Energy-Efficient and Adversarially Robust Machine Learning with Selective Dynamic Band Filtering

Neha Nagarkar
nnagarka@gmu.edu
George Mason University
Fairfax, Virginia, USA

Khaled Khasawneh
kkhasawn@gmu.edu
George Mason University
Fairfax, Virginia, USA

Setareh Rafatirad
srafatirad@ucdavis.edu
University of California Davis
Davis, California, USA

Avesta Sasan
asasan@gmu.edu
George Mason University
Fairfax, Virginia, USA

Houman Homayoun
hhomayoun@ucdavis.edu
University of California Davis
Davis, California, USA

Sai Manoj Pudukotai Dinakarrao
spudokot@gmu.edu
George Mason University
Fairfax, Virginia, USA

ABSTRACT

The popularity of neural networks is increasing day-by-day. Traditional machine learning solutions, such as image recognition, object detection, are being replaced by deep learning solutions because of their vigorous performance in computer vision. Despite their superior performance in these applications, neural networks are prone to adversarial attacks. Adversarial attack is the process of using adversarial samples as an input to the neural network which causes the network to misclassify, eventually degrading overall performance. Thus, it becomes very important to maintain their robustness by identifying, analyzing, and eliminating the cause of their vulnerability. In this paper, we introduce a technique to determine the most sensitive frequency band of input samples and filter the noise from this band to shield the network against adversarial attack. First, we decompose the input sample into four different frequency components and then, identify the sensitive component by measuring the change in behavior of the pretrained network on normal frequency band and that on frequency band with added noise (frequency band of an adversary). Next, we exploit this vulnerable component to assist the network in tackling the adversaries through noise filtering. Thereby, enhancing the neural networks' performance and defending against the adversarial attack. The low frequency component was the most vulnerable and mitigating the noise from this band significantly improved the accuracy of Convolutional Neural Networks (CNN) along with that of state-of-art networks against adversarial attacks such as Fast Gradient Sign Method (FGSM), DeepFool (DF) and other techniques. The proposed technique showed performance enhancement from 85% to 95% classification accuracy for ResNet50.

GLSVLSI '21, June 22–25, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8393-6/21/06…$15.00
https://doi.org/10.1145/3453688.3461756

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Object detection**; **Object recognition**; **Object identification**; **Supervised learning by classification**.

KEYWORDS

Adversarial Attacks, Computer Vision, Machine Learning, Robustness, Vulnerability

ACM Reference Format:
Neha Nagarkar, Khaled Khasawneh, Setareh Rafatirad, Avesta Sasan, Houman Homayoun, and Sai Manoj Pudukotai Dinakarrao. 2021. Energy-Efficient and Adversarially Robust Machine Learning with Selective Dynamic Band Filtering. In *Proceedings of the Great Lakes Symposium on VLSI 2021 (GLSVLSI '21), June 22–25, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3453688.3461756

1 INTRODUCTION

Image classification using convolutional neural networks has come a long way since its introduction. Today, we see many state-of-the-art classifiers are being designed and published, all of these having classification accuracy in the range of 90%. For instance, the state-of-the-art network ResNet50 [12] gives 92% (top-5) accuracy on ImageNet [14]. Moreover, deep learning does not require any expert knowledge or features extraction. There are many research groups working on improving the network architecture, or designing entirely new architectures to enhance the network performance since, we are adapting the state-of-the-art network architectures in real world applications. In this case, one of the major concern is the robustness of these models. These network architectures based on convolutional layers perform well, that is, they classify correctly until the input provided is noise-free. However, if some selected pixels of the input image are changed while keeping the majority of the input image pixels unchanged such that the input image looks the same for human eye, the network cannot classify this modified image correctly. The perturbed image is the adversary of the normal input image. The state-of-the-art classifiers [13] [12] fail to classify the image correctly even if the image is slightly perturbed. This indicates, the neural networks are vulnerable to adversaries and this can be a crucial problem if such classifiers are deployed in real world applications.

To understand this more clearly, let us consider an example of real world application such as advanced driver-assistance systems (ADAS) [8]. The system uses road traffic markings for recognizing the lane with the help of these state-of-art classifiers, when these markings were perturbed, the system used this information to set the vehicle in opposite lane [1]. Such aberrant behavior of the vehicle caused by perturbed input could be hazardous. Hence, adversarial machine learning is a vital research domain in which many scientist and research are working on this issue to come up with defenses against adversaries [4]. One such defense mechanism is introduced in this paper which would be explained later in the paper.

Deep neural networks perform well, but we don't have the total knowledge of the reasoning behind their performance [17]. There are studies which interpret the deep neural networks, by by creating adversarial attacks [9]. By inspecting the frequency bands of adversarial samples along with normal samples, we propose to understand the working of neural network and get some insights on which frequency band influences the network's outcome the most, which can be helpful in increasing robustness of the neural networks. This can aid in improving the interpretability of neural networks. This technique of sensitivity analysis is explained and the results are discussed in this paper.

2 UNDERSTANDING OF ADVERSARIAL ATTACKS

Adversarial attacks are adversarial samples fed to the network to degrade its performance. A normal image of digit '7' from MNIST dataset is shown along with its adversary in the figure 1. The digit '7' on the left is the original image from MNIST database and the one on the right is its adversary. As shown, both the images look the same to human eye, but the one on right is slightly perturbed by the Carlini Wagner (CW) adversarial attack.

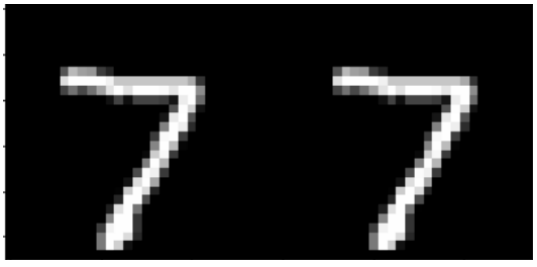

Figure 1: (from left) original sample and its adversary

A neural network classifier trained and tested on MNIST [18] dataset with an overall accuracy of 98.52% was used for testing both of these images. The classifier classified the left image in Figure 1 correctly as '7' but, classified the image on right as '2'. This shows that the slightest modification in the input sample causes the good performing classifier to misclassify. Moreover, changing the parameters of the attack, output of the classifier can be modified and targeted to a particular class. In other words, despite their performance, the neural networks can be fooled by crafting targeted adversarial attacks. Attackers can generate adversarial commands to fool the automatic speech recognition systems and voice controllable systems such as Amazon Alexa, Microsoft Cortana [17]. There are two categories of adversarial attacks namely, (a) poisoning attacks and (b) evasive attacks. Poisoning attacks are the ones which are used during training phase whereas the evasive attacks are targeted adversarial attacks used for testing phase [7]. This work focuses on the use of evasive attacks since many of the applications use pretrained networks and they require defense against the adversaries during inference stage. The next sub-section discuss the crafting of adversarial attacks which are used in later part of the paper for sensitivity analysis.

2.1 Fast gradient Sign Method (FGSM)

This is the most simple yet very efficient method and uses the gradients of the neural network to craft an adversarial sample [16] [6]. The original sample is perturbed by using gradient of the loss with respect to the original sample that maximizes the loss. The advantage of this technique is its low complexity and it takes less time to generate adversaries. This method is mathematically expressed as follows.

$$adv\_x = x + \epsilon * sign(\nabla_x J(\theta, x, y)) \tag{1}$$

where adv_x is the adversarial image and x is the original input image. ϵ is the multiplier which can be gradually increased to increase the perturbations in the original image until it is misclassified. ∇_x is the gradient, J is the loss and θ represents parameters of the classifier model.

2.2 Jacobian Saliency Map Attack (JSMA)

This method causes the classifier to misclassify the adversarial image to a specified target class by saturating a few pixels to their maximum or minimum value. JSMA [15] can be used as targeted as well as untargeted attack. Initially, the saliency maps were developed with the notion of visualizing the classifier's prediction process. The map rates each pixel $x_{(i)}$ on how influential it is in making the classifier to classify the input to a particular class $c = argmax_{c'} f(x)_{(c')}$, where $f(x)$ is the softmax probabilities vector predicted by the classifier. The formula for saliency map is is given as follows.

$$S^+(x_{(i)}, c) = \begin{cases} 0, \text{if } \frac{\partial f(x)_{(c)}}{\partial x_{(i)}} < 0 \text{ or } \sum_{c' \neq c} \frac{\partial f(x)_{(c')}}{\partial x_{(i)}} > 0 \\ -\frac{\partial f(x)_{(c)}}{\partial x_{(i)}} \cdot \sum_{c' \neq c} \frac{\partial f(x)_{(c')}}{\partial x_{(i)}}, \text{otherwise} \end{cases} \tag{2}$$

The saliency map can be exploited by targeting a class t which is different from actual class label c for a given sample x. When some pixels are increased such that $S^+(x_{(i)}, c = t)$, then the perturbed image x' will have higher prediction confidence $f(x')_{(t)}$ for class $t \neq c$, which will result in misclassification.

2.3 DeepFool (DF)

DeepFool is another simple algorithm which can be used to fool deep networks through adversarial perturbations. This is an untargeted method which makes use of L_2 distance metric. Assuming, a classifier has c classses and x is the input to the classifier, for each class there is a hyper-plane, that is, the classifier is linear and depending on the placement of x in the space, the class for the

input is decided. The algorithm finds the closest hyper-plane to x and projects x on the hyper-plane thereby, propelling it beyond the plane and misclassifying the input with minimal perturbations. Since neural networks are non-linear, the algorithm tries to find a solution and repeat the process until it obtains an adversarial sample. The exact formulation of this attack can be referred from the work [13], for the readers who are interested in the mathematical aspect of this attack due to the advanced nature of the formulation.

2.4 Carlini-Wagner Attack (CW)

The CW attack is the most recent adversarial attacks and it is known to defeat the defense mechanisms such as defensive distillation [6]. The CW attack finds adversarial samples against multiple defenses iteratively using a specially chosen loss function which gives lower perturbations as compared to other attacks. The CW attack is much slower than other attacks. It covers a range of attacks based on norms which are created through same optimization framework, thereby resulting in 3 powerful attacks, that are designed using L_0, L_2, and L_∞ norms. We use L_2 attack in this work, for which, the distortion in the input, that is, δ is given as follows.

$$\delta_i^* = \frac{1}{2}(tanh(\omega_i + 1)) - x_i \tag{3}$$

Next, the δ_i^* which is an unrestricted perturbation is optimized over ω to find an adversarial instance. Due to scope of this work, we limit the explanation and for details, one can refer to [6] .

Until now, we have seen the generation of adversarial attacks and Carlini et al, [6] show us vulnerability of the neural networks towards adversarial samples. Our proposed work will help in overcoming this vulnerability to improve the robustness of neural networks which is explained in next section.

3 PROPOSED BAND-FILTERING BASED ADVERSARIAL DEFENSE

To defend the network against adversarial attack, that is, to improve the performance of deep networks on adversarial data, we need to enhance the classification accuracy of the network and that can be achieved by filtering noise from the most sensitive band of the adversarial image. For this purpose, we need to identify the most sensitive frequency band of the input image which affects the performance of the network, in other words, we want to determine the frequency band that determines the classification accuracy of the network. Then, filter this vulnerable band to improve classification accuracy. Firstly, the image needs to be divided into its frequency components we need to find all the frequency bands of an input image to perform sensitivity analysis. We are using Discrete Wavelet Transform (DWT) for decomposing the input image into 4 different frequency components and Inverse Discrete Wavelet Transform (IDWT) for reconstructing the image from frequency components. Before going into the process, we will provide a brief overview about DWT and IDWT since it is an essential part of the technique. The overall technique is shown in Figure 2

3.1 Discrete Wavelet Transform (DWT)

Discrete Wavelet Transform (DWT) [3] [5] is basically decomposition of image signals into a set of independent spatially oriented frequency channels. When an image is passed through two complementary filters, a set of two signals along with approximation and details is returned which are nothing but the four different frequency channels and process is called as decomposition. This set of components can be put back together to form the original image without any loss of information. DWT can decompose an image into a sequence of different spatial resolution images, for example, a 2D image can be have N level decomposition and can have $3N+1$ different frequency bands such as Low Low (LL), Low High (LH), High Low (HL) and High High (HH) as shown in Figure 3 [11] below. The 1-level decomposition can be represented mathematically [10] as, 1. For LL Band

$$W_\phi\left(j_o, m, n\right) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\phi_{j_o,m,n}(x,y) \tag{4}$$

2. For LH, HL, HH Band

$$W_\psi^k(j,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\psi_{j,m,n}^k(x,y) \tag{5}$$

Where $k = H, V, D$ for LH, HL and HH respectively.

3.2 Discrete Wavelet Transform(IDWT)

In the previous section, we got an overview of DWT. Now, we will understand Inverse DWT which is exactly reverse of DWT. IDWT is reconstruction of image from its frequency components without any information loss. Here, all the components are pieced together to get the original image. In this work, we are just using the IDWT for reconstructing the image and considering the objective of the work, it is not required to understand the process details. However, the mathematical representation [10] is given below. Here, the previously decomposed wavelets are combined together to reconstruct the image.

$$\begin{aligned} f(x,y) &= \frac{1}{\sqrt{MN}} \sum_m \sum_n W_\phi\left(j_0, m, n\right) \phi_{j_o,m,n}(x,y) \\ &+ \frac{1}{\sqrt{MN}} \sum_{i=H,V,D} \sum_{j=j_0}^{\infty} \sum_m \sum_n W_\psi^i(j,m,n)\psi_{j,m,n}^i(x,y) \end{aligned} \tag{6}$$

3.3 Band Replacement

We saw in earlier section that DWT is used to decompose the image into its frequency components. Thus, we obtain 4 different frequency components, namely, Low Low (LL), Low High (LH), High Low (HL) and High High (HH) of each image from testing samples as well as adversarial samples. Next, we selectively test for the vulnerable component of the image by replacing the bands of adversarial image by that of a normal image one-by-one. For the purpose of this work, we have considered MNIST [18] and CIFAR-10 [2] datasets. Since, our experiment uses evasive adversarial attacks, we need to identify and clean the most vulnerable frequency band before testing. This is done by replacing the frequency band of adversarial sample by that of normal sample. For instance, we replace LL component of adversarial image by LL component of normal image and then reconstruct the image using these components through IDWT. we use the mathematical formulation mentioned earlier to decompose the image into 4 bands. Then while recombining the 4 parts, we replace the one part of adversarial that of noise-free image. The process can be mathematically represented

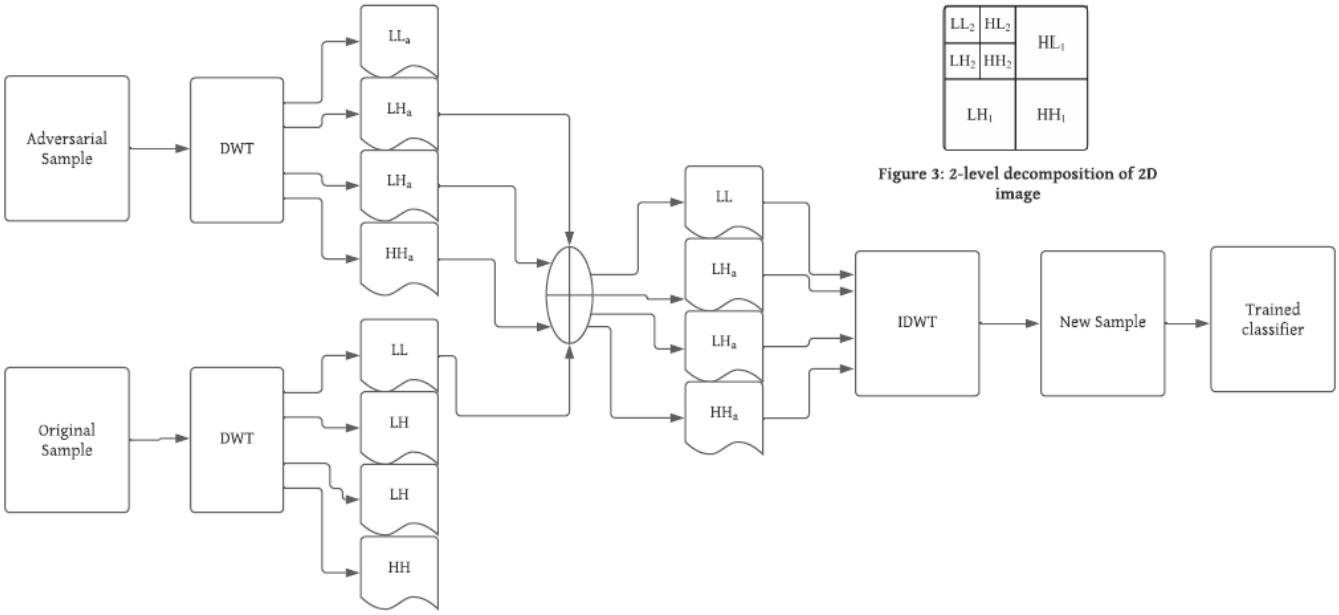

Figure 2: Technique Overview

as,

$$f(x,y) = [\frac{1}{\sqrt{MN}} \sum_m \sum_n W_{\phi adv}(j_0,m,n)\, \phi_{j_{o,m,n}}(x,y)| \frac{1}{\sqrt{MN}} \sum_m \sum_n W_{\phi}(j_0,m,n)\, \phi_{j_{o,m,n}}(x,y)] + \frac{1}{\sqrt{MN}} \sum_{i=H,V,D} \sum_{j=j_0}^{\infty} \sum_m \sum_n {W_{\psi}}^{i}_{adv}(j,m,n) \psi^i_{j,m,n}(x,y) \quad (7)$$

where $W_{\phi adv}(j_0,m,n)$ is LL band for adversarial image , $W_{\phi a}dv(j_0,m,n)$ is LL Band for original image and $[x|y]$ represents x is replaced by y. Similarly, for other 3 bands, the mathematical formulation is given as,

$$f(x,y) = \frac{1}{\sqrt{MN}} \sum_m \sum_n W_{\phi adv}(j_0,m,n)\, \phi_{j_{o,m,n}}(x,y) + [\frac{1}{\sqrt{MN}} \sum_{i=H,V,D} \sum_{j=j_0}^{\infty} \sum_m \sum_n {W_{\psi}}^{i}_{adv}(j,m,n) \psi^i_{j,m,n}(x,y)| \frac{1}{\sqrt{MN}} \sum_{i=H,V,D} \sum_{j=j_0}^{\infty} \sum_m \sum_n W^i_{\psi}(j,m,n) \psi^i_{j,m,n}(x,y)] \quad (8)$$

Where ${W_{\psi}}^{i}_{adv}(j,m,n)$ is for a band of adversarial image and $W^i_{\psi}(j,m,n)$ for band of normal image, i=H, V, D for LH,HL,HH bands respectively.

This process is repeated for all images of the testing set and we obtain a new set which has one normal component, that is, LL and all other other adversarial components. This newly obtained set is tested on the pretrained network to obtain the accuracy. Similarly, this process is repeated for all the components to obtain the accuracy. It is clear that each set has one normal component and other 3 are adversarial components, that means, one component is clean/ filtered and other components are perturbed. Later, the accuracy from testing of all the new sets are compared, and the one with highest accuracy is considered to be the most vulnerable component. This deduction is based on the fact that, each of the 4 new set has one normal or clean component and 3 components with added noise, therefore, whichever set performs better than other 3, that is, set which has higher accuracy as compared to other sets, has the most vulnerable frequency band noise-free. Here, higher accuracy means, accuracy higher than other 3 as well as accuracy higher than the overall adversarial accuracy (accuracy when all the components are adversarial). The next subsection discusses the impact of band replacement on the state-of-art networks trained on MNIST as well as CIFAR-10 datasets.

4 EXPERIMENTAL RESULTS

4.1 Simulation Setup

We have used MNIST and CIFAR-10 datasets for training and testing of CNN as well as deep neural networks. We will start with MNIST data. MNIST dataset consists of 60000 training and 10000 testing samples. In MNIST, each sample is 2-dimensional of size 28x28. We trained a CNN with the training samples and then tested the accuracy of the network using the test samples. Earlier, we saw the performance of deep networks on CIFAR-10 data before and after the adversarial attack. Similarly, for this CNN trained on MNIST data we compare the accuracy of the network before and after the attack. The classification accuracy of the CNN before the attack is 98.52% and after the attack it reduced to 9.46% for the FGSM attack. Next, we perform band replacement simulation for MNIST data. The results for this simulation are discussed in next part of this section.

Another part of this simulation is CIFAR-10 data which is the main part of our work. CIFAR-10 data consists of 50000 training images and 10000 testing images where each image is 3-dimensional of size 32x32. Firstly, we train the network using training samples. The adversarial set of data is obtained by crafting an attack on pretrained network by methods explained earlier in this paper and

generating adversaries for each image from the testing set of CIFAR-10 data. With this simulation, we want to check if band replacement improves the classification accuracy of the deep networks.

4.2 Performance Analysis

First, we will analyze the performance of networks on MNIST data. As mentioned earlier, the MNIST data is 2-dimensional, we have performed the band replacement for a CNN with adversaries generated from JSMA, FGSM, DeepFool and CW attacks. The results of the experiment are shown in Table 1.

Table 1: Impact of Band replacement on Adversarial Samples of MNIST Data

	Frequency Band	Accuracy (%)			
		JSMA	FGSM	DF	CW
Accuracy before attack		98.52	98.52	98.52	98.52
Accuracy after attack		79.14	9.46	10.82	82.89
1-Band replacement	LL	98.04	74.45	94.31	98.37
	LH	85.07	11.81	10.18	86.82
	HL	84.45	11.81	10.99	86.93
	HH	81.42	10.63	10.74	84.6
2-Band replacement	LL and LH	98.24	90.32	97.97	98.52
	HL and HH	86.35	12.75	12.99	88.58
	LH and HH	86.82	12.85	10.29	88.56
	LL and HL	98.30	89.78	97.6	98.52
	LL and HH	98.13	80.76	94.69	98.48
	HL and LH	89.39	14.44	10.39	90.18

As observed from Table 1, the accuracy of the network is affected significantly by adversarial attacks. However, band replacement improves the classification performance notably. When LL component of adversarial image is replaced by that of normal we see a significant rise in the classification accuracy. The performance boost is almost equivalent to the performance of the network before the adversarial attack. Similarly, When LL and LH of adversarial sample are replaced by that of normal image, the accuracy is close to the accuracy before the attack. Thus, one can say that the low frequency components carry significant information which is utilized by the network for classification purpose and we can improve the performance of the CNN by cleaning these frequency bands.

Table 2 shows the performance of various networks on 1-band replacement for CIFAR-10 data.

From Table 2, it is clearly evident that the lower frequency bands are the most sensitive bands. Dataset obtained after replacing LL band has the highest accuracy of all the the other sets. That means, a deep network performs better when the lower frequency bands of the input sample are noise free. Here, the lower frequency bands include both LL and LH, since they have higher accuracies which are close in range than other two. Table 2 shows results obtained after substituting one noise added frequency component by cleaner frequency component, and we observed that two bands LL and LH performed better to enhance the performance of the network on adversaries.

Table 2: Impact of Frequency Band replacement on Adversarial Samples

Network	Frequency Band	Accuracy (%)			
		JSMA	FGSM	DF	CW
CNN	LL	41.48	17.72	42.29	42.06
	LH	30.71	16.38	30.83	30.80
	HL	27.29	13.48	29.06	28.50
	HH	20.86	14.85	21.57	21.16
ResNet50	LL	90.39	85.6	90.6	92.82
	LH	89.70	84.85	89.44	92.22
	HL	85.8	84.38	88.72	89.21
	HH	86.08	84.49	88.68	89.39
DenseNet121	LL	93.38	83.72	93.64	92.57
	LH	92.40	84.27	92.59	91.36
	HL	91.18	82.01	91.70	87.90
	HH	91.49	82.01	91.91	87.22
VGG16	LL	91.43	86.69	91.57	91.58
	LH	90.01	86.58	90.11	90.12
	HL	91.14	84.96	91.49	91.45
	HH	89.23	85.11	89.38	89.35

Now, we present if we can enhance the performance of the network and make the classifier more robust by subsituting 2 bands at a time, which can prove our earlier statement that both the lower frequency bands are most vulnerable bands of an input image. To perform this experiment, we repeat the procedure explained in section 3.3. Here, instead of replacing one frequency component, we substitute two frequency components. For instance, LL and LH of adversarial image are substituted by LL and LH of normal image and then a new image is reconstructed using IDWT. Replicating the same steps for all the images will give us a new set where lower frequency bands are clean. This process is repeated for different combinations of frequency components to obtain different datasets. They are tested on a pretrained network to obtain different accuracies. These classification accuracies are compared in Table 3 which presents the results for 2-band replacement simulation.

It can be inferred from Table 3 that, after substituting all the lower bands, that is, LL and LH bands of adversarial image by that of normal image, we get improved classification accuracy of any network. This means, when the perturbed lower frequency bands are replaced by denoised lower frequency components, the performance of the classifier is enhanced and we get higher accuracy

Table 3: Performance of CIFAR-10 adversarial samples on 2-Band replacement

	Accuracy (%)															
	CNN				ResNet50				DenseNet121				VGG16			
Frequency Band	JSMA	FGSM	DF	CW	JSMA	FGSM	DF	CW	JSMA	FGSM	DF	CW	JSMA	FGSM	DF	CW
LL and LH	61.93	45.22	62.09	62.16	95.09	92.43	94.36	94.90	96.10	93.89	96.21	96.11	94.54	92.81	94.57	94.57
HL and HH	23.95	16.71	24.36	24.18	87.59	85.05	89.22	90.36	92.54	82.06	92.83	88.97	89.11	85.67	89.25	89.23
LH and HH	45.87	17.81	46.28	46.28	90.73	85.50	90.42	93.20	94.26	85.28	94.41	93.06	92.76	87.26	92.82	92.81
LL and HL	48.34	18.18	49.18	49.08	91.21	86.30	91.36	93.44	94.39	84.40	94.55	93.51	92.68	87.03	92.81	92.81
LL and HH	40.13	19.61	40.71	40.75	91.58	86.45	91.61	93.66	93.34	84.52	94.47	93.52	91.77	87.22	91.86	91.85
HL and LH	35.43	16.8	35.84	35.81	90.48	85.27	90.27	93.08	93.72	84.82	93.88	92.51	91.31	86.95	91.42	91.41

which is close to the accuracy obtained from normal input samples. To explain the results more clearly, we can take an example of the ResNet50 network. The accuracy of this network is 95.8% with normal images and this accuracy reduces to 85.28% for JSMA adversaries. However, after experiment I, when there is one clean frequency component in JSMA adversaries, the accuracy of the same network is improved to 90.36% for LL component and 89.70% for LH component which can be seen in Table 2. This is an enhancement in the performance of the network after adversarial attack which means the network was able to combat adversaries. After further exploiting this vulnerability, that is, after experiment II, we observe that if combination of frequency components are replaced by their cleaner version, the classification accuracy is improved significantly. From Table 3, the accuracy of ResNet50 for JSMA adversaries is 95.09% for LL and LH component, that is, when lower frequency bands of JSMA adversaries are replaced by that of normal noise free samples, the performs of ResNet50 on JSMA adversaries is as good as its performs on normal images which is 95.09% from Table 3.

5 CONCLUSION

To summarize our work, we saw the impact of adversarial attacks on CNN as well as some of the state-of-art networks, some attacks had severe impact on network performance which reduced the performance significantly, however some were mild. However, we observed an enhancement in the performance of the same networks through frequency band replacement simulation. Replacing the low frequency components of the adversarial sample significantly improved the classification accuracy. After looking at the results, we can conclude that the low frequency bands are most vulnerable for deep neural networks and denoising the same can enhance the robustness of network by improving classification accuracy. In other words, the network can be defended against adversarial attack by denoising the low frequency components of an input.

REFERENCES

[1] 2019. *Tencent Keen Security Lab: Experimental Security Research of Tesla Autopilot.* Retrieved March 03, 2019 from https://keenlab.tencent.com/en/2019/03/29/Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/

[2] Vinod Nair Alex Krizhevsky and Geoffrey Hinton. 2008. *The CIFAR-10 dataset.*

[3] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. 1992. Image coding using wavelet transform. *IEEE Transactions on Image Processing* 1, 2 (1992), 205–220. https://doi.org/10.1109/83.136597

[4] Marzieh Ashrafiamiri, Sai Manoj Pudukotai Dinakarrao, Amir Hosein Afandizadeh Zargari, Minjun Seo, Fadi Kurdahi, and Houman Homayoun. 2020. R2AD: Randomization and Reconstructor-Based Adversarial Defense on Deep Neural Network. In *Proceedings of the 2020 ACM/IEEE Workshop on Machine Learning for CAD (MLCAD '20).* Association for Computing Machinery, New York, NY, USA, 21–26. https://doi.org/10.1145/3380446.3430628

[5] R.H. Bamberger and M.J.T. Smith. 1992. A filter bank for the directional decomposition of images: theory and design. *IEEE Transactions on Signal Processing* 40, 4 (1992), 882–893. https://doi.org/10.1109/78.127960

[6] Nicholas Carlini and David Wagner. 2016. Towards Evaluating the Robustness of Neural Networks. arXiv:cs.CR/1608.04644

[7] Sai Manoj P D, Sairaj Amberkar, Setareh Rafatirad, and Houman Homayoun. 2018. Efficient Utilization of Adversarial Training towards Robust Machine Learners and Its Analysis. In *Proceedings of the International Conference on Computer-Aided Design (ICCAD '18).* Association for Computing Machinery, New York, NY, USA, Article 78, 6 pages. https://doi.org/10.1145/3240765.3267502

[8] Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. 2020. An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models. *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (2020), 1–10.

[9] Sai Manoj Pudukotai Dinakarrao, Sairaj Amberkar, Sahil Bhat, Abhijitt Dhavlle, Hossein Sayadi, Avesta Sasan, Houman Homayoun, and Setareh Rafatirad. 2019. Adversarial Attack on Microarchitectural Events Based Malware Detectors. In *Proceedings of the 56th Annual Design Automation Conference 2019 (DAC '19).* Association for Computing Machinery. https://doi.org/10.1145/3316781.3317762

[10] Mohammed Gulam Ahamad D.Ravichandran, Ramesh Nimmatoori. 2016. Mathematical Representations of 1D, 2D and 3D Wavelet Transform for Image Coding. *International Journal On Advanced Computer Theory And Engineering (IJACTE)* 5 (2016). Issue 3.

[11] Chih-Hsien Hsia, Jen-Shiun Chiang, and Jing-Ming Guo. 2011. *Multiple Moving Objects Detection and Tracking Using Discrete Wavelet Transform.*

[12] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. 2015. Deep Residual Learning for Image Recognition. arXiv:cs.CV/1512.03385

[13] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. 2574–2582. https://doi.org/10.1109/CVPR.2016.282

[14] Hao Su Jonathan Krause Sanjeev Satheesh Sean Ma Zhiheng Huang Andrej Karpathy Aditya Khosla Michael Bernstein Alexander C. Berg Li Fei-Fei Olga Russakovsky, Jia Deng. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (Dec. 2015), 36–44. https://doi.org/10.1007/s11263-015-0816-y

[15] Anqi Xu Rey Wiyatno. 2018. Maximal Jacobian-based Saliency Map Attack. arXiv:cs.LG/1808.07945

[16] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. 2018. *Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII.* 644–661. https://doi.org/10.1007/978-3-030-01258-8_39

[17] Qile Zhu Xiaolin Li Xiaoyong Yuan, Pan He. 2017. Adversarial Examples: Attacks and Defenses for Deep Learning. arXiv:cs.LG/1712.07107

[18] Corinna Cortes Yann LeCun and Christopher J.C. Burges. 1998. *THE MNIST DATABASE of handwritten digits.* http://data.pymvpa.org/datasets/mnist/