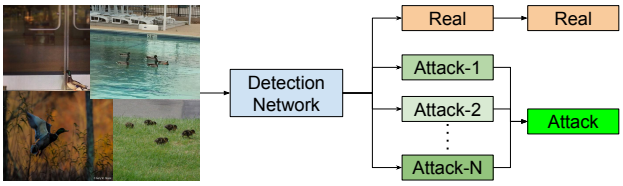

Figure 5. Adversarial examples detection architecture utilizing 'N' attack classes to act as auxiliary information. At the time of testing, if the image is classified in any of the attack categories is classified as adversarial else termed as real.

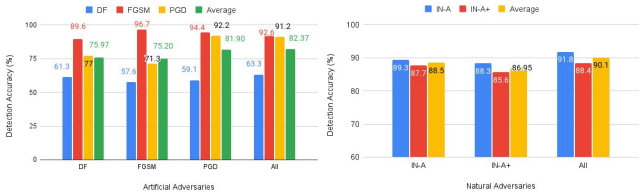

Figure 6. Adversarial examples detection accuracy (%) when individual and all (of same category only) adversaries of natural and artificial adversaries are used for training.

In terms of performance of the individual CNN architecture deployed for adversarial examples detection, VGG yields the best average detection AUC value of 93.02% followed by the DenseNet121 with an AUC value of 92.63%.

5.2. Unseen Adversary Detection

Compared to seen adversary performance, the performance under unseen adversary sees a significant drop as expected. The analysis can be broken down into two parts: (i) unseen adversary detection but belonging to the same broad category and (ii) unseen adversary coming unseen broad category as well. We are going to discuss the performance obtained using VGG; however, similar observation has been observed across different networks. In unseen adversaries belonging to the same category, the detection networks are found robust but the vulnerability of the detection networks increases as soon as the adversaries from the different categories occur. For instance, when the VGG network is trained on IN-A and tested on IN-A+, it yields 93.66% detection AUC which is even better when the IN-A+ is trained and tested in the seen setting. A similar higher generalization is observed when the IN-A+ trained detector is evaluated on the IN-A images. It reflects the possible connection between both the natural adversaries. Interestingly, natural adversaries are found close to semantic adversarial as compared to artificial adversaries. The phenomena can also be observed in the reverse case, where the adversarial examples detection network trained on the semantic adversaries yields better performance on natural adversaries as compared to the artificial adversaries. It can be thought from the fact that both these adversaries utilize some form of natural statistics to find the adversarial examples. Among the natural adversaries, the IN-A+ adversarial images are hard to detect as compared to IN-A as also observed in the seen adversary training-testing setting. In terms of artificial perturbations, the detectors trained on them show higher detection performance on semantic adversaries in comparison to natural adversarial images. The above-discussed findings are reported in Table 1.

5.3. N-Way Attack Supervision

We assert that in place of training the classifier individually on each adversarial attack algorithm, can we use all adversaries of maybe the same category to learn '$N+1$' class classification architecture. Where N belongs to the number of attack classes, for example, natural adversaries have two classes and $+1$ stands for real class. Hence as shown in Figure 5, in the case of all-natural adversaries, 3 class classification architecture is trained. The motivation is in place of wasting the attack information can use that extra information as auxiliary information to enhance the detection performance. The results reported in Figure 6 reflect that such intuition is beneficial where the average attack detection performance shows improvement. For instance, when all-natural adversaries are used for training, the average detection accuracy on the natural adversarial examples is 90.1% which is 1.6% and 3.15% better when the detector is trained on IN-A and IN-A+ images only, respectively. A similar observation has been noticed in other categories of the adversarial examples described in this research. For instance, in the case of artificial adversaries, when all attack class images are used for training, it yields at least 0.47% and up to 7.17% higher average detection accuracy.

6. Conclusion

Artificial adversarial examples have shown tremendous stealthy nature in fooling almost 'any' deep learning classifier whether CNN, reinforcement learning, or vision transformer. On top of that, the recent surge in the development of natural adversarial examples has complicated the robustness research. In this research, for the first time, we have explored a possible robustness connection between the natural and artificial adversarial examples. For that, we have developed the adversarial examples detection network and performed an experimental analysis using several CNN architectures. Through the extensive experiments, we found that some adversarial attacks are highly simple to detect, and on the other hand few are hard to defend even in seen training-testing settings. Our study also reveals that natural and semantic adversaries share characteristics that make the detection algorithm trained on one generalized on another. In the future, we plan to extend the proposed study both towards expanding the dataset and developing a sophisticated algorithm to enhance the detection performance.

References

[1] Akshay Agarwal, Gaurav Goswami, Mayank Vatsa, Richa Singh, and Nalini K. Ratha. Damad: Database, attack, and model agnostic adversarial perturbation detector. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2021. 2

[2] Akshay Agarwal, Richa Singh, and Mayank Vatsa. The role of 'sign' and 'direction' of gradient on the performance of cnn. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2748, 2754, June 2020. 1, 2, 3

[3] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2018. 2

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

[5] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Cognitive data augmentation for adversarial defense via pixel masking. *Pattern Recognition Letters*, 146:244–251, 2021. 2

[6] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Intelligent and adaptive mixup technique for adversarial robustness. In *IEEE International Conference on Image Processing*, pages 824–828, 2021. 2

[7] Ahmed Aldahdooh, Wassim Hamidouche, Sid Ahmed Fezza, and Olivier Déforges. Adversarial example detection for dnn models: a review and experimental comparison. *Artificial Intelligence Review*, pages 1–60, 2022. 3

[8] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020. 2

[9] Divyam Anshumaan, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Wavetransform: Crafting adversarial examples via input decomposition. In *European Conference on Computer Vision*, pages 152–168, 2020. 2

[10] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018. 2

[11] Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *Asian Conference on Computer Vision*, 2020. 2

[12] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM workshop on Artificial Intelligence and Security*, pages 3–14, 2017. 2

[13] Saheb Chhabra, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Attack agnostic adversarial defense via visual imperceptible bound. In *IEEE International Conference on Pattern Recognition*, pages 5302–5309, 2021. 2

[14] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 4

[15] Sanghyuk Chun, Seong Joon Oh, Sangdoo Yun, Dongyoon Han, Junsuk Choe, and Youngjoon Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. *arXiv preprint arXiv:2003.03879*, 2020. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[17] Salah Ud Din, Naveed Akhtar, Shahzad Younis, Faisal Shafait, Atif Mansoor, and Muhammad Shafique. Steganographic universal adversarial perturbations. *Pattern Recognition Letters*, 135:146–152, 2020. 2

[18] Lianli Gao, Zijie Huang, Jingkuan Song, Yang Yang, and Heng Tao Shen. Push & pull: Transferable adversarial examples with attentive attack. *IEEE Transactions on Multimedia*, 2021. 2

[19] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. *arXiv preprint arXiv:2003.08937*, 2020. 2

[20] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Deepring: Protecting deep neural network with blockchain. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 2019. 2

[21] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini K. Ratha. Dndnet: Reconfiguring cnn for adversarial robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 2020. 2

[22] Akhil Goel, Anirudh Singh, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2018. 2

[23] Adrian Goldwaser and Michael Thielscher. Deep reinforcement learning for general game playing. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 1701–1708, 2020. 1

[24] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2, 3

[25] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *International Journal of Computer Vision*, 127(6):719–742, 2019. 1

[26] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1

[27] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 2

[28] Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Bing Yu, Wei Feng, and Yang Liu. Watch out! motion is blurring the vision of your deep neural networks. *Advances in*

Neural Information Processing Systems, 33:975–985, 2020. 1

[29] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 2, 3

[30] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018. 1, 3

[31] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 4

[32] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 4

[33] Tianjin Huang, Vlado Menkovski, Yulong Pei, YuHao Wang, and Mykola Pechenizkiy. Direction-aggregated attack for transferable adversarial examples. *arXiv preprint arXiv:2104.09172*, 2021. 2

[34] Wei Jiang, Zhiyuan He, Jinyu Zhan, Weijia Pan, and Deepak Adhikari. Research progress and challenges on application-driven adversarial examples: A survey. *ACM Transactions on Cyber-Physical Systems*, 5(4):1–25, 2021. 3

[35] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1

[36] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018. 1, 2

[37] Xiao Li, Jianmin Li, Ting Dai, Jie Shi, Jun Zhu, and Xiaolin Hu. Rethinking natural adversarial examples for classification models. *arXiv preprint arXiv:2102.11731*, 2021. 1, 2, 3, 4

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 3

[39] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017. 1

[40] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. 2, 3

[41] Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2019. 2

[42] Anibal Pedraza, Oscar Deniz, and Gloria Bueno. Really natural adversarial examples. *International Journal of Machine Learning and Cybernetics*, pages 1–13, 2021. 2

[43] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019. 2

[44] Alex Serban, Erik Poll, and Joost Visser. Adversarial examples on object recognition: A comprehensive survey. *ACM Computing Surveys*, 53(3):1–38, 2020. 3

[45] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019. 2

[46] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 5636–5643, 2020. 2

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[48] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13583–13589, 2020. 3

[49] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Asymptotic behavior of adversarial training in binary classification. *arXiv preprint arXiv:2010.13275*, 2020. 2

[50] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018. 1

[51] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *IEEE/CVF International Conference on Computer Vision*, pages 7639–7648, 2021. 2

[52] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. *arXiv preprint arXiv:2109.04176*, 2021. 2

[53] Mingfu Xue, Chengxiang Yuan, Can He, Jian Wang, and Weiqiang Liu. Naturalae: Natural and robust physical adversarial examples for object detectors. *Journal of Information Security and Applications*, 57:102694, 2021. 2

[54] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[55] Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. *arXiv preprint arXiv:1901.04684*, 2019. 2