Noise is Inside Me! Generating Adversarial Perturbations with Noise Derived from Natural Filters

Akshay Agarwal[1], Mayank Vatsa[2], Richa Singh[2], and Nalini K. Ratha[3]
[1]IIIT-Delhi, India; [2]IIT Jodhpur, India; [3]IBM TJ Watson Research Center, USA
[1]akshaya@iiitd.ac.in; [2]{richa, mvatsa}@iitj.ac.in; [3]ratha@us.ibm.com

Abstract

Deep learning solutions are vulnerable to adversarial perturbations and can lead a "frog" image to be misclassified as a "deer" or random pattern into "guitar". Adversarial attack generation algorithms generally utilize the knowledge of database and CNN model to craft the noise. In this research, we present a novel scheme termed as Camera Inspired Perturbations to generate adversarial noise. The proposed approach relies on the noise embedded in the image due to environmental factors or camera noise incorporated. We extract these noise patterns using image filtering algorithms and incorporate them into images to generate adversarial images. Unlike most of the existing algorithms that require learning of noise, the proposed adversarial noise can be applied in real-time. It is model-agnostic and can be utilized to fool multiple deep learning classifiers on various databases. The effectiveness of the proposed approach is evaluated on five different databases with five different convolutional neural networks such as ResNet-50, VGG-16, and VGG-Face. The proposed attack reduces the classification accuracy of every network, for instance, the performance of VGG-16 on the Tiny ImageNet database is reduced by more than 33%. The robustness of the proposed adversarial noise is also evaluated against different adversarial defense algorithms.

1. Introduction

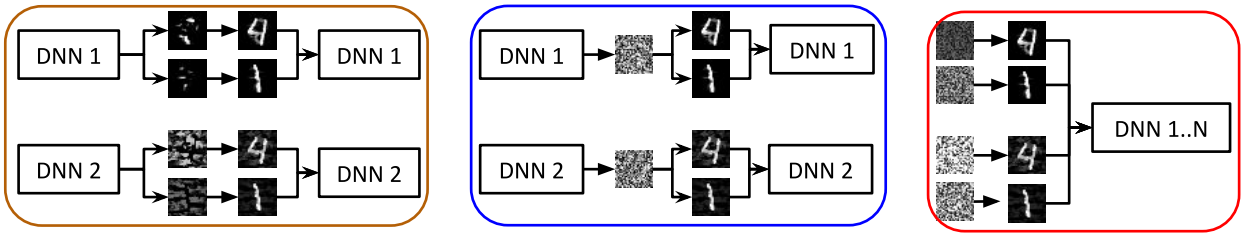
The vulnerability of convolutional neural networks (CNNs) against adversarial attacks raises several issues regarding the reliability of these networks. Some of the possible reasons for adversarial vulnerability are: (i) sharing of the spatial structure in input pixel domain between the weights of convolutional layer and (ii) high bias of the CNN models towards texture and shape of the input [12, 13]. A small modification in the input space might alter the spatial structure of the CNN filters which in turn gets escalated deeper in the networks. Goswami et al. [18, 19] have shown the filter responses of CNN models and observed that some neurons are sensitive towards hand-crafted adversaries such as random lines.

To attack a model, most of the existing attack generation algorithms (including gradient and optimization based attacks [17, 48]) require information about the Deep Neural Network (DNN) model in consideration, such as the parameters and gradient or logit layer information. Firstly, this information is difficult to achieve and secondly, this leads the adversaries to be specific to the models. Therefore, extending these attacks to multiple models is computationally expensive.

Inspired by these findings and the sensitivity of neural network models towards input pixels, this research proposes a novel method of generating the adversarial examples. In this research, we pose the question *whether the knowledge of only the input image can be used to create adversaries*. Image filtering operations such as Laplace and wavelet transform are useful in extracting the texture/shape related information such as edges and object structure. We utilize these operations to obtain the noise from the image itself for adversarial example generation. The proposed adversarial generation algorithm is a real-world attack algorithm where no knowledge of classifier is required to fool it. Figure 1 summarizes the difference in the concept of the proposed adversarial attack algorithm with most of the existing algorithms.

1.1. Related Work

Existing adversarial noise generation algorithms can be broadly grouped into two categories: (i) unique perturbation and (ii) generalized perturbation. Unique image perturbations are the ones where, for each image, a noise vector is learned to fool a CNN classifier. On the other hand, a generalized perturbation is a single noise vector that can be used on multiple images to fool the classifier. Szegedy et al. [48] proposed the first unique perturbation algorithm, where the authors generated the imperceptible noise using L-BFGS technique. Since then, several researches have proposed various unique perturbation algorithms such as Fast



(a) Existing DNN and Image specific Adversary

(b) Existing DNN and Database specific Adversary

(c) Proposed Image specific Adversary (DNN Agnostic)

Figure 1: Comparing the proposed DNN agnostic adversarial algorithm and existing adversarial approaches. The proposed algorithm does not require DNN model training for generating the adversary and same adversary can be used to fool multiple DNN models. First part (i.e., before images) shows the generation of adversarial noise from DNNs and second part (i.e., after images) shows the fooling of DNN models after noise manipulation.

Gradient Sign Method (FGSM) [17] and iterative method of FGSM [25]. Biggio et al. [6] proposed gradient based test time evasion attack. Papernot et al. [34] presented adversarial attack by modifying the most salient pixels in the image. Su et al. [47] have demonstrated that alteration of a single pixel can also fool the deep classification models into making wrong classification with high confidence. Carlini and Wagner [7] presented three different kinds of attacks with the minimization of l_0, l_∞, and l_2 norm of the loss function. Similar to this, Chen et al. [8] estimated the gradient from the targeted model for crafting an adversarial attack. Zhao et al. [55] proposed the optimization of zeroth order norm without leveraging the gradient of the network. Sharif et el. [44] proposed the adversarial generative networks to craft the noise. The computation of above mentioned adversarial attacks is based on a different norm of the loss function and is image specific. Mopuri et al. [33] and Moosavi-Dezfooli et al. [31] proposed a generalized perturbation vector to fool deep classifiers on *any* image based on a particular distribution. The model works for different images of a database but it is not database agnostic.

Other than the above-discussed attacks, Sabour et al. [42] have presented modifications in the internal layers of DNN as a potential adversary. The aim is to make the internal representation of adversarial images similar to the original images. Apart from these attacks, multiple adversarial defense algorithms are also presented in the literature [3, 14–16, 39, 46, 53].

1.2. Research Contributions

The proposed adversarial attack algorithm generates image specific noise pattern that is agnostic to DNN models and image database characteristics. In other words, the proposed algorithm is independent of the knowledge of the classification model being attacked and can generate multiple noise patterns for a single image of *'any'* database in real-time. The key contributions of this research are:

1. a novel class of real-time adversarial generation algorithm termed as *CiPer* (camera inspired perturbation) is proposed which is based on the inherent noise present during data acquisition. Figure 2 shows the original, noise, and adversarial examples created using the proposed algorithm on the CIFAR-10 database [24].

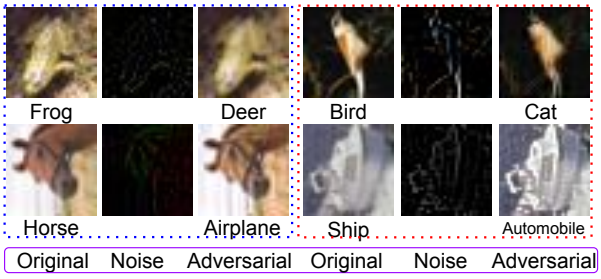


Figure 2: Adversarial images with their corresponding noise generated from the images using the proposed CiPer-S algorithm. Class labels of original and adversarial images are mentioned at the bottom of the respective images. The images are misclassified with atleast $\mathbf{85}\%$ confidence.

2. The proposed algorithm is generic in two aspects - it is database-agnostic and model-agnostic. The generalizability of CiPer is demonstrated on four databases including Tiny ImageNet[1], CIFAR-10 and CIFAR100, and five CNN models including ResNet-50 [22], VGG-16 [45] and VGG-Face [36].

3. The resiliency of CiPer is demonstrated against multiple state-of-the-art (SOTA) adversarial defense approaches: (i) adversarial training [25, 49], (ii) wavelet denoising based on Bayes thresholding, (iii) Bilateral filtering [38], (iv) Total Variance Minimization [41] [20], and (v) Defense GAN [43].

[1] https://tiny-imagenet.herokuapp.com/

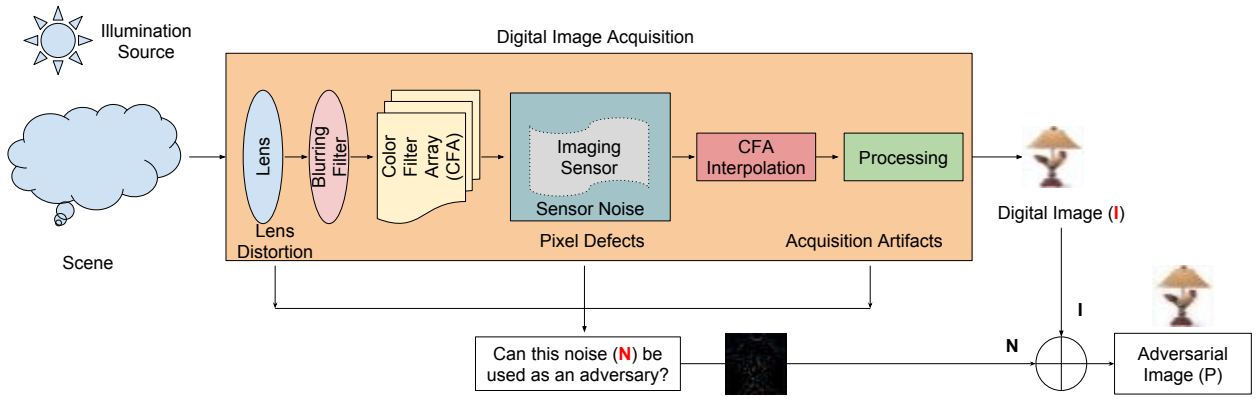


Figure 3: Illustrating the potential sources of adversarial noise that may be present in an image due to environment and sensor imperfections.

2. CiPer: Proposed Adversarial Attack

As shown in Figure 3, while capturing the image of a scene/object from a digital camera, it passes through a lens, a blurring filter, and a color filter array before being captured on an imaging sensor. The digital version of the scene/object is then processed through color interpolation, gamma correction, color correction, and white balance adjustment. Finally, the digital image is stored in the camera memory. This entire process introduces inevitably hidden noise/imperfections in images. These imperfections might be due to environmental noise or camera noise. Even if the image is captured in perfect conditions, there can be multiple sources of camera noise such as the photonic noise and sensor noise. Photonic noise is a random statistical noise and pattern noise is the deterministic component of the image pixels, which remains roughly the same for multiple images of the same scene. These sensors and pattern noises are effective enough and have been explored in the field of camera/sensor identification to identify the source from which a particular image is captured [1, 2, 27] and for Steganalysis [5]. In this research, we propose to extract this information and utilize for adversarial noise generation. This inherent noise is extracted using multiple image filtering/enhancement algorithms to generate adversarial examples to fool DNN models.

2.1. Generating Adversarial Example

The input image is first filtered using different kinds of filtering techniques. The difference image (N) is then computed as the absolute change between the clean image (I) and the filtered image (I'). The difference image can be considered as noise, which might be present due to the camera or environment, or can also be treated as high-frequency information such as edges and structure. The extracted noise (N) can either be added or subtracted from the clean image (I) to make it an adversarial image (P). Mathematically, the adversarial noise (N) can be represented as:

$$N = I - F(I) \tag{1}$$

where, F is a linear or non-linear filtering operation to be performed on input image I. For example, the Gaussian filter F can be defined as:

$$\begin{aligned} F_g(n_1, n_2) &= e^{-}(n_1^2, n_2^2)/2\sigma^2 \\ F(n_1, n_2) &= F_g(n_1, n_2)/\sum_{n1}\sum_{n_2} F_g \end{aligned} \tag{2}$$

where, n_1, n_2, and σ are the size and standard deviation of the filter. The adversarial examples can be generated by solving the following optimization problem:

$$P = I \oplus (\varphi \cdot N), \qquad s.t.\ P\ and\ I \in [0, 1] \tag{3}$$

where, $\oplus$ denotes the addition or subtraction operation. φ represents the strength parameter which controls the amount of noise to be added to make it imperceptible. The noise N might not be unique and can be generated through different image filtering operations. Let C be the deep learning classifier which outputs continuous probability values corresponding to each training class by minimizing the loss function. The aim is to fool this classifier so that $C(P) \neq l$, where l represents the true class of I.

In this paper, φ ranges from 0 to 1. Similar to existing algorithms such as L-BFGS and FGSM, the adversarial examples can be generated by line-searching the $\varphi > 0$ parameter. Due to the extraction of noise inherited during the time of image acquisition from the camera, the proposed adversary is termed as **CiPer** (**Camera Inspired Perturbations**). The proposed algorithm has two variants: **(i)** the algorithm with the addition operator is referred to as **CiPer-P** and **(ii)** with subtraction as **CiPer-S**. Addition is performed when the difference image is treated as noise inherent in images during acquisition while subtraction is performed to remove high-frequency information, i.e., edge or structure. One of the advantages of the proposed algorithm is the high

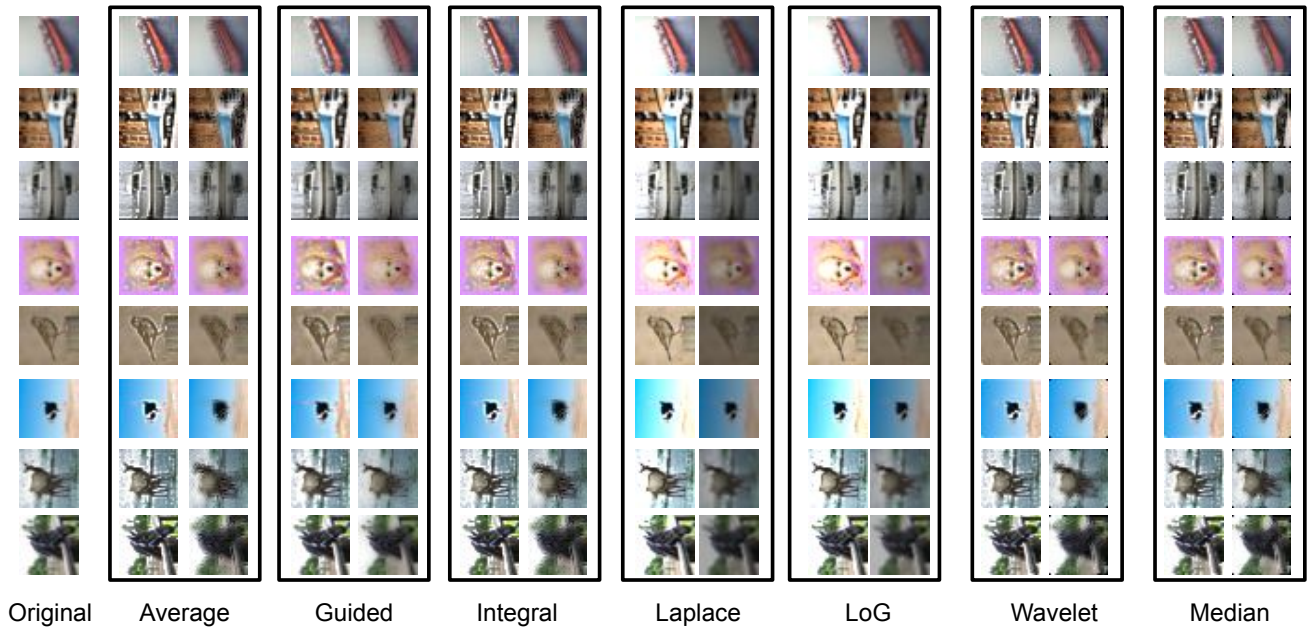


Figure 4: Sample adversarial examples generated using various filtering operations on CIFAR-10. The first column shows the original images. Thereafter, the first column of each block comprises the adversarial images generated via the proposed CiPer-P algorithm and the second column of each block demonstrates the adversarial images generated via the proposed CiPer-S algorithm.

efficiency in attacking black-box DNN models. Currently, the algorithm operates for untargeted attacks[2].

2.2. Image Filtering Techniques

In this research, various filtering algorithms ranging from spatial to wavelet domain are used to extract the adversarial noise (N). Each filter encodes different information and hence by using different filters, multiple noise vectors can be learnt for a single image. For instance, median filter helps in eliminating salt-and-pepper noise or impulsive noise. Extracting noise using median filter can generate adversary image with salt-and-pepper or impulse noise.

Laplace Filtering [35]: It is a derivative filter which is generally used to find the most frequently altered information in the image i.e., edges. The noise affects the high frequency information more as compared to the low frequency counterpart and hence filtering with Laplacian can help in better extracting the noise information.

LoG Filtering [35]: Derivative filters are sensitive to noise and thus lead to extraction of undesired information that might not be useful to generate the adversary. Therefore, the image is first smoothed using a Gaussian filter followed by applying the Laplacian filter.

$$LoG(x,y) = \frac{-1}{\pi\sigma^4}[1 - \frac{x^2+y^2}{2\sigma^2}]e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (4)$$

[2]A targeted attack is defined where the image is misclassified to one of the particular target class, whereas untargeted attack is defined as the classification of the image in the class other than its original label.

The filter F to generate the noise can be calculated as:

$$F_g(n_1,n_2) = e^{-}(n_1^2,n_2^2)/2\sigma^2$$
$$F(n_1,n_2) = (n_1^2+n_2^2-2\sigma^2)F_g(n_1,n_2)/\sigma^4 \sum_{n1}\sum_{n_2} F_g \quad (5)$$

The response of LoG is zero in the region of uniform intensities and returns +ve or -ve at regions with sharp changes, e.g. edges.

In addition to the filtering techniques discussed above, the following filtering algorithms are also used: **(i) Average:** pixel values in the center of each image patch (3×3) are replaced with the mean value of its neighborhood, **(ii) Guided [21]:** performs smoothing while preserving edge information. The algorithm does not suffer from gradient reversal artifacts and transfer structure information to the filtered output image, **(iii) Integral:** in integral image a pixel value is defined as the sum of the pixel values above and to the left of it. The advantage of integral filtering is the fast and non-uniform filtering in the image, **(iv) Wavelet transform:** provides four sub-bands: low-frequency and three high-frequency sub-bands in the horizontal, vertical, and diagonal directions, respectively. We have performed single-level decomposition followed by adaptive thresholding on each of the high-frequency components. The filtered wavelet sub-bands are reconstructed back and further enhanced using Gaussian and median filtering techniques,

Table 1: Number of original and adversarial test images of each database used to perform the experiments. n represents the number of filtering algorithms and 2 is for CiPer-P and CiPer-S.

Database	Original	Adversarial
CIFAR-10 [24]	10,000	10,000×(2n)
CIFAR-100 [24]	10,000	10,000×(2n)
Tiny ImageNet[3]	10,000	10,000×(2n)
MNIST [26]	10,000	10,000×(2n)
LFW [23]	13,143	13,143×(2n)

and **(v) Median:** it is one of the popular non-linear filtering techniques to reduce the impulse noise. The median value is expected to be robust to outliers.

Figure 4 shows the adversarial images generated after addition or subtraction of these noises using the proposed CiPer algorithm. The addition of the noise in the original image mostly altered the objects pixels as compared to the background. The phenomena can be easily verified while capturing the object from the camera; the noise affects the object pixels due to the reflection of objects. Similarly, the subtraction of noise shows the reduction in edge information.

3. Experimental Results and Analysis

To demonstrate the generalizability and robustness of the proposed algorithm, we have computed results with multiple models and databases. The next subsection summarizes the databases and algorithms along with the implementation details. This is followed by the results of the proposed CiPer-S and CiPer-P attacks with different filters. Finally, we demonstrate the resiliency of the proposed attacks to state-of-the-art defense algorithms.

3.1. Experimental Setup

Databases: Five databases are used for evaluation: MNIST [26], CIFAR-10 [24], CIFAR-100 [24], Tiny ImageNet[4], and Labeled Faces in the Wild (LFW) [23]. The results are demonstrated for three different classification tasks: digit classification, object recognition, and face attribute identification. Attribute classification experiments are performed with two attributes, pale skin and smile. The statistics of the number of testing and corresponding adversarial images are given in Table 1. CNN models are trained on the training set of each database used while evaluation is performed on the original testing and adversarial testing sets.

Implementation Details of CNN Models: To demonstrate the effect of the proposed CiPer adversarial algorithm and justify the adversarial impact, we have used three SOTA pre-trained CNN models: ResNet-50 [22], VGG-16 [45] and VGG-Face [36] fine-tuned for object recognition and face attribute identification, respectively. ResNet-50 model is trained from scratch with the learning rate and batch size set to 10^{-3} and 32 respectively. The model is trained for 10 epochs.

[4]https://tiny-imagenet.herokuapp.com/

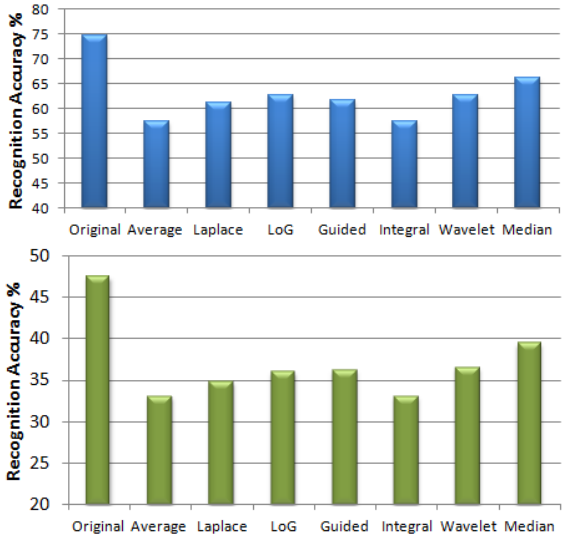


Figure 5: Results of VGG-16 [45] based object classification on the original and CiPer-S images generated from **CIFAR-10** (Top) and **CIFAR-100** (Bottom).

Apart from using SOTA CNN models we have trained two different CNN models: one shallow (#1) and one deep (#2). The shallow model contains 4 convolutional blocks, 3 dropout blocks, and 2 dense layers, whereas the deep model contains 13 convolutional blocks each followed by batch-normalization layer, 7 dropout blocks, and 2 dense layers. The models are trained with cross-entropy loss using the Adam optimizer. The learning rate and weight decay are fixed to 0.0001 and $1e^{-6}$. For training, we use 200 epochs for CIFAR-100 and 100 epochs for CIFAR-10 with a batch size of 32. CNN models are trained on the clean (i.e., unperturbed) training set of each database and evaluated on both clean and adversarial testing sets.

3.2. Results of CiPer Attack

The results are presented according to the classification task.

Object Classification: On the original images from CIFAR-10 and CIFAR-100 databases, VGG-16 [45] yields classification accuracy of 74.8% and 47.4%, respectively. It is observed that the adversarial examples based on Laplace and Integral filters, which can extract the texture and edge information, are the most effective in fooling the CNNs. For example, the integral filtering based CiPer-S attack reduces the accuracy of VGG-16 by 23.3% and 30.4% on adversarial test set compared to clean test set of CIFAR-10 and CIFAR-100, respectively. The results of each filter on CI-

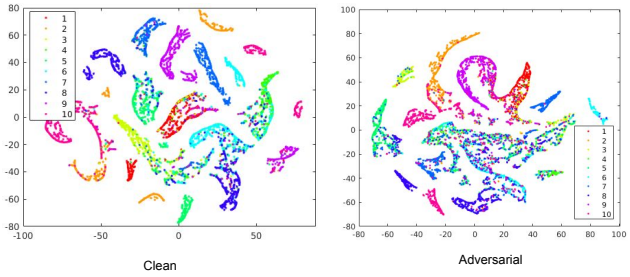


Figure 6: Score Distribution of clean and CiPer-S adversarial images of **CIFAR-10** database using VGG-16.[2] Due to the proposed attack overlap between the scores of different classes increased which leads to misclassification in the network.

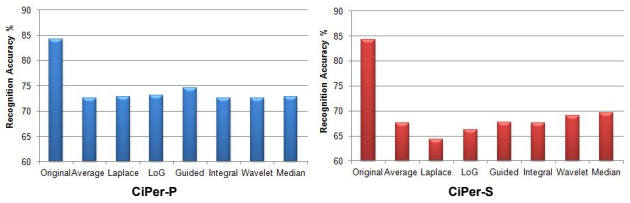


Figure 7: Results of ResNet-50 [22] based object classification on the CiPer-P and CiPer-S images generated from **CIFAR-10**.

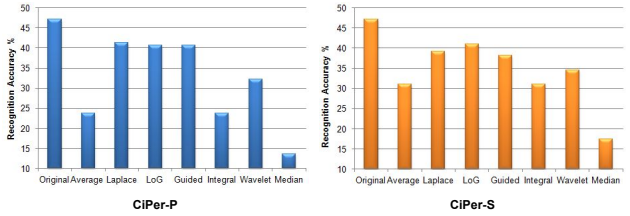


Figure 8: Results of VGG-16 [45] based object classification on the CiPer-P (left) and CiPer-S (right) adversarial images generated on **Tiny ImageNet**.

FAR databases are shown in Figure 5[5]. Similar reductions are observed on the CIFAR-100 database as well. Figure 6 shows the score distribution computed using VGG-16 on original and CiPer adversarial CIFAR-10 images. The score distribution shows the increase in the overlap between the images of various classes due to the CiPer adversary thus leading to misclassification.

ResNet-50 [22] model yields 84.2% accuracy on the original CIFAR-10 database (Figure 7) and CiPer-S with Laplacian filtering reduces the performance by 19.9%. The reduction in recognition performance shows the effectiveness of the proposed adversarial algorithm in fooling state-of-the-art deep learning models. On the Tiny ImageNet database, the VGG-16 model yields 47.2% recognition accuracy. The results reported in Figure 8 show that median filtering has maximum impact on the performance, and the accuracy reduces by more than 33% and 29% with CiPer-P and CiPer-S, respectively.

[5]Due to space constraints graphs corresponding to the best CiPer algorithm are reported.

Table 2: Object classification accuracy (%) on original and adversarial images generated using the proposed CiPer algorithm.

Data Type		CIFAR-10		CIFAR-100	
		Model 1	Model 2	Model 1	Model 2
Original		**82.0**	**88.5**	**53.3**	**65.1**
CiPer-P	Average Filtered	74.6	82.9	38.8	53.4
	Laplace Filtered	79.5	86.8	46.5	61.0
	LoG Filtered	67.6	86.5	46.7	60.7
	Guided Filtered	78.8	86.5	45.6	59.3
	Integral Filtered	74.6	82.9	38.8	53.4
	Wavelet Filtered	77.3	85.0	43.3	57.3
	Median Filtered	78.4	85.5	45.2	57.9
CiPer-S	Average Filtered	71.0	67.8	40.7	**34.8**
	Laplace Filtered	**65.2**	68.8	**36.3**	40.4
	LoG Filtered	76.1	70.8	38.7	40.9
	Guided Filtered	74.4	78.8	44.0	48.8
	Integral Filtered	71.0	**67.8**	40.7	**34.8**
	Wavelet Filtered	72.6	73.5	44.0	41.4
	Median Filtered	75.3	78.2	47.0	48.2

Results with Model #1 (Shallow) and Model #2 (Deep): The results of these models are summarized in Table 2. On the original images from CIFAR-10 and CIFAR-100, CNN model #1 yields 82.0% and 53.3% accuracy, respectively. As shown in Table 2, custom model #1 has shown high vulnerability against texture and edge-based filtering. The Laplace filtering based attack reduces the accuracy on CIFAR-10 and CIFAR-100 databases by up to $\sim 30\%$.

Model #2 yields 88.5% and 65.1% accuracy on clean images of CIFAR-10 and CIFAR-100, respectively. With integral filtered CiPer-S adversarial images, the relative performance reduces by 23.3% and 46.5%, respectively. Other than object recognition performance on CIFAR databases, the proposed attack reduces the digit recognition accuracy by at least 14% on the MNIST database. The deeper CNN model (i.e., #2) shows higher sensitivity towards CiPer-S adversarial images and average/integral filter with CiPer-S is observed to be the most effective perturbation.

The vulnerability of SOTA deep models shows that noise generated through image filtering operations is model agnostic, i.e., it is inherently transferable across multiple models [10]. One possible reason for the success of this attack could be the that filtering generally reduces the edge information in the images. It also shows the sensitivity of CNN models towards high frequency information present in the input image.

Attribute Classification: Following the protocol defined by Chhabra et al. [9], with original images, VGG-Face yields 75.1% and 66.2% accuracy for smiling and pale skin attributes, respectively (Figure 9). The maximum impact is observed with guided filtering perturbation, and the accuracy on adversarial examples reduces by 6.92% and 48.9% compared to clean images for smile and pale skin classification, respectively. On the LFW database, guided along with average and integral filtering have the maximum impact on

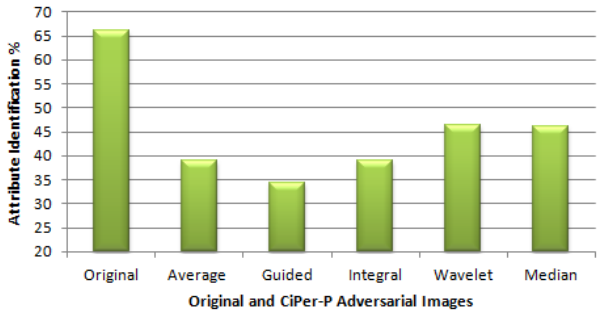


Figure 9: Facial attribute perturbation results on **LFW** using VGG-Face.

the performance.

Visual quality of CiPer: We also perform the visual quality assessment to show that CiPer perturbation does not affect the quality of CIFAR images, while affecting the classification performance of deep models. Two no-reference image quality metrics are used to analyze the visual quality of perturbed images: (i) Naturalness Image Quality Evaluator (NIQE) [29] and (ii) Blind Image Spatial Quality Evaluator (BRISQUE) [30]. *Lower* value of these metrics indicate better perceptual quality of the image. The NIQE scores remain approximately the same (in the range of 18.85–18.87) for original as well as different kinds of CiPer adversarial images. The BRISQUE quality scores either remain same or improve on adversarial images compared to original images. For example, for the CiPer generated images, the BRISQUE scores change to 41.97 and 58.63, compared to 43.44 and 59.06 on original images. This interestingly showcases higher perceptual quality of adversarial images.

To further understand the visual perception of attacked images, we computed the Average Structural Similarity Index (SSIM) that measures the similarity between two images [40, 52]. Values close to 1 indicate higher similarity between two images. On the TinyImageNet database, average SSIM of CiPer-P adversarial (integral) images is 0.89. Further, the norm of the CiPer attack can also be controlled as shown in Equation 3.

Analysis Regarding Filter Selection: Since multiple adversarial images can be generated by using different filters, we also evaluated the effectiveness of the filters. It is observed that every filter is not as effective in crafting the adversarial noise and the performance reduction is more for some filters compared to others. We have observed that texture based filters such as Laplace and LoG, and smoothing based filters such as average and median are useful for generating adversarial noise. The integral filter has shown consistent performance across all databases and CNN models, including deep and shallow networks.

Comparison with Random Noise and Universal Attack: We compared the performance of CiPer with the quasi-imperceptible Gaussian noise corrupted images. With VGG16, the performance on CIFAR-10 and CIFAR-100 databases reduces by 3% (noise variance = 0.0005) and 4.5% (noise variance = 0.001), respectively. On the other hand, the proposed CiPer attack reduces the accuracy by 17.4% and 14.4%, respectively. On CIFAR100 with Model 2, the Gaussian attack (noise var. = 0.001) reduces the performance by 9.5% whereas, the CiPer attack reduces it by 30.3%.

Table 3: Resiliency of the proposed CiPer adversarial generation algorithms using wavelet filtering with Bayesian thresholding for object classification. Recognition rate further decreases in the presence of a defense mechanism against CiPer especially on CIFAR-10 and CIFAR-100 highlighting the resilience of CiPer algorithms. The values here should be compared with Table 2 to understand the effectiveness of defense algorithm.

Data Type		CIFAR-10		CIFAR-100	
		Model 1	Model 2	Model 1	Model 2
Original		**82.0**	**88.5**	**53.3**	**65.1**
CiPer-P	Average Filtered	71.2	78.5	41.6	49.5
	Laplace Filtered	75.2	82.1	46.1	54.7
	LoG Filtered	63.5	81.9	47.0	55.0
	Guided Filtered	73.2	79.6	45.3	53.2
	Integral Filtered	71.2	78.5	41.6	49.5
	Wavelet Filtered	72.6	78.9	45.2	51.0
	Median Filtered	72.9	78.5	46.2	49.5
CiPer-S	Average Filtered	**62.0**	**61.7**	38.2	**31.9**
	Laplace Filtered	62.7	68.6	**32.5**	36.6
	LoG Filtered	75.4	69.5	34.6	37.5
	Guided Filtered	67.1	73.1	39.9	42.4
	Integral Filtered	**62.0**	**61.7**	38.2	**31.9**
	Wavelet Filtered	64.6	67.5	39.8	37.9
	Median Filtered	69.1	72.6	43.3	44.4

We also compared the performance with universal perturbation [31]. We followed the same protocol for both the attacks and observed that universal attack reduces the accuracy of a CNN by a similar amount as CiPer, when noise is learned using a training set of the same database. For example, the CIFAR-10 trained universal perturbation noise, which has a similar magnitude as CiPer, reduces the accuracy of model # 2 on CIFAR-10 by 30%. However, when the CIFAR-10 noise is used on CIFAR-100, the accuracy drops by 10%, whereas CiPer leads to 33.2% reduction. This shows that the universal perturbation algorithm is not database agnostic, whereas CiPer is database agnostic.

3.3. Resiliency under Adversarial Defense

For evaluating the resiliency of the proposed attack, two different kinds of adversarial defense algorithms are used: **(i)** adversarial training and **(ii)** mitigating the effect of adversarial noise through filtering.

In the literature, the most robust adversarial defense is based on adversarial training [49] where the adversarial images are used for retraining the network. We have observed that when the models are trained using CiPer-S images, it performs well on the adversarial images generated using CiPer-S images but fails for CiPer-P images and vice

versa (i.e., lacks generalizability). For example, when the VGG-16 network is re-trained using CiPer-P generated using Laplace filtering, it fails to significantly improve the accuracy on CiPer-S adversarial images generated using average and integral filtering (i.e., improvement of 1–2% only). The results are observed across each deep model, filtering, and generation technique (i.e., CiPer-P and CiPer-S).

Adversarial training using existing state-of-the-art attacks such as FGSM [17], DeepFool [32], and Saliency [34] attacks are also performed. The adversarially retrained model does not show any performance improvement compared to the original model. For example, the performance of DeepFool retrained VGG-16 model is 20% lower than the original model. The significant drawbacks of adversarial training are computationally intensive nature, lower generalizability [37,43], vulnerable to new attacks [11,54], and open other serious threats [28] such as privacy.

Other strong defenses are based on filtering techniques such as wavelet and bilateral [20, 38]. The resiliency of CiPer is therefore tested against wavelet denoising based on Bayes thresholding, Bilateral filtering [38], Total Variance Minimization [20], [41], and Defense GAN [43]. The proposed CiPer algorithm is based on manipulation of inevitably inherent noise in the images through filtering. Therefore, it might be the question of concern: whether this manipulated noise can be removed using a robust image filtering operation? To show the strength of the proposed algorithm, wavelet denoising with Bayesian thresholding is applied on the adversarial images as follows: The noisy signal is first decomposed using selected wavelet filter up to l levels. the decomposed detailed coefficients of noisy signal at each level are filtered using the threshold computed by Bayesian rule, and the modified detailed coefficients are combined with original low-frequency components to reconstruct the noise-free signal. The results related to these findings are reported in Table 3. It is interesting to observe that under wavelet filtering based defense, the accuracy further reduces except for LoG filtering and the possible reason is further reduction of high frequency information. In case of adversarial defense through wavelet filtering on CIFAR-100, the recognition accuracy under Laplace and LoG CiPer-S adversary reduces to 32.5% and 34.6% from 36.3% and 38.7%, respectively.

Other than wavelet denoising, bilateral filtering by Ratzlaff et al. [38], total variance minimization (TVM) denoising by Guo et al. [20] and Defense GAN by Samangouei et al. [43] are found to be ineffective defenses against the proposed CiPer algorithms. Bilateral filtered images reduce the recognition accuracy in the range of 4.7%–10.6% which further reduces in the range of 12.4%–18.4% when denoised using bilateral filtering across all three databases. In the original paper, the Defense GAN algorithm is evaluated for MNIST database and hence not claimed effective for CIFAR databases. Similar argument is given by Athalye et al. [4] regrading its effectiveness on CIFAR-10. On using Defense GAN to counter the proposed attack, CiPer attack is able to retain the fooling rate on each CNN model including VGG-16. Most of the existing defense algorithms show the effectiveness against learning based adversaries which find and perturb the most salient pixels for classification. The proposed attack does not utilize any classification related information of the images and hence might be able to fool the existing defense algorithms. Another reason could be that inherent nature of these noise which we have used might be treated as *natural* and existing defenses do not provide any attention. The filtering based defenses are found to be ineffective to recover the recognition performance because further filtering of CiPer adversarial images reduces the information which might be relevant for recognition.

4. Conclusion

In this research, we propose **CiPer**: a novel adversarial noise generation algorithm which is both database and model agnostic. CiPer identifies *Camera Inspired Perturbations* in images using various image filtering operations. These perturbations are inherently present as noise in images, and CiPer utilizes them to fool the classifier. The performance of the algorithm is evaluated on multiple databases and CNN models, and the results show significant reduction in performance. The analysis using the results of CiPer-S might help in generating strong adversaries using the removal of high-frequency from images and in increasing the robustness of CNNs by looking at what it learns. The adversarial images generated using CiPer are not only effective in fooling SOTA CNNs but also resilient to multiple defense approaches including image enhancement and generative models. In future, we plan to explore simultaneous implementations of multiple filtering algorithms to generate adversarial noise, as the noise present in images might be due to various sources. Further, image watermarking literature can be explored to embed imperceptible adversarial noise in different regions of the images [50,51]

5. Acknowledgement

A. Agarwal was partially supported by the Visvesvaraya PhD Fellowship. M. Vatsa is partially supported through the Swarnajayanti Fellowship by the Government of India.

References

[1] A. Agarwal, R. Keshari, M. Wadhwa, M. Vijh, C. Parmar, R. Singh, and M. Vatsa. Iris sensor identification in multi-camera environment. *Information Fusion*, 45:333–345, 2019. 3

[2] A. Agarwal, R. Singh, and M. Vatsa. Fingerprint sensor classification via mélange of handcrafted features. In *IEEE ICPR*, pages 3001–3006, 2016. 3

[3] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? *IEEE BTAS*, pages 1–7, 2018. 2

[4] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ICML*, pages 274–283, 2018. 8

[5] I. Avcibas, N. Memon, and B. Sankur. Steganalysis using image quality metrics. *IEEE TIP*, 12(2):221–229, 2003. 3

[6] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *ECML-KDD*, pages 387–402, 2013. 2

[7] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *IEEE S&P*, pages 39–57, 2017. 2

[8] P. Chen, Y. Sharma, H. Zhang, J. Yi, and C. Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. *AAAI*, pages 10–17, 2018. 2

[9] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta. Anonymizing k-facial attributes via adversarial perturbations. *IJCAI*, pages 656–662, 2018. 6

[10] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. *USENIX Security*, pages 321–328, 2019. 6

[11] A. Galloway, T. Tanay, and G. W Taylor. Adversarial training versus weight decay. *arXiv preprint arXiv:1804.03308*, 2019. 8

[12] L. A Gatys, A. S Ecker, and M. Bethge. Texture and art with deep neural networks. *Current opinion in neurobiology*, 46:178–186, 2017. 1

[13] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2019. 1

[14] A. Goel, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha. DeepRing: Protecting deep neural network with blockchain. *CVPRW*, 2019. 2

[15] A. Goel, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha. Securing CNN model and biometric template using blockchain. *IEEE BTAS*, pages 1–6, 2019. 2

[16] A. Goel, A. Singh, A. Agarwal, M. Vatsa, and R. Singh. Smartbox: Benchmarking adversarial detection and mitigation algorithms with application to face recognition. *IEEE BTAS*, pages 1–6, 2018. 2

[17] I. J Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015. 1, 2, 8

[18] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *IJCV*, 127:719–742, 2019. 1

[19] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. *AAAI*, pages 6829–6836, 2018. 1

[20] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten. Countering adversarial images using input transformations. *ICLR*, 2018. 2, 8

[21] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE TPAMI*, (6):1397–1409, 2013. 4

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 5, 6

[23] G. B Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 5

[24] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 2, 5

[25] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *ICLR*, 2017. 2

[26] Y. LeCun, C. Cortes, and CJ Burges. MNIST handwritten digit database. 2, 2010. 5

[27] J. Lukas, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE TIFS*, 1(2):205–214, 2006. 3

[28] F.A Mejia, P. Gamble, Z. Hampel-Arias, M. Lomnitz, N. Lopatina, L. Tindall, and M. A. Barrios. Robust or private? adversarial training makes models more vulnerable to privacy attacks. *arXiv preprint arXiv:1906.06449*, 2019. 8

[29] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012. 7

[30] A. Mittal, R. Soundararajan, and A. C Bovik. Making a "completely blind" image quality analyzer. *IEEE SPL*, 20(3):209–212, 2013. 7

[31] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *CVPR*, pages 1765–1773, 2017. 2, 7

[32] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016. 8

[33] K. R. Mopuri, U. Garg, and R V. Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *BMVC*, 2017. 2

[34] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. *IEEE European S&P*, pages 372–387, 2016. 2, 8

[35] S. Paris, S. W Hasinoff, and J. Kautz. Local laplacian filters: edge-aware image processing with a laplacian pyramid. *ACM TOG*, 30(4):1–12, 2011. 4

[36] O. M Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *BMVC*, 1(3):6, 2015. 2, 5

[37] A. Raghunathan, S. M Xie, F. Yang, J. C Duchi, and P. Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019. 8

[38] N. Ratzlaff and L. Fuxin. Unifying bilateral filtering and adversarial training for robust neural networks. *arXiv preprint arXiv:1804.01635*, 2018. 2, 8

[39] K. Ren, T. Zheng, Z. Qin, and X. Liu. Adversarial attacks and defenses in deep learning. *Engineering*, pages 1–15, 2020. 2

[40] A. Rozsa, E. M Rudd, and T. E Boult. Adversarial diversity and hard positive generation. In *IEEE CVPRW*, pages 25–32, 2016. 7

[41] L. I Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 2, 8

[42] S. Sabour, Y. Cao, F. Faghri, and D. J Fleet. Adversarial manipulation of deep representations. *arXiv preprint arXiv:1511.05122*, 2015. 2

[43] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ICLR*, 2018. 2, 8

[44] M. Sharif, S. Bhagavatula, L. Bauer, and M. K Reiter. A general framework for adversarial examples with objectives. *ACM TOPS*, 22(3):16:1–16:30, 2019. 2

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 5, 6

[46] R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa. On the robustness of face recognition algorithms against attacks and bias. *AAAI*, 2020. 2

[47] J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE TEC*, 23(5):828–841, 2019. 2

[48] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[49] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *ICLR*, 2018. 2, 7

[50] M. Vatsa, R. Singh, P. Mitra, and A. Noore. Digital watermarking based secure multimodal biometric system. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2983–2987, 2004. 8

[51] M. Vatsa, R. Singh, and A. Noore. Feature based rdwt watermarking for multimodal biometric system. *Image and Vision Computing*, 27(3):293 – 304, 2009. 8

[52] Z. Wang, A. C Bovik, H. R Sheikh, and E. P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 7

[53] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE TNNLS*, 30(9):2805–2824, 2019. 2

[54] H. Zhang, H. Chen, Z. Song, D. Boning, I. S Dhillon, and C. Hsieh. The limitations of adversarial training and the blind-spot attack. *ICLR*, 2019. 8

[55] P. Zhao, S. Liu, P. Chen, N. Hoang, K. Xu, B. Kailkhura, and X. Lin. On the design of black-box adversarial examples by leveraging gradient-free optimization and operator splitting method. *ICCV*, pages 121–130, 2019. 2