iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.59950

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Efficient Algorithmic Models for Improved Adversarial Attacks and Defenses in Deep Learning Based Image Recognition Models

Manasa V[1], Meenakshi T[2], Pratisha F[3], Rachana K[4], Sarala D V[5]
Department of Computer Science, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India

Abstract: ***This paper delves into the critical concern of ensuring the safety as well as the reliability of models for deep learning amidst the escalating landscape of adversarial attacks. Techniques like FGSM, DeepFool and PGD pose substantial threats by manipulating input data, leading to erroneous outcomes within machine learning systems. Addressing this challenge head-on, our study introduces an innovative model explicitly engineered to counteract these adversarial threats. Our model specifically focuses on combating the notorious attack algorithms of FGSM, DeepFool and PGD by implementing robust defense mechanisms such as Adversarial Training and GANs. Through meticulous evaluations spanning diverse datasets, including CIFAR and MNIST, our model's efficacy in defending against these sophisticated attacks was rigorously assessed. The empirical results underscore the resilience of our model, showcasing its effectiveness in fortifying deep learning frameworks against hostile intrusions across varied datasets. Our research contributes crucial insights and formidable defense mechanisms, augmenting the security, trust, and reliability of these systems, even amidst the complex challenges posed by adversarial manipulations.***

I. INTRODUCTION

Deep learning models have improved image recognition for applications like object detection and classification. Nevertheless, it has also been demonstrated that these models are susceptible to adversarial attacks, which use carefully crafted inputs to trick the model into making false predictions. The security of image recognition systems is put at risk by adversarial attacks, which may be used to incorrectly identify a variety of objects, manipulate traffic signs, or even cause autonomous vehicles to make dangerous decisions. In this work, we use several datasets and techniques to offer an overview of adversarial attacks and defenses for image recognition models.

Adversarial attacks are a technique used in Machine Learning to fool or misguide a model with malicious inputs. Black box and white box attacks are the major types of attacks. Black box attacks occur when the attackers do not have information about the targeted model and have no access to its architecture, parameters, or gradients. White box attacks occur when the attackers have all access to the targeted model and information of its architecture, parameters, and gradients as well. The Fast Gradient Sign Method (FGSM), Deep Fool,Projected Gradient Descent (PGD), and One Pixel Attack are a few examples of adversarial attacks.

An approach for defending machine learning models from adversarial attacks is called adversarial defense. The goal is to make a machine learning model more robust to adversarial attacks. Adversarial Training involves a process where we train a model on a combination of clean and adversarial perturbed data. This process helps the model learn to recognize adversarial examples during the training and improve its generalization to unseen data. Input preprocessing modifies the input data before it reaches the model to make it more robust to adversarial perturbations. Techniques include input normalization, randomization, and spatial transformations. To increase overall accuracy and resilience, the ensemble approach will attempt to integrate predictions from several models. Adversarial attacks are often model-specific, and an ensemble of diverse models may be more resistant to adversarial manipulation. Feature Squeezing reduces the precision of input features to make the model less sensitive to small perturbations. By quantizing or smoothing input features, the impact of adversarial noise can be minimized.

Adversarial attacks and defense mechanisms have many applications across the domains. The security of models is threatened by adversarial attacks, yet building strong defenses is essential to maintaining the system's integrity. Some applications include Image Recognition and classification systems with real-world implications, such as misclassifying objects in autonomous vehicles or security surveillance systems. Attacks on Natural Language Processing (NLP) models have the potential to tamper with spam detection or any other text-based application, hence influencing decision-making.

©IJRASET: All Rights are Reserved

In the field of healthcare, the attacks on medical imaging models could result in misdiagnosis, affecting the accuracy of disease detection in applications like radiology or pathology. Biometric systems such as facial recognition or fingerprint recognition can contribute to security in applications such as access control or authentication. The interplay between adversarial attacks and defenses presents a dynamic and evolving landscape, demanding continuous efforts to protect models or systems from malicious manipulation and make them accountable and effective across various domains.

II. LITERATURE REVIEW

The authors of this paper, Haibo Zhang et al.[1] address the pressing concern surrounding the vulnerability of deep learning models to adversarial attacks. To mitigate these threats, they propose a novel approach centered around pre-denoising all input images. This method involves adding a purification layer before the classification model, leveraging the fundamental architecture of Conditional Generative Adversarial Networks (cGANs).By integrating image perception loss into the Pix2pix algorithm, their technique achieves more effective image recovery, ensuring that noise-attacked images are restored to a level closely resembling the original, thereby upholding the accuracy of classification results. Their experimental findings showcase the efficacy of their approach, revealing a remarkable 20.22% increase in recovery accuracy compared to existing state-of-the-art methods. The proposed defense mechanism successfully recovers noisy images swiftly, safeguarding the integrity of classification processes within deep learning systems.

This research, conducted by the authors Dai et al.[2] confront the vulnerability of deep neural networks to imperceptible adversarial examples. Recognizing limitations in existing defensive methods like ComDefend, which rely heavily on large training dataset priors, they highlight the challenge of generalization with biased image statistics in adversarial examples. Motivated by the robustness of the deep image prior in capturing rich image statistics from individual images, the authors introduce DIPDefend. Leveraging a Deep Image Prior (DIP) generator, this method tailors defenses to fit individual adversarial inputs. Their findings reveal a learning pattern where early stages focus on robust feature acquisition against perturbations, followed by incorporation of non-robust features. Experimental results demonstrate DIP Defend's superiority over state-of-the-art defenses against both white-box and black-box adversarial attacks. Their adaptive strategy mitigates overfitting concerns, offering applicability across various scenarios, especially in real-world settings with limited labeled data.

Similarly the authors of this paper, Hwajung Yoo et al.[3] address the vulnerability of deep learning-based biometric authentication systems, particularly in fingerprint authentication, to adversarial attacks. They propose a novel defense method against adversarial fingerprint attacks. Their approach incorporates the Deep Image Prior (DIP) mechanism, known for its superior performance in reconstructing images without prior training or large datasets. Their method aims to eliminate adversarial perturbations within fingerprint images, reconstructing them to closely resemble the original fingerprint image by leveraging Deep Image Prior. Their results exhibit robust defense performance against diverse adversarial fingerprint attacks across various datasets encompassing different sensor variations, shapes, and materials of fingerprint images. Furthermore, their method outperforms other image reconstruction techniques, showcasing its efficacy in eliminating adversarial perturbations while intricately reconstructing fingerprint images without necessitating extensive training data.

The authors Pranpaveen Laykaviriyakul et al.[4] address the growing concern in deep learning, where well-designed adversarial samples deceive models without human perception, posing a significant threat to the reliability of these systems, especially in critical applications. To counter this issue, the authors propose a novel defense approach aimed at enhancing the robustness of deep learning models. Their strategy involves filtering perturbation noise in adversarial samples before prediction. They introduce a defense framework based on DiscoGANs, aiming to understand the interplay between attacker and defender characteristics. Attacker models generate adversarial samples from the training data, while the defender model reconstructs original samples from these adversarial samples. Both frameworks engage in a competitive training process to enhance defense capabilities. The proposed Collaborative Defense-GAN exhibits promising results, bolstering the robustness against both white-box and black-box attacks. It outperforms popular defense mechanisms on MNIST, fashion MNIST, and CIFAR 10 datasets, showcasing superior performance and resilience across various attack types and computational efficiency.

In the paper presented by Yamina Mohamed Ben Ali [5] the author proposes a novel approach that explores the use of the Smell Bees Optimization Algorithm (SBOA) to generate adversarial examples and evaluates their impact on deep learning networks. The methodology used was to deceive deep learning models by causing misclassifications and focuses on image classification using a new algorithm called StegFool using an optimization algorithm. The results demonstrated that it achieved satisfactory results in some cases but encountered challenges, particularly when dealing with complex deep learning networks.

In the paper written by Arka Ghosh, et al., [6] the authors have focused on the generation of adversarial perturbations for deep neural networks, making these attacks more practical and effective in real world applications. Methodology used was the DEceit algorithm which is designed for constructing effective universal pixel-restricted perturbations using black-box feedback from a target neural network. The outcome of this was that DEceit demonstrates its effectiveness by outperforming its competitors in terms of Fooling Rate.

Similarly, in the paper written by Tao Bai, et al., [7] they have elaborated on the vulnerability of deep learning image-based recognition systems, particularly when deployed on mobile devices. The problem addressed is patch-based attacks where a part of the original image is replaced with an adversarial patch to cause prediction errors in well-trained target models. GAN-based approach was used to generate inconspicuous adversarial patches (IAPs) with only one single image as the training data. The outcome of this paper was that IAP proved to be effective in reducing detection risks during qualitative and quantitative evaluations.

The authors Abebaw Alem et al.[8] considered evaluating and comparing deep learning models for LCLU classification in remote sensing imagery. The Deep Learning Models Performance Evaluations for Remote Sensed Image Classification which was done in the year 2022. They used Convolutional Neural Network Feature Extractor (CNN-FE) for developing a deep learning model from scratch and Transfer Learning (TL) for utilizing pre-trained deep learning models and fine-tuning them for LCLU classification. The final outcome was the fine-tuned deep learning model is expected to achieve profound accuracy results on the UCM dataset. This research will provide valuable insights into classification in remote sensing, environmental monitoring.

The authors Yi Ding, Fuyuan Tan et.al.[9] did research in order to focus on understanding the differences between adversarial and clean examples and exploring defense mechanisms. The CAM algorithm implements the global average pooling before the final output layer on the CNN architecture. It can enhance the visual explanation of the deep learning model. The adversarial training could enhance the classification accuracy and robustness of the original classification network. Manipulating pixel values in the salient regions resulted in significant fluctuations in the confidence values of clean examples.

The authors Jia Wang et al.[10] studied the principles behind adversarial examples which, crafted with small malicious perturbations, would mislead the deep neural network (DNN) model to output wrong prediction results. These small perturbations are imperceptible to humans. The existence of adversarial examples poses great threat to the robustness of DNN-based models. develop their countermeasures. This paper surveys and summarizes the recent advances in attack and defense methods extensively and in detail, analyzes and compares the pros and cons of various attack and defense schemes. Several attacks have been explored such as Attacks based on gradients like Fast gradient sign method (FGSM), Momentum iterative attack, Deepfool and Attacks based on optimization or Attacks by miscellaneous methods, attacks on applications. Several defense methods were researched also like Proactive defense: enhanced deep learning model, Randomization, Information masking, and Detection-based defense.

The authors Kui Ren et al.[11] describe a few research efforts on the defense techniques, which cover the broad frontier in the field. With the rapid developments of artificial intelligence (AI) and deep learning (DL) techniques, it is critical to ensure the security and robustness of the deployed algorithms. Recently, the security vulnerability of DL algorithms to adversarial samples has been widely recognized. The fabricated samples can lead to various misbehaviors of the DL models while being perceived as benign by humans. Successful implementations of adversarial attacks in real physical-world scenarios further demonstrate their practicality. Hence, adversarial attack and defense techniques have attracted increasing attention from both machine learning and security communities and have become a hot research topic in recent years. In this paper, the theoretical foundations, algorithms, and applications of adversarial attack techniques have been introduced.

The authors Samer Y. Khamaiseh et al.[12] have thoroughly reviewed the most recent and state-of-the-art adversarial attack methods by providing an in-depth analysis and explanation of the working process of these attacks. They have thoroughly reviewed the most recent and state-of-the-art adversarial attack methods by providing an in-depth analysis and explanation of the working process of these attacks. A comprehensive review of the most recent defense mechanisms and discuss their effectiveness in defending DNNs against adversarial attacks has been provided. There are two different attack scenarios, attacks during training phase and attacks during testing phase. Several attack strategies like White box attacks, Carlini and Wagner attacks are reviewed. Defense strategies like Modifications made to the ANN or to the training have been put into practice.

The authors Naveed Akhtar and Ajmal Mian[13] presented the first comprehensive survey on adversarial attacks on deep learning in computer vision. They have reviewed the works that design adversarial attacks, analyzed the existence of such attacks and proposed defenses against them. To emphasize that adversarial attacks are possible in practical conditions, they have separately reviewed the contributions that evaluate adversarial attacks in real-world scenarios. Finally, drawing on the reviewed literature, they provided a broader outlook of this research direction by exploring several adversarial attacks such as attacks for classification, FGSM, One pixel attack, etc. and even attacks in the real world like road sign attack.

©IJRASET: All Rights are Reserved

Several defense measures have been put into practice such as brute force adversarial training, defensive distillation, etc. Finally general observations regarding this research have been made such as that adversarial vulnerability is a general phenomenon.

The authors Hongsong Chen et al.[14], took FGSM and DeepFool methods for generating adversarial samples which are white-box attacks. Security Issues and Defensive Approaches in Deep Learning Frameworks in this they start with a description of the framework of deep learning algorithms and a detailed analysis of attacks and vulnerabilities in them. Future research focuses on the continuous evolution of attacks and defenses, the widespread existence of adversarial samples for improving robustness, the need for high parallel computing power in deep learning, and the challenges of translating well-performing neural networks from experimental to application stage due to the complexity of the real world.

The authors Ximeng Liu et al.[15], in this paper first describe the potential risks of DL and then reviewed the two types of attack: model extraction attack and model inversion attack in DL. Regarding the defense methods of security, we describe the defense approach from three aspects: pre-processing, improving model robustness, and malware detection. The rise of DL is to rely on the vast quantities of data, which is also accompanied by the risk of privacy leakage. DL has been extensively applied in a variety of application domains such as speech recognition, medical diagnosis, but the recent security and privacy issues of DL have raised concerns.

III. PRELIMINARY

A. Adversarial Attacks

Adversarial attacks is an attempt to misclassify the data by manipulating the input. The attack adds some noise to the input data called adversarial perturbations, and the sample obtained after adding noise is called adversarial sample. The inputs to a deep learning algorithm can be images, text, audio or numeric vectors. A variety of adversarial attack techniques exist, including DeepFool, Projected Gradient Descent (PGD), and the Fast Gradient Sign Method (FGSM). The most basic attack is called FGSM, and it may be used to change an image's pixel values. Through analysis of the decision boundaries and a search for the smallest modifications necessary for misclassification, DeepFool iteratively computes minimum perturbations. PGD maximizes the prediction error of the model by applying tiny adjustments repeatedly.

B. Datasets

MNIST and CIFAR-10 datasets are used in this paper. MNIST (Modified National Institute of Standards and Technology) is a dataset of handwritten digits(0-9) commonly used for digit recognition tasks with 60,000 training images and 10,000 testing images. MNIST dataset consists of images that are preprocessed and flattened into 784-dimensional vectors(28x28 pixels). Many machine learning libraries and frameworks such as TensorFlow provide built-in functions to download and load the MNIST dataset. CIFAR-10 (Canadian Institute For Advanced Research) consists of natural colored images of objects that belong to 10 different classes including airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. It consists of 60,000 32x32 colorful pictures total, with 6,000 pictures in each class. The CIFAR-10 dataset can also be accessed and downloaded using PyTorch's inbuilt function. CIFAR-10 poses greater challenges compared to MNIST due to colored images and complex objects. This allows us to assess the model's ability to handle increased complexity. Both the datasets are widely used in image classification, deep learning models.

C. Architectural Diagram

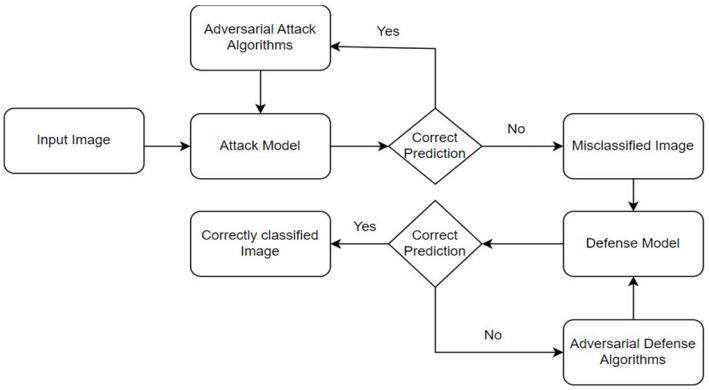


Figure 1. Architectural Diagram

D. Models

1) Convolutional Neural Network (CNN)

It is a special type of artificial neural network which accepts images as inputs. They are distinguished from other networks by their performance with image, speech or audio signal inputs. CNNs have three main layers. First layer is the convolutional layer, the core building block of CNN and it is in this layer that the majority of the computation happens. The input to the convolutional layer is input data and a filter. Filter is also known as kernel and it is a feature detector. Second layer is the pooling layer which performs dimensionality reduction thereby reducing the number of parameters in the input. There are two steps in pooling which can be applied in a mutually exclusive fashion i.e., max pooling and average pooling. Third layer is the fully connected layer which helps to perform the task of classification based on the feature extraction from the previous layers and filters. Convolutional layer and pooling layer apply ReLU (Rectified Linear Unit) activation function whereas the fully connected layer uses SoftMax activation function.

The idea behind activation functions is to introduce non linearity into the neural network so that it learns complex functions. ReLU is the most used activation function in deep learning models. It returns 0 if it receives any negative input and for any positive input it returns the original value. The SoftMax function is mostly used as the activation function in the output layer of a neural network, where it normalizes the output values to sum to one.

2) Extreme Gradient Boosting (XGBoost)

XGBoost has become one of the most popular machine learning algorithms due to its high performance and versatility. It is a type of ensemble learning method, which means it combines multiple weak learners to create a stronger learner. In XGBoost's case, the weak learners are decision trees, which are simple and interpretable models that can make predictions based on a series of splits in the data.

A technique called gradient boosting is used to train the ensemble of decision trees. Gradient boosting works by iteratively adding new decision trees to the ensemble, where each tree is trained to minimize the error of the previous trees. This process is repeated until the ensemble reaches a desired level of performance. XGBoost has several advantages over other machine learning algorithms like high performance, scalability, and interpretability.

3) Residual Network (ResNet)

Residual Neural Network (ResNet) is a deep learning architecture that introduced a novel concept of "skip connections" to address the vanishing gradient problem in deep neural networks. This technique allows the gradients to flow more efficiently through the network, enabling training of deeper and more accurate models. ResNet formulates the learning process as a residual function, where each layer learns the incremental change from the input to the output. This formulation simplifies the training process and encourages the network to focus on learning meaningful features rather than simply copying the input. It often employs gated activation units, such as ReLU (Rectified Linear Unit), to introduce non-linearity into the network. These activation functions allow the network to learn complex patterns and relationships in the data.

IV. PROPOSED METHODOLOGY

A. Adversarial Attacks

1) Fast Gradient Sign Method (FGSM)

To assess the durability of machine learning models, especially deep neural networks, adversarial examples are created using the Fast Gradient Sign method (FGSM). In 2014, Lan Goodfellow and his associates presented it. Making a perturbation to the input data that maximizes the model's undetectable misclassification is the main objective of support vector machine learning (FGSM). A demonstration of neural network M, and input information x. Let's let J be the negative work of the demonstration. The FGSM irritant (δ) will be calculated as $\delta = \text{sign}(\nabla x J(M, x, y)$ where:

ϵ = magnitude of perception,

$\nabla x J(M,x,y)$ = gradient loss function with respect input data x,

y = true label of the input x,

The adversarial example (xadv) is created by adding perturbation to the original input: $xadv=x+\delta$

2) DeepFool

An adversarial attack method called DeepFool aims to produce perturbations that trick deep neural networks. In contrast to other attack techniques, DeepFool does not depend on the neural network's gradient. Instead, by linearizing the decision boundaries, it determines the least perturbation needed to misclassify the model.

For a given input x, the decision boundary can be linearized using the first order Taylor expansion. For a single class classification problem, the linearization can be represented as follows: $f(x)=f(x0)+\nabla f(x0)T(x-x0)$ where:

f(x) = output of the neural network,

x0 = current input,

$\nabla f(x0)$ = gradient of the neural network's output with respect to the input at x0.

The perturbation (δ) is then calculated by finding the least L2-norm solution of the following linear equation: $\min\|\delta\|2$ subject to $f(x0)+\nabla f(x0)T\delta!=f(x')$where:

x′ = closest decision boundary point in terms of L2 norm,

$\nabla f(x0)$ = gradient of the neural network's output with respect to the input at x0,

An updated competitive example(xadv) is obtained by adding disturbance to the original input: $xadv=x+\delta$

These steps are repeated until a predetermined number of iterations is reached or the model misclassifies.

3) Projected Gradient Descent (PGD)

An adversarial attack known as a "projected gradient design attack" tries to create subtle alterations to the input data that lead a trained model to anticipate things incorrectly. It is an iterative process that involves calculating the loss function's gradient in relation to the input data again and taking smaller steps in the direction of a negative gradient. Until the adversarial example is produced or the maximum number of iterations is achieved, the process is repeated. The general formula for the L∞ norm, which is the most common norm used for PGD attacks

$x' = x + \alpha * sign(\nabla J(x))$ where,

x = original input data ,

x' = perturbed input data ,

α = step size ,

$\nabla J(x)$ = gradient of the loss function with respect to the input data ,

sign(x) = sign of the vector x

B. Adversarial Defenses

1) Adversarial Training

Adversarial examples are incorporated into the model's training process as part of the adversarial training technique, which defends against adversarial attacks. The model gains the ability to make more reliable predictions by being exposed to a variety of hostile cases. Using adversarial attack algorithms like PGD or FGSM, create adversarial examples for the target model. To the initial training model, add the adversarial examples that were produced in the previous phase. The model is trained using an augmented training dataset. Assess how resilient the trained model is against fresh adversarial samples.

2) Adversarial Perturbation Elimination - GAN (APE – GAN)

A generator and a discriminator are two competing neural networks that make up a GAN, a kind of artificial intelligence. The discriminator attempts to distinguish between created and actual data, while the generator produces realistic data that closely mimics the training data. In order to learn a mapping from the adversarial samples back to the original, clean inputs, APE-GAN uses a GAN. Using this mapping, the original input is restored after adversarial perturbations caused by noise are eliminated from the data.

V. RESULTS AND DISCUSSIONS

A. Performance Measurement

In our research we have made use of two datasets, MNIST and Cifar-10. Two models, CNN(Convolution Neural Networks) and XGBoost(Extreme Gradient Boosting), have been trained on the MNIST dataset consisting of handwritten images. We have made use of 12 CNN layers which has given an accuracy on the test data as 99.08% while XGBoost has given an accuracy of 95.93% on the same.

When the FGSM attack was applied to the models, the result was perturbed images that were crafted to deceive the model. The process involves taking the gradients of the loss with respect to the input, and then modifying the input in the direction that maximizes the loss. The images are perturbed such that the model misclassified the image while keeping the perturbation small and imperceptible to the human eye. On the other hand, in Deepfool attack, the data is distorted such that the perturbations will be entirely imperceptible. However sometimes, the perturbations might become visible to a human observer. Both attacks are successful in fooling both the XGBoost as well as CNN models but ultimately it is observed that the Deepfool attack is more effective at producing adversarial examples with fewer perturbations and higher misclassification rates. In order to counter these attacks, we implemented two defenses, Adversarial Training and APE-GAN. Post adversarial training, the accuracy was observed to have increased from 0.5856 to 0.9864 while the loss decreased from 4.1891 to 0.04627. Evaluating on adversarial examples: Loss: 7.177410225267522e-06, Accuracy: 1.0. After applying APE-GAN defense we observe that the GAN model has stabilized with low loss of 0.02343 and an accuracy of 1.0.

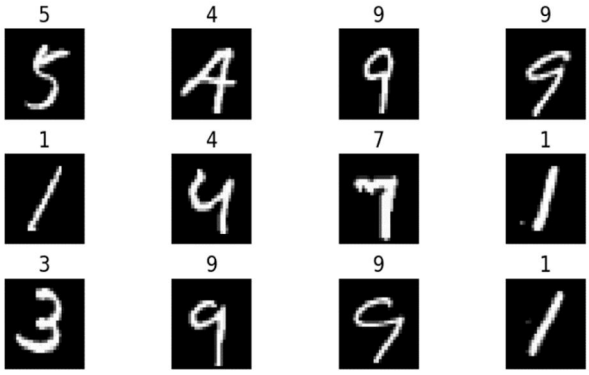


Figure 2. Images sample from MNIST dataset

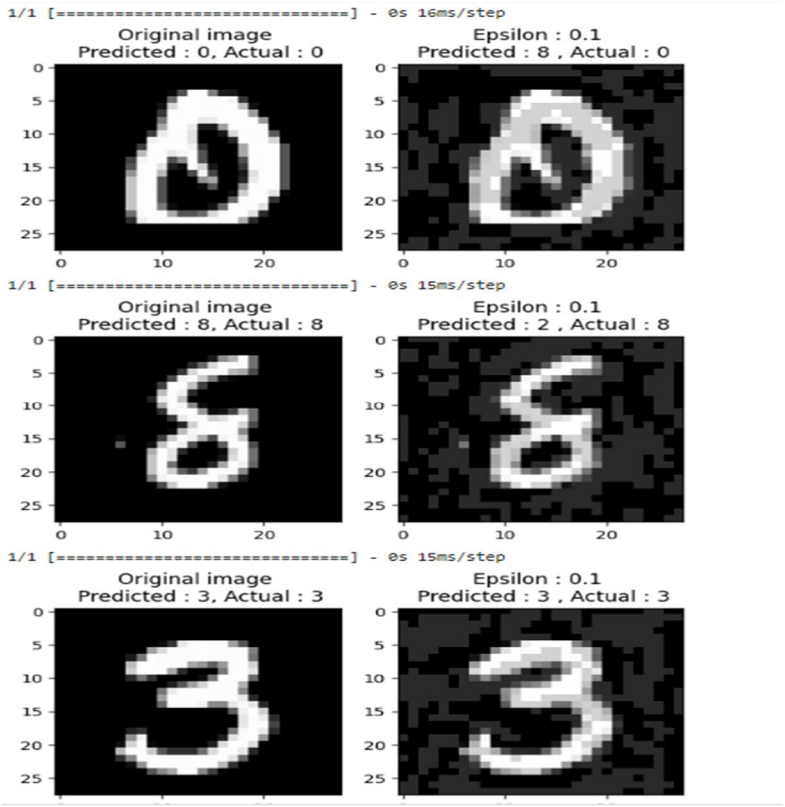


Figure 3. FGSM Attack on MNIST

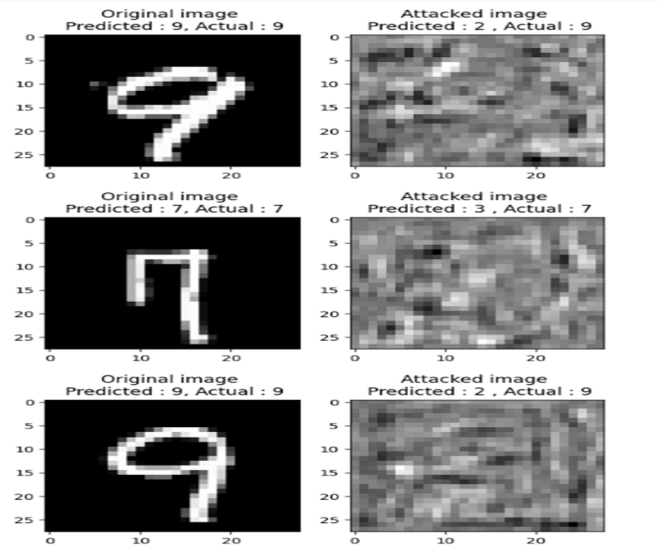
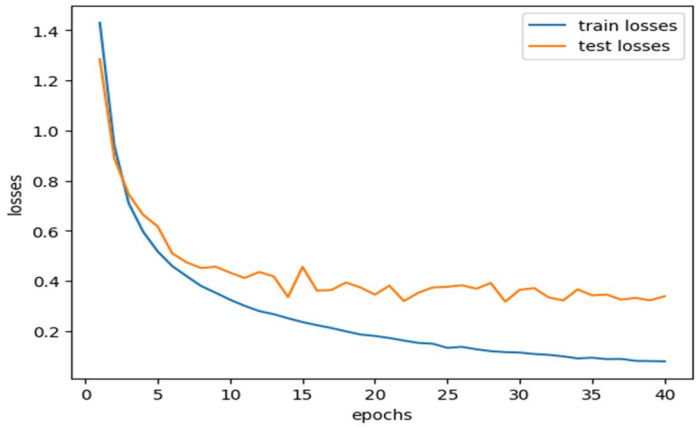


Figure 4. Deepfool Attack on MNIST

The resnet model has been trained on the Cifar-10 dataset where the accuracy of the network on the test images is 90%. When performing Adversarial Training with FGSM, Accuracy of the network on unperturbed test images is 89%.Training losses of the adversarially-trained model are higher than training losses of the naturally-trained model, which is intuitive since the adversarially-trained model is trained against adversarial examples, which makes it harder for the model to label these perturbed inputs correctly and results in higher errors. The loss of the naturally-trained model on test data is higher than the training loss, since test data is unseen by the model, resulting in higher error in classification. However, the loss of the adversarially-trained model on test data is lower than the corresponding training loss. This is probably because the test instances are not adversarial (in contrast to training data) and that the model has learned to extract important and useful features, thus performing better on test images. We see that epsilon and the accuracy are inversely proportional. If one increases, then the other will decrease. In targeted attacks, we want the model to misclassify its input to the given target class. Therefore, instead of just maximizing the loss of the true label, we maximize the loss of the true label and minimize the loss for the alternative label. We observe accuracy of naturally trained model against FGSM_targeted attack is 28.38% and accuracy of adversarially trained model against FGSM_targeted attack is 88.08%. The accuracy of the adversarially-trained model is much higher than the naturally-trained model since the adversarially-trained model has become more robust and therefore has higher accuracy against adversarial examples. In Adversarial Training with PGD accuracy of the network on the test images is 91%. The accuracy of net_pgd against PGD attack with iterations 3 is 90.36%, 7 is 88.85% and 12 is 87.81%., the accuracy decreases. Stronger attacks are produced, which will in turn make it harder for the model to label them correctly. Therefore, the accuracy of the model decreases.



Fi 5Figure 5. Training and test loss of Resnet18 model on CIFAR-10

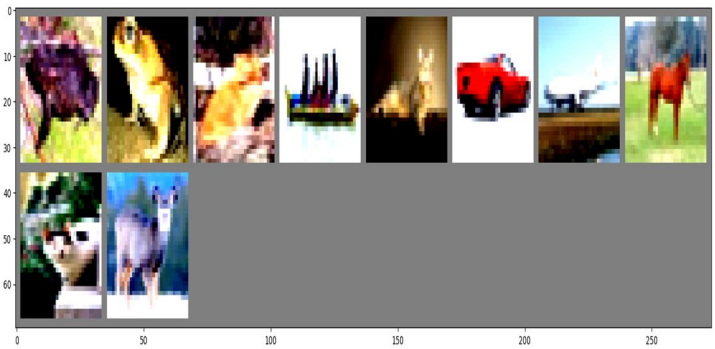


Figure 6. Images sample from dataset.

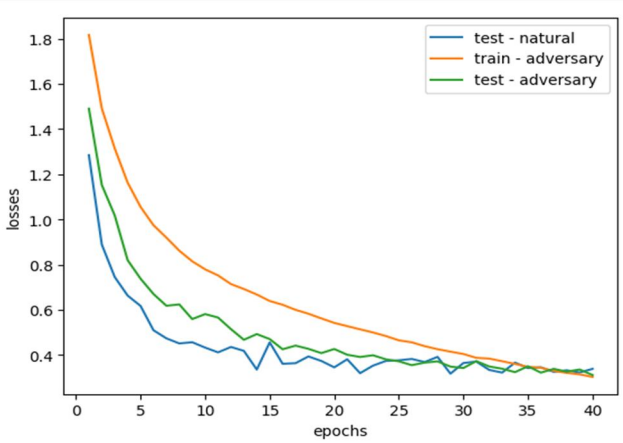


Figure 7. Comparing naturally-trained and adversarially-trained models.

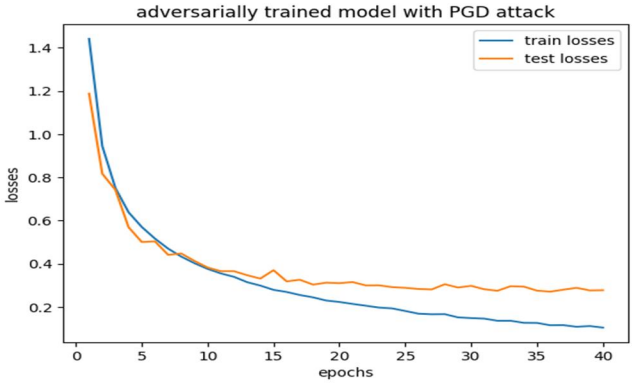


Figure 8. Adversarially trained model with PGD Attack

VI. CONCLUSION

This paper is dedicated to addressing the critical challenge of enhancing the security of models of deep learning, particularly those handling image data. Its primary objectives are to protect sensitive information, safeguard intellectual property, ensure compliance with regulations, and ultimately support the security landscape of deep learning in image-related applications.

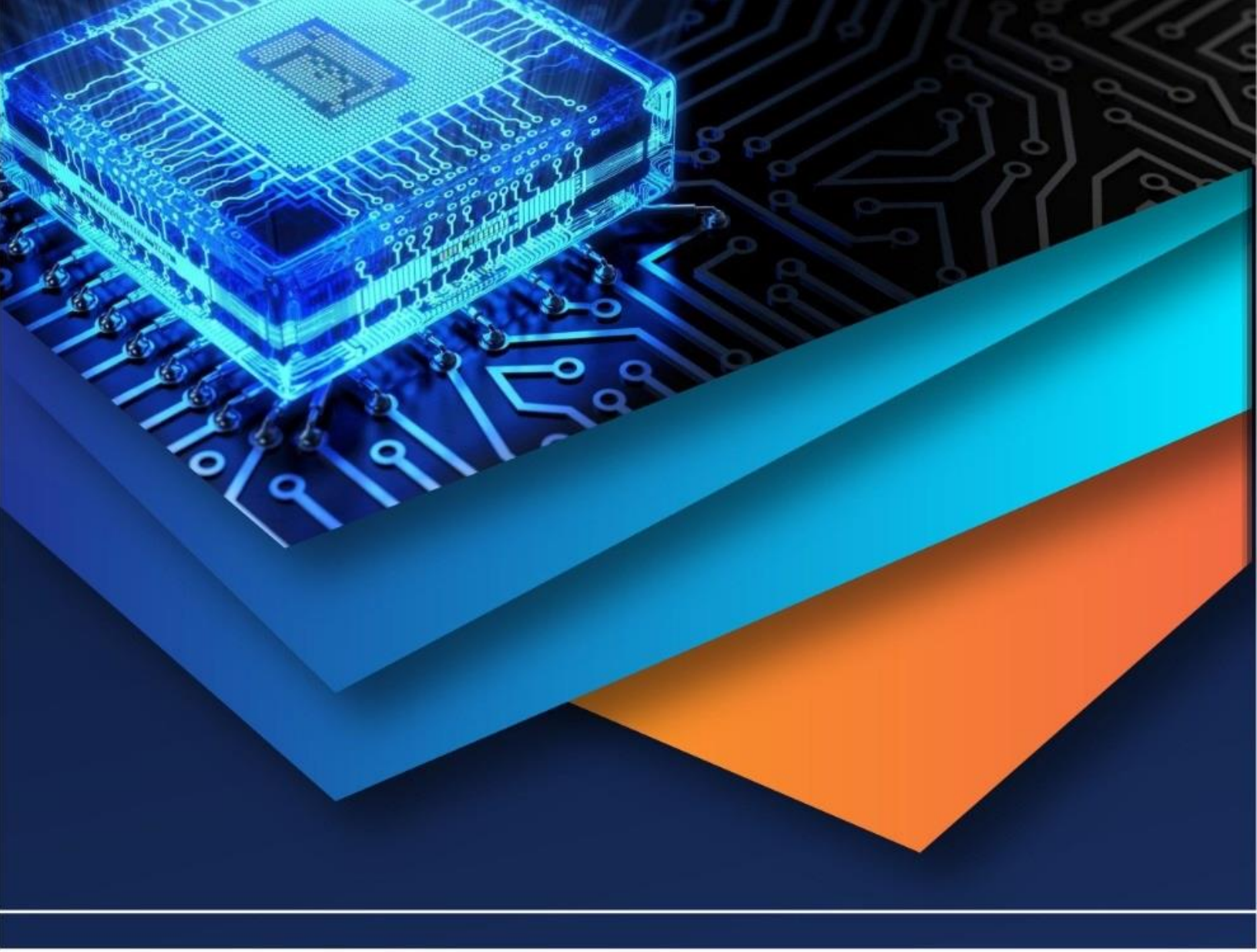
The two-fold approach involves the development of efficient attacking algorithms to assess the vulnerabilities of deep learning frameworks across several various image datasets. Simultaneously, it focuses on implementing robust defensive algorithms to counter these attacks and evaluates their effectiveness. We have achieved this through several attacks like FGSM, Deepfool, PDG and several defenses like Adversarial training on FGSM and PDG and APE-GAN and models like XGBOOST, RESNET and CNN.

In summary, this project represents a vital step towards fortifying the security of deep learning in image-centric applications. It aligns with the broader mission of building a secure and trustworthy AI ecosystem, instilling confidence in AI technologies' responsible integration into our lives.

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

REFERENCES

[1] K. Sakurai," tentative Generative Adversarial Network- Grounded Image Denoising for Defending Against inimical Attack,"Department of Information Science and Technology, Graduate School of Information Science and Electrical Engineering,Kyushu University, Fukuoka819-0395, Japan, 2021.
[2] Y. Feng,B. Chen,J. Lu, andS.-T. Xia," Deep image prior grounded defense against inimical exemplifications," College of Computer Science and Software Engineering, Shenzhen University, China, 2021.
[3] AH. YooP.M. Hong,T. Kim,J.W. Yoon, andY.K. Lee," Defending Against inimical Point Attacks Grounded on Deep Image previous,"Dept. Comput.Eng., HongikUniv., Seoul 04066, Republic of Korea, 2023.
[4] E. Phaisangittisagul," Collaborative Defense- GAN for guarding inimical attacks on bracket systems," Department of Electrical Engineering, Faculty of Engineering, Kasetsart University, Thailand, 2023.
[5] Yamina Mohamed Ben Ali, " inimical attacks on deep literacy networks in image bracket grounded on Smell notions Optimization Algorithm " Computer Science Department, University of Badji Mokhtar, Annaba, BP 12, 23000, Algeria,
[6] Arka Ghosh, Sankha Subhra Mullick, Shounak Datta, Swagatam Das b, Asit Kr. Das, Rammohan Mallipeddi," A black- box inimical attack strategy with malleable sparsity and generalizability for deep image classifiers", 2021.
[7] Tao Bai, Jinqi Luo, and Jun Zhao, Member, IEEE," invisible inimical Patches for Image- Recognition Systems on Mobile Devices," 2022.
[8] Abebaw Alem, Shailender Kumar, " Deep Learning Models Performance Evaluations for Remote tasted Image Bracket ", Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India,
[9] Yi Ding, Fuyuan Tan, Ji Geng, Zhen Qin,, Mingsheng Cao, Kim- Kwang Raymond Choo, " Interpreting Universal Adversarial illustration Attacks on Image Classification Models ", 2023.
[10] Jia Wang, Chengyu Wang, Qiuzhen Lin, Chengwen Luo, Chao Wu, Jianqiang Li, " inimical attacks and defenses in deep learning for image recognition ", Shenzhen University, China, 2022.
[11] Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu, " inimical Attacks and Defenses in Deep Learning ", Institute of Cyberspace Research, Zhejiang University, China, 2019
[12] Samer Y Khamaiseh, Derek Bagagem, Abdullah Al- Alaj, Mathew Mancino, and Hakam W. Alomari, " inimical Deep Learning A check on inimical Attacks and Defense Mechanisms on Image Bracket ", Department of Computer Science and Software Engineering, Miami University Oxford, 2022
[13] Naveed Akhtar and Ajmal Mian, " trouble of inimical Attacks on Deep Learning in Computer Vision A Survey ",Department of Computer Science and Software Engineering, The University of Western Australia, Australia, 2018
[14] Hongsong Chen, Yongpeng Zhang, Yongrui Cao, Jing Xie," Security Issues and Protective Approaches in Deep Learning Frameworks ", 2021.
[15] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, Jinbo Xiong, Zuobin Ying, AthanasiosV. Vasilakos, "sequestration and Security Issues in Deep Learning A Survey ", Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, 97187 Luleå, SwedIn this paper, describe the implicit pitfalls of DL .



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)