ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

Comprehensive Review on Advanced Adversarial Attack and Defense Strategies in Deep Neural Network

Oliver Smith, Anderson Brown

University of Western Australia, Crawley WA 6009, Australia

DOI: **https://doi.org/10.51584/IJRIAS.2023.8418**

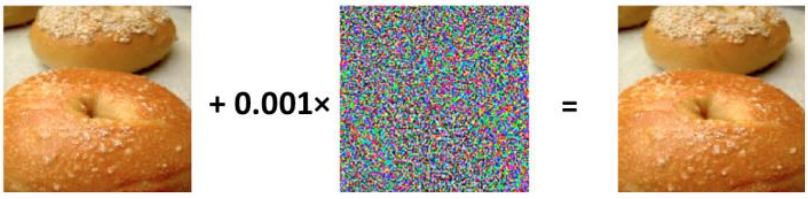
Received: 06 March 2023; Accepted: 16 March 2023; Published: 5 May 2023

Abstract. In adversarial machine learning, attackers add carefully crafted perturbations to input, where the perturbations are almost imperceptible to humans, but can cause models to make wrong predictions. In this paper, we did comprehensive review of some of the most recent research, advancement and discoveries on adversarial attack, adversarial sampling generation, the potency or effectiveness of each of the existing attack methods, we also did comprehensive review on some of the most recent research, advancement and discoveries on adversarial defense strategies, the effectiveness of each defense methods, and finally we did comparison on effectiveness and potency of different adversarial attack and defense methods. We came to conclusion that adversarial attack will mainly be blackbox for the foreseeable future since attacker has limited or no knowledge of gradient use for NN model, we also concluded that as dataset becomes more complex, so will be increase in demand for scalable adversarial defense strategy to mitigate or combat attack, and we strongly recommended that any neural network model with or without defense strategy should regularly be revisited, with the source code continuously updated at regular interval to check for any vulnerability against newer attack.

Keywords: Adversarial sampling, adversarial example, adversarial training, deep Neural Network, adversarial defense, neural network robustness

I. Introduction

It is now possible to achieve state-of-the-art performance in various artificial intelligence tasks such as speech translation, image classification, game, and machine translation [27],[28],[29]. Despite the magnitude of the success accomplished in the application of artificial intelligence for state-of-the-art performance, machine learning models remains vulnerable to adversarial attack. By simply adding perturbation to original input, the input becomes adversarial sampling which fool a classifier by making it to misclassify [25],[26] thereby causing machine learning model to make a wrong decision, though the input might still be the same to the human eye but the addition of perturbation makes it entirely something different to machine learning model and these ultimately cause the model to misclassify and eventually lead to misleading decision. Techniques to protect models against adversarial input are called adversarial defense methods. Inputs that make machine learning model to misclassify or make wrong prediction are called adversarial input or sampling. Hence, an adversarial example is an input to a machine learning model that is purposely designed to cause a model to make a mistake in its predictions despite resembling a valid input to a human.



In figure 1, the image on the left-hand side is the original image while the one on the right hand-side is the adversarial sampling which was generated by adding small perturbation to the original image. Although the two images look the same to human eye, feeding the image on the right hand-side (adversarial sample) to the classifier will make it to misclassify as a result of the addition of small imperceptible perturbation to it. The image on the left-hand side was correctly classify as bagel which is the ground truth while the image on the right-hand side falsely classifies as piano whereas the actual ground truth is bagel, in short. the essence of an adversarial sample is to fool the model to make it to misclassify. And so, the robustness of a classifier is a measure of its ability to withstand diverse adversarial attack. While neural network model is very much vulnerable to adversarial attach, the attach is not only peculiar with neural network as recent research had shown that other

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

models like tree-based model, and logistic regression are also vulnerable to adversarial attack [30].

Looking at the vulnerability of machine learning model to adversarial attack which makes it to misclassify and take wrong decision coupled with the application of such system in critical areas like healthcare systems, flight control systems, self-driving car, recommendation system, real estate, online banking, loan accessibility etc. which impacts our daily life. It is easier to know why serious attention is being put on adversarial attack and defense of machine learning model. Studies and existing works on adversarial machine learning are categorized into two; the first category is based on the efficiency and methods in which adversarial samples can be efficiently generated and then used to attack a model, this category is based on how adversarial samples can be effectively generated, different methods of generating them, the impact of the attack on the model and how they can be effectively used to attack a model. The second category is the defense against such attack, this is concerned with different strategies to defend against adversarial attack. Here comes the robustness of the machine learning model, as the defense of a neural network model to adversarial attack is a measure of its robustness.

In this paper, we did comprehensive review of some of the most recent research, advancement and discoveries on adversarial attack, adversarial sampling generation, the potency or effectiveness of each of the existing attack methods, we also did comprehensive review on some of the most recent research, advancement and discoveries on adversarial defense strategies, the effectiveness of each defense methods, and finally we did comparison on effectiveness and potency of different adversarial attack and defense methods. We also wrote some python code to carry out some of the existing strategies for adversarial attack and defense which we uploaded to GitHub to be publicly available for readers to explore. Because many of the review work on adversarial attack and defense was in the area of computer vision and natural language processing [36],[37],[38],[39],[40],[68] we based ours on neural network model which is known to be very vulnerable to adversarial attack and re used for state-of-the-art performance.

For this review, we use cifer-10 dataset which has 60,000 images (50,000 for training, 10,000 for validation) split into ten categories, with each category of training images containing 5000 images each and each category of validation images containing 1000 images each. We wrote our code in python to carry out each of the existing adversarial attack and defense methods on jupyter notebook idea, and make result comparation. Although some research work had been done on adversarial machine learning by using discrete data like discrete models, KNN classifiers, and natural language processing [31],[32],[33],[34],[35]. Some additional review work on adversarial attack and defense on machine learning had been done but from different perpective like Serban [36] reviewed was in the field of object detection (Computer vision), while others [37],[38],[39],[40],[66],[67] based there review on natural language processing, computer vision, but our comprehensive review is based on artificial neural network which is very vulnerable to adversarial attack and commonly use for state-of-the-art performance.

Adversarial Techniques

Adversarial attack can be referred to as processes which involves the generation of adversarial sample from original sample by addition of perturbation to the original sample before feeding it a model with the intension of making the model to misclassify (make the wrong decision) Figure 1.

Supposed we have a neural network classifier model such that it requires X_0 which is the original sample input for accurate prediction, in adversarial attack we can add a small perturbation δ to the original sample before feeding it to the classifier model. The presence of the small perturbation in the input will make the model to misclassify it. The new sample after the addition of small perturbation to the original image is adversarial sample such that:

$Y = X_0 + \delta$, such that

Y is the adversarial sample

X_0 is the original image

δ is the small perturbation

The newly generated adversarial sample will look the same to the human eye because the added perturbation is not visible to the human eye but the model will see it as something else due to the presence of perturbation and therefore classify it as something else thereby causing the model to make wrong decision. For an attacker who want to attack a deployed neural network model in production or development stage, there are two kinds of goal, namely: Targeted goal and untargeted goal.

For an attacker with targeted goal in mind, the attack will only have any chance of success if and only if the adversarial sample is classified as the targeted class. This implies that the attacker should have some insight about the model. let assume we are using cifar-10 dataset and the targeted class is ship, the attack will only be successful if the adversarial sample is labelled as ship else the

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

attack will fail to fool the model into misclassification. As for the untargeted attack, the attacker does not need any insight into how the model work, his aim is just to fool the model into misclassification, and in this case, the attack can be categorized as successful if and only if the model misclassifies it as something else. let assume, we are using a cifar-10 dataset which has 10 classes of images and generate adversarial sampling by adding small perturbation inform of noise to an original ship image, because it is an untargeted attack, the attack is successful if the model misclassifies it to be any of the other nine (9) classes of image except "ship" which is the original ground truth.

All the strategies and techniques for carrying out adversarial attack on a neural network model are group into three (3) categories, namely:

(1) gradient-based,

(2) score-based

(3) decision based.

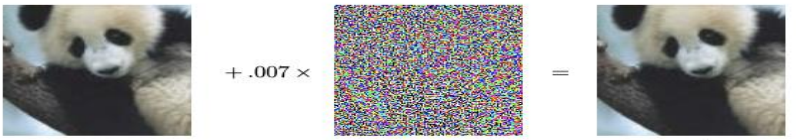
While it is true that all the above stated categories of attack can carry out both targeted and untargeted attack, any attack method will definitely fall into any of the three stated categories of attack, and can eventually lead into either white box attack or black box attack depending on the availability and level of information about the model such as model structure, parameter, weight, etc. that is available to the attacker or that the attacker have access to. It is Whitebox attack if the attacker have access to structure of the model, parameter, weight, label and so on but Blackbox attack if the attacker does not have access to any information about the model, in order word if the attack was carried out without the attacking having any insight about the model.

Gradient-Based Attack

In gradient based attack, we want to maximize the loss, and this can be achieved by leveraging on the loss with respect to the input. The higher we adjust the gradient such as to increase the loss, the more guarantee that we have that it will be misclassify and the attack will be successful. This is type of attack requires the attacker to know the gradient, and so it is white box attack. Most of the existing methods of adversarial attack falls into this category such as Fast Gradient Sign Method (FGSH), gradient-based evasion attack, projected gradient descent, and Calini and Wagner, and adversarial patch attack as will see in the next section.

Gradient-based evasion attack

In gradient base evasion attack, a perturbed image which seems like untampered to human eyes is made to be misclassified by neural network model.



We can carry out this type of attack by trial-and-error method as we don't know in advance, the exact data manipulation [65] that will break the model and make it to classify. Let say we want to probe the boundaries of a machine learning model designed to filter out spam emails, it is possible for us to experiment by sending different emails to see what gets through. And so, a model has been trained for certain words like "momentum", and now we want to make an to make exceptions for emails that contains other words, if we want to attack, we can craft email with enough extraneous words which will eventually make the model to misclassify it.

Fast Gradient Sign Method (FGSM)

Let assume we want to produce an adversarial sample x' = x + η such that x' is misclassified by the neural network. For us to make x' and x produce different outputs, η should be greater than the precision of the features. Let's represent pixel of an image by 8 bits, and we want any information below 1/255 of the dynamic range to be discarded. we can achieve this by η :

$$\boldsymbol{\eta} = \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)$$

Here, the input is perturbed with the gradient of the loss with respect to the input which gradually increases magnitude of the loss until the input is eventually misclassified.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

where J is the cost function used to train the neural network, θ is the parameter of NN model, x is the input to the model and and the target is represented by y which is associated with x. While ε decides both the size and sign of each and every element of the perturbation vector which might be matrix or tensor which are being determined by the sign of the input gradient. Here, we just have to linearize the cost function and find the perturbation that maximizes the cost subject to an L∞ constraint. This technique causes varieties of models to misclassify input and are also faster than other methods

Projected Gradient Descent (PGD)

PGD initializes the sample to a random point in the ball of interest which is being decided by the L∞ norm and does random restarts. This applies the same step as FGSM multiple times with a small step size while at the same time cliping the pixel values of intermediate results of each step to ensure that they are in an ε-neighbourhood of the original image.

Carlini and Wagner (C&W) attack

Berkeley, Nicholas Carlini and David Wagner in 2016 propose a faster and more robust method to generate adversarial examples.[11]The attack proposed by Carlini and Wagner begins with trying to solve a difficult non-linear optimization equation:

$$\min(||\delta||_p) \text{ subject to } C(x+\delta) = t, x+\delta \in [0,1]^n$$

However instead of directly the above equation, Carlini and Wagner propose using a new function f such that:

$$C(x+\delta) = t \iff f(x+\delta) \leq 0$$

This condenses the first equation to the two problems below:

$$\min(||\delta||_p) \text{ subject to } f(x+\delta) \leq 0, x+\delta \in [0,1]^n$$

$$\min(||\delta||_p + c \cdot f(x+\delta)), x+\delta \in [0,1]^n$$

Carlini and Wagner then propose the use of the below function in place of f using z, a function that determines class probabilities for given input x. When substituted in, this equation can be thought of as finding a target class that is more confident than the next likeliest class by some constant amount:

$$f(x) = ([\max_{i \neq t} Z(x)_i] - Z(x)_t)^+$$

With the use of stochastically gradient descent, we can use the above equation to produce a very strong adversarial sample especially when we compare it to fast gradient sign method which can effectively bypass a defensive distillation technique which was previously proposed for adversarial defense [11],[41],[42],[43].

Adversarial patch attack

Adversarial patch can be devised to fool a machine learning models. They work by causing physical obstruction in an image or by randomizing images with algorithm. Since computer vision models are trained on images that are straight forward. It is inevitable that any alteration to the input image can make the model to misclassify depending on the severity of the alteration. We could define a patch function p corresponding to every transformation t ∈ T which applies the transformed patch onto the image where Θ refers to the pixel-wise Hadamard product, and the final adversarial perturbed image $\hat{x}$ must satisfy $\hat{x} = p_t(x; ž)$ in order to trained patch ž and some t ∈ T.

For us in order to train patch ž, we could use a variant of the Expectation over Transformations (EOT) framework of Athalye et al.[26]. Let assume a family of transformations T, a distance metric d in the transformed space, and the objective is to find a perturbed image $\hat{x}$ satisfying the adversarial patch exploits the way machine learning model are trained for image classification by producing more salient inputs than real world objects. Such salient input are misclassify when feed to a machine learning model.

Scored-Based Attack

Score-based attack is based on the output score and so does not require access to gradient by the attacker. In 2017, Chen [12] proposed a method on how gradient could be estimated from the outputted score information and therefore create adversarial sample from the estimated gradient while in 2018, llyas et al. [44] leverages natural evolusionary strategy to estimate the gradient and generate adversarial example from the estimated gradient. Score -based attack can be further sub-categorised into gradient-

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

approximation based methods and Non-gradient approximation based method.

gradient-approximation based methods:

There are two stages when using this method, the first is the actual estimation of either the gradient of sign of the gradient which might be positive (+) or negative (-), while the second stage is the actual generation of adversarial sample from the estimated gradient or sign of the gradient (+/-). In 2017, Chen at al. [12] uses zero order optimization-based method to estimate the gradient of the loss with respect to the input, and then uses Calini and Wagner to generate adversarial sample from the estimated gradient. The time used to estimate the gradients grows along with the dimension, and so when the dimension of the input is large, series of techniques were introduced to scale-up the estimation. In addition, some other methods are introduced for efficient estimation of the gradient and generation of adversarial samples based on the efficiently estimated gradient as seen in NES attack [45] and Bandits Attack [46]

Non-gradient approximation-based method

This method does not utilize or rely on generation of adversarial sample from estimated gradient, as Li et al. [50] uses gaussian Blackbox adversarial attack to search for further adversarial samples after modelling the adversarial sample by Gaussian distribution, and this clearly confirm the possibility of getting several adversarial samples from the same input by different attack techniques. A good example is GenAttack by Alzantot et al. [47], Blackbox attack by Guo et al. [48], and square attack by Andriushchenko et al. [49].

Decision-Based Attack

In 2017, Papernot et al. [51] introduced a transfer method that requires just the observation of the label predicted by the model. The logic here is to train a substitute model with label similar to the target model. Rather than attacking the target model directly with trial and error, the substitute model is attack instead. For this method, the boundary attack was proposed by Brendel et al. in 2018 [52] which is based on the search for adversarial sample by random walking on the decision boundary.

There is also further extension to boundary attack, many pf which were proposed to increase the performance and efficiency of the attack, such as the work done by Brunner et al. in 2019, [53], Chen and Jordan in 2019, [54]; Chen et al. in 2020, [55] and Guo et al. in 2020, [56] all of which are transfer and random-walk based on the decision boundary. Other works on decision-based attack which are not transfer based were done by Ilyas et al. in 2018[44] Cheng et al. in 2020, [57] and Guo et al. in 2019, [58] as the goal of decision-based attack is to generate many adversarial samples from the predicted labels returned by the target model.

II. Current Defense Method:

Current adversarial defense methods are in three major categories:

1. Adversarial training

This is a brute force supervised learning method whereas many adversarial examples as possible are fed into the model and explicitly labeled as threatening. This is the same approach the typical antivirus software used on personal computers employs, with multiple updates every day. While quite effective, it requires continuous maintenance to stay abreast of new threats and also still suffers from the fundamental problem that it can only stop something that has already happened from occurring again.

2. Randomization

Many recent defenses resort to randomization schemes for mitigating the effects of adversarial perturbations in the input/feature domain [23]. The intuition behind this type of defense is that DNNs are always robust to random perturbations. A randomization-based defense attempts to randomize the adversarial effects into random effects, which are not a concern for most DNNs. Randomization-based defenses have achieved comparable performance under black-box and gray-box settings, but in the white-box setting, the EoT method [2] can compromise most of them by considering the randomization process in the attack process.

Below are the several typical randomization-based defenses and which introduces their performance against various attacks in different settings.

2.1. Random input transformation

Xie et al. [1] utilize two random transformations—random resizing and padding—to mitigate the adversarial effects at the inference time. Random resizing refers to resizing the input images to a random size before feeding them into neural network model while random padding refers to padding of zeros around the input images in a random manner [1]. This mechanism achieves a remarkable performance under black-box adversarial settings. However, under the white-box setting, this mechanism was compromised by the EoT method [2]. Specifically, by approximating the gradient using an ensemble of 30 randomly resized and padded images, EoT

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

can reduce the accuracy to 0 with 8/255 perturbations. In addition, Guo et al. [3] apply image transformations with randomness such as bit-depth reduction, JPEG compression, total variance minimization, and image quilting before feeding the image to a CNN. This defense method resists 60% of strong gray-box and 90% of strong black-box adversarial samples generated by a variety of major attack methods. However, it is also compromised by the EoT method [2].

2.2. Random noising

Liu et al. [4] proposed to defend adversarial perturbations by a random noising mechanism called random self-ensemble (RSE). RSE adds a noise layer before each convolution layer in both training and testing phases, and ensembles the prediction results over the random noises to stabilize the neural network's prediction. Lecuyer et al. [5] view the random noising defensive mechanism from the angle of differential privacy (DP) [6] and propose a DP-based defense called PixelDP. PixelDP incorporates a DP noising layer inside DNN to enforce DP bounds on the variation of the distribution over its predictions of the inputs with small, norm-based perturbations. PixelDP can be used to defend / attacks using Laplacian/Gaussian DP mechanisms, [7] further propose to directly add random noise to pixels of adversarial examples before classification, in order to eliminate the effects of adversarial perturbations. Following the theory of Rényi divergence, it proves that this simple method can upper-bound the size of the adversarial perturbation it is robust to.

2.3. Random feature pruning

Dhillon et al. [8] present a method called stochastic activation pruning (SAP) to protect pre-trained networks against adversarial samples by stochastically pruning a subset of the activations in each layer and preferentially retaining activations with larger magnitudes. After activation pruning, SAP scales up the surviving activations to normalize the inputs of each layer. However, on CIFAR-10, EoT [2] can also reduce the accuracy of SAP to 0 with 8/255 adversarial perturbations. Luo et al. [9] introduce a new CNN structure by randomly masking the feature maps output from the convolutional layers. By randomly masking the output features, each filter only extracts the features from partial positions. The authors claim that this assists the filters in learning features distributing consistently with respect to the mask pattern; hence, the CNN can capture more information on the spatial structures of local features.

III. Denoising

Previous works point out two directions to design such a defense: input denoising and feature denoising. The first direction attempts to partially or fully remove the adversarial perturbations from the inputs, and the second direction attempts to alleviate the effects of adversarial perturbations on high-level features learned by DNNs.

Below are some of the input denoising methods for adversarial defense

3.1. Conventional input rectification

In order to mitigate the adversarial effects, Xu et al. [10] first utilize two squeezing (denoising) methods—bit-reduction and image-blurring—to reduce the degrees of freedom and remove the adversarial perturbations, Adversarial sample detection is realized by comparing the model predictions on the original and squeezed images. If the original and squeezed inputs produce substantially different outputs from the model, the original input is likely to be an adversarial sample. Xu et al. [11] further show that the feature-squeezing methods proposed [10] can mitigate the C&W attack. However, He et al. [12] demonstrate that feature squeezing is still vulnerable to an adaptive knowledgeable adversary. It adopts the loss as the adversarial loss. After each step of the optimization procedure, an intermediate image is available from the optimizer. The reduced-color-depth version of this intermediate image is checked by the detection system proposed by Xu et al. [10]. Such an optimization procedure runs multiple times, and all the intermediate adversarial samples that can bypass Xu's system are aggregated. This whole adaptive attack can break the input squeezing system with perturbations much smaller than those claimed in [10]. Moreover, Sharma and Chen [13] also show that EAD can bypass the input squeezing system with increasing adversary strength.

3.2. GAN-based input cleansing

Defense-GAN and adversarial perturbation elimination GAN (APE-GAN) are two typical algorithms among all these works. Defense-GAN [14] trains a generator to model the distribution of benign images. In the testing stage, Defense-GAN cleanses an adversarial input by searching for an image close to the adversarial input in its learned distribution, and feed this benign image into the classifier. This strategy can be used to defend against various adversarial attacks. Currently, the most effective attack scheme against Defense-GAN is based on backward pass differentiable approximation [17], which can reduce its accuracy to 55% with 0.005 adversarial perturbations. APE-GAN [15] directly learns a generator to cleanse an adversarial sample by using it as input, and outputs a benign counterpart. Although APE-GAN achieves a good performance in the testbed [15], the adaptive white-box attack [16] can easily defeat APE-GAN.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

3.3. Auto encoder-based input denoising

MagNet is a defensive system which includes a detector and a reformer [18]. In MagNet, an auto-encoder is used to learn the manifold of benign samples. The detector distinguishes the benign and adversarial samples based on the relationships between those samples and the learned manifold. The reformer is designed to rectify the adversarial samples into benign samples. The effectiveness of MagNet against a variety of adversarial attacks under gray-box and black-box settings, such as FGSM, BIM, and C&W is already shown. However, Carlini and Wagner [24] demonstrate that MagNet is vulnerable to the transferable adversarial samples generated by the attack.

3.4. Feature denoising

Liao and Wagner [19] propose a high-level representation guided denoiser (HGD) to polish the features affected by the adversarial perturbations. Instead of denoising on the pixel level, HGD trains a denoising u-net [20] using a feature-level loss function to minimize the feature-level difference between benign and adversarial samples. Despite the effectiveness under black-box settings, HGD is compromised by a PGD adversary under a white-box setting [21]. [22] design a block for learning several denoising operations to rectify the features learned by intermediate layers in DNNs. The modified PGD adversarially trained network ranked first place in the adversarial defense track of CAAD 2018. Despite the remarkable success, the contribution of the feature-denoising block to network robustness is not compared with PGD adversarial training, since the PGD adversarial trained baseline can also achieve nearly 50% accuracy under white-box PGD attacks, and the denoising block only improves the accuracy of this baseline by 3%.

Experimentation

In this section, we did actual implementation of adversarial attack on deep neural network model in which each of Gradient-based: Fast Gradient Sign Method [26] and Projected Gradient Descent [59], Score-based: ZOO [12] and square attack [49], Decision-based: Boundary attack [52], OPT attack [61] and Sign-OPT [60] categories of adversarial attack were implemented in python using Jupiter notebook integrated development environment (IDE). We used the actual python implementation to show how various forms of adversarial attack fool DNN model and compare their result. For the experiment, we use cifar-10 dataset which contains 60,000 (50,000 training dataset and 10,000 validation dataset) 32 by 32-dimensional image spread across 10 classes of images in which each classis of image in the training dataset contains 5,000 images and each classis of images in the validation dataset contains 1000 images. We use the training images for our actual experiment and then use the validation datasets for validation and evaluation to arrive at our conclusion.

Attack Implementation on Neural Network Model

We developed and trained a neural network model from the scratch using different convolusional layer on cifar10 datasets, the model was trained across the 10 classes of images in cifar10 dataset with 50000 images for training dataset and 10000 images for validation dataset.

In calculating the loss, we used categorical cross entropy because the target classes are more than two

keras.losses.CategoricalCrossentropy(from_logits, label_smoothing, reduction, name="categorical_crossentropy",)

And then, we created a custom loss function and added it to the network by writing a function to compute the loss and passing the function as parameter in keras .compile method

```
from keras import Sequential

from keras.layers import Dense

model=Sequential ()

model.add (Dense (64, kernel_initializer='uniform',input_shape=(20,)))

model.compile (loss='categorical_crossentropy',optimizer='adam',activation='softmax')

def lossFunction(actual,prediction):

loss=(prediction-actual)* (prediction-actual) return loss

model.compile(loss=custom_loss_function,optimizer='adam')
```

When there was no attack, we got 98% for our prediction accuracy but in the presence of adversarial attack i.e., when we fed the model with adversarial sample, the accuracy got reduced to 33%. And then, we move ahead to compare the potency and effect of

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

the different attack method by using different method to form adversarial sample from the original image which was fed to the model. When forming our adversarial sample to attack the model, we use the following three forms of attack.

• Gradient-based: FGSM [26] and PGD [59].

• Score-based: ZOO [12] and Square attack [49].

• Decision-based: Boundary attack [52], OPT attack [61] and Sign-OPT [60].

The results of attacking the two classification models on CIFAR10 with gradient-based methods are shown in Table 1. -- the attack strength and its distortion level. Hence, as the value of -- increases, so si the attack strength, and the fact that PGD uses the process of optimization to search for attack while FGSM runs optimization for just one step makes PGD attack stringer than FGSM as seen in table 1.

Table 1: Comparison of Attack for Gradient-Based Attack.

Grad. Based Adversarial Method	Orig. Accuracy without Attack	€ = 0.01	€ = 0.02	€ = 0.03
FGSM	91.25	54.22	49.36	45.46
PGD	91.25	9.12	4.53	1.34

Our result for score-based attack method on CIFAR10 dataset are shown on Table 2, and while the maximum number of searches use for estimation of gradient is dependent on the maximum number of iterations, the more we increase the maximum number of iterations, the more accuracy our gradient estimation will be. Hence, increasing the number of maximum iterations gives a better performance. When we use the same number of iterations for both Square-based and ZOO attack, we found that the square based attack performed better than ZOO attack and this can be seen in table 2

Table 2: Comparison of Attack for Score-Based Attack.

Score-Based attack	Orig. Accuracy without Attack	Max.iter= 100	Max.iter= 200	Max.iter= 300
ZOO	94.51	93.24	90.41	75.36
Square	94.51	9.12	4.53	1.34

Experiment on Defense Method

We carried out the following adversarial defense method on our newly developed neural network model.

• Adversarial training: Madry's Adv [59] and TRADES [58].

• Projection: ER-CLA [62].

• Randomization: RSE [62] and Ran. Smooth [63].

We discovered that, as the strength of the attack increases when we increase the epsilon, the accuracy keeps on dropping except for randomization defense method for which the attack is evenly randomized. We tried to confirm if RSE and Madry's adversarial training

perform in a similar way, and to see if TRADES performs best when attack is strong as claimed by Yao et al [64] for which our experiment confirms their findings and result, we also find that while adversarial training is a good method of adversarial defense, combining it with ER-Classifier can increase the efficiency in defending against adversarial attack. We finally agreed that neural network model with defense strategy performs better and efficiently than a model without any defense strategy.

IV. Conclusion

That fact that, most of the time, the attacker does not know about the gradient of the model to attack, we can conclude that gradient or score-based attack method will not be feasible for an external attacker without knowing or the ability to estimate the gradient,

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

especially when we consider the fact that there is limitation to the number of queries an outside attacker can make without knowing the gradient. For this reason, we had come to the conclusion that adversarial attacker will mainly be Blackbox for the foreseeable future at least until when there is mechanism in which an outside attacker can estimate the gradient of the model before gradient-based or score-based attack could be feasible.

We also concluded that, as dataset becomes more complex, so will be the demand for efficient, and scalable adversarial defense method to combat or mitigate against the effect of adversarial attack on neural network model. Since, there is no perfect system, we are of the opinion that any adversarial defense system can be bypass, beaten, or become vulnerable to newer attack over time, and hence no adversarial defense system is perfect.

We strongly recommend that, for deployed neural network model in testing phase, development phase, or deployment phase with defense strategy whether randomization, adversarial training or projection. The model should regularly be revisited, with the source code continuously updated at regular interval to check for any vulnerability against newer attack.

References

1. Xie C, Wang J, Zhang Z, Ren Z, Yuille A. Mitigating adversarial effects through randomization. 2017. arXiv: 1711.01991.
2. Athalye A, Engstrom L, Ilya A, Kwok K. Synthesizing robust adversarial examples. 2017. arXiv:1707.07397.
3. Guo C, Rana M, Cisse M, van der Maaten L. Countering adversarial images using input transformations. 2017. arXiv: 1711.00117.
4. Liu X, Cheng M, Zhang H, Hsieh CJ. Towards robust neural networks via random self-ensemble. In: Proceedings of the 2018 European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany; 2018. p. 369–85.
5. Lecuyer M, Atlidakis V, Geambasu R, Hsu D, Jana S. Certified robustness to adversarial examples with differential privacy. 2018. arXiv:1802.03471v4.
6. Dwork C, Lei J. Differential privacy and robust statistics. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing; 2009 May 31– Jun 2; Bethesda, MD, USA; 2009. p. 371–80.
7. Li B, Chen C, Wang W, Carin L. Certified adversarial robustness with additive noise. 2018. arXiv: 1809.03113v6.
8. Dhillon GS, Azizzadenesheli K, Lipton ZC, Bernstein J, Kossaifi J, Khanna A, et al. Stochastic activation pruning for robust adversarial defense. 2018. arXiv: 1803.01442.
9. Luo T, Cai T, Zhang M, Chen S, Wang L. Random mask: towards robust convolutional neural networks. In: ICLR 2019 Conference; 2019 Apr 30; New Orleans, LA, USA; 2019.
10. Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks. 2017. arXiv: 1704.01155.
11. Xu W, Evans D, Qi Y. Feature squeezing mitigates and detects Carlini/Wagner adversarial examples. 2017. arXiv: 1705.10686.
12. He W, Wei J, Chen X, Carlini N, Song D. Adversarial example defenses: ensembles of weak defenses are not strong. 2017. arXiv: 1706.04701.
13. Sharma Y, Chen PY. Bypassing feature squeezing by increasing adversary strength. 2018. arXiv:1803.09868.
14. Samangouei P, Kabkab M, Chellappa R. Defense-GAN: protecting classifiers against adversarial attacks using generative models. 2018. arXiv:1805.06605.
15. Shen S, Jin G, Gao K, Zhang Y. APE-GAN: adversarial perturbation elimination with GAN. 2017. arXiv: 1707.05474.
16. Carlini N, Wagner D. MagNet and "efficient defenses against adversarial attacks" are not robust to adversarial examples. 2017. arXiv:1711.08478.
17. Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. 2018. arXiv:1802.00420.
18. Meng D, Chen H. MagNet: a two-pronged defense against adversarial examples. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security; 2017 Oct 30–Nov 3; New York, NY, USA; 2017. p. 135–47.
19. Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J. Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018. p. 1778–87.
20. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015 Oct 5–9; Munich, Germany; 2015. p. 234–41.
21. Athalye A, Carlini N. On the robustness of the CVPR 2018 white-box adversarial example defenses. 2018. arXiv:1804.03286.
22. Xie C, Wu Y, van der Maaten L, Yuille A, He K. Feature denoising for improving adversarial robustness. 2018.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

arXiv:1812.03411.
23. Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu, Adversarial Attacks and Defenses in Deep Learning, Engineering,Volume 6, Issue 3,2020,Pages 346-360,ISSN 2095-8099, https://doi.org/10.1016/j.eng.2019.12.012.
24. Rokach L. Decision forest: twenty years of research. Inf Fusion. 2016;27:111–25.
25. Biggio, B. et al. (2013). Evasion Attacks against Machine Learning at Test Time. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science (), vol 8190. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40994-3_25
26. Goodfellow I, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations, 2015.
27. Simonyan, K. and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
28. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016) Mastering the game of go with deep neural networks and tree search. Nature, 529, 484–489.
29. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.
30. Chen, H., Zhang, H., Si, S., Li, Y., Boning, D. and Hsieh, C.-J. (2019) Robustness verification of tree-based models. In Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc.
31. Jia, R. and Liang, P. (2017) Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2021–2031. Copenhagen, Denmark: Association for Computational Linguistics.
32. Samanta, S. and Mehta, S. (2017) Towards crafting text adversarial samples. arXiv preprint arXiv:1707.02812.
33. Gao, J., Lanchantin, J., Soffa, M. L. and Qi, Y. (2018) Black-box generation of adversarial text sequences to evade deep learning classifiers. In 2018 IEEE Security and Privacy Workshops (SPW), 50–56.
34. Cheng, M., Yi, J., Chen, P.-Y., Zhang, H. and Hsieh, C.-J. (2020b) Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 3601–3608.
35. Yang, P., Chen, J., Hsieh, C.-J., Wang, J.-L. and Jordan, M. I. (2020b) Greedy attack and Gumbel attack: Generating adversarial examples for discrete data. Journal of Machine Learning Research, 21, 1–36.
36. Serban, A., Poll, E. and Visser, J. (2020) Adversarial examples on object recognition: A comprehensive survey. ACM Computing Surveys (CSUR), 53, 1–38.
37. Qiu, S., Liu, Q., Zhou, S. and Wu, C. (2019) Review of artificial intelligence adversarial attack and defense technologies. Applied Sciences, 9, 909.
38. Yuan, X., He, P., Zhu, Q. and Li, X. (2019) Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems, 30, 2805–2824.
39. Ren, K., Zheng, T., Qin, Z. and Liu, X. (2020) Adversarial attacks and defenses in deep learning. Engineering, 6, 346–360.
40. Xu, H., Ma, Y., Liu, H.-C., Deb, D., Liu, H., Tang, J.-L. and Jain, A. K. (2020) Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing, 17, 151–178.
41. "carlini wagner attack". richardjordan.com. Retrieved 2021-10-23.
42. Plotz, Mike (2018-11-26). "Paper Summary: Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods". Medium. Retrieved 2021-10-23.
43. Wang, Xinran; Xiang, Yu; Gao, Jun; Ding, Jie (2020-09-13). "Information Laundering for Model Privacy". arXiv:2009.06112 [cs.CR].
44. Ilyas, A., Engstrom, L., Athalye, A. and Lin, J. (2018) Black-box adversarial attacks with limited queries and information. In International Conference on Machine Learning, 2137–2146.
45. Ilyas, A., Engstrom, L. and Madry, A. (2019) Prior convictions: Black-box adversarial attacks with bandits and priors. In International Conference on Learning Representations.
46. Jalal, A., Ilyas, A., Daskalakis, C. and Dimakis, A. G. (2017) The robust manifold defense: Adversarial training using generative models. arXiv preprint arXiv:1712.09196
47. Alzantot, M., Sharma, Y., Chakraborty, S. and Srivastava, M. (2018) Genattack: Practical blackbox attacks with gradient-free optimization. arXiv preprint arXiv:1805.11090.
48. Guo, C., Gardner, J., You, Y., Wilson, A. G. and Weinberger, K. (2019a) Simple black-box adversarial attacks. In

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS | Volume VIII Issue IV April 2023

Proceedings of the 36th International Conference on Machine Learning (eds. K. Chaudhuri and R. Salakhutdinov), vol. 97 of Proceedings of Machine Learning Research, 2484–2493. PMLR.

49. Andriushchenko, M., Croce, F., Flammarion, N. and Hein, M. (2020) Square attack: a queryefficient black-box adversarial attack via random search. In European Conference on Computer Vision, 484–501.
50. Li, Y., Li, L., Wang, L., Zhang, T. and Gong, B. (2019) Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In International Conference on Machine Learning, 3866–3876. PMLR.
51. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B. and Swami, A. (2017) Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 506–519.
52. Brendel, W., Rauber, J. and Bethge, M. (2018) Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. International Conference on Learning Representations.
53. Brunner, T., Diehl, F., Le, M. T. and Knoll, A. (2019) Guessing smart: Biased sampling for efficient black-box adversarial attacks. In Proceedings of the IEEE International Conference on Computer Vision, 4958–4966.
54. Chen, J. and Jordan, M. I. (2019) Boundary attack++: Query-efficient decision-based adversarial attack. arXiv preprint arXiv:1904.02144.
55. Chen, J., Jordan, M. I. and Wainwright, M. J. (2020) Hopskipjumpattack: A query-efficient decision-based attack. In 2020 IEEE Symposium on Security and Privacy (SP), 1277–1294.
56. Guo, C., Frank, J. S. and Weinberger, K. Q. (2020) Low frequency adversarial perturbation. In Proceedings of the 35th Uncertainty in Artificial Intelligence Conference (eds. R. P. Adams and V. Gogate), vol. 115 of Proceedings of Machine Learning Research, 1127–1137. Tel Aviv, Israel: PMLR.
57. Cheng, M., Yi, J., Chen, P.-Y., Zhang, H. and Hsieh, C.-J. (2020b) Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 3601–3608.
58. Guo, Y., Yan, Z. and Zhang, C. (2019b) Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. In Advances in Neural Information Processing Systems (eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alch´e-Buc, E. Fox and R. Garnett), vol. 32. Curran Associates, Inc.
59. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2018) Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations
60. Cheng, M., Singh, S., Chen, P. H., Chen, P.-Y., Liu, S. and Hsieh, C.-J. (2020a) Sign-opt: A queryefficient hard-label adversarial attack. In International Conference on Learning Representations.
61. Cheng, M., Le, T., Chen, P.-Y., Zhang, H., Yi, J. and Hsieh, C.-J. (2019a) Query-efficient hard-label black-box attack: An optimization-based approach. In International Conference on Learning Representations. URL https://openreview.net/forum?id=rJlk6iRqKX
62. Li, Y., Min, M. R., Yu, W., Hsieh, C.-J., Lee, T. C. M. and Kruus, E. (2018b) Optimal transport classifier: Defending against adversarial attacks by regularized deep embedding. arXiv preprint arXiv:1811.07950
63. Cohen, J., Rosenfeld, E. and Kolter, Z. (2019) Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning, 1310–1320.
64. Li, Y., Cheng, M., Hsieh, C., & Lee, T.C. (2021). A Review of Adversarial Attack and Defense for Classification Methods. The American Statistician, 76, 329 - 345.
65. Ige, T., & Adewale, S. (2022a). Implementation of data mining on a secure cloud computing over a web API using supervised machine learning algorithm. International Journal of Advanced Computer Science and Applications, 13(5), 1–4. https://doi.org/10.14569/IJACSA.2022.0130501
66. Ige, T., & Adewale, S. (2022b). AI powered anti-cyber bullying system using machine learning algorithm of multinomial naïve Bayes and optimized linear support vector machine. International Journal of Advanced Computer Science and Applications, 13(5), 5–9. https://doi.org/10.14569/IJACSA.2022.0130502
67. Ige, T., Kolade, A., Kolade, O. (2023). Enhancing Border Security and Countering Terrorism Through Computer Vision: A Field of Artificial Intelligence. In: Silhavy, R., Silhavy, P., Prokopova, Z. (eds) Data Science and Algorithms in Systems. CoMeSySo 2022. Lecture Notes in Networks and Systems, vol 597. Springer, Cham. https://doi.org/10.1007/978-3-031-21438-7_54
68. Tosin Ige, William Marfo, Justin Tonkinson, Sikiru Adewale and Bolanle Hafiz Matti, "Adversarial Sampling for Fairness Testing in Deep Neural Network" International Journal of Advanced Computer Science and Applications (IJACSA), 14(2), 2023. http://dx.doi.org/10.14569/IJACSA.2023.0140202