Robustness Against Gradient based Attacks through Cost Effective Network Fine-Tuning

Akshay Agarwal[1], Nalini Ratha[2], Richa Singh[3], and Mayank Vatsa[3]
[1]IISER Bhopal, India, [2]University at Buffalo, USA, and [3]IIT Jodhpur, India
akagarawal@iiserb.ac.in, nratha@buffalo.edu, {richa, mvatsa}@iitj.ac.in

Abstract

Adversarial perturbations aim to modify the image pixels in an imperceptible manner such that the CNN classifier misclassifies an image, whereas humans can predict the original class. Several defense algorithms against adversarial attacks are proposed in the literature, such as binary classification which aims to detect adversarial examples, and network retraining using perturbed images. The challenge with the adversarial detection approach is that once the perturbed samples are detected, they are discarded, and the system requires fresh input. On the other hand, adversarial training requires the generation of adversarial images for data augmentation and hence is computationally demanding. It is well known that training a deep CNN architecture is resource-intensive, and therefore retraining again from scratch is not feasible in resource-constrained scenarios. We propose computationally efficient fine-tuning of pre-trained networks to increase their robustness against the prevalent gradient-based attacks. The proposed fine-tuning is performed in a complete black-box fashion, where we do not know the training setting such as optimizer, batch size, and learning rate used in the training of the network. Extensive experiments using multiple CNN architectures such as VGG and ResNet show that the proposed fine-tuning provides significant robustness against various widespread gradient attacks.

1. Introduction

Building an adversarially robust network requires huge computational resources. For instance, eight layers convolutional neural network (CNN) which can obtain $\sim$ 16% error on ImageNet [19] requires 1.4 GFlop operations, and 152 layers require 22.6 GFLOP functions for $\sim$ 3.5% error. Similarly, on four M40 GPUs, the ResNet-18 model requires 2.5 days for training, whereas the ResNet-152 requires 1.5 weeks of training time [18]. Making these resource-hungry networks adversarially robust via adversarial training is an arduous task.

As shown in Figure 1, the motivation of the proposed research is to provide a mechanism to enhance the adversarial robustness of deep networks even in absence of large computational resources through cost-effective fine-tuning. The proposed approach also provides a trade-off between natural accuracy and adversarial robustness. The proposed algorithm fine-tunes the network through novel data augmentation for 15-20 minutes on a 1080 Ti GPU machine, on which the pre-trained network takes a couple of days to a week. For network fine-tuning, data augmentation is performed by perturbing the local regions of images through various transformation functions and data corruption. To showcase that the proposed approach provides security against popular gradient-based attacks, extensive experiments using several pre-trained networks on multiple databases are used. The key contributions of this research are:

- We propose a novel data augmentation technique to increase the pre-trained network's adversarial robustness in a computationally efficient manner. The proposed technique can be applied to any pre-trained network in a black-box fashion, i.e., parametric details are not required for fine-tuning;
- The effect of different parameters in the proposed data augmentation is analyzed concerning the trade-off between the robustness and natural accuracy of the network;
- Extensive experiments and comparisons with existing algorithms are performed which highlight that the proposed algorithm yields significantly higher performance.

2. Related Work

With the introduction of a simple l_p norm minimization-based attack referred to as L-BFGS [42], several popular adversarial attacks are proposed. Adversarial attacks aim to minimize the perturbation norm in such a way as to fool only the deep learning classifier while maintaining the decision of human examiners intact. The popular and effective

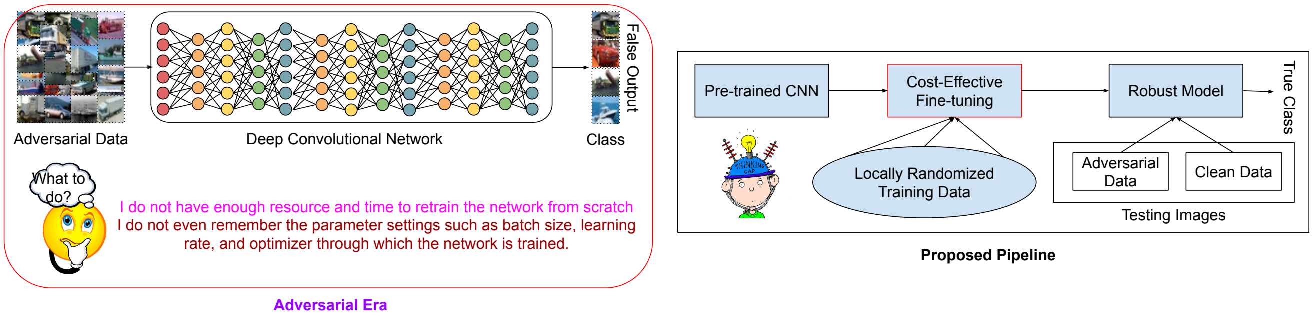
Adversarial Data
Deep Convolutional Network
False Output
Class
What to do?
I do not have enough resource and time to retrain the network from scratch
I do not even remember the parameter settings such as batch size, learning rate, and optimizer through which the network is trained.
Adversarial Era
Pre-trained CNN
Cost-Effective Fine-tuning
Robust Model
True Class
Locally Randomized Training Data
Adversarial Data
Clean Data
Testing Images
Proposed Pipeline

Figure 1. Motivation of the proposed research. The resource-constrained scenario lacks enough training resources, time, and parametric details of the pre-trained network. The proposed research provides a solution through simple network fine-tuning through randomized data augmentation to increase the adversarial robustness in a white-box setting.

adversarial attacks in literature can be broadly grouped into gradient-based [11], logits based [14], classifier-based [33], universal perturbations [32], and black-box [2, 23]. Out of these attacks, gradient-based attacks are the most effective in dropping the target model's accuracy. The gradient attacks can also be used as a backdoor attack reflecting their high potential in restricting the deployment of deep neural networks [6].

Several defense algorithms are also presented to ameliorate the effect by looking at the significant limitations of deep classifiers against adversaries. Existing defense strategies can be broadly divided into the following categories: (i) binary detectors, (ii) input modification or network modification, (iii) adversarial training, and (iv) certified defenses. The binary detectors trained a separate classifier(s) to decide whether an image is clean or perturbed [24, 31]. Based on the classifier's decision, the image is either discarded or passed to the network for further processing. The input modification or network modification-based defenses apply some processing on the network filters or input images to mitigate the effect of adversarial noise [21, 23]. However, most of the existing defenses based on an external classifier, input, or network manipulation are proven ineffective under white-box attack setting [12] where an attacker has complete access to the target network and its defense mechanism. Another defense that is in a nascent stage is to provide a certificate of robustness. Certified robustness is defined as the defense, which states that in certain limits of the network or input, no adversarial examples exist. However, most of these certified defenses provide slight robustness against gray-scale images. Further, Ghiasi et al. [20] have shown that certified defenses can also be broken.

Out of the existing defenses, adversarial training is found to be the most robust defense along with generalized detection-based defenses [1, 5]. Adversarial training refers to the scenario where the network is trained using both clean and their adversarial counterpart. [22] have first performed the adversarial training with the adversarial examples generated using the gradient method. Later, several studies have been proposed to train the network using strong adversaries or ensemble adversarial examples [39]. However, the adversarial training-based defense has a few limitations: (i) computationally expensive because of the generation of adversarial images, (ii) affects generalization, and (iii) vulnerable to unseen adversarial attacks not used in the training [49]. A discussion on the existing adversarial examples generation and defense works can also be found in the survey papers [10]. The recent studies open a new direction of defense which is either based on denoising the input [37], transfer-learning [15], augmentation [36], retraining of the network [28], and universality [3, 4, 9]. The above studies inspire us to build a generic and cost-effective solution based on data augmentation and fine-tuning of vulnerable models.

3. Proposed Fine-tuning for Adversarial Robustness

The classifiers which are trained on a large number of images share a strong relationship between the input and output. As shown by Goswami et al. [24] and Szegedy et al. [42], there might be individual filters that can be affected by the specific properties of the input or adversarial pattern such as stroke in the upper round portion of a digit image, spiky flowers or color of the input. In addition, they have also observed that adversarial examples modify the local regions which are also highlighted in the filter responses [23]. The proposed data augmentation is inspired by this understanding of the adversarial examples and filter maps of CNN. The proposed algorithm minimizes the dependency of the network on the input data and adds a non-deterministic function layer between the input and the first layer of the system.

Specifically, for fine-tuning a pretrained CNN architecture, corresponding to each input x_i a random function is selected. The function alters the local patch of size w starting from the location (m, n). We have used a square size patch around the center of an image. The modified input image is then augmented with the original image and passed to the network for fine-tuning. The objective function with such randomized data augmentation can be reported as $min_\theta \sum_{i=1}^{N} \mathcal{L}(\phi((x_i, x_j), \theta), y)$ where, ϕ is the CNN classifier function mapping input to class output and N is the total number of original training images. x_i and x_j are the clean and modified versions of the input. θ represents the parameters of the network optimized concerning input and output class label y. The entire network's fine-tuning is performed similarly as the network is trained using gradient descent optimization. For a particular sample, a randomized function modifying the local region in an image remains the same throughout the fine-tuning process, i.e., it does not change with the epoch or batch. The randomization is applied only at the time of training; the testing is performed on the clean or adversarially modified images.

In the proposed data augmentation, a randomized function q is chosen and applied to the local region of an input image x_i. To increase the randomness, multiple functions are used, and randomly few of them are selected. The functions selected to modify the local patches include Gaussian blur, various types of noises such as Gaussian noise, Salt&Pepper noise, Speckle noise, Poisson noise, mild rotation of regions, pixel nullification, and horizontal and vertical pixel flip. The proposed data augmentation can be connected to two concepts: (i) dropout and drop connect [41] and (ii) vulnerability of CNNs against patch-based attacks. It is observed that the network trained with all clean neurons might lead to good accuracy on training examples but less robust on the testing set [8, 35]. Dropout and drop connect help in increasing the randomness in the system and regularizing the network. We take advantage of this concept in fine-tuning the pretrained networks. In the proposed setting, instead of perturbing the structure, we manipulate the input at a local level through randomly selected functions on a particular image. Another reason for choosing the manipulation of local image patches for augmentation can be seen from the vulnerability of CNNs against patch-based attacks [29, 47]. These works show the importance of local regions in the decision-making process of CNNs, therefore reducing their dependency can increase the robustness.

3.1. Functions Used for Randomization in the Proposed Data Augmentation

1. Rotation: The local region of an image is rotated 10 degrees in the clockwise direction. This operation helps make the deep classifiers robust to handle the geometric transformation, which might be present in the natural images. The missing values, if generated through rotation are filled using bicubic interpolation;

2. Translation: Another common transformation in the real world is the translation of different parts of an image at different locations in different images. The missing values are filled with the mean of the patch shifted/translated 2 pixels in the horizontal and vertical direction;

3. Gaussian Noise: We have applied the Gaussian noise with mean value 0 and variance 0.05;

4. Poisson Noise: The Poisson distributed noise is generated from the information of the local patch itself. For example, if the pixel intensity input value is k, then the output value is generated from the Poisson distribution using k as the mean value;

5. Salt&Pepper Noise: It acts as a switch between the input and output, i.e., the input value is replaced with either zero or max value of the data type. Each pixel value is assigned with a probability value uniformly generated between (0,1) to assign a value to a pixel. Suppose the probability value lies in the range [0,d/2]. In that case, the input pixel is assigned 0 in the output image, and if the probability value lies in the range [d/2,d], the maximum value of the image data type is assigned. For the probability value in the range [d,1], the pixel value remains unchanged. d is the density of the noise, and the value of 0.5 is used in this research;

6. Speckle Noise: It is a multiplicative noise, i.e.,

$$Image_{out} = Image_{in} + n * Image_{in}$$

   where, n is the random uniform noise, generated from the specific mean and variance. $Image_{out}$ and $Image_{in}$ are the noisy and clean image, respectively. We have used 0 mean and 0.05 variance value;

7. Gaussian Blur: In this operation, an image is blurred using a Gaussian kernel with a standard deviation of 1.5. It can help in reducing the sensitivity of the deep classifier on high-frequency information as it is one of the critical components for generalizability [44] and can be used for adversarial perturbations [7, 11];

8. Pixel Masking: Some parts of the image may get corrupted or missed out because of occlusion, and camera angle; therefore, an exemplary system should make the correct decision even in such cases. To evaluate, we have masked the local patch(es) of an image to give partial information while learning the network;

9. Flipping: The image parts might be deformed or presented in other forms, such as objects flipped upside down. We have incorporated this by adding horizontal and vertical flips of local patches.

Table 1. Configuration of the custom CNNs.

CNN	Configuration
F-MNIST$_{conv}$	Conv($8 \times 8 \times 64$), ReLU, Conv($6 \times 6 \times 128$), ReLU, Conv($5 \times 5 \times 128$), ReLU, Fully Connected(10), SoftMax
CIFAR$_{conv}$	Conv($5 \times 5 \times 32$), ReLU, MaxPool(3×3), Conv($8 \times 8 \times 64$), ReLU, AvgPool(3×3), Conv($8 \times 8 \times 64$), ReLU, AvgPool(3×3), Fully Connected(64), Fully Connected(10), SoftMax

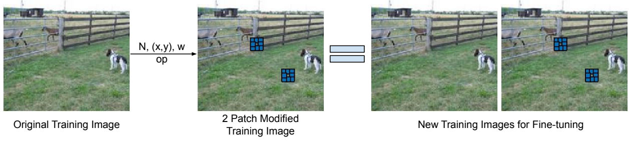
N, (x,y), w
op
Original Training Image
2 Patch Modified Training Image
New Training Images for Fine-tuning

Figure 2. Shows the generation of augmented data for proposed network fine-tuning. The parameters for data augmentation are location (x,y) in the image to be modified, size (w) of the patch, and number (N) of patches. op represents the randomized operation to be applied. The original and local region modified is combined to find the new training data for fine-tuning.

Table 2. Attack parameters for each database.

Attack	Parameters
CIFAR-10	
PGD-v1	ϵ = 0.1, α=0.01, Iterations = 100, Restarts = 10
PGD-v2	ϵ = 0.03, α=0.01, Iterations = 100, Restarts = 10
FGSM-v1	ϵ = 0.1
FGSM-v2	ϵ = 0.03
IFGSM-v1	ϵ = 0.1, α=0.01, Iterations = 100
IFGSM-v2	ϵ = 0.03, α=0.01, Iterations = 100
F-MNIST	
PGD-v1	ϵ = 0.3, α=0.01, Iterations = 100, Restarts = 0
PGD-v2	ϵ = 0.3, α=0.01, Iterations = 100, Restarts = 10
FGSM	ϵ = 0.3
IFGSM	ϵ = 0.3, α=0.01, Iterations = 100
ImageNet	
PGD-v11	ϵ = 0.1, α=0.01, Iterations = 100, Restarts = 0
PGD-v12	ϵ = 0.1, α=0.01, Iterations = 100, Restarts = 10
PGD-v21	ϵ = 0.03, α=0.01, Iterations = 100, Restarts = 0
PGD-v22	ϵ = 0.03, α=0.01, Iterations = 100, Restarts = 10
FGSM-v1	ϵ = 0.1
FGSM-v2	ϵ = 0.03
IFGSM-v1	ϵ = 0.1, α=0.01, Iterations = 100
IFGSM-v2	ϵ = 0.03, α=0.01, Iterations = 100

4. Experimental Results and Analysis

In this section, first, we present the details of databases, networks, and attacks used to showcase the adversarial resiliency of the proposed defense in the white-box setting. Later, the results and analysis related to adversarial defense are provided.

4.1. Evaluation Settings

Databases: We have used three databases namely Fashion-MNIST (F-MNIST) [46], CIFAR-10 [30], and a subset of ImageNet [19].

CNN: Multiple CNN models, including state-of-the-art models for object recognition such as VGG [40] and ResNet [25] along with custom models, are used. The configuration of CIFAR$_{conv}$ and F-MNIST$_{conv}$ are given in Table 1.

Adversarial Attacks: Three challenging gradient-based adversarial attacks are selected for extensive experimentation: (i) Fast Gradient Sign Method (FGSM), (ii) Iterative FGSM (IFGSM), and (iii) Projected Gradient Descent (PGD). Recent studies [34, 38, 44] have shown attack detection methods can defend against optimization attacks such as C&W l_2 [13] and classifier-based adversaries such as DeepFool [33], however, fail significantly for gradient-based attacks. The parameters used in the generation of adversarial samples are given in Table 2.

4.2. Results

In the proposed randomized data augmentation, the two important parameters are the number of patches and the patch size. We have either modified a single patch or two patches. We have tried to select these patches near the center of the image. The analysis will provide an estimate of how big and small of a patch can be selected for perturbation and the number of patches to be used to increase the randomization in the network. In the experiments, the perturbation of patch size 7×7 and 11×11 with random function is referred to as n-P_k. $n \in (7, 11)$ represents the size of the patch, and P_k represents the k number of patches modified through functions chosen randomly. First, the results on the CIFAR-10 dataset are reported using multiple pre-trained CNN models along with the comparison with existing defense algorithms and adversarial training. Later the experimental analysis of the F-MNIST dataset is provided. Finally, to showcase the practicality of the defense, experiments on high-resolution images from the subset of ImageNet [19] are also performed.

4.3. Results and Analysis on CIFAR-10

To demonstrate the generalizability of the proposed fine-tuning, multiple CNNs are used on the CIFAR-10 database, including VGG-16 and ResNet. The proposed fine-tuning aims to not only defend the models from adversarial attacks but also retain or improve the accuracy of clean natural images. The VGG and ResNet yield 83.91% and 91.81% clas-

Table 3. Clean and adversarial examples accuracy of the pre-trained and fine-tuned (defended) CNN models under white-box attack setting on CIFAR-10. Best viewed in color.

CNN	Clean/Attack	Undefended	Defended			
			7-P_1	11-P_1	7-P_2	11-P_2
VGG-16	Natural	**83.91**	**86.71**	84.01	82.14	82.54
	PGD-v1	**0.87**	**60.73**	50.27	54.81	52.42
	PGD-v2	**0.57**	**70.40**	65.32	49.06	55.65
	FGSM-v1	**3.35**	**64.58**	59.28	61.26	58.12
	FGSM-v2	**7.49**	**67.64**	67.13	62.29	55.89
	IFGSM-v1	**2.61**	**64.36**	53.85	61.02	68.09
	IFGSM-v2	**2.75**	55.61	**66.33**	59.60	61.81
ResNet	Natural	**91.81**	**91.40**	90.57	90.40	90.27
	PGD-v1	**0.53**	15.17	13.43	15.33	**18.18**
	PGD-v2	**0.0**	**43.14**	41.14	43.10	40.92
	FGSM-v1	**6.54**	**50.96**	48.39	47.26	48.73
	FGSM-v2	**6.48**	**48.99**	48.19	47.80	45.37
	IFGSM-v1	**2.20**	**51.00**	47.25	48.04	45.84
	IFGSM-v2	**2.20**	**49.86**	47.54	47.94	45.48

Table 4. Comparison of the proposed fine-tuning based defense with several existing defenses such as EMPIR [38] and adversarial training (AT) on clean and various adversarial attack images. The results are reported on CIFAR-10 using CIFAR$_{conv}$.

Data	FGSM-AT	PGD-AT	EMPIR	EMPIR-AT	Proposed
Clean	73.62	73.55	72.56	73.62	**74.74**
FGSM	41.58	12.45	20.45	31.67	**44.35**
IFGSM	12.92	10.97	24.59	29.55	**45.31**
PGD	11.24	8.52	13.55	14.74	**44.26**

sification accuracy on the clean test set of CIFAR-10, respectively. As expected, both the pre-trained models' performance dropped significantly when adversarial samples are processed.

Table 3 shows the experimental results on CIFAR-10 using VGG and ResNet CNNs. Under the PGD attack with a strength of perturbation $\epsilon = 0.03$ and $\epsilon = 0.1$, the pre-trained VGG model's accuracy degrades to 0.87%. The proposed fine-tuning, which is performed under a black-box setting, i.e., without knowledge of the parameters used in the pre-training, can increase the robustness of the pre-trained model in a white-box attack setting. The fine-tuned model shows an improvement of 59.86 (60.73 from 0.87) when higher strength PGD attack is applied in a white-box setting. The ResNet model yields better recognition performance than VGG but is observed to be highly susceptible to PGD attacks as well. The accuracy of the model degrades up to 0.53% from 91.81%. However, the fine-tuning defense boosts the performance significantly. For example, for a PGD attack with $\epsilon = 0.03$, the fine-tuned model's performance is 43.14% as compared to 0.0% of the undefended pre-trained model.

Comparison with adversarial training: Adversarial training is one of the most vigorous defenses in the literature and is based on the augmentation of images while training the deep classifier. Therefore, it is the best fit for comparing with the performance of the proposed fine-tuning-based defense. The experiments are reported using the CIFAR-10 database and VGG network; however, similar trends are observed across other networks. The networks are adversarially trained individually using all three attacks used in the paper. The results reported in Table 5 show that even the adversarially trained models are highly vulnerable to attacks. It is interesting to note that while the adversarial training uses the same parameter settings such as optimizer and batch size as the pre-training of the model and even adversarial perturbations still lack performance compared to the proposed defense.

Comparison with existing defenses: In addition to the comparison with adversarial training, a comparison with recent work termed EMPIR [38] is also performed. In this comparison, the CIFAR$_{conv}$, along with the attacks' parameters are chosen as provided in the original EMPIR paper. The comparison with EMPIR, along with adversarial training, is provided in Table 4. The proposed fine-tuning not only outperforms the algorithms in comparison to adversarial robustness but also yields higher accuracy on clean images. The EMPIR, which is found significantly useful on C&W l_2 attack, fails on the PGD attack; whereas, the proposed fine-tuning restores the performance with a large margin. The performance of the proposed defense on PGD with EMPIR is 36.71% better, and as compared to adversarially trained models, it is at least 35.52% higher.

We further increased the complexity of the PGD attack ($\epsilon = 0.03$ and $\alpha = 0.01$) by running it for 100 iterations and multiple random restarts (5, 10, and 20). The proposed

Table 5. Comparison of the proposed defense with existing adversarial defense and data augmentation algorithms including adversarial training (AT). The experiments are performed on CIFAR-10 using the VGG model under a white-box attack setting. The existing algorithms are: hidden space [34], L2L-DA [27], high frequency [44], mixup [50], manifold mixup [43], and erasing [52].

Attack	Hidden Space	L2L-DA	High Frequency	Mixup	Manifold Mixup	Erasing	FGSM AT	IFGSM AT	PGD AT	Proposed (7-P_1)
FGSM-v2	47.70	45.77	35.60	28.12	29.26	21.45	39.96	43.67	36.30	**70.40**
IFGSM-v2	32.60	50.26	22.04	24.56	29.71	18.62	35.08	44.64	34.63	**67.41**
PGD-v2	27.20	39.69	33.70	21.76	24.20	19.89	32.12	34.31	37.93	**65.61**

Table 6. Adversarial robustness using F-MNIST$_{conv}$ under white-box attack setting on Fashion-MNIST.

Clean/Attack	Undefended	Defended			
		7-P_1	11-P_1	7-P_2	11-P_2
Clean	**91.49**	**90.47**	89.27	90.21	89.86
PGD-v1	**0.41**	**88.09**	85.42	83.02	77.63
PGD-v2	**0.01**	**52.69**	40.29	47.34	34.92
FGSM	**1.97**	**87.85**	75.13	83.46	73.46
IFGSM	**1.11**	**86.65**	83.61	82.38	74.64

Table 7. Adversarial robustness under white-box attack setting on ImageNet. Best viewed in color.

Clean/Attack	Undefended	Defended	
		P_1	P_2
Natural	**84.73**	**84.86**	84.37
PGD-v11	**0.37**	**42.94**	43.94
PGD-v12	**0.12**	**36.16**	35.06
PGD-v21	**0.37**	**44.16**	43.59
PGD-v22	**0.60**	**42.95**	41.15
FGSM-v1	**6.90**	**44.61**	42.16
FGSM-v2	**1.47**	**45.31**	42.54
IFGSM-v1	**5.37**	**44.02**	42.76
IFGSM-v2	**1.37**	**44.64**	42.95

defense can retain the accuracy to at least 44.88%, which is $\sim 36\%$ better than the pre-trained CIFAR$_{conv}$ model. The proposed security can resist even the higher perturbation ($\epsilon = 0.1$ and $\alpha = 0.01$) computed from 100 iterations and multiple random restarts (5, 10, and 20), and yields $\sim 35\%$ better performance than the pre-trained model.

Comparison with other defenses such as restriction of CNN hidden space [34], L2L-DA [27] and data augmentation-based techniques such as Mixup [50], Manifold mixup [43], and random erasing [52] are also performed. The results of these existing methods and the proposed fine-tuning of VGG on CIFAR-10 are given in Table 5. The proposed algorithm can surpass each algorithm in a computationally efficient manner. The attacks are performed using standard parameters with $\epsilon = 0.03$ for each attack and $\alpha = 0.01$ for iterative attacks.

4.4. Results and Analysis on Fashion-MNIST

Another popular database explored in the literature is the Fashion-MNIST database to showcase robustness against adversarial attacks. The experiments are performed using the F-MNIST$_{conv}$ and multiple attacks (with $\epsilon = 0.3$), including iterative PGD and IFGSM variants. The results are reported in Table 6. Clean test images show the classification performance of 91.49%. The proposed fine-tuning shows a slight drop in performance; however, able to increase the robustness of the network significantly. The PGD attack with ten random restarts reduces the performance of the undefended model from 91.49% to 0.01%. On the other hand, in the fine-tuned model, the attack can reduce the performance up to 52.69% only. In the case of another challenging iterative FGSM attack, the fine-tuned model shows 85.54% higher performance than the undefended pre-trained model. The high robustness can also be noticed against other variants of FGSM and PGD attacks. Similar to CIFAR-10, we have observed that single patch modification with small size (7×7) shows higher robustness than multiple and large patches.

4.5. Results on High-Resolution Images

While the above two databases are extensively explored for adversarial defense, they contain images of low resolution. Therefore, to further evaluate the effectiveness of the proposed defense, a higher resolution subset of ImageNet [19] is also used. The attacks on undefended pre-trained and the proposed fine-tuned models are performed and results are reported in Table 7. Through these experiments, we have observed that the proposed fine-tuning-based defense is effective against a wide range of gradient attacks on high-resolution images as well. Experiments to validate the effectiveness of the proposed fine-tuning with PGD attacks are also performed using XceptionNet [16] and MobileNet [26]. The XceptionNet and MobileNet models yield 80.53% and 85.66% accuracy on clean images, respectively; however, fine-tuning improves the performance by at-least 2%. On the other hand, in the case of adversarial robustness, higher robustness as compared to VGG is noticed. The networks can achieve at least 67.90% accuracy on PGD examples ($\epsilon = 0.03$ and $\alpha = 0.01$).

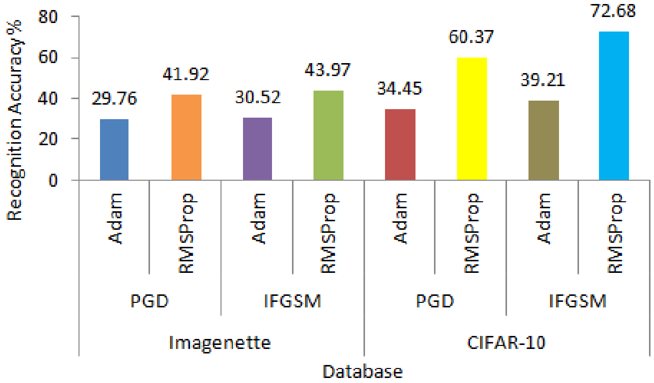

Figure 3. Comparison of the Adam (used in pre-training and fine-tuning) and RMSProp (used in fine-tuning and hence different from pre-training) optimizer on the performance of fine-tuning for white-box adversarial robustness of VGG.

Table 8. Time to train a classifier using PGD adversarial training (AT) and proposed algorithm (including pre-training + fine-tuning). The reduction in computational cost is also reported to showcase the impact. For each algorithm, values are reported using the same machine configuration.

Database	CIFAR10		F-MNIST	ImageNet
Network	VGG	ResNet	F-MNIST	VGG
AT (minutes)	$\sim$245	$\sim$180	$\sim$196	$\sim$380
Proposed (minutes)	$\sim$20	$\sim$28	$\sim$39	$\sim$21
Reduction from AT	12.2x	6.4x	5.0x	18.1x

Table 9. Proposed defense (%) on AutoAttack [17].

Database	CNN	Undefended	Proposed Defense
CIFAR10	VGG16	0.23	**67.86**
	ResNet50	0.16	**44.59**
F-MNIST	Wide-ResNet	0.0	**54.62**
	F-MNIST$_{conv}$	0.0	**86.70**
ImageNet	VGG16	0.80	**59.76**
	Wide-ResNet	0.0	**47.69**

Table 10. Robustness (%) effect of the proposed data augmentation when used while training the network and when it is used for finetuning on the ImageNet database. In both scenarios, the proposed approach can surpass the model trained from scratch using clean images.

Attack	Pretrained (Trained from scratch using)		M1 Fine-tuned using proposed augmentation
	Clean Images only (M1)	Proposed Data Augmentation (M2)	
FGSM	4.47	46.07	**58.13**
IFGSM	3.37	45.24	**58.75**
PGD	0.05	42.75	**58.05**

5. Ablation Studies

Role of optimizer: We have also conducted an ablation study on the role of network optimizers in fine-tuning for adversarial robustness. To demonstrate, two optimizer settings are used for fine-tuning: (i) Adam, which is also used for pre-training the networks, and (ii) RMSprop, which is different from the optimizer used for pre-training. From Figure 3, it is interesting to observe that when a different optimizer is used, the CNN models show higher robustness against adversarial attacks. The analysis is reported using 10 iterative PGD (ϵ = 0.03, α = 0.01, and 10 restarts) and 100-step FGSM (ϵ = 0.03, α = 0.01) attack.

Computational and Performance Gain: Comparison with adversarial training (AT), establishes the advantage of the proposed defense over one of the most robust defenses both in terms of computational complexity (Table 8) and accuracy (Figure 4). The proposed *simple* and *effective* fine-tuning approach outperforms state-of-the-art AT defenses by a significant margin. The recent research [28] (similar broad concept of finetuning) for robustness claims to reduce the complexity by $\sim$ 10x compared to AT. The proposed approach not only further reduces the time by $\sim$ 8x but also obtains better robustness on both clean and adversarial examples (Figure 4) as compared to [28].

Practicality: As compared to several adversarial learning (Figure 4 (a)), the higher performance of the proposed algorithm on the clean images makes the proposed solution a practical solution. In the literature, it is assumed that the system in evaluation is expecting either 100% adversarial images or 0% adversarial images, while this is not the case in the real world. The adversarial examples might come in some proportion, therefore, the accuracy of natural examples must not suffer for such a low number of adversarial examples. In other words, while improving the adversarial robustness we should not lose the accuracy of natural examples and almost every adversarial learning defense fails in this context. Here, the major advantage of the proposed defense is the clean examples accuracy which is much better than existing adversarial defenses.

Strong Evaluation: We have also evaluated the robustness of the proposed defense against recent and reliable attack [17]. The proposed defense can handle the latest reliable attack and aims to evaluate the strong future prospective defense. Table 9 shows the success of the proposed defense and makes it an ideal choice for the real world by being effective against strong attacks and computationally efficient. We have also performed additional experiments to further evaluate the robustness of the proposed defense against another state-of-the-art attack namely BPDA [12]. We have observed that the success rate of BPDA (obfuscated gradients) attack reduces by 45% on the Wide-ResNet model with ImageNet database as compared to the undefended model. Apart from that, the comparisons with recent and state-of-the-art adversarial training based defenses (Figure 4)) establishes the strength of the proposed defense.

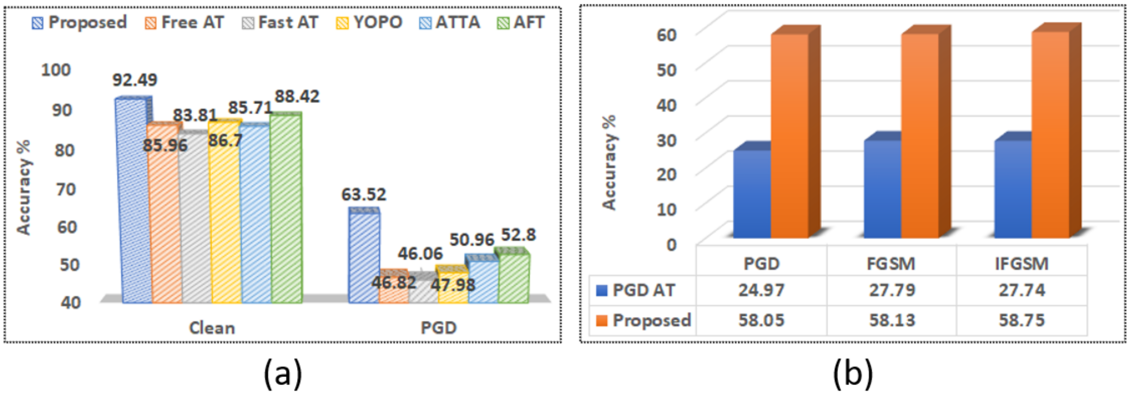

Figure 4. (a) Robustness (%) of the proposed defense on CIFAR10 database using Wide-ResNet along with comparison using Free AT [39], Fast AT [45], YOPO [48], ATTA [51], and AFT [28] using same attack parameters. (b) Robustness (%) on **ImageNet** database using VGG16 along with PGD adversarial training (AT) on three standard attacks.

Table 11. Ablation (%) study in terms of the number of epochs for proposed finetuning.

Database	Epochs					
	10	20	30	40	50	60
CIFAR	31.7	41.7	42.8	45.7	**52.1**	50.9
FMNIST	21.6	35.1	46.0	53.5	60.8	**67.8**
ImageNet	32.6	43.1	51.0	54.9	56.8	**58.1**

Fine-tune vs. Pre-train Networks using Proposed Augmentation: The proposed defense aims to improve the robustness of the pre-trained network through fine-tuning. We also showcase, the effect of proposed data augmentation both when used at the time of training the network from scratch or through fine-tuning (Table 10). Our results indicate that whether the proposed data augmentation either used for fine-tuning the pre-trained (already trained) networks or used to train the network from scratch shows higher adversarial robustness than undefended networks. The recent preliminary research [15] verifies our idea of fine-tuning for better robustness. However, the authors have not evaluated the re-trained model against white-box settings and multiple challenging adversaries. The proposed fine-tuning defense surpasses the above re-training defense [15] by a significant margin of at least 70%.

Effect of Fine-tuning Epochs: We have also performed an ablation study in terms of the number of epochs (Table 11). The higher the number of epochs the better the robustness of the augmentation. However, there is always a trade-off between robustness with increased epochs and computational cost.

6. Conclusion

The popularity and effectiveness of deep neural networks are continuously growing. Most of these highly accurate classifiers are trained on a large amount of data and for a couple of days to a couple of months; therefore, changing the classifier's structure completely or retraining from scratch for adversarial robustness might not be an effective solution. This research proposes a cost-effective fine-tuning-based defense against multiple gradient-based attacks in a white-box setting. Inspired by the adversarial modification at the local level of the images, local patch modification-based data augmentation is performed using several randomly selected functions. The proposed defense takes a few minutes (Table 8) only to boost the performance. The extensive experiments performed using multiple datasets of varying resolutions showcase the effectiveness of the proposed algorithm. The proposed fine-tuning can also surpass the existing defense algorithms that are computationally demanding. The proposed defense can be seen as a step towards robustness against adversarial attacks in a computationally feasible manner without modifying the classifier's underlying architecture.

References

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

[2] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Crafting adversarial perturbations via transformed image component swapping. *IEEE Transactions on Image Processing*, 31:7338–7349, 2022. 2

[3] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Exploring robustness connection between artificial and natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 179–186, 2022. 2

[4] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Benchmarking robustness beyond lp norm adversaries. In *Computer Vision–ECCV 2022 Workshops: Tel*

Aviv, Israel, October 23–27, 2022, Proceedings, Part I, pages 342–359. Springer, 2023. 2

[5] 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

[6] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Ibattack: Being cautious about data labels. *IEEE Transactions on Artificial Intelligence*, pages 1–10, 2022. 2

[7] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Noise is inside me! generating adversarial perturbations with noise derived from natural filters. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2020. 3

[8] 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. 3

[9] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Parameter agnostic stacked wavelet transformer for detecting singularities. *Information Fusion*, 95:415–425, 2023. 2

[10] Ahmed Aldahdooh, Wassim Hamidouche, Sid Ahmed Fezza, and Olivier Déforges. Adversarial example detection for dnn models: A review and experimental comparison. *Artificial Intelligence Review*, 55(6):4403–4462, 2022. 2

[11] Divyam Anshuman, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Wavetransform: Crafting adversarial examples via input decomposition. *IEEE European Conference on Computer Vision Workshop*, 2020. 2, 3

[12] 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, 2018. 2, 7

[13] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 4

[14] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. *AAAI Conference on Artificial Intelligence*, 2018. 2

[15] Ting-Wu Chin, Cha Zhang, and Diana Marculescu. Renofeation: A simple transfer learning method for improved adversarial robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3243–3252, 2021. 2, 8

[16] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 6

[17] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020. 7

[18] Bill Dally. Efficient methods and hardware for deep learning. In *Proceedings of the Workshop on Trends in Machine-Learning (and impact on computer architecture)*, page 1, 2017. 1

[19] 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. 1, 4, 6

[20] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. Breaking certified defenses: semantic adversarial examples with spoofed robustness certificates. In *International Conference on Learning Representations*, 2020. 2

[21] Akhil Goel, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. DNDNet: Reconfiguring CNN for adversarial robustness. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2020. 2

[22] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015. 2

[23] 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(6-7):719–742, 2019. 2

[24] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. *AAAI Conference on Artificial Intelligence*, 2018. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[26] 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. 6

[27] Yunseok Jang, Tianchen Zhao, Seunghoon Hong, and Honglak Lee. Adversarial defense via learning to generate diverse attacks. In *IEEE International Conference on Computer Vision*, pages 2740–2749, 2019. 6

[28] Ahmadreza Jeddi, Mohammad Javad Shafiee, and Alexander Wong. A simple fine-tuning is all you need: Towards robust deep learning via adversarial fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2021. 2, 7, 8

[29] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*, pages 2507–2515, 2018. 3

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[31] Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, and Nenghai Yu. Detection based defense against adversarial examples from the steganalysis point of view. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4825–4834, 2019. 2

[32] 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. 2

[33] 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, 4

[34] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *IEEE International Conference on Computer Vision*, pages 3385–3394, 2019. 4, 6

[35] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8571–8580, 2018. 3

[36] Han Qiu, Yi Zeng, Tianwei Zhang, Yong Jiang, and Meikang Qiu. Fencebox: A platform for defeating adversarial examples with data augmentation techniques. *arXiv preprint arXiv:2012.01701*, 2020. 2

[37] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *arXiv preprint arXiv:2003.01908*, 2020. 2

[38] Sanchari Sen, Balaraman Ravindran, and Anand Raghunathan. Empir: Ensembles of mixed precision deep networks for increased robustness against adversarial attacks. In *International Conference on Learning Representations*, 2020. 4, 5

[39] 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! In *Advances in Neural Information Processing Systems*, pages 3358–3369, 2019. 2, 8

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. 4

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 3

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014. 1, 2

[43] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. *International Conference on Machine Learning*, 2019. 6

[44] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 3, 4, 6

[45] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019. 8

[46] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 4

[47] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. *arXiv preprint arXiv:2004.05682*, 2020. 3

[48] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. 2019. 8

[49] H. Zhang, H. Chen, Z. Song, D. Boning, I. S Dhillon, and C. Hsieh. The limitations of adversarial training and the blind-spot attack. *International Conference on Learning Representations*, 2019. 2

[50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 6

[51] Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. Efficient adversarial training with transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1181–1190, 2020. 8

[52] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *AAAI conference on artificial intelligence*, 2020. 6