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

Akshay Agarwal¹, Mayank Vatsa², Richa Singh²

¹IIIT-Delhi, India  
²IIT Jodhpur, India  
akshaya@iiit.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 increase 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 modify 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.

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.

Figure 1: Proposed CNN fine-tuning with a randomized sampling-based data augmentation. $\lambda$ 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 $\lambda$. $(x, y)$ indicating the coordinates in the patch in an image. $\oplus$ is the elementary operation between image patch and $q$. $\otimes$ is the concatenation of original and modified input for fine-tuning.
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), (ii) iterative FGSM, and (iii) projected gradient descent (PGD) (Madry et al. 2018). The models are trained using an adaptive learning rate with an initial value of \(10^{-3}\) for VGG and \(10^{-4}\) 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

<table>
<thead>
<tr>
<th>Attack</th>
<th>Parameters</th>
<th>Undefended</th>
<th>Proposed Defended</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-MNIST</td>
<td>$\epsilon = 0.3$, $\alpha = 0.01$</td>
<td>11.41</td>
<td>87.51</td>
</tr>
<tr>
<td>IFGSM</td>
<td>$\epsilon = 0.3$, $\alpha = 0.01$</td>
<td>39.96</td>
<td>58.21</td>
</tr>
<tr>
<td>PGD-10</td>
<td>$\epsilon = 0.3$, $\alpha = 0.01$</td>
<td>43.67</td>
<td>53.72</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>$\epsilon = 0.3$, $\alpha = 0.01$</td>
<td>44.64</td>
<td>53.72</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Attack</th>
<th>FGSM</th>
<th>IFGSM</th>
<th>PGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGSM</td>
<td>39.96</td>
<td>43.67</td>
<td>36.30</td>
</tr>
<tr>
<td>IFGSM</td>
<td>35.08</td>
<td>44.64</td>
<td>34.63</td>
</tr>
<tr>
<td>PGD</td>
<td>32.12</td>
<td>34.31</td>
<td>37.93</td>
</tr>
</tbody>
</table>

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

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\%.

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


