Pattern Recognition Letters 146 (2021) 244–251

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Cognitive data augmentation for adversarial defense via pixel masking

Akshay Agarwal[a,*], Mayank Vatsa[b], Richa Singh[b], Nalini Ratha[c]

[a] IIIT-Delhi, India
[b] IIT Jodhpur, India
[c] University at Buffalo-SUNY, USA

ARTICLE INFO

Article history:
Received 15 July 2020
Revised 14 December 2020
Accepted 21 January 2021
Available online 4 February 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Adversarial attacks
Deep learning
Data augmentation

ABSTRACT

The vulnerability of deep networks towards adversarial perturbations has motivated the researchers to design detection and mitigation algorithms. Inspired by the dropout and dropconnect algorithms as well as augmentation techniques, this paper presents "PixelMask" based data augmentation as an efficient method of reducing the sensitivity of convolutional neural networks (CNNs) towards adversarial attacks. In the proposed approach, samples generated using PixelMask are used as augmented data, which helps in learning robust CNN models. Experiments performed with multiple databases and architectures show that the proposed *PixelMask* based data augmentation approach improves the classification performance on adversarially perturbed images. The proposed defense mechanism can be applied effectively for different adversarial attacks and can easily be combined with any deep neural network (DNN) architecture to increase the robustness. The effectiveness of the proposed defense is demonstrated in gray-box, white-box, and unseen train-test attack scenarios. For example, on the CIFAR-10 database under adaptive attack (i.e., projected gradient descent), the proposed PixelMask is able to improve the recognition performance of CNN by at-least 22.69%. Another advantage of the proposed algorithm over several existing defense algorithms is that the proposed defense either is able to retain or increase the classification accuracy of clean examples.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Advancements in deep learning paradigm have seen several successful applications such as object/person recognition, autonomous driving, and image forensics [14,32]. However, several researchers have demonstrated their susceptibility to *adversarial noise based perturbations*. Generally, these noise/patterns are learned through a particular optimization function to minimize the performance of a base deep learning model. Existing adversarial noise generation algorithms perturb the images either by manipulating the pixels randomly in the entire image or at some specific locations [4,11,13,27,39]. Fig. 1(a) shows an example of an adversarial attack which is able to fool a CNN model.

One possible reason of singularities in convolutional neural networks is the high dependency on training data distribution, which might lead to overfittng and misclassifying out-of-distribution samples [19]. Adversarial perturbation attacks exploit this singularity and pose a challenge towards building generalizable deep learning models.

Several defense algorithms are proposed to mitigate the effect of adversarial perturbation and increase the robustness of deep learning models. Existing defense approaches can be broadly grouped into (i) external classifier based [1,19–21], (ii) input processing based [33,38,46], and (iii) network manipulation or retraining [9,10,12,17,35,44,49]. Among the existing defenses, adversarial training is one of the most robust approaches for seen attacks in the literature. However, the adversarial retraining-based defenses are ineffective against the unseen attack and raise privacy issue [30,50]. This technique also shows that retraining of network with some intelligently created noisy data can help in increasing the robustness against adversarial attacks. Furthermore, data augmentation techniques have shown effectiveness in increasing the robustness of deep learning models [36]. The details of existing algorithms can also be found in the survey papers [31,37,47].

Inspired by this, in this research, we propose a novel image augmentation approach termed as "PixelMask" to mitigate the effect of adversarial attacks and bridge the gap due to this singularity in deep learning models. Fig. 1(b) and 1(c) illustrate the proposed PixelMask based data augmentation for training a deep

* Corresponding author.
E-mail addresses: akshaya@iiitd.ac.in (A. Agarwal), nratha@buffalo.edu (N. Ratha).

https://doi.org/10.1016/j.patrec.2021.01.032
0167-8655/© 2021 Elsevier B.V. All rights reserved.

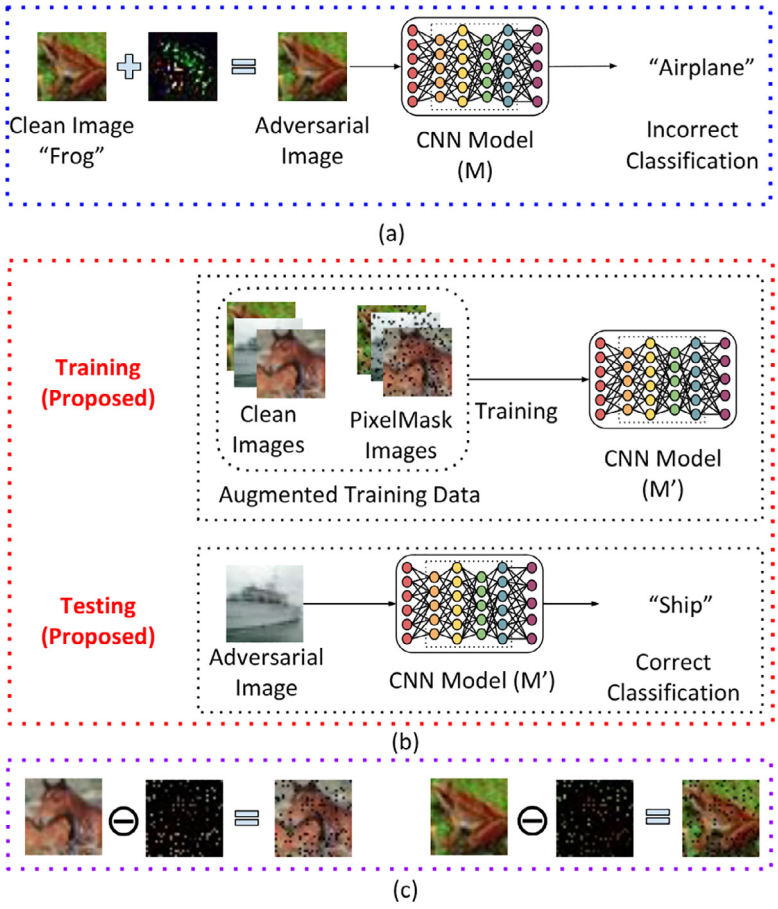

Fig. 1. (a) Traditional models are vulnerable to imperceptible adversarial attacks. (b) A CNN model trained using the proposed PixelMask based data augmentation. (c) PixelMask sample images (here, $\ominus$ represents the process of creating PixelMask images).

model. PixelMask is inspired by the concept of dropout and dropconnect, where the hidden variable/nodes or the connection between the nodes are masked with the motivation of increased robustness against missing or corrupted information. In the proposed PixelMask, a specific percentage of pixels of an input image are masked (or switched off) to create samples for data augmentation. Training with these augmented samples along with original (clean) samples helps in ensuring that the target model (i) does not entirely rely on the original or clean data and (ii) noisy/corrupted samples or samples with missing information help increase the robustness and generalizability of the CNN models.

The key highlights of this research can be summarized as follows:

- We propose PixelMask as a data augmentation technique which works by randomly dropping or masking a small percentage of the input image pixels to create "noisy or corrupted" samples,
- Extensive experiments on CIFAR-10 and CIFAR-100 databases are performed with different adversarial attacks, for gray-box, white-box, and unseen train-test attack scenarios, to evaluate the effectiveness of the proposed defense algorithm,
- Comparison with existing data augmentation techniques, including *mixup* [51], *manifold mixup* [42], *random erasing* [52], and *Cutout* [6], and defense by [16] show that the proposed PixelMask augmentation based defense outperforms other strong data augmentation techniques,
- The proposed PixelMask not only improves the performance on adversarial examples but also retains or improves the accuracy with clean images i.e., it can also act as general data augmentation technique for the CNN.

2. Proposed adversarial defense using pixelmask

Deep learning algorithms require a large amount of data for effective learning. However, the data used in training may not exhibit every properties of data captured in real-world scenarios. For example, the training data might be captured in a controlled environment whereas the testing data might be captured in a real-world, uncontrolled situation. These inconsistencies between the training and testing data lead to a difference in the training and testing accuracies. Data augmentation is a popular choice to address this problem of generalizability. Some popular data augmentation techniques include color transformation, geometric transformations, data mixup, adversarial training, generative adversarial networks, and noise data augmentation [29,36].

The motivation of the proposed approach is to reduce the dependency of the network on deterministic input information. The random dropping of pixels from input images makes the CNN model to look for non-deterministic information while learning parameters. The non-deterministic nature helps in ensuring that the network is not relying on a strong relationship between input image and output classification probabilities. It also makes attacking the models challenging because of difficulty in finding which image features to perturb for misclassification due to lower class-feature relations. The effect of adversarial perturbation after passing through random PixelMask should significantly be reduced.

In the proposed formulation of **PixelMask**, a random mask vector is generated containing the index of the pixel location in the 'input image' (i.e., layer 0).

$$y = f(W(m \star x)) \tag{1}$$

where, x is the input vector, and m is the random vector containing the pixel indexes. For masking or dropping $n\%$ pixels, randomly selected locations in m are set to 0, while other location values are set to 1. 0 denotes masking/dropping of a pixel at that location while 1 means preserve the pixel value at the index. The size of m and x is equal to the dimension of the input image.

The proposed PixelMask can be viewed as a regularization technique where some pixel information is suppressed while learning useful features. In the real world, while capturing and transferring digital images, some information may be corrupted or masked. Further, the information can intentionally be deleted/corrupted to fool the model. While it is hard to train a model with all possible variations of corruptions/noise, at the time of training a model, PixelMask approach can be used to augment the clean/original data with the corrupted samples. For data augmentation using PixelMask, augmented image set is generated after applying PixelMask with a specific n value (such as 5% or 10%) to all the training images. To increase the robustness of the CNN models, the models are trained using the training set augmented with PixelMask augmented set[1].

3. Experimental setup

This section describes the databases, attacks, and CNN models used for demonstrating the effectiveness of the proposed algorithm.

Databases: As suggested by Carlini and Wagner [2], the robustness of the defense algorithms must be evaluated on complex databases such as CIFAR [18]. Therefore, in this research, we have used the CIFAR-10 and CIFAR-100 databases for experiments. The CIFAR databases contain images of size $32 \times 32 \times 3$ pertaining to 10 and 100 object classes. Both the databases contain 50,000 training and 10,000 testing images. To train the classification model, the training set is utilized, and classification results are reported on the testing set. For defense training, data augmentation set of 50,000 images are generated and combined with 50,000 images from the original/clean training set. Testing is performed on 10,000 clean test images and 10,000 adversarial images (for each attack - discussed later in this section). Adversarial defense results are also reported on Fashion-MNIST database (i.e., grayscale images) [45] in the supplementary file.

CNNs: For classification, two different CNN models are used: (i) a CNN-10 with 3 blocks (base network) and (ii) ResNet-18 [15]. Based on the standard practice in literature [3,4,28,43], the base network is selected to show generalizability across models. The first and second blocks of the CNN-10 contain two convolutional layers where the second convolutional layer is followed by max-pooling and dropout of 0.25. The third block contains a dense layer followed by a dropout value of 0.5 and a final classification layer. The number of convolutional filters in the first and second blocks are 32 and 64, respectively. The dense layer contains 512 neurons, and the classification layer contains nodes equal to the number of classes in the database. The second model used for evaluation is ResNet-18 [15]. For each of the experiments, the CNN-10 is trained for 50 epochs while the ResNet model is trained for 20 epochs with 'Adam' optimizer and batch size is set to 16.

Adversarial Attack Algorithms [2] : The proposed defense algorithm is evaluated on the following attacks including fast gradient sign method (FGSM) [11], projected gradient descent (PGD) [22], iterative optimization-based attacks such as DeepFool [24] and universal adversarial perturbation [25].

Algorithms for comparison[3]: Several data augmentation based defense algorithms along with adversarial training based approaches are implemented for comparisons. A brief summary of each defense algorithm is described below:

- Adversarial training: [22] The adversarial images generated using the target model are augmented in the training set for retraining the network.

- Gaussian noise based data augmentation: Experiments with different variance values such as 0.001, 0.002, and 0.005 (for Gaussian noise) are added to the training images. In the experimental results, they are referred to as 001GN, 002GN, and 005GN.

- Enhanced image augmentation: Image quality plays a significant role in the classification of an image. Based on that, we have first enhanced the images using contrast stretching and sharpening operations. Later, these enhanced images are used for data augmentation for retraining the CNN model.

- Mixup: zhang et al. [51] and Verma et al. [42] have presented the mix-up based data augmentation techniques to smoothen out the decision boundary of the neural network.

- Random erasing and Cutout: Zhong et al. [52] have performed the data augmentation by replacing certain rectangular portion of an image either by mean value of ImageNet [5] or with random values. Cutout proposed by DeVries and Taylor [6] is very similar to random erasing approach.

Another closest defense from the proposed PixelMask defense is based on dropout. Feinman et al. [7] have proposed the defense based on dropping the nodes of CNN at the rate of 0.5 both during training and testing. However, the dropout defense is not shown to be effective on complex attacks such as optimization attacks [2] with challenging databases including CIFAR-10. Further, Tompson et al. [41] have shown that dropout itself is not effective when applied on convolutional layers.

4. Performance evaluation and analysis

PixelMask is evaluated on the CIFAR-10 and CIFAR-100 databases with multiple adversarial attacks in the following conditions:

1. **Gray-Box:** Assumes that the attacker might have access to the target model but does not know the defense strategy. In the proposed research, adversarial images in gray-box setting are generated using the vulnerable model itself;
2. **White-Box:** Assumes that the attacker has complete access to both the target model and defense algorithm. Hence, the adversarial images are generated from the defended model, i.e., in the case of the proposed defense, the PixelMask trained model is used for adversary generation;
3. **Unseen train-test:** Simulates the scenario where one type of adversarial images are used for training while testing is performed on another type of adversarial attack. The proposed defense does not utilize the attack information; hence, it is always unseen attack testing.

First, we discuss the results corresponding to the gray-box attack, followed by white-box attacks and unseen train-test. In this research, different parameters related to the FGSM, PGD, and UAP attacks are considered to evaluate the effectiveness of the proposed defense algorithms. The additional results with clean images are presented followed by the performance with varying drop rates. In the end, comparative results with popular data augmentation techniques followed by the analysis using the combination of PixelMask with other data augmentations are discussed.

[1] We will release the clear formatting code to the public or ones interested in comparison with the work.

[2] Adversarial robustness toolbox developed by [26] is used for adversarial examples generation.

[3] **Publicly available codes are used for implementation.**

Table 1
Classification accuracy (%) of **CNN-10** on the CIFAR-10 database under **gray-box setting**. The accuracy of original model on clean examples is 59.15%. Results with 5% PixelMask (best performing) are reported. Results of two best performing defense algorithms along with baseline, i.e., no defense are highlighted.

Attack	Parameters	Attacked -	Adver.	Defense via Data Augmentation					
		No Defense	Training	Contrast	Sharpen	001GN	002GN	005GN	PixelMask
FGSM	$\epsilon = 0.05$	**32.66**	**57.32**	51.17	51.70	51.85	52.20	51.31	**52.74**
	$\epsilon = 0.07$	**32.26**	**53.94**	45.02	44.49	45.02	44.20	41.74	**47.89**
	$\epsilon = 0.10$	**25.29**	**52.68**	43.76	43.96	45.42	44.17	40.33	**46.75**
UAP	δ=0.4, ϵ=0.10	**36.43**	**52.27**	30.89	40.47	35.48	40.24	36.79	**44.07**
	δ=0.4, ϵ=0.15	**28.45**	**47.46**	24.43	26.68	29.23	33.46	31.41	**36.61**
	δ=0.5, ϵ=0.10	**35.69**	**52.40**	39.59	43.29	41.15	42.56	43.24	**45.60**
	δ=0.5, ϵ=0.15	**29.52**	**49.28**	33.93	39.49	35.93	38.33	37.22	**36.80**
DeepFool	Default	**53.51**	**59.02**	55.15	55.81	56.92	56.66	54.46	**57.50**
PGD	$\epsilon = 0.05$, ϵ_{step}= 0.03	**27.57**	**56.08**	49.99	50.49	48.66	49.33	50.39	**50.26**
	$\epsilon = 0.03$, ϵ_{step}= 0.01	**24.41**	**56.19**	52.29	53.61	53.13	53.08	53.70	**54.61**

Table 2
Classification accuracy (%) of **CNN-10** on the CIFAR-100 database under **gray-box setting**. The accuracy of original model on clean examples is 25.03%. Results with 5% PixelMask (best performing) are reported. Results of two best performing defense algorithms along with baseline, i.e., no defense are highlighted.

Attack	Parameters	Attacked -	Adver.	Defense via Data Augmentation					
		No Defense	Training	Contrast	Sharpen	001GN	002GN	005GN	PixelMask
FGSM	$\epsilon = 0.05$	**14.89**	**28.01**	23.65	23.60	24.46	24.21	23.40	**25.36**
	$\epsilon = 0.07$	**11.67**	**25.39**	19.29	19.36	20.00	20.32	19.36	**20.45**
	$\epsilon = 0.10$	**09.03**	**23.66**	15.97	16.09	16.38	15.62	16.33	**17.24**
UAP	δ=0.4, ϵ=0.10	**17.39**	**21.21**	17.93	19.26	19.09	18.90	20.46	**22.27**
	δ=0.4, ϵ=0.15	**08.53**	**14.28**	11.35	11.66	11.47	10.79	11.34	**12.32**
	δ=0.5, ϵ=0.10	**16.98**	**21.48**	18.14	18.76	18.18	18.33	18.70	**19.61**
	δ=0.5, ϵ=0.15	**09.30**	**18.51**	12.92	14.43	12.92	12.53	13.14	**16.02**
DeepFool	Default	**22.65**	**30.14**	28.46	28.27	28.32	28.31	28.15	**29.81**
PGD	$\epsilon = 0.05$, ϵ_{step}= 0.03	**11.66**	**26.92**	21.15	21.12	21.23	20.74	20.80	**21.36**
	$\epsilon = 0.03$, ϵ_{step}= 0.01	**16.58**	**26.15**	23.33	22.83	24.09	23.82	23.52	**24.66**

4.1. Performance in different attack settings

Gray-Box: The gray-box attack being one of the most practical attack scenarios, is explored in detail on the both databases using each attack. The results with gray-box attack using the CNN-10 on CIFAR-10 and CIFAR-100 databases are shown in Tables 1 and 2. For the FGSM attack, ϵ values of 0.05, 0.07, and 0.10 are used to generate the adversarial examples. Original (non-perturbed) images yield an accuracy of 59.15%, which is reduced to 32.66% with FGSM attack (0.05), which is a drop of 27%. With the proposed defense algorithm of 5% PixelMask, the performance increases by at least 15% across all strengths of FGSM. In the same scenario, the adversarial training algorithm shows an improvement of 20%. The results are reported in Table 1. Gaussian noise based augmentation algorithm also increases the robustness of the CNN-10 model. Similar to FGSM, the PixelMask and image enhanced based defense increases the classification accuracy on the UAP and deepfool adversarial images (Table 1). On the CIFAR-100 database, for example, the FGSM attack with $\epsilon = 0.05$ reduces the accuracy of CNN-10 from 27.17% to 14.89% (Table 2). The 5% PixelMask improves the accuracy up to 25.36%, whereas, the adversarial training increases up to 28.01% (Table 2). Similarly, for other attacks such as FGSM, UAP, Deepfool, and PGD attack and their parametric variants, the PixelMask and image enhanced defense is found to be effective. The complex optimization based attack i.e., PGD with $\epsilon = 0.03$ and $\epsilon_{step} = 0.01$ reduces the recognition accuracy up to 24.41% from 57.04%. The proposed adversarial defense algorithm demonstrates improvement in the classification performance of 30.07%.

To show that the proposed defense can be combined with any CNN model, experiments with ResNet-18 model are also performed on both CIFAR-10 and CIFAR-100 databases. Results with FGSM attack on CIFAR-10 and CIFAR-100 databases are reported in Table 3. The performance with 5% PixelMask augmentation shows the robustness on both the databases under several adversarial perturbations. The proposed defense is also found significant for another adversarially vulnerable CNN architecture, namely wide-ResNet [48]. On challenging attacks such as DeepFool and C&W l_2 attack, the proposed PixelMask defense is highly effective. From the results, we can infer that the reduced dependency on the clean pixel values is vital for adversarial robustness. The adversarial robustness and vulnerability of the wide-ResNet model are reported in Table 4.

The success of the proposed PixelMask on the complex adversarial attacks such as C&W l_2 and DeepFool shows its efficacy in handling minute adversaries. Under gray-box setting, key observations are summarized as follows: **(i)** lower strength PixelMask yields greater boost in the performance as compared to higher strength PixelMask; **(ii)** the proposed PixelMask improves the performance on challenging iterative attack i.e., PGD by atleast 22.42% on the CIFAR-10 database (Table 1); **(iii)** on CIFAR-100 database under lower strength image-agnostic perturbation ($\epsilon = 0.4$, ϵ_{step}= 0.10), the proposed PixelMask even better than adversarial training (Table 2); **(iv)** in most of the cases, PixelMask either performs comparable or better than adversarial training and other data augmentation techniques; and **(v)** the significant drawbacks of adversarial training are lower generalizability [30], vulnerable to attacks [8,50], and open other serious threats such as privacy [23].

White-Box: The white-box attack is generally not feasible to obtain in the real world scenario; however, we have performed the experiments in such a situation where the attacker has complete access to the defense and target model and can use this information to craft the adversarial attack. The results corresponding to the FGSM, IFGSM, and PGD attacks are reported in Table 5. Apart from that, we have also performed the comparison with EMPIR by Sen et al. [34] and PGD adversarial training [22]. Adversarial training is considered one of the rigorous defense in the literature be-

Table 3
Classification accuracy (%) of the no defense and defended **ResNet-18** model using data augmentation through the proposed PixelMask under **gray-box setting** on the CIFAR-10 and CIFAR-100 databases. The accuracy of original model on CIFAR-10 and CIFAR-100 clean examples is 85.34% and 63.23%, respectively. Results with 5% PixelMask (best performing) are reported. Results of two best performing defense algorithms along with baseline, i.e., no defense are highlighted.

Database	Attack	Params.	Attacked - No Defense	Adver. Training	Defense via Data Augmentation					
					Contrast	Sharpen	001GN	002GN	005GN	PixelMask
CIFAR-10	FGSM	$\epsilon = 0.03$	**26.48**	**45.36**	36.82	38.82	38.12	37.78	36.93	**41.17**
	IFGSM	$\epsilon = 0.03$	**22.20**	**35.06**	36.82	37.57	37.72	36.87	36.19	**43.42**
	DeepFool	Default	**6.29**	**25.06**	16.82	17.76	17.20	16.74	16.92	**25.42**
	PGD	$\epsilon = 0.03$	**22.20**	**35.06**	26.49	26.71	27.97	28.23	26.19	**42.56**
	CW l_2	Default	**5.30**	**55.06**	46.29	47.70	47.37	46.56	46.89	**67.40**
CIFAR-100	FGSM	$\epsilon = 0.03$	**13.63**	**49.02**	35.07	35.45	36.11	35.99	37.03	**48.14**
	IFGSM	$\epsilon = 0.03$	**11.06**	**39.15**	24.71	25.52	26.88	25.06	27.34	**38.48**
	DeepFool	$\epsilon = 0.03$	**7.61**	**19.28**	14.11	16.27	15.18	14.06	16.92	**18.67**
	PGD	$\epsilon = 0.03$	**10.28**	**47.24**	33.10	34.27	34.71	35.63	35.34	**47.11**
	CW l_2	$\epsilon = 0.03$	**2.26**	**54.12**	44.45	46.27	43.16	47.71	46.44	**56.79**

Table 4
CIFAR-10 classification accuracy (%) of the original and proposed PixelMask defended **Wide-ResNet-40-10** under **'gray-box'** setting.

Attack	No- Defense	PixelMask Defense			
		5%	10%	15%	20%
PGD-40	**11.54**	18.90	24.31	**24.88**	24.60
DeepFool	**13.89**	50.60	51.97	**52.34**	52.07
FGSM	**10.89**	26.18	29.82	31.00	**32.33**
IFSM-40	**11.54**	18.91	23.89	**24.98**	23.89
C&W l_2	**11.70**	**77.40**	76.30	77.00	76.60

Table 5
Comparison of the proposed defense with existing (exist.) EMPIR [34] and PGD adversarial training (PGD Adv. Train) on CIFAR-10 classification accuracy (%) under **'white-box'** setting.

Algorithm		Clean	FGSM	IFGSM	PGD
No-defense		**74.54**	**10.28**	**11.97**	**10.69**
Exist.	EMPIR	72.56	20.45	24.59	13.55
	PGD Adv. Train	73.55	12.45	10.97	8.52
Proposed	PixelMask 5%	**72.86**	25.23	26.52	25.78
	PixelMask 10%	71.79	**26.58**	25.90	**27.36**
	PixelMask 15%	72.04	26.33	**26.77**	26.02
	PixelMask 20%	71.04	24.52	25.72	25.91

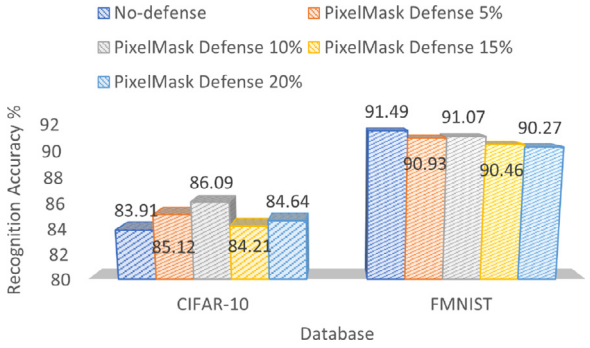

Fig. 2. Clean images accuracy (%) of CIFAR-10 using **VGG-16** and Fashion MNIST using **FMNISTconv**. The results are reported using no-defense, i.e., the original model and PixelMask defense models with varying drop %.

cause of availability of adversarial examples in the training set itself. The model, which shows 74.54% accuracy on the clean images of CIFAR-10 similar to previous deep models, is highly adversarially vulnerable. On each attack, the proposed PixelMask defense can surpass both EMPIR and adversarial training defense. The EMPIR defense is highly ineffective against PGD, commonly referred to as a first-order universal adversary. Even in such challenging scenarios, the proposed PixelMask yields performance at-least twice than the original undefended and defense models found using EMPIR and adversarial training. The results can be found in Table 5. The results under white-box setting using VGG-16 on CIFAR-10 and custom CNN on Fashion MNIST are also provided in the supplementary file.

Unseen train-test: In the complex scenario when different types of adversarial images are used in training and testing, the ineffectiveness of adversarial training defense is visible. For example, **(i)** when FGSM $\epsilon = 0.05$ based images are used for adversarial training, the accuracy on UAP attack with $\delta = 0.5$ and $\epsilon = 0.15$ increases by 0.23%, whereas the proposed PixelMask defense shows the improvement of 2.62%, **(ii)** when UAP with $\delta = 0.5$ and $\epsilon = 0.15$ is used for training against FGSM $\epsilon = 0.10$ attack, the adversarial training shows no improvement while the accuracy of PixelMask increase by 2.69%.

The above sets of experiments related to gray-box and white-box scenarios show the effectiveness of the proposed defense approach. In challenging conditions, the proposed defense performs better than adversarial training based defense. The critical drawback of adversarial training is lower generalizability against the unseen attack and its computationally intensive nature because of adversarial examples generation of each attack.

Accuracy on Clean Examples: It is interesting to note that the proposed defense can improve the classification performance on adversarial images and enhance the recognition accuracy of original/clean images. The performance of the proposed PixelMask with varying mask rate on clean images of CIFAR-10 and F-MNIST [45] are reported in Fig. 2. The 10% PixelMask can improve the recognition accuracy on CIFAR-10 from 83.91% to 86.09%. On the F-MNIST database, the proposed PixelMask can retain the accuracy of FMNIST_{conv} model on clean images through a slight drop of 0.42%. Similar performance improvement has been observed on the CIFAR-100 database. The recognition accuracy of the PixelMask retrained ResNet-18 model on clean images on CIFAR-100 database increases by 2.74% from the un-defended ResNet-18 model. On the other hand, it is also observed that adversarial training yields lower accuracy on clean images than PixelMask. This result demonstrates the PixelMask is robust against adversarial examples and it does not sacrifice the accuracy of clean test images.

4.2. Ablation with varying masking rates

We have performed the experiments with 2%, 3%, 5%, 10%, 15%, and 20% PixelMask. With 2% and 3% masking, we have not noticed significant robustness against adversarial attacks. In general, the best performance improvement is achieved at a lower Pixel-

Table 6
CIFAR-10 and Fashion MNIST classification accuracy (%) with varying PixelMask under **'gray-box'** and **'white-box'** setting.

CNN	Pixel Mask	PGD		FGSM		IFGSM	
		Gray	White	Gray	White	Gray	White
ResNet-50	5%	34.84	**30.82**	36.83	**31.27**	34.22	**31.07**
	10%	42.56	26.26	41.17	28.79	43.42	27.46
	15%	46.75	26.97	**44.25**	27.64	45.97	27.08
	20%	**48.00**	25.27	43.00	25.45	**47.58**	24.88
$CIFAR_{conv}$	5%	11.79	25.78	13.41	25.23	11.57	**26.77**
	10%	12.09	**27.36**	13.36	**26.58**	12.17	25.90
	15%	**12.89**	26.02	**14.08**	26.33	**12.58**	26.52
	20%	12.26	25.91	14.59	24.52	12.86	25.72
$FMNIST_{conv}$	5%	47.90	**86.41**	16.37	**77.43**	46.48	**82.08**
	10%	47.58	84.88	16.52	77.21	43.98	73.16
	15%	**48.59**	72.35	**17.13**	76.84	**47.67**	75.54
	20%	44.00	75.19	16.12	75.75	40.76	75.73

Mask rate in comparison to a higher under the white-box setting. In contrast, under the gray-box setting, the higher PixelMask rate performance is better than the lower PixelMask rate. We have not observed any significant improvement by going beyond 20% under the gray-box setting. However, the performance of the networks drops significantly under the white-box setting with a higher PixelMask rate. The ablation study on CIFAR-10 and F-MNIST with varying PixelMask under different adversarial attacks and classification networks are shown in Table 6. We have also observed that applying PixelMask in random locations yields higher performance than applying it in a specified location. For instance, when 5% PixelMask is applied at the center of the image, the performance is about 1% lower than applying 5% PixelMask at random locations.

4.3. Ablation with varying pixelmask augmented images

The proposed PixelMask defense performs the data augmentation for retraining the CNN models for possible robustness. For boosting, an equal number of images are generated with PixelMask as contained in the original training set of the database. Therefore, an ablation study has been conducted to check how many images are actually required to boost the adversarial performance of the models. The experiments are conducted using a PGD attack on two databases, namely CIFAR-10 using the VGG-16 model and Fashion MNIST (FMNIST) using $FMNIST_{conv}$. The results are reported in Fig. 3.

The original training set of CIFAR-10 consists of 50,000 images; hence the five-fold experiments are performed where, in each fold 10k images are added iteratively. Under both gray-box and white-box settings, as the number of augmentation images increases, the adversarial robustness also increases. Similar phenomena are observed on the FMNIST database, where 3 fold augmentation has been performed. It is observed that with the increase in the number of images for augmentation, the adversarial robustness performance of the networks improves. For example, the PixelMask trained model yields 6.74% better accuracy when 50k images are used for augmentation compared to 10k images.

4.4. Comparison with existing data augmentations

Recently, several data augmentation algorithms such as Mixup [51], Manifold Mixup (M-Mixup) [42], Random Erasing [52], and Cutout [6] are proposed. Originally, these algorithms are not evaluated with respect to adversarial attacks. In this paper, we have compared the proposed PixelMask approach with some existing augmentation techniques. Under the same experimental protocol, Table 7 reports the results on the CIFAR-10 database using CNN-10 model under **gray-box** setting. The results clearly show that the proposed PixelMask based augmentation algorithm yields higher performance and more resiliency towards adversarial attacks compared to existing data augmentation approaches. For example, on the FGSM attack (ϵ=0.05), *Mixup* and *Erasing* based data augmentation techniques are able to improve the performance by 7% and 6%, respectively. On the other hand, the proposed PixelMask based data augmentation improves the performance by 20%. As mentioned earlier, random erasing and cutout based data augmentation approaches are very similar and in our experiments, we have observed that cutout based approach yields slightly lower performance compared to random erasing based approach. For instance, under FGSM attack, cutout shows 0.2% lower classification accuracy compared to random erasing approach. In a similar fashion, the proposed PixelMask outperforms *Mixup* in white-box setting. The increment in classification accuracy using *Manifold Mixup* (which is a variant of *Mixup*) and Erasing based augmentation, is also lower compared to PixelMask in all the settings.

We have also compared the performance of PixelMask with recently proposed subsampled based network training [16] using PGD attack ($\epsilon = 0.03$). Hosseini et al. [16] have used a similar concept but not as a data augmentation technique; instead converted the original training images into subsampled images by dropping 90% pixel information. In the gray-box setting, the robust ResNet-18 model using PixelMask and [16] on CIFAR-10 yield 64.30% and 31.71% accuracy, respectively. On the other hand, in the white-box setting, the algorithm by [16] suffers drastically and the accuracy reduces to 8.41%, which is 26% lower than the PixelMask. Further, the proposed PixelMask improves the accuracy by 2.6% on clean images as well, while [16] suffers 13% drop.

We have also observed that by combining the PixelMask with adversarial training, the accuracy on CIFAR-100 and CIFAR-10 databases increases by $2-3\%$ and $1-2\%$ across attacks, respectively. However, no significant improvement has been noticed with Gaussian noise based data augmentation. We have tested PixelMask with different batch sizes (such as 16, 32, and 64) and different learning rates (such as 0.001 to 0.000001). However, no significant difference has been observed in the classification accuracy.

4.5. General observations

The problem of adversarial robustness can be seen from the point of view of the following limitations: (i) the deep neural network which is believed to be highly non-linear is linear in the local regions, (ii) overfitting towards the distribution of the training images [11,40]. The earlier research shows that the augmentation of training images with their corresponding adversarial counterparts can improve the adversarial robustness of the networks [22,35]. The proposed research aims to reduce the high depen-

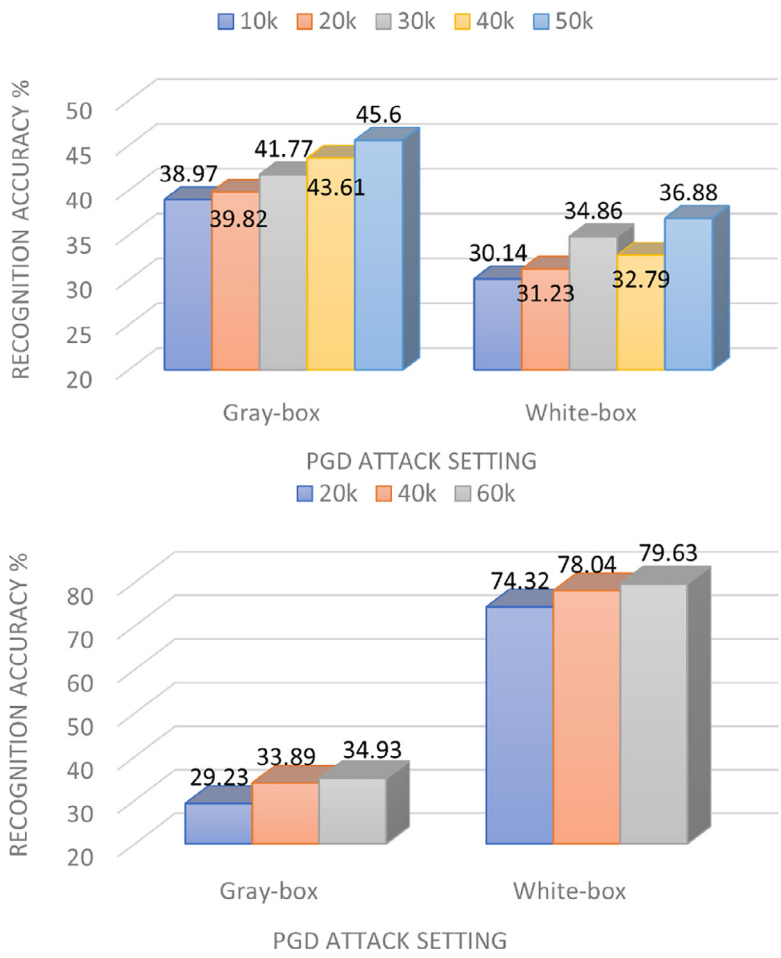

Fig. 3. PGD adversarial robustness performance (%) under varying amount of images used for augmentation. The results are reported using VGG-16 on CIFAR-10 (top) and FMNISTconv on Fashion MNIST (bottom).

Table 7
Comparing the proposed PixelMask adversarial mitigation performance with other data augmentation algorithms. The performance is reported on CIFAR-10 database using CNN-10 model under **gray-box** setting.

Attack Algorithm	Attacked -	Mitigation Accuracy (%) via			
	No Defense (%)	Mixup	M-Mixup	Erasing	PixelMask (5%)
FGSM ($\epsilon = 0.05$)	**32.66**	39.66	40.50	40.34	**52.70**
UAP ($\delta = 0.4$, $\epsilon = 0.10$)	**36.43**	37.23	39.20	39.33	**44.07**
DeepFool	**53.51**	56.22	56.66	53.72	**57.50**
PGD ($\epsilon = 0.03$, $\epsilon_{step} = 0.01$)	**24.41**	49.21	50.00	49.00	**54.48**

dency of the system on data by masking the regions in images. We assert that corrupting the pixel structure and making them unavailable at the time of training can boost the generalization of the network even if some pixels are perturbed. The results verify this phenomenon where pixels' corruption either through adversarial noise or masking helps improve the adversarial robustness better than other transformations.

5. Conclusion

The vulnerability of machine learning models demands practical defense algorithms that are not only able to protect from the current attacks but also future attacks. In this research, we proposed the defense based on PixelMask based data augmentation to secure the CNN model. The proposed defense does not utilize the knowledge of the adversarial attack or the target CNN model. Hence, it can be applied against different attacks to secure any CNN model. PixelMask defense is also compared with one of the most effective defense based on training the network with the adversarial images itself. The proposed and existing defenses are evaluated on two benchmark object recognition databases CIFAR-10 and CIFAR-100 under multiple adversarial attack algorithms. PixelMask performs better than the adversarial training based defense in challenging scenarios such as unseen training-testing attack conditions. While a significant drawback of the adversarial training based defense is the requirement of the knowledge of adversarial examples and ineffectiveness against unseen attacks, the proposed algorithm can increase the robustness of CNN models in the gray-box and white-box attacks. In the case of no adversary, the proposed PixelMask can act as a data augmentation technique.

Declaration of Competing Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgment

Akshay Agarwal is partly supported by the Visvesvaraya PhD Fellowship. Richa Singh and Mayank Vatsa were partially supported through a research grant from MeitY, India. Mayank Vatsa is also partially supported through Swarnajayanti Fellowship by the Government of India.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patrec.2021.01.032

References

[1] A. Agarwal, R. Singh, M. Vatsa, N. Ratha, Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? IEEE BTAS (2018) 1–7.
[2] N. Carlini, D. Wagner, Adversarial examples are not easily detected: Bypassing ten detection methods, in: AISec Workshop, 2017a, pp. 3–14.
[3] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: S&P, 2017b, pp. 39–57.
[4] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, C.-J. Hsieh, EAD: Elastic-net attacks to deep neural networks via adversarial examples, AAAI (2018) 10–17.
[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255.
[6] T. DeVries, G.W. Taylor, Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552 (2017).
[7] R. Feinman, R.R. Curtin, S. Shintre, A.B. Gardner, Detecting adversarial samples from artifacts, arXiv preprint arXiv:1703.00410 (2017).
[8] A. Galloway, T. Tanay, G.W. Taylor, Adversarial training versus weight decay, arXiv preprint arXiv:1804.03308 (2019).
[9] A. Goel, A. Agarwal, M. Vatsa, R. Singh, N. Ratha, Deepring: protecting deep neural network with blockchain, IEEE CVPRW (2019a).
[10] A. Goel, A. Agarwal, M. Vatsa, R. Singh, N. Ratha, Securing CNN model and biometric template using blockchain, IEEE BTAS (2019b) 1–6.
[11] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, ICLR (2015).
[12] G. Goswami, A. Agarwal, N. Ratha, R. Singh, M. Vatsa, Detecting and mitigating adversarial perturbations for robust face recognition, IJCV 127 (6–7) (2019) 719–742.
[13] G. Goswami, N. Ratha, A. Agarwal, R. Singh, M. Vatsa, Unravelling robustness of deep learning based face recognition against adversarial attacks, in: AAAI, 2018, pp. 6829–6836.
[14] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving, J. Field Rob. 37 (3) (2020) 362–386.
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[16] H. Hosseini, S. Kannan, R. Poovendran, Dropping pixels for adversarial robustness, in: IEEE CVPRW, 2019.
[17] X. Jia, X. Wei, X. Cao, H. Foroosh, Comdefend: An efficient image compression model to defend adversarial examples, in: IEEE CVPR, 2019, pp. 6084–6092.
[18] A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report, Citeseer, 2009.
[19] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, in: NIPS, 2018, pp. 7167–7177.
[20] J. Liu, W. Zhang, Y. Zhang, D. Hou, Y. Liu, H. Zha, N. Yu, Detection based defense against adversarial examples from the steganalysis point of view, in: IEEE CVPR, 2019, pp. 4825–4834.
[21] J. Lu, T. Issaranon, D. Forsyth, Safetynet: Detecting and rejecting adversarial examples robustly, in: IEEE ICCV, 2017, pp. 446–454.
[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, ICLR (2018).
[23] F.A. Mejia, P. Gamble, Z. Hampel-Arias, M. Lomnitz, N. Lopatina, L. Tindall, M.A. Barrios, Robust or private? adversarial training makes models more vulnerable to privacy attacks, arXiv print arXiv:1906.06449 (2019).
[24] S. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: A simple and accurate method to fool deep neural networks, in: CVPR, 2016, pp. 2574–2582.
[25] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, CVPR (2017) 1765–1773.
[26] M.-I. Nicolae, M. Sinn, M.N. Tran, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, B. Edwards, Adversarial robustness toolbox v0.8.0, CoRR 1807.01069 (2018).
[27] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: Euro S&P, 2016a, pp. 372–387.
[28] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in: S&P, 2016b, pp. 582–597.
[29] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, arXiv preprint arXiv:1712.04621 (2017).
[30] A. Raghunathan, S.M. Xie, F. Yang, J.C. Duchi, P. Liang, Adversarial training can hurt generalization, arXiv preprint arXiv:1906.06032 (2019).
[31] K. Ren, T. Zheng, Z. Qin, X. Liu, Adversarial attacks and defenses in deep learning, Engineering 6 (3) (2020) 346–360.
[32] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: IEEE ICCV, 2019, pp. 1–11.
[33] P. Samangouei, M. Kabkab, R. Chellappa, Defense-gan: protecting classifiers against adversarial attacks using generative models, ICLR (2018).
[34] S. Sen, B. Ravindran, A. Raghunathan, EMPIR: Ensembles of mixed precision deep networks for increased robustness against adversarial attacks, in: ICLR, 2020.
[35] A. Shafahi, M. Najibi, M.A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L.S. Davis, G. Taylor, T. Goldstein, Adversarial training for free!, in: NeurIPS, 2019, pp. 3353–3364.
[36] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 60.
[37] R. Singh, A. Agarwal, M. Singh, S. Nagpal, M. Vatsa, On the robustness of face recognition algorithms against attacks and bias, AAAI (2020).
[38] Y. Song, T. Kim, S. Nowozin, S. Ermon, N. Kushman, Pixeldefend: leveraging generative models to understand and defend against adversarial examples, ICLR (2017).
[39] J. Su, D.V. Vargas, K. Sakurai, One pixel attack for fooling deep neural networks, TEC 23 (5) (2019) 828–841.
[40] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, ICLR (2014).
[41] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: IEEE CVPR, 2015, pp. 648–656.
[42] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, A. Courville, D. Lopez–Paz, Y. Bengio, Manifold mixup: better representations by interpolating hidden states, ICML (2019) 6438–6447.
[43] X. Wang, S. Wang, P.-Y. Chen, Y. Wang, B. Kulis, X. Lin, P. Chin, Protecting neural networks with hierarchical random switching: towards better robustness-accuracy trade-off for stochastic defenses, IJCAI (2019) 6013–6019.
[44] E. Wong, L. Rice, J.Z. Kolter, Fast is better than free: revisiting adversarial training, arXiv preprint arXiv:2001.03994 (2020).
[45] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, arXiv:1708.07747 (2017).
[46] C. Xie, J. Wang, Z. Zhang, Z. Ren, A. Yuille, Mitigating adversarial effects through randomization, ICLR (2018).
[47] X. Yuan, P. He, Q. Zhu, X. Li, Adversarial examples: attacks and defenses for deep learning, TNNLS 30 (9) (2019) 2805–2824.
[48] S. Zagoruyko, N. Komodakis, Wide residual networks, arXiv preprint arXiv:1605.07146 (2016).
[49] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, B. Dong, You only propagate once: painless adversarial training using maximal principle, arXiv preprint arXiv:1905.00877 2 (2019a).
[50] H. Zhang, H. Chen, Z. Song, D. Boning, I.S. Dhillon, C.-J. Hsieh, The limitations of adversarial training and the blind-spot attack, ICLR (2019b).
[51] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: beyond empirical risk minimization, ICLR (2018).
[52] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, arXiv preprint arXiv:1708.04896 (2017).