International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 12 | Dec 2024 www.irjet.net p-ISSN: 2395-0072

Defending Against Adversarial Attacks Using the Guided Filter: A Robust Spatial Smoothing Approach

Mujeeb Ullah Daudzai[1], Zhang Xinyou[2]

[1]MSc Student, School of Computer Science and Artificial Intelligence, Department of Computer Science, Southwest Jiaotong University, Chengdu, China
[2]Associate Professor, School of Computer Science and Artificial Intelligence, Department of Computer Science, Southwest Jiaotong University, Chengdu, China

---

Abstract - *Adversarial attacks create a substantial threat to the robustness and reliability of deep learning models, necessitating the creation of effective and efficient mitigation strategies. This study seeks to assess the vulnerability of neural networks to adversarial attacks and presents the Guided Filter, a Spatial Smoothing technique, as a potential countermeasure. The proposed defense is evaluated using three datasets: ImageNet Subset, CIFAR-10, and MNIST, against various adversarial attack strategies, including single-step (L∞ Fast Gradient Sign Method), iterative (L∞ and L2 Projected Gradient Descent), and optimization-based (L2 Carlini & Wagner) attacks. Experimental results indicate that the Guided Filter significantly improves adversarial robustness, increasing robust accuracy across all datasets, even under extreme perturbation levels. For example, on the ImageNet Subset, robust accuracy increased from 0.00% (without defense) to 73.52% (with defense) under L2 C&W attacks, and on CIFAR-10, accuracy against L∞ PGD attacks surpassed 87.29% at elevated epsilon values. The Guided Filter alleviated adversarial perturbations and occasionally enhanced baseline model performance by reducing input noise. This research shows that the Guided Filter is a lightweight, computationally efficient, and effective technique for mitigating adversarial perturbations, making it appropriate for practical application. Future directions include assessing its robustness to adaptive attacks and exploring its applicability in real-time scenarios to improve its significance in advancing strong and resilient AI & DNNs systems.*

Key Words: Adversarial Attacks; Spatial Smoothing; Guided Filter; DNNs; FGSM; PGD; C&W

1. INTRODUCTION

Deep neural networks have achieved significant success in vision and speech recognition tasks; nevertheless, they display paradoxical characteristics, particularly their susceptibility to adversarial attacks. These attacks include the introduction of unnoticeable perturbations to input images, resulting in misclassification by the network. Studies demonstrate that little perturbations can profoundly interfere with the categorization process, exposing the discontinuities in the learned mappings of neural networks. Furthermore, the characteristics of these adversarial cases are not purely random; identical perturbations can induce misclassifications across diverse networks trained on distinct datasets. This highlights the necessity for efficient defence strategies to improve model robustness. Recent methodologies indicate that integrating adversarial examples into training enhances generalization, paving the way for more robust network architectures, Szegedy et al. 2013 [1].

In neural networks and machine learning, models' susceptibility to adversarial examples is a major obstacle that may compromise their efficacy and dependability. Adversarial examples, which are minor alterations of accurately categorized inputs, can deceive models into generating incorrect outputs with high confidence, raising issues regarding how resistant they are in diverse applications. Previous research has specifically indicated that these adversarial weaknesses frequently arise from the linear characteristics of models functioning in high-dimensional environments, rather than the nonlinearity and overfitting that are typically referenced. This comprehension offers a chance to investigate efficient adversarial training methods that can function as a type of regularization, Goodfellow et al. 2015 [2].

Our research involves employing a variety of datasets, including CIFAR-10, MNIST, and ImageNet Subset, in conjunction with multiple untargeted adversarial attack techniques, such as L∞ FGSM, L∞ PGD, L2 PGD, and L2 C&W attack, to facilitate a thorough examination of model performance under adversarial circumstances. To tackle these issues, applying Spatial Smoothing techniques, especially Guided Filters, has surfaced as an effective defence strategy, showing the capability to improve model robustness against adversarial perturbations. The results of these studies demonstrate that integrating such defence mechanisms can substantially enhance the robustness of deep learning models, underscoring the necessity for ongoing investigation in adversarial machine learning.

2. LITERATURE REVIEW

The study emphasizes the increasing risk of adversarial attacks on deep neural networks, especially in image

recognition applications. FGSM is a prominent and straightforward technique for creating adversarial examples. This method used gradient information from the target model to generate perturbations, though it has shortcomings, including the creation of visually different perturbations that are frequently detectable by humans. The research highlights the significance of Convolutional Neural Networks (CNNs), specifically the EfficientNet design, recognized for its excellent scaling in depth, width, and resolution. The suggested defense mechanism, Entropy Aware Spatial Smoothing, employs entropy as a statistical metric to evaluate the unpredictability of incoming images, facilitating the distinction between natural images and those perhaps influenced by adversary attacks. This method employs a sophisticated Spatial Smoothing strategy to reduce adversarial perturbations while preserving the authenticity of real image information, thus improving the model's resilience. The assessment is based on studies using a malarial cell image dataset, highlighting the relevance of the suggested defense mechanism in medical image classification contexts and emphasizing the need for dependable and secure deep learning systems in critical situations, Muthuraman et al. 2024 [3].

The study offers an extensive analysis of adversarial attacks and defenses in medical imaging, outlining a technique that encompasses a systematic review of attack strategies and their respective defensive measures. The study emphasizes the categorization of medical images into distinct groups, concentrating on the manner in which adversarial attacks distort inputs to provoke inaccurate classifications. The results of this research demonstrate that adversarial attacks, especially those employing techniques such as the FGSM and Adaptive Mask Segmentation Attack, considerably affect the accuracy of deep learning models. The research emphasizes the necessity of integrating adverse images into training datasets to strengthen model robustness, illustrating that ensemble-based defense strategies can improve classification precision against adversarial disturbances. The findings advocate for robust defenses to guarantee the reliability and accuracy of segmentation models used in healthcare, highlighting the necessity for continuous research to address developing issues in the field, Muoka et al. 2023 [4].

This study evaluates adversarial defenses using datasets like CIFAR-10 and ImageNet, highlighting the influence of adaptive attacks on defense strategies. It underscores the efficacy of an adaptive attack that integrates projected gradient descent (PGD) with expectation over transformations (EOT) to markedly diminish the accuracy of several defenses. The concept entails formulating loss functions to produce adversarial cases by maximizing misclassification while regulating perturbation magnitude. The results indicate that numerous defenses are still susceptible, frequently diminishing model accuracy to levels comparable to undefended classifiers, hence underscoring the need for ongoing enhancement in protection against adaptive attack techniques. This highlights the persistent difficulty of creating dependable defenses in adversarial machine learning, Tramer et al. 2020 [5]. The study provides a comprehensive evaluation of adversarial attacks and countermeasures, classifying attacks according to the attacker's knowledge and objectives. It analyzes diverse approaches, including adversarial training, which seeks to enhance model resilience by incorporating adversarial cases during the training process. The survey classifies significant attacks, such as confidence reduction white-box attacks and targeted black-box attacks, emphasizing their effects on deep neural networks. It also examines prevalent datasets used for evaluation, including MNIST, CIFAR-10, and ImageNet, while addressing the issues related to attaining precise accuracy when defenses are included. The findings demonstrate the importance of sophisticated methods designed to improve model robustness against adversarial attacks and the persistent necessity for efficient evaluation metrics to assess the efficacy of diverse defenses across multiple datasets. The research emphasizes the necessity of examining real-world events, advocating for the exploration of varied datasets to improve comprehension of adversarial effects and the efficacy of suggested solutions, Costa et al. 2024 [6].

Various mechanisms have been suggested to counter adversarial attacks. However, despite the existing attacks and defenses, there is no guarantee regarding the effective robustness of these networks or their reliability in critical domains, highlighting the necessity for deep neural networks to be inherently robust or easily updatable upon discovering new vulnerabilities. This supports the current work, whose primary contributions are outlined as follows:

- In this study, we provide a thorough examination of recent adversarial attacks on popular image datasets, namely CIFAR-10, MNIST, and ImageNet Subset. Our research examines several untargeted attack methodologies, including L∞ PGD, L∞ FGSM, L2 PGD, and L2 C&W attacks. We also use Guided Filters to examine their efficacy against a state-of-the-art Spatial Smoothing defense mechanism.
- We classify adversarial attacks according to their efficacy and robustness, emphasizing the unique attributes of each attack category. Additionally, we present data illustrating the differences in attack success rates of adversarial examples against the networks involved.
- Furthermore, we examine the metrics typically employed to assess the efficacy of both adversarial attacks and defenses, presenting our findings on the outcomes obtained across the specified

datasets. Our findings demonstrate considerable enhancements in model robustness through the application of the Spatial Smoothing technique, indicating promising avenues for future research in adversarial defenses.

3. OBJECTIVE

This research effort aims to evaluate the robustness of neural networks to adversarial attacks by applying and analyzing several attack tactics on popular datasets, such as CIFAR-10, MNIST, and an ImageNet Subset. The study investigates several untargeted adversarial methods, specifically L∞ FGSM, L∞ PGD, L2 PGD, and L2 C&W attacks, to illustrate the vulnerabilities present in deep learning models. This study aims to improve defensive strategies against these attacks by using Spatial Smoothing techniques, specifically Guided Filters which have demonstrated significant efficacy in strengthening model robustness. The research seeks to deliver substantial insights into adversarial robustness and establish new performance standards for defense mechanisms in machine learning through the development of a comprehensive evaluation methodology.

4. METHODOLOGY

This section provides a detailed summary of the used datasets, the adversarial attacks executed on these datasets, the deep neural network models employed in this study, and the application of the suggested Robust Spatial Smoothing Technique for defending against adversarial attacks.

4.1 Datasets Used

4.1.1 The CIFAR-10 Dataset

The CIFAR-10 dataset is frequently used in machine learning and computer vision, having 60,000 color images of 32x32 pixels, categorized into 10 classes, each containing 6,000 images. The categories encompass airplane, vehicle, bird, cat, deer, dog, frog, horse, ship, and truck. This dataset includes 50,000 images designated for training and 10,000 for testing, providing a manageable yet diverse collection that poses a significant challenge for classification algorithms, especially CNNs, Krizhevsky, 2009 [7].

4.1.2 The MNIST Dataset

This dataset is a valuable resource in machine learning and computer vision, often used for training various image processing systems. The collection includes 70,000 grayscale images of handwritten digits (0-9), each measuring 28x28 pixels. The dataset contains 60,000 training photographs and 10,000 test images, rendering it appropriate for assessing classification systems. MNIST has significantly contributed to the development and evaluation of several machine learning techniques and neural network architectures because to its simplicity and accessibility, Lecun et al. 1998 [8].

4.1.3 The ImageNet Dataset

ImageNet is a major resource in computer vision, developed for extensive picture classification applications. This ImageNet Subset validation dataset has 1,000 classes, each illustrated by 5 photographs, concluding in a total of 5,000 images. This small dataset is essential for systematically testing and verifying models, enabling researchers to assess performance across different object categories. This subset's images are typically scaled to 224x224 pixels and include 3 color channels (RGB), giving a total of 150,528 parameters per image. This format has become standard for many deep learning applications like CNNs, and is essential for the advancement of the field, Deng et al. 2009 [9].

4.2 Adversarial Attacks

4.2.1 FGSM

FGSM is a well-established adversarial attack technique that targets machine learning models, especially neural networks. Ian Goodfellow and his colleagues developed FGSM. This study focuses on the L∞ FGSM, an effective technique for generating adversarial examples by perturbing input data. FGSM generates adversarial examples using the gradient of the loss function to modify pixel values, aiming to cause misclassification while preserving minimal perturbations. The L∞ variant precisely limits these alterations to a maximum permissible threshold for any individual pixel adjustment, rendering it a practical option for assessing model resilience, Deng et al. 2009b [9].

$$x + \varepsilon sgn\left(\nabla_x L(\theta, x, y)\right)$$

Fig -1: FGSM Equation

4.2.2 PGD

PGD attack is a powerful adversarial technique that enhances the FGSM by iteratively optimizing adversarial samples, hence increasing the distortion of input data to expose model vulnerabilities. In my research, I used both the L∞ and L2 varieties of PGD because of their unique advantages. The L∞ PGD variation employs a focused strategy by limiting perturbations to a maximum permissible alteration per pixel, hence enhancing the probability of misclassification while preserving the adversarial features of the input. The L2 PGD variant, on the other hand, allows for more subtle modifications by evaluating the total distance of changes across all pixels.

This may hinder the identification of adversarial instances. Employing both methods facilitates a comprehensive assessment of a model's robustness by evaluating its performance against both severe and nuanced adversarial alterations, Madry et al., 2017 [10].

$$x^{t+1} = \prod_{x+s} \left(x^t + \propto sgn(\nabla_x L(\theta, x, y))\right)$$

Fig -2: PGD Equation

4.2.3 C&W

The L2 C&W untargeted attack is an efficient technique for generating adversarial samples that lead classifiers to generate incorrect results without focusing on a particular class. In my research, I selected this technique due to its ability to implement small alterations that preserve a significant level of visual similarity to the source images while effectively bypassing many model protections. The L2 variation explicitly seeks to minimize the L2 distance between the original and adversarial scenarios, hence facilitating accurate control of the applied distortion. The attack employs an optimization-driven approach, aiming to identify perturbations that minimize a specified objective function. The objective function is carefully constructed to achieve enhanced success rates in generating adversarial examples while conforming to the established restrictions. The optimization process of the L2 C&W attack is typically performed through iterative methods, progressively refining the perturbation until an appropriate adversarial sample is obtained, Carlini & Wagner, 2017 [11].

4.3 Neural Networks Architectures

4.3.1 Resnet-50

Microsoft researchers developed ResNet-50. ResNet-50, He et al. 2015 [12] enhances deep neural network training by addressing the degradation problem. Containing 50 layers with a bottleneck design, it employs residual learning using a combination of 1x1 and 3x3 convolutions, resulting in reduced computational expenses while improving accuracy. This architecture markedly mitigates vanishing gradient issues and has led to its extensive adoption in deep learning applications.

4.3.2 Resnet-18

Microsoft researchers developed ResNet-18. ResNet-18, He et al. 2015 [12] is a convolutional neural network architecture formulated to effectively train deep networks. The architecture consists of 18 layers, including a 7x7 convolutional layer with 64 filters, a 3x3 max pooling layer, and four residual blocks, each containing two 3x3 convolutional layers with 64, 128, 256, and 512 filters, respectively, in addition to batch normalization and ReLU activations. The architecture consists of a 7x7 average pooling layer and a fully connected layer. ResNet-18 uses shortcut connections to mitigate the issue of vanishing gradients, improving gradient propagation during training and demonstrating the effectiveness of residual learning in deep networks.

4.3.3 LeNet

LeNet, developed by, Lecun et al., 1998, [8], and his colleagues in the late 1980s, is one of the earliest convolutional neural network architectures, containing seven layers: an input layer, a convolutional layer (5x5, 6 filters), an average pooling layer, a second convolutional layer (5x5, 16 filters), another average pooling layer, a fully connected layer, and an output layer. Originally designed for the identification of handwritten digits, it gained attention for adversarial attacks when, Szegedy et al., 2013 [1], demonstrated its susceptibility to adversarial attacks, showing how minimal alterations in input could lead to significant misclassifications.

4.4 Spatial Smoothing and Guided Filter

4.4.1 Spatial Smoothing

Spatial Smoothing is an important technique in image processing aimed at reducing noise and enhancing visual quality by averaging pixel values within a defined neighborhood, thus eliminating high-frequency noise while preserving essential structures. This technique is necessary for applications like image enhancement, object recognition, and computer vision, where maintaining edge integrity is imperative. The main objective of Spatial smoothing is to balance noise reduction with detail preservation, employing methods including Gaussian Smoothing, median filtering, and bilateral filtering, each possessing unique benefits and drawbacks. However, traditional methods may struggle with images containing significant edge information, leading to blurriness and diminished clarity. To resolve these challenges, advanced techniques such as Guided Filtering have been developed, enhancing noise reduction while preserving critical features.

4.4.2 Guided Filter

The Guided Filter is an advanced edge-preserving smoothing technique designed to enhance image quality and decrease noise. It operates through using a reference image that guides the filtering procedure to maintain critical structural details. The process starts by padding the input and reference images to ensure uniform application. It additionally computes local averages and variances within a designated neighborhood to derive linear coefficients that signify the relationship between the guided picture and the input. Applying these coefficients to the guiding picture, the Guided Filter

generates a smoothed output that effectively reduces noise while preserving edges and details. This property makes the Guided Filter particularly beneficial in various image processing applications, enabling improved analysis and interpretation of visual data.

$$q(x) = \left[\frac{cov(I,p)}{var(I+e)}\right].I\,(x) + \left[\frac{1}{(2r+1)^2}\sum_x 1 \in N_{(x)}\, p(x') - \frac{cov(I,p)}{var(I)+e} \cdot \frac{1}{(2r+1)^2}\sum_x 1 \in N_{(x)}\, I(x')\right]$$

Fig -3: Guided Filter Equation

The formula for q(x) regarding adversarial attacks and perturbations including local statistics from the guiding image I and the input image P to produce a robust output. The expression (I)+ ∈ stabilizes the computation by avoiding division by zero and reducing the impact of small perturbations, so assuring that the variance of the guiding image favorably influences the result. The sum of this modified variance and I(x) enhances the impact of constant guidance, despite the existence of adversarial noise. The computation of the local mean of p inside a neighborhood N(x) reduces temporary adversarial effects, while enhancing the fundamental output. Furthermore, the result is modified by deducting a component that reflects the covariance between I and p, thereby tuning the output to mitigate distortions from adversarial alterations. This method preserves the integrity of picture features, improving robustness to perturbations while assuring that the final output q(x) accurately represents the desired attributes of both images.

5. Experiments

5.1 Resnet-50, Resnet-18, LeNet

In our experiments, we used ResNet-50, ResNet-18, and LeNet, three common convolutional neural network designs. Kaiming created ResNet-50 and ResNet-18, He et al., 2015 [12]. A residual learning architecture helps him and his colleagues construct extraordinarily deep networks, optimizing gradient flow and improving classification performance. However, Lecun et al., 1998, [8] and his colleagues constructed LeNet, an early convolutional neural network for handwritten digit recognition, to demonstrate convolutional layers and pooling. We integrated the architectural aspects of these three models into our study to assess their classification accuracy and efficacy in numerous experimental situations.

5.2 FGSM, PGD and C&W Attacks

5.2.1 L∞ FGSM

The FGSM was introduced in, Goodfellow et al., 2015 [2]. One of the oldest and most popular adversarial attack methods. It generates adversarial examples by computing the loss function gradient for the input image and altering the input to maximize this loss. The L∞ variant of FGSM limits perturbation to a specific epsilon threshold, minimizing image alterations but still misleading the model. While regular FGSM may not limit perturbations under specified norms, the L∞ version significantly controls image distortions, preserving visual resemblance. This makes it ideal for controlled model robustness evaluation. This study generated adversarial examples from all images in the testing dataset using L∞ FGSM and evaluated their impact on model performance. Its computational efficiency and simplicity made it suitable for this work, allowing speedy and comprehensive dataset testing.

5.2.2 L∞ and L2 Descent (PGD)

A popular method for developing adversarial examples through repeated modifications is PGD, Madry et al., 2017 [10]. The L∞ PGD version uses the gradient of the loss function to make slight, controlled alterations to the input image, keeping them within a set limit (epsilon). This generates powerful adversarial samples that fool the model while preserving the images. L2 PGD measures image changes using the L2 standard, allowing for smaller alterations that may be less obvious. I used L∞ and L2 PGD to assess the model's response to adversarial attacks in this study. These approaches are better than FGSM because their iterative approach often generates more effective attacks. I used these methods to analyze the model's flaws and design stronger defenses against opposed threats.

5.2.3 L2 C&W Attack

The L2 C&W approach, proposed by Carlini and Wagner, Carlini & Wagner, 2017, Carlini & Wagner, 2016 [11], [14], generates advanced adversarial samples that deceive machine learning algorithms. Optimizing a loss function to provide small perturbations allows this attack to manipulate model output. The L2 variant minimizes the perturbation's L2 norm to misclassify the original image with little alterations. I included the L2 C&W attack in this research because it produces extremely effective adversarial samples that are harder to identify than simpler methods. The optimization-based approach ensures that adversarial perturbations are effective and subtle, which is essential for validating the model's robustness to complex attacks. I used this method to evaluate the model's vulnerabilities and inform protection methods against adversarial threats.

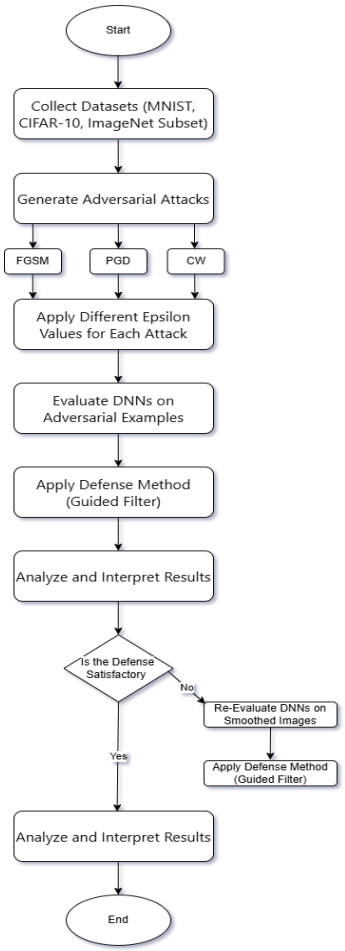

Fig -4: Experiments Flowchart

5.3 Defense

In order to mitigate these adversarial attacks, we used Spatial Smoothing, specifically through the Guided Filter. By employing a localized smoothing approach that maintains edge properties while reducing noise from adversarial changes, this technique reduces the impact of adversarial perturbations. To increase the model's resistance to the applied attacks, we used the Guided Filter.

6. RESULTS AND DISCUSSIONS

6.1 ImageNet Dataset Results

The top-1 accuracy of ResNet-50 on this dataset is 76.06%. The optimized ResNet-50 model was assessed on an ImageNet Subset dataset of 5,000 images, including a normalization procedure for standardization with designated mean and standard deviation values. The model employed a sequential architecture, integrating the normalization layer with a pre-trained ResNet-50. A validation dataset including 2,500 randomly picked photos was established, with preprocessing that included scaling to 224x224 pixels and conversion to tensors. The dataset was imported into a DataLoader for optimized evaluation. The cross-entropy loss was employed for performance evaluation, and model robustness was examined against L∞ PGD, L∞ FGSM, L2 PGD, and L2 C&W adversarial attacks across multiple epsilon values, each producing distinct outcomes. To alleviate these concerns, Spatial Smoothing was used, particularly through the Guided Filter to improve robustness against disturbances.

Table -1: ImageNet Dataset Results

Attack Type	Epsilon Size	Accuracy	Loss Value
L∞ FGSM	1/255	11.12%	0.4180
	2/255	6.52%	0.4787
	4/255	5.04%	0.4854
	8/255	6.40%	0.4490
	16/255	8.28%	0.4004
	32/255	6.64%	0.3951
L∞ FGSM & Guided Filter	1/255	73.48%	0.0695
	2/255	73.32%	0.0697
	4/255	73.16%	0.0700
	8/255	72.44%	0.0719
	16/255	70.12%	0.0782
	32/255	66.00%	0.0907
L∞ PGD	1/255	0.64%	1.1201
	2/255	0.08%	1.7398
	4/255	0.04%	2.2789
	8/255	0.00%	2.6505
	16/255	0.00%	2.7968
	32/255	0.00%	2.7309
L∞ PGD & Guided Filter	1/255	73.60%	0.0699
	2/255	73.64%	0.0699
	4/255	73.20%	0.0699
	8/255	72.76%	0.0703
	16/255	71.84%	0.0731
	32/255	70.52%	0.0785
L2 PGD	0.25	22.64%	0.3174
	0.50	5.80%	0.6418
	1.00	0.72%	1.1411
	2.00	0.08%	1.7180
	4.00	0.04%	2.2628
	8.00	0.00%	2.6874
L2 PGD & Guided Filter	0.25	73.52%	0.0693
	0.50	73.60%	0.0695
	1.00	73.60%	0.0695
	2.00	73.52%	0.0694
	4.00	73.52%	0.0698
	8.00	73.04%	0.0711
L2 C&W	100	0.00%	0.1548
L2 C&W & Guided Filter	100	73.52%	0.0694

6.1.1 ImageNet Dataset Findings

Without the Guided Filter, models on ImageNet Subset experienced dramatic performance degradation as attack strengths increased. Under L∞ FGSM, robust accuracy decreased from 11.12% at 1/255 to 6.64% at 32/255, and L∞ PGD and L2 PGD attacks made the model nearly worthless, with robust accuracy falling to 0.00% at comparatively low epsilon values (e.g., 8/255). The L2 C&W attack showed a comparable pattern, entirely compromising the model with 0.00% accuracy. Nonetheless, the application of the Guided Filter resulted in significant enhancements across all categories of attacks. As an instance, the L∞ PGD applying the Guided Filter attained a robust accuracy of 73.60% at 1/255, sustaining 70.52% even at high epsilon values (e.g., 32/255), whereas the L2 C&W attack improved robust accuracy from 0.00% to 73.52%. Loss values remained minimal across every circumstance with the filter applied, illustrating its stabilizing effect.

6.2 CIFAR-10 Dataset Results

The basic architecture of ResNet-18 achieves a top-1 accuracy of 95.28%. The ResNet-18 model was assessed using the CIFAR-10 dataset, including color pictures of 32x32 pixels across 10 distinct classes. A normalization procedure was implemented to standardize the input photos during preprocessing. The dataset employed PyTorch's integrated CIFAR-10 resources, with the images transformed into tensors to enhance training efficiency. A DataLoader was established with a batch size of 64 for effective data management. The model's efficacy was assessed using the cross-entropy loss function. Additionally, its resilience against adversarial attacks, L∞ PGD, L∞ FGSM, and L2 PGD, was evaluated by examining different epsilon values to assess their impact on performance. To improve the model's resilience against adversarial threats, Spatial Smoothing techniques were used, especially the Guided Filter, to mitigate the impact of perturbations.

Table -2: CIFAR-10 Dataset Results

Attack Type	Epsilon Size	Accuracy	Loss
L∞ FGSM	1/255	67.40%	0.0338
	2/255	58.06%	0.0493
	4/255	52.55%	0.0574
	8/255	48.17%	0.0616
	16/255	35.25%	0.0754
	32/255	16.78%	0.1068
L∞ FGSM & Guided Filter	1/255	93.45%	0.2951
	2/255	93.41%	0.2949
	4/255	93.20%	0.3038
	8/255	92.18%	0.3566
	16/255	87.44%	0.5950
	32/255	74.36%	1.3332
L∞ PGD	1/255	51.22%	0.0624
	2/255	26.94%	0.1344
	4/255	8.39%	0.2116
	8/255	0.91%	0.2651
	16/255	0.07%	0.2920
	32/255	0.01%	0.3143
L∞ PGD & Guided Filter	1/255	93.41%	0.2965
	2/255	93.42%	0.2971
	4/255	93.41%	0.3018
	8/255	93.02%	0.3148
	16/255	91.79%	0.3753
	32/255	87.29%	0.6026
L2 PGD	0.25	31.77%	0.1169
	0.50	13.11%	0.1930
	1.00	2.40%	0.2593
	2.00	0.16%	0.2988
	4.00	0.00%	0.3239
	8.00	0.00%	0.3577
L2 PGD & Guided Filter	0.25	93.34%	0.2968
	0.50	93.33%	0.2989
	1.00	93.14%	0.3097
	2.00	92.23%	0.3630
	4.00	88.19%	0.5583
	8.00	75.09%	1.2642

6.2.1 CIFAR-10 Dataset Findings

On the CIFAR-10 dataset, attacks such as L∞ FGSM and L∞ PGD led to severe performance drops without defenses. Robust accuracy decreased from 67.40% to 16.78% with increasing epsilon values for L∞ FGSM and from 51.22% to 0.01% for L∞ PGD. A similar pattern was noted with L2 PGD, where the starting accuracy of 31.77% decreased to 0.00% at elevated epsilon values, underscoring significant model vulnerabilities. The Guided Filter significantly enhanced robust accuracy: L∞ FGSM maintained accuracy at almost 93% for lower epsilon values and 74.36% for 32/255, whereas L∞ PGD achieved 93% robust accuracy for smaller epsilons and above 87.29% for higher perturbations. Likewise, L2 PGD using the Guided Filter showed robust performance, start at 93.34% at 0.25 epsilon and sustaining above 75% even at 8.00 epsilon.

6.3 MNIST Dataset Results

MNIST fundamental architecture, the top-1 accuracy of LeNet is 98.99%. The LeNet model was evaluated using the MNIST dataset, consisting of 28x28 pixel grayscale images of handwritten numbers. The model used a normalization procedure to standardize input photos. The

training and test datasets were generated via a custom Dataset class that imported photos from CSV files, standardizing the pixel values to the interval [0, 1]. A DataLoader was employed to efficiently batch the data with a size of 32. The model's efficacy was evaluated by cross-entropy loss. Furthermore, adversarial robustness was evaluated against FGSM and PGD attacks, examining the impact of different epsilon values on performance. A Spatial Smoothing defense technique was employed to improve robustness against adversarial threats, specifically through the use of the Guided Filter.

Table -3: MNIST Dataset Results

Attack Type	Epsilon Size	Accuracy	Loss
L∞ FGSM	1/255	98.86%	0.0006
	2/255	98.65%	0.0007
	4/255	98.19%	0.0009
	8/255	96.85%	0.0015
	16/255	91.71%	0.0042
	32/255	60.88%	0.0221
L∞ FGSM & Guided Filter	1/255	99.01%	0.0320
	2/255	99.01%	0.0320
	4/255	99.03%	0.0324
	8/255	98.98%	0.0330
	16/255	98.89%	0.0353
	32/255	98.60%	0.0453
L∞ PGD	1/255	99.64%	0.0004
	2/255	99.50%	0.0005
	4/255	99.13%	0.0008
	8/255	97.64%	0.0019
	16/255	90.74%	0.0080
	32/255	44.78%	0.0630
L∞ PGD & Guided Filter	1/255	99.71%	0.0132
	2/255	99.70%	0.0134
	4/255	99.70%	0.0137
	8/255	99.68%	0.0146
	16/255	99.58%	0.0173
	32/255	99.25%	0.0272
L2 PGD	0.25	98.97%	0.0009
	0.50	96.95%	0.0024
	1.00	86.11%	0.0124
	2.00	26.25%	0.1084
	4.00	0.10%	0.4697
	8.00	0.00%	0.9533
L2 PGD & Guided Filter	0.25	99.70%	0.0004
	0.50	99.70%	0.0004
	1.00	99.68%	0.0005
	2.00	99.57%	0.0006
	4.00	99.07%	0.0011
	8.00	94.91%	0.0050

6.3.1 MNIST Dataset Findings

On MNIST, the observed degradation without defenses followed a similar pattern. Under L∞ FGSM, accuracy decreased from 98.86% at 1/255 to 60.88% at 32/255, while L∞ PGD similarly dropped from 99.64% to 44.78%. L2 PGD attacks were especially destructive, decreasing robust accuracy from 98.97% at 0.25 epsilon to 0.00% at 8.00 epsilon. The Guided Filter significantly enhanced results, with L∞ FGSM sustaining a strong accuracy of over 99.00% for all epsilon levels, decreasing slightly to 98.60% at 32/255. In the case of L∞ PGD, robust accuracy continually above 99.25%, even at elevated epsilon values, whereas L2 PGD demonstrated comparable performance, attaining 99.70% at 0.25 epsilon and sustaining a commendable 94.91% at 8.00 epsilon, accompanied by minimal loss values.

7. CONCLUSION

This research methodically examined the adversarial robustness of neural networks across three datasets (ImageNet Subset, CIFAR-10, and MNIST) under several attack, including L∞ FGSM, L∞ PGD, L2 PGD, and L2 C&W attacks. The research primarily aimed to assess the performance degradation that generated by adversarial perturbations and the efficacy of the Guided Filter, a Spatial Smoothing method, in addressing these impacts. This study emphasizes the significance of preprocessing defenses, such as the Guided Filter, in effectively mitigating adversarial vulnerabilities. Achieving robust AI systems requires the integration of preprocessing approaches with other strategies, including adversarial training, input transformations, or feature denoising, to establish comprehensive and layered defenses. The Guided Filter functions as an effective initial safeguard, enhancing model-level approaches for adversarial robustness.

8. LIMITATIONS AND FUTURE DIRECTIONS

While this study demonstrated the effectiveness of the Guided Filter across diverse datasets and attacks, several limitations and future research directions are worth highlighting:

- **Dataset Complexity**

  While the Guided Filter performed well on MNIST, CIFAR-10, and the Subset of ImageNet, future work is needed to evaluate its scalability and efficacy on more challenging datasets with greater diversity and complexity, such as ImageNet-Adv or COCO.

- **Stronger Adaptive Attacks**

  Standard adversarial attacks were assessed in the current study, but future research should examine

how robust the Guided Filter is against adaptive attacks that are specifically built to go over defenses.

- **Real-Time Applications**

  To ensure the Guided Filter is suitable for deployment in time-sensitive or resource-constrained environments, its efficacy in real-time contexts must be assessed by analyzing its computational overhead and latency.

9. REFERENCES

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013b). Intriguing properties of neural networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1312.6199

[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations.https://ai.google/research/pubs/pub43405

[3] Muthuraman, A. A., Kathir, R. M., Pragadeesh, T., & Rajam, V. M. A. (2024b). Entropy Aware Spatial Smoothing in Removal of Perturbation as a Defense Against Adversarial Attacks. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1–8. https://doi.org/10.1109/icccnt61001.2024.10725664

[4] Muoka, G. W., Yi, D., Ukwuoma, C. C., Mutale, A., Ejiyi, C. J., Mzee, A. K., Gyarteng, E. S. A., Alqahtani, A., & Al-Antari, M. A. (2023). A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense. Mathematics, 11(20), 4272. https://doi.org/10.3390/math11204272

[5] Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On Adaptive Attacks to Adversarial Example Defenses. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2002.08347

[6] Costa, J. C., Roxo, T., Proença, H., & Inácio, P. R. M. (2024). How Deep Learning Sees the World: A Survey on Adversarial Attacks & Defenses. IEEE Access, 12, 61113–61136. https://doi.org/10.1109/access.2024.3395118

[7] Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[8] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

[9] Deng, J., Dong, W., Socher, R., Li, L., Li, N. K., & Fei-Fei, N. L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2009.5206848

[10] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv (Cornell University). http://arxiv.org/pdf/1706.06083.pdf

[11] Carlini, N., & Wagner, D. (2017). Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1705.07263

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1512.03385

[13] Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial Examples in the Physical World. In Chapman and Hall/CRC eBooks (pp. 99–112). https://doi.org/10.1201/9781351251389-8

[14] Carlini, N., & Wagner, D. (2016). Towards Evaluating the Robustness of Neural Networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1608.04644

BIOGRAPHIES

Mujeeb Ullah Daudzai
Master's student with primary research interests in artificial intelligence, cybersecurity, machine learning, and big data.
mujeebdaudzai888@gmail.com

Zhang Xinyou
Ph.D. Associate Professor with primary research focuses on distributed computing applications, network security, and artificial intelligence.
xyzhang@swjtu.edu.cn