Role of Optimizer on Network Fine-tuning for Adversarial Robustness (Student Abstract)

Akshay Agarwal[1], **Mayank Vatsa**[2], **Richa Singh**[2]
[1]IIIT-Delhi, India
[2]IIT Jodhpur, India
akshaya@iiitd.ac.in, {mvatsa, richa}@iitj.ac.in

Abstract

The solutions proposed in the literature for adversarial robustness are either not effective against the challenging gradient-based attacks or are computationally demanding, such as adversarial training. Adversarial training or network training based data augmentation shows the potential to increase the adversarial robustness. While the training seems compelling, it is not feasible for resource-constrained institutions, especially academia, to train the network from scratch multiple times. The two fold contributions are: (i) providing an effective solution against white-box adversarial attacks via network fine-tuning steps and (ii) observing the role of different optimizers towards robustness. Extensive experiments are performed on a range of databases, including Fashion-MNIST and a subset of ImageNet. It is found that the few steps of network fine-tuning effectively increases the robustness of both shallow and deep architectures. To know other interesting observations, especially regarding the role of the optimizer, refer to the paper.

Introduction

Training of deep neural networks (DNNs) requires a large number of computational resources and time; therefore, modifying the architecture or training them from scratch is not a feasible solution against adversarial perturbations (Goswami et al. 2019; Agarwal et al. 2020; Singh et al. 2020). This research aims to provide a solution for adversarial robustness without putting an extensive burden on the resource-constrained institutions. A novel data augmentation is performed by adding the random transformations within the local patches of the images. The modification of local patches is inspired by the findings, which show that individual neurons are sensitive towards specific parts of an image (Goodfellow, Shlens, and Szegedy 2015); additionally, adversarial perturbations modifies the images at a local level. Apart from that, the patch-based adversarial attacks also show the importance of local regions in the decision of the network (Yang et al. 2020). In brief, this research's contributions are: (i) a few-step network fine-tuning is proposed by augmenting the training set of a database with the locally modified images and (ii) comparison with complex defense algorithms including adversarial training.

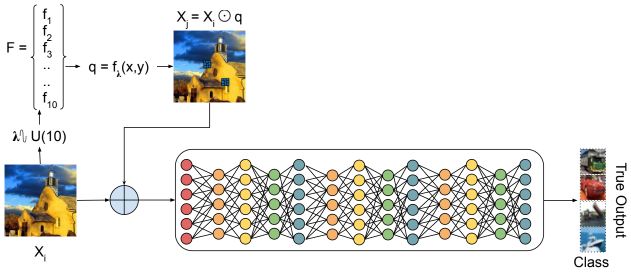

Figure 1: Proposed CNN fine-tuning with a randomized sampling-based data augmentation. λ is the random number $\mathcal{U} \in (1, K)$ uniformly generated to select the function to apply on an image patch. q is the randomized function selected from the pool of operations K based on λ. (x,y) indicating the coordinates in the patch in an image. $\odot$ is the elementary operation between image patch and q. $\oplus$ is the concatenation of original and modified input for fine-tuning.

Proposed Adversarial Robustness through Fine-tuning

In this research, to provide the adversarial robustness, fine-tuning is performed by inducing the randomness in the data. Generally, the deep classifiers are trained using a large amount of database and, therefore, share a strong dependency between the distribution of the training images. But when out of distribution data or digitally perturbed data comes, the network fails to behave as expected. In this research, we try to reduce this dependency between the input and output of the system. To do so, the local patches of an image are perturbed/modified using multiple functions randomly selected. The function chosen depicts the possible variation in the real-world testing images, including transformations such as rotation, translation, and flipping, noise, and occlusion via pixel masking.

As shown in Figure 1, patch(es) of size $w \times w$ is(are) selected around the center of an image. A function selected using the variable drawn from a uniform distribution is applied on the patch(es), and the modified image is combined with its clean image for network fine-tuning.

Attack	Parameters	Undefended	Proposed Defended	
			Adam	RMSProp
F-MNIST (Clean Accuracy = 91.49%)				
FGSM	$\epsilon = 0.3$	11.41	**87.51**	85.73
IFGSM	$\epsilon = 0.3, \alpha = 0.01$	09.85	**88.21**	84.38
PGD-10	$\epsilon = 0.3, \alpha = 0.01$	03.06	**53.72**	50.08
CIFAR-10 (Clean Accuracy = 83.91%)				
FGSM	$\epsilon = 0.03$	17.49	43.69	**65.93**
IFGSM	$\epsilon = 0.03, \alpha = 0.01$	14.70	39.21	**72.68**
PGD-10	$\epsilon = 0.03, \alpha = 0.01$	14.59	34.45	**60.37**
Imagenette (Clean Accuracy = 84.73%)				
FGSM	$\epsilon = 0.03$	17.47	32.23	**43.58**
IFGSM	$\epsilon = 0.03, \alpha = 0.01$	15.37	30.52	**43.97**
PGD-10	$\epsilon = 0.03, \alpha = 0.01$	14.60	29.76	**41.92**

Table 1: Performance (%) of undefended and multiple optimizer fine-tuned CNNs under *white-box* attacks.

Experimental Setup

We have used three databases namely: Fashion MNIST (F-MNIST) (Xiao, Rasul, and Vollgraf 2017), CIFAR-10 (Krizhevsky and Hinton 2009), and subset of Imagenet namely Imagenette[1]. This research aims to provide the robustness against challenging gradient-based attacks: (i) fast gradient sign method (FGSM) (Goodfellow, Shlens, and Szegedy 2015), (ii) iterative FGSM, and (iii) projected gradient descent (PGD) (Madry et al. 2018). The results of the proposed defense are reported using VGG-16 model (Simonyan and Zisserman 2015) for object databases. For F-MNIST, a custom CNN model of 3 *conv* layers, each followed by ReLU, is used for recognition. VGG and custom models are trained using an adaptive learning rate with an initial value of $1e^{-4}$ and $1e^{-3}$, respectively. Whereas the fine-tuning is performed using Adam and RMSProp optimizer with a fixed learning rate of $1e^{-4}$ for VGG and $1e^{-3}$ for custom CNN. The VGG model is trained for 200 epochs and took around 20 hours on a 1080ti GPU machine. In comparison, fine-tuning is performed for 20 minutes.

Experimental Results and Analysis

The results of the adversarial sensitivity and robustness of each database are reported in Table 1. On each database, proposed fine-tuning can increase the robustness of CNNs against each attack. The custom model's accuracy decreases from 91.49% to 3.06% when an iterative PGD attack is applied. When the network is fine-tuned using the proposed data augmentation scheme, accuracy improves significantly to 53.72%. Similar robustness has been observed on the CIFAR-10 and Imagenette database. The defended model fine-tuned using different optimizers (RMSProp) from the pre-trained model (Adam) shows higher robustness on the challenging object recognition database. The fine-tuned model is optimized with different hyper-parameters, such as batch size and learning rate. The finding indicates that for proposed fine-tuning, the exact parameters of the network are not needed. The comparison of the proposed fine-tuning has also been performed with adversarial training, and the results are reported in Table 2. The proposed fine-tuning based robustness surpasses the computationally expensive adversarial training with a significant margin on each attack. The proposed defense not only improves the adversarial robustness but also either retains or increases the performance on clean images, where the majority of the existing algorithms fail. On the CIFAR-10 database, the performance of VGG improves by 3%, whereas on the Imagenette database, it increases by 0.2%.

Attack	FGSM AT	IFGSM AT	PGD AT	Proposed
FGSM	39.96	43.67	36.30	**65.93**
IFGSM	35.08	44.64	34.63	**72.68**
PGD	32.12	34.31	37.93	**60.37**

Table 2: Comparison (%) of the proposed defense with adversarial training (AT) using VGG on CIFAR-10 database.

Conclusion and Future Work

In this research, we present a cost-effective network fine-tuning based defense for CNNs against multiple attacks. The findings show that we do not always need complex resource extensive algorithms for adversarial robustness. Rather, the networks can be fine-tuned by increasing the intelligent randomness without even knowing the hyper-parameters of the pre-trained undefended network. In the future, the selection of patches and transformations can be learned.

Acknowledgment

The research is partially supported by the Visvesvaraya Ph.D. Fellowship, Swarnajayanti Fellowship by the Government of India (GoI), and a research grant from MeitY, GoI.

References

Agarwal, A.; Vatsa, M.; Singh, R.; and Ratha, N. 2020. Noise is Inside Me! Generating Adversarial Perturbations with Noise Derived from Natural Filters. In *IEEE CVPRW*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *ICLR*, 1–11.

Goswami, G.; Agarwal, A.; Ratha, N.; Singh, R.; and Vatsa, M. 2019. Detecting and mitigating adversarial perturbations for robust face recognition. *IJCV* 127(6-7): 719–742.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Thesis, University of Toronto* 32–35.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 1–10.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional network for large-scale image recognition. In *ICLR*, 1–14.

Singh, R.; Agarwal, A.; Singh, M.; Nagpal, S.; and Vatsa, M. 2020. On the robustness of face recognition algorithms against attacks and bias. In *AAAI*, 3583–13589.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* .

Yang, C.; Kortylewski, A.; Xie, C.; Cao, Y.; and Yuille, A. 2020. PatchAttack: A Black-box Texture-based Attack with Reinforcement Learning. *arXiv preprint:2004.05682* .

[1] https://github.com/fastai/imagenette