Effect of cascading attacks on an ensemble defense

Eiram Mahera Sheikh
Saarland University
Saarbrucken, Germany
eish00001@stud.uni-saarland.de

Shayari Bhattacharjee
Saarland University
Saarbrucken, Germany
shbh00002@stud.uni-saarland.de

Shravan Swaminathan
Saarland University
Saarbrucken, Germany
shsw00001@stud.uni-saarland.de

***Abstract*—Deep learning (DL) models are being widely adopted for security-sensitive applications like autonomous driving, facial recognition, etc. Exploring the vulnerability of such models have become an emergent topic. [4] has proposed a strategy involving an ensemble of substitute models for black-box attacks. They have also proposed a technique of augmenting the training data using perturbations generated by an ensemble of substitute models. [1] have empirically proved the effectiveness of this technique. However, we have observed that in all the related papers, researchers have only used a single attack method at a time. In this paper we introduce a new strategy that uses a cascade of attack methods to generate adversarial data. We demonstrate that our proposed technique leads to a stronger attack and defense.**

***Index Terms*—Adversarial Attacks, Adversarial Training, Black-box attack, Deep Learning**

I. Introduction

Deep Learning methods have revolutionized many domains in the present age. From autonomous vehicles to human robot cooperation projects which seemed quite far fetched a few years back, now seem quite possible and today, we can see various successful systems built with the backbone of neural network.

Research on the security and privacy aspects of Machine Learning (ML) models have attracted wide attention recently. [3] have comprehensively explored the threats for ML models. Within an adversarial framework, attacks can be classified as either white-box (internal knowledge of the model available) or black-box. More often, an attacker does not have knowledge of the model parameters and hence black-box attacks have become popular.

Although it is difficult for an attacker to acquire information on the decision boundary of target models, it has been proven that adversarial samples generated using a random model can be used to attack the target model. Most black-box adversaries leverage this transferability to craft adversarial samples using a substitute model. However, using a single substitute model is not very effective, an ensemble of substitutes enhances the transfer capability of perturbed samples, thus making the attack stronger. This technique can be extended to training the target model using data generated by an ensemble of substitutes.

Prior works have made use of a single attack method for crafting adversarial samples using an ensemble of substitutes. We perturbed images using a small perturbation budget (epsilon=0.1) by cascading them through four attack methods. We observe that these images have lesser noise than the ones generated using a larger perturbation budget (epsilon=0.4) with a single attack method. This leads us to believing that each attack method is optimizing over the output of previous one resulting in an image that can attack more strongly. We performed experiments to demonstrate that cascading attack methods are more powerful for black-box attacks. We also performed adversarial training on our target model by augmenting the dataset with adversarial images generated using cascading attack methods on an ensemble of substitute models.

The report has been organised as follows. The section II focuses on the preliminary background required for the implementation of the technique proposed. The proposed problem formulation is discussed in section III. Section IV focuses on the results of the attack and defense strategy and the paper is concluded by section V.

II. Preliminary Background

In this section, we briefly introduce the different attack methods that have been used in our experiments. We also give a short description of the adversarial training and attack technique employing an ensemble of substitutes. Refer the original papers for a more detailed understanding.

A. Adversarial Attacks

1) Fast Gradient Sign Method(FGSM): In this type of attack, the attacker uses the gradient on the input image to maximise the loss. It can be mathematically formalised as:

$$adv\_x = x + \epsilon * sgn(\nabla_x J(\theta, x, y))$$

where adv_x refers to adversarial image, x is original image, y is output label, ϵ is perturbation, θ is model parameters and J is the loss.

2) Projected Gradient Method(PGD): PGD is considered to be a white box attack. It is similar to FGSM and BIM , the difference lies in the fact that PGD initialises from a random point and does random restarts.

3) Carlini and Wagner Method(CW): In this type of attack, the adversarial samples are generated from the following optimisation constraint:

$$minimizeD(x, x + \delta)$$
$$suchthatC(x + \delta) = t$$
$$x + \delta \in [0, 1]^n$$

Here the D refers to distance metric (L_0, L_2 or L_∞) and C is the model being used.

4) Deepfool: This attack, focuses on the minimal perturbation required to fool the classifier by obtained the projection of the input data to its closest hyperplane by using the following equation:

$$\hat{r} = -\frac{f(x_0)}{||w||_2^2} * w$$

where x_0 is the input, classifier is f, w refers to the gradient and the $\hat{r}$ is the perturbation.

5) Fast Adaptive Boundary(FAB): This is also considered as an white-box attack.This attacks aims to modify the input by using L_p-norm based perturbation, where p refers to $p \in \{1, 2, \infty\}$

6) Basic Iterative Method(BIM): BIM is considered as an imporvement of FGSM. This suggests applying FGSM multiple times and is also called IFGSM.

$$X_0^{adv} = X$$

$$X_{N+1}^{adv} = Clip_{X,\epsilon}\{X_N^{adv} + \alpha sign(\nabla_X J(X_N^{adv}, y_{true}))\}$$

here, α is used as 1, ie only 1 pixel is changed per step.

B. Substitute Ensemble

Adversarial samples are generated using an ensemble of substitute models as depicted below:

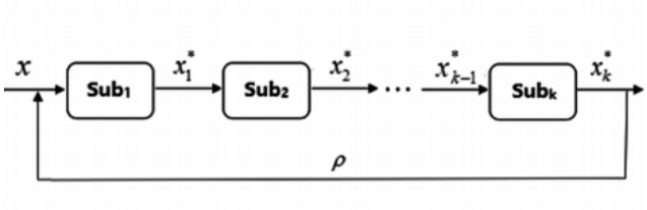

Fig. 1. Substitutes Ensemble Strategy

The strategy involves creating an ensemble of substitute models. The first model receives a clean image and each subsequent model receives a perturbed image from the previous model. Each one of them adds some noise to their input image using a simple single-step attack method and passes it on to the next model. Output from the last model is used as an adversarial sample.

1) Adversarial Attack: The adversarial samples generated using the above described procedure is used to attack the target model.

2) Adversarial Training: The adversarial samples generated using the above described procedure is added to the training data of the target model. The target model is then trained to make it more robust.

C. Neural Network Architectures

Neural Network Architectures plays a major role when it comes to evasion attacks, as the attacker tweaks various parameters during training to fool the system. In the proposed, the following types of neural network architectures are used for cascaded ensemble attack approach.

1) Multi-Layer Perceptron/ Feed Forward Network: This network uses two hidden layers and dropout which helps to avoid the cases of overfitting.

The first the image is of 784 tensors, which is then forwarded to the hidden layer, followed by dropout. The activation function of the model is relu.

The difference between the two networks lies in the hidden layer (512 and 500 respectively),batch size(10 and 100) and learning rate(0.01 and 0.001).

2) Resnet-50: This model utilises resnet-50 as a pretrained backbone, followed by 2d convolution with a stride=(2,2) and padding=(3,3) which is followed by a fully connected network which takes 2048 input and gives out 10 outputs.

III. Proposed Approach

In this section we describe our proposed technique. This section is divided into two subsections. The first subsection, represents the Cascading Attacks Methods where we discuss the strategy of using cascading attacks on ensemble of substitute models. Followed by that, we discuss about the adversarial training which utilises the cascading attacks strategy to get adversarial examples which are then used to train the classifier which improves the classification accuracy after attack.

A. Cascading Attack Methods

We choose a random set of simple attack methods. A clean image is passed through the first attack method which adds noise to it and outputs a perturbed image. This image is then passed through another attack method from the set, which adds more noise to it and outputs the resulting image. We repeat this procedure for all the methods in our set. The output image from the last method is collected as an adversarial sample.

Fig. 2. Cascading Attacks Methods

The figure 3 gives an illustration of our motivation for cascading an image through multiple attack methods. Using an epsilon value of 0.4 results in large perturbations. However, using an epsilon value of 0.1 and cascading the images through four different attack methods results in comparatively lesser perturbations.

B. The Attack Framework

Figure 4 resembles our proposed attack strategy. We use an ensemble of k substitute models and a set of n attack methods to generate adversarial images. These images are then used to attack the target model.

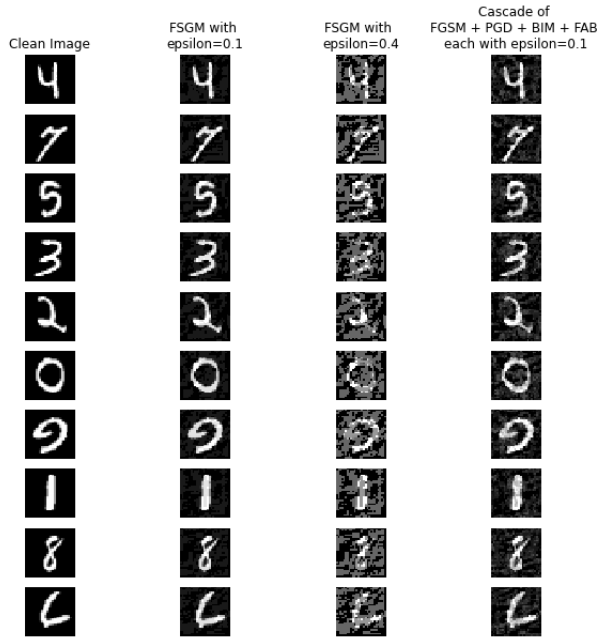

Fig. 3. Effect of cascaded attack methods

Fig. 4. Attack based on an ensemble of k substitute models using n attack methods

C. Adversarial Training

For the adversarial training of target model, we propose generating a set of adversarial images using the same strategy as described above. These images are then added to the training dataset for the target model and the model is trained to make it more robust against adversarial black-box attacks.

IV. Results and Discussion

In this section, we discuss the results we obtained by applying the proposed approach on MNIST dataset. All the computations for the attack part was performed on M1 Pro microprocessor with 16GB RAM and the defence part was performed on i7(8th gen) microprocessor with 16GB RAM. In this section, we first discuss about the dataset we used and the results for the attack followed by results for the adversarial training. The complete code repo can be found on $https://github.com/ESS-MLCS/Project$.

A. MNIST Dataset

MNIST[2] is a gray scale handwritten images data set with data labels from 0 to 9. It has a training set of 60,000 examples, and a test set of 10,000 examples. It was proposed by Lecun et al.

B. Preparing the models

The target model and all the substitute models were trained and tested on the MNIST dataset. Details related for their test accuracy is listed in the tables in below subsections.

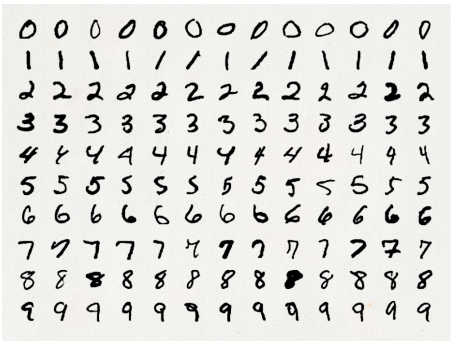

Fig. 5. MNIST Dataset

Model	Train Accuracy(%)	Test Accuracy(%)
Target Model	93.49	98.36
Multilayer Perceptron network	89.78	93.55
Feed-forward neural network	98.94	97.88
Resnet50	98.88	98.46

TABLE I
MNIST Classifier Values

C. Attack Results

In this subsection, we discuss the results we obtained after the cascaded ensemble attacks on a MNIST classifier.

First we visualise the effects of cascaded ensemble attack in Fig. 3, in comparison to the normal single attacks on MNIST dataset. We can visualise that when we use FGSM with $\epsilon = 0.1$ we don't observe much noise and the classifier accuracy remains satisfactory, however when we increase the perturbation $\epsilon = 0.14$, we notice a significant noise in the image and the classifier fails. However, when we use cascaded ensemble attack, we can observe that the noise is not visible, however the accuracy drops by a significant amount. Thus, these types of attacks are even difficult to be recognised by the humans, and these can pose problem specifically when in comes to applications in the safety-critical systems.

In the Table II, we visualise the difference in the test accuracy of various types of cascaded ensemble attack combinations. Out of all attacks, we can see that the combination Deepfool + FGSM + CW + PGD is the most effective attack which reduced the classifier accuracy drastically to 3.43% in comparison to single attacks like FGSM and PGD which still have the classification accuracy above 90%.

D. Adversarial Training Results

In this section, we used adversarial samples generated as described in Section III.C and trained the model with them to improve the performance of the system when attacked.

In Table III, the accuracy of the classifier is shown for various cascaded attacks when trained with a combination of FGSM, PGD, FAB, CW adversarial samples. We notice that the average improvement for all the types of attacks comes under the range of 3-23%, where the average improvement is about 13% which we consider to be quite significant.

Similarly in Table IV, we visualise the accuracy of classifier on various cascaded attacks when trained with a combination

Attack methods	Test Accuracy(%)
FGSM	91.26
PGD	92.30
Deepfool	97.79
CW	98.36
FGSM + PGD	36.06
Deepfool + FGSM	72.58
FGSM + CW	91.19
Deepfool + PGD	76.37
CW + PGD	92.16
Deepfool + CW	97.75
Deepfool + FGSM + PGD	3.77
FGSM + CW + PGD	35.43
Deepfool + FGSM + CW	72.56
Deepfool + CW + PGD	75.55
Deepfool + FGSM + CW + PGD	3.43

TABLE II
TEST ACCURACY OF THE TARGET MODEL WHEN ATTACKED WITH CASCADED ENSEMBLE ATTACKS

Attack methods	Test Accuracy(%)
FGSM	93.98
PGD	94.26
Deepfool	97.29
CW	96.97
FGSM + PGD	58.29
Deepfool + FGSM	88.73
FGSM + CW	93.98
Deepfool + PGD	89.20
CW + PGD	94.01
Deepfool + CW	97.26
Deepfool + FGSM + PGD	13.35
FGSM + CW + PGD	58.01
Deepfool + FGSM + CW	88.70
Deepfool + CW + PGD	88.98
Deepfool + FGSM + CW + PGD	13.25

TABLE III
PERFORMANCE OF THE TARGET MODEL AFTER ADVERSARIAL TRAINING USING AN ENSEMBLE OF MODELS AND A COMBINATION OF FGSM, PGD, FAB, CW ATTACKS

of FGSM, PGD, BIM, CW adversarial samples. Here we observe about 2-3% more improvement than the results with FGSM, PGD, FAB, CW adversarial samples in Table III.

In each of these cases, we can observe a significant improvement in the performance in comparison to the undefended model (Table II vs Table III and Table II vs Table IV). The lowest prediction accuracy obtained on the untrained model is 3.43% (Table II) and this attack method on the better adversarial trained model gives a 14.72% accuracy (Table IV). This improvement is quite significant however still renders the adversarial trained model unusable for many applications. Therefore, the proposed method of attack was still successful despite training the model with adversarial data.

V. CONCLUSION AND FUTURE SCOPE

In this report, we first propose a cascaded attack based on an ensemble of substitute models and prove that such an attack is successful in lowering the accuracy whilst keeping the data visually unperturbed as opposed to single attack methods with larger perturbations. Using such adversarial datasets to train the model to make it more robust against similar black-box

Attack methods	Test Accuracy(%)
FGSM	96.06
PGD	96.12
Deepfool	97.81
CW	97.41
FGSM + PGD	63.52
Deepfool + FGSM	93.58
FGSM + CW	96.08
Deepfool + PGD	93.62
CW + PGD	96.26
Deepfool + CW	97.76
Deepfool + FGSM + PGD	15.19
FGSM + CW + PGD	62.90
Deepfool + FGSM + CW	93.58
Deepfool + CW + PGD	93.69
Deepfool + FGSM + CW + PGD	14.72

TABLE IV
PERFORMANCE OF THE TARGET MODEL AFTER ADVERSARIAL TRAINING USING AN ENSEMBLE OF MODELS AND A COMBINATION OF FGSM, PGD, BIM, CW ATTACKS

attacks proved to be useful as shown in our results. For certain cases, we observed that the improvement after the adversarial training was not sufficient enough to mark the success of the defense strategy. Nevertheless, the visible improvement in the accuracy is a strong indicator of future exploration in this direction.

REFERENCES

[1] Jie Hang, KeJi Han, and Yun Li. Delving into diversity in substitute ensembles and transferability of adversarial examples. In *International Conference on Neural Information Processing*, pages 175–187. Springer, 2018.

[2] Y. LECUN. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*. URL `https://ci.nii.ac.jp/naid/10027939599/en/`.

[3] Nicolas Papernot, Patrick D. McDaniel, Arunesh Sinha, and Michael P. Wellman. Towards the science of security and privacy in machine learning. *CoRR*, abs/1611.03814, 2016. URL `http://arxiv.org/abs/1611.03814`.

[4] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.