International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Exploring Magnitude Perturbation in Adversarial Attack & Defense

Bhasha Anjaria[1*], Jaimeel Shah[2]

Submitted: 25/11/2023 **Revised**: 05/01/2024 **Accepted**: 15/01/2024

Abstract: Adversarial attacks pose a significant threat to the robustness and security of machine learning models. In recent years, researchers have focused on developing defense mechanisms to mitigate the impact of adversarial attacks. One such avenue of investigation involves examining the effects of magnitude perturbation on the success rate and effectiveness of these attacks, as well as evaluating the performance of various defense strategies in countering them. In this study, present a comprehensive analysis of the impact of magnitude perturbation in adversarial attacks and the effectiveness of defense mechanisms against them. The investigation of the influence of perturbation magnitudes on the success rate and transferability of attacks across different models and datasets. Furthermore, evaluation the performance of state-of-the-art defense mechanisms under varying perturbation strengths. The experimental results reveal intriguing insights into the behavior of adversarial attacks and the efficacy of defense mechanisms. The observation shows that increasing the magnitude of perturbations can significantly amplify the success rate of attacks, rendering models more vulnerable. Additionally, demonstrate that certain defense mechanisms exhibit varying levels of resilience against different perturbation magnitudes, shedding light on their limitations and strengths. The findings of this study contribute to a deeper understanding of the role of magnitude perturbation in adversarial attacks and the effectiveness of defense mechanisms. This knowledge can aid in the development of robust defense strategies and provide valuable insights for enhancing the security of machine learning systems in the face of adversarial threats.

Keywords: *Magnitude perturbation; Adversarial attacks; Defense mechanisms; Robustness Security; Deep learning models; Success rate; Perturbation strengths; Attack effectiveness; Defense strategy evaluation; Limitations; Strengths; Security enhancement.*

1. Introduction

Adversarial attacks have emerged as a critical concern for the robustness and security of machine learning models [1,2]. These attacks involve carefully crafted perturbations to input data, leading to misclassification or incorrect predictions by the targeted models. In response, researchers have been developing defense mechanisms to mitigate the impact of these attacks. However, the effectiveness of these defenses can be influenced by the magnitude of the perturbations applied during the attack process [3,4,5].

This paper aims to provide a comprehensive study of the effects of magnitude perturbation in adversarial attacks and the corresponding defense mechanisms. The investigate the role of perturbation strengths in terms of their impact on attack success rates, transferability, and the performance of defense strategies.

In experimental analysis, conduct various attack scenarios using different magnitudes of perturbation on diverse machine learning models and datasets [6,7,8]. Then measure the success rates of the attacks and evaluate the transferability of the perturbed examples across different models [10]. Moreover, assess the performance of state-of-the-art defense mechanisms under varying perturbation strengths. The results of this study reveal intriguing findings. It was observed in fig 1 that increasing the magnitude of perturbations significantly enhances the success rates of adversarial attacks, making models more vulnerable. This highlights the critical role of perturbation strength in determining the severity of attacks. Furthermore, find that defense mechanisms exhibit varying levels of resilience against different magnitudes of perturbation, shedding light on their limitations and strengths [11].

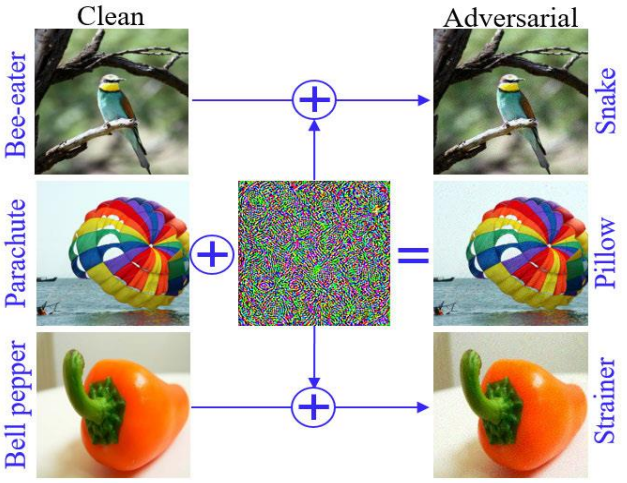


Fig 1. Adversarial Attack Magnitude Perturbation [1]

This paper contributes to a deeper understanding of the impact of magnitude perturbation in adversarial attacks and

[1] *Faculty of Engineering and Technology, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India,*
Email ID: bhasha.anjaria21316@paruluniversity.ac.in,
ORCID ID : 0000-0002-0562-4336
[2] *Faculty of Engineering and Technology, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujarat, India*
Email ID: jaimeel.shah@paruluniversity.ac.in,
ORCID ID : 0000-0002-1543-750X
** Corresponding Author Email: bhasha.anjaria21316@paruluniversity.ac.in*

the corresponding defense mechanisms. The insights gained from this study can aid in the development of more robust defense strategies and inform the enhancement of security measures for machine learning systems. By exploring the interplay between perturbation magnitude, attack success rates, and defense effectiveness, this research aim to provide a valuable resource for researchers and practitioners working in the field of adversarial machine learning.

2. Materials and Methods

If you are using Word, use either the Microsoft Equation Editor or the MathType add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation or MathType Equation). "Float over text" should not be selected.

2.1. Datasets

The MINIST dataset [1] is a widely used benchmark dataset in the field of computer vision and machine learning. It consists of a collection of 60,000 grayscale images of handwritten digits from 0 to 9, with 10,000 additional images reserved for testing purposes. Each image in the dataset is a 28x28 pixel square, providing a total of 784 pixels per image. The dataset is well-balanced, with an equal number of examples for each digit class.

The CelebA dataset [11] is a large-scale face attributes dataset that comprises over 200,000 celebrity images. Each image in the dataset contains various annotations, including facial landmarks, identities, and attribute labels such as gender, age, and presence of accessories like eyeglasses or hats. The images exhibit diverse facial expressions, poses, and lighting conditions, making it a valuable resource for research in facial analysis and recognition tasks.

The Fashion dataset [16], often referred to as the Fashion-MNIST dataset, is a benchmark dataset designed as a drop-in replacement for the original MNIST dataset. It consists of 70,000 grayscale images of fashion items categorized into ten classes, including T-shirts, dresses, shoes, and more. Each image in the dataset is a 28x28 pixel square, similar to the MINIST dataset, providing a total of 784 pixels per image.

2.2. Adversarial Attacks

Adversarial attacks can be classified into two main types: white-box attacks and black-box attacks. These categories describe the level of knowledge the attacker has about the target model and its internal workings. Additionally, there are specific attack methods, such as FGSM, PGD, and Deep Fool, which are commonly used in adversarial attacks. Let's explore each concept in detail:

2.2.1. White-box attacks:

White-box attacks [1] occur when the attacker has complete knowledge of the target model, including its architecture, parameters, and training data. This information allows the attacker to design sophisticated and tailored adversarial examples that exploit the vulnerabilities of the model. The attacker has direct access to the gradients and can optimize the perturbations for maximum effectiveness.

Examples of white-box attacks include:

Fast Gradient Sign Method (FGSM): FGSM [1,16] is a popular white-box attack. It computes the gradients of the loss function with respect to the input and perturbs the input by a small step in the direction that maximizes the loss. The resulting adversarial example can cause the model to misclassify the input.

Projected Gradient Descent (PGD): PGD [14] is an iterative white-box attack. It performs multiple iterations of the FGSM attack with a small step size, projecting the perturbed input back into an epsilon neighborhood around the original input at each iteration. This iterative process improves the effectiveness of the attack by refining the perturbations.

Deep Fool: Deep Fool [16] is another white-box attack that aims to minimize the Euclidean distance between the original input and an adversarial example within a given Lp norm constraint. It iteratively calculates the minimum perturbation required to move the input across the decision boundary of the model.

2.2.2. Black-box attacks:

Black-box attacks [2] occur when the attacker has limited or no knowledge about the target model. They do not have access to the model's architecture, parameters, or training data. Instead, the attacker can only interact with the model by providing inputs and observing the corresponding outputs. Black-box attacks often rely on transferability, which refers to the ability of adversarial examples generated on one model to also deceive other models.

Examples of black-box attacks include:

Zeroth Order Optimization (ZOO) [19]: ZOO is a black-box attack that approximates the gradients by querying the target model multiple times with carefully crafted inputs and observing the corresponding outputs. It then utilizes these approximate gradients to generate adversarial examples.

Boundary Attack [5,8]: The boundary attack is a query-based black-box attack. It starts with an initial random input and iteratively moves towards the decision boundary of the target model by performing queries. It adapts the input based on the model's responses until it finds an adversarial example.

Genetic Algorithm-based attacks [9,12]: Genetic Algorithm (GA) based attacks are evolutionary optimization methods that utilize a population-based search strategy to generate adversarial examples. These attacks modify the input iteratively using mutation and crossover operations inspired

by the principles of natural evolution.

These examples represent a subset of the many attack methods used in white-box and black-box settings. Adversarial attacks pose a significant challenge to the robustness of machine learning models, and defense strategies are continually being developed to mitigate their impact.

2.3. Defense Strategy

2.3.1. Filtering

Filtering-based adversarial attack defense strategies refer to approaches that aim to mitigate the impact of adversarial attacks by employing filtering mechanisms to identify and filter out potentially adversarial inputs. These strategies focus on detecting and rejecting inputs that exhibit characteristics of adversarial perturbations, thereby protecting the underlying machine learning models from making erroneous predictions [4].

One common filtering-based defense strategy is based on input preprocessing techniques. This approach involves applying preprocessing steps to incoming inputs before they are fed into the model for prediction. These preprocessing steps can include noise reduction, image denoising, or spatial filtering techniques. By reducing the impact of perturbations or filtering out suspicious patterns, these techniques help to enhance the robustness of the model against adversarial attacks [13].

Another filtering-based defense strategy involves utilizing anomaly detection algorithms. These algorithms aim to identify inputs that deviate significantly from the expected distribution of normal or benign samples. By leveraging statistical or machine learning techniques, these algorithms can detect adversarial inputs that exhibit anomalous properties and flag them as potentially malicious. This enables the system to reject or subject such inputs to further scrutiny before making predictions [15].

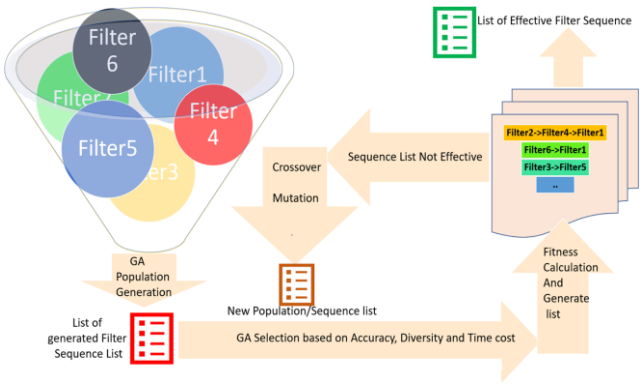


Fig 2. Filtering Based Approach

As shown in fig 2. Filtering-based defense strategies can also utilize thresholding mechanisms. These mechanisms involve setting specific thresholds for certain features or properties of inputs. Inputs that exceed these thresholds are deemed potentially adversarial and are filtered out or subjected to additional scrutiny. For example, in image classification tasks, a threshold can be set on the magnitude of pixel changes, and inputs that exceed this threshold are considered adversarial and rejected.

Furthermore, ensemble-based approaches can be employed as filtering-based defense strategies. Ensemble models combine multiple models or classifiers to make predictions, and by leveraging the diversity of these models, they can detect and filter out adversarial inputs. The disagreement among the ensemble members can indicate the presence of adversarial perturbations, allowing the defense system to reject such inputs.

However, it is important to note that filtering-based defense strategies may not provide foolproof protection against adversarial attacks. Adversarial attacks are constantly evolving, and attackers can adapt their strategies to bypass filtering mechanisms. Therefore, it is crucial to regularly evaluate and update the defense mechanisms to ensure their effectiveness against emerging attack techniques.

In summary, filtering-based adversarial attack defense strategies employ various techniques such as input preprocessing, anomaly detection, thresholding, and ensemble-based approaches to identify and filter out potentially adversarial inputs. While these strategies can enhance the robustness of machine learning models, they should be combined with other defense techniques and undergo continuous evaluation to counter evolving adversarial threats effectively.

2.3.2. Encoder-Decoder

Encoder-decoder-based adversarial attack defense strategies refer to defense mechanisms that leverage the concept of encoder-decoder architectures to protect machine learning models from adversarial attacks [3]. These strategies aim to reconstruct the original input from the potentially perturbed or adversarial input, effectively removing the adversarial perturbations and restoring the integrity of the input before it is fed into the model for prediction.

In this defense strategy, an encoder-decoder architecture is utilized, where the encoder is responsible for encoding the input data into a latent representation, and the decoder reconstructs the original input from this latent representation [7]. The key idea is that the adversarial perturbations, which are designed to cause misclassification or erroneous predictions, would be distorted or eliminated during the reconstruction process.

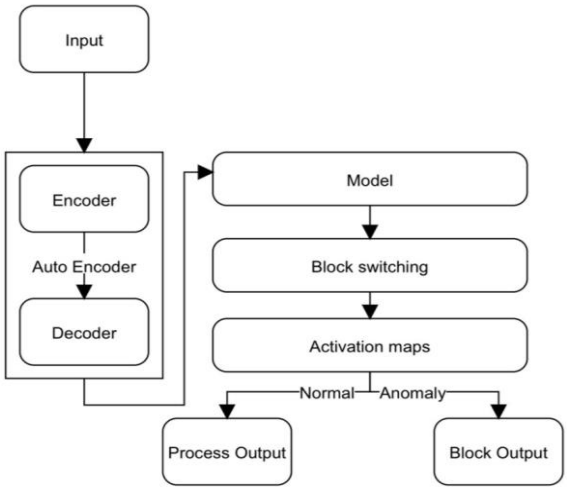


Fig 3. Encoder-Decoder [23]

As shown in fig 3. The encoder-decoder-based defense strategy typically involves the following steps:

Training: The encoder-decoder architecture is trained on a dataset consisting of clean, non-adversarial inputs. The encoder learns to extract a meaningful representation of the input data, while the decoder learns to reconstruct the original input from this representation.

Adversarial input reconstruction: When an adversarial input is encountered, it is passed through the encoder to obtain its latent representation. This latent representation is then fed into the decoder, which reconstructs the original input. The reconstruction process aims to remove or diminish the impact of adversarial perturbations, effectively restoring the input to its clean state.

Reconstructed input evaluation: The reconstructed input is evaluated by the machine learning model for prediction. Since the adversarial perturbations have ideally been eliminated or significantly reduced during the reconstruction process, the model can make predictions based on the restored, clean input, which is expected to be more reliable and accurate [6\.

Encoder-decoder-based defense strategies provide a mechanism to counter adversarial attacks by leveraging the reconstruction capability of the decoder to eliminate adversarial perturbations. By restoring the integrity of the input, these strategies aim to enhance the robustness of the model and reduce its vulnerability to adversarial attacks.

However, it is important to note that encoder-decoder-based defense strategies may have limitations. Adversarial attacks are continually evolving, and attackers can adapt their techniques to bypass reconstruction-based defenses. Furthermore, these defense strategies may introduce additional computational overhead due to the encoder-decoder architecture and the reconstruction process [24].

In summary, encoder-decoder-based adversarial attack defense strategies employ an encoder-decoder architecture to reconstruct the original input from the adversarial or perturbed input, aiming to remove or diminish the impact of adversarial perturbations. These strategies offer a promising approach to enhance the robustness of machine learning models against adversarial attacks, but they should be combined with other defense techniques and undergo rigorous evaluation to effectively counter evolving attack methods.

2.3.3. Deep Learning Model

CNN model-based adversarial attack defense strategies refer to defense mechanisms as shown in fig 4. It utilizes convolutional neural network (CNN) models to protect against adversarial attacks. These strategies aim to enhance the robustness of CNN models by incorporating specific design principles or techniques that can mitigate the impact of adversarial perturbations [1].

CNN model-based defense strategies typically involve the following approaches:

Adversarial training: This approach involves augmenting the training process of the CNN model with adversarial examples. Adversarial examples are generated by applying carefully crafted perturbations to the original input data [16]. By training the CNN model on a combination of clean and adversarial examples, the model can learn to recognize and handle adversarial perturbations, thereby improving its robustness against attacks.

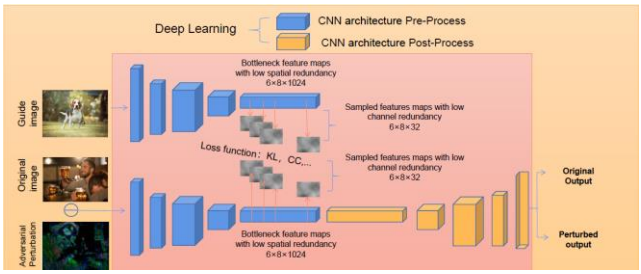


Fig 4. Deep Learning [1]

Defensive distillation: Defensive distillation is a technique that involves training a CNN model using softened probabilities instead of hard labels. It introduces a temperature parameter during training, which smooths the predicted probabilities across different classes [17]. This smoothing effect can make the model more resilient to small perturbations introduced by adversarial attacks.

Feature squeezing: Feature squeezing aims to reduce the vulnerability of CNN models to adversarial attacks by reducing the dimensionality of input data. This can be achieved by quantizing or reducing the bit-depth of input images, which effectively removes fine-grained details that might be exploited by adversarial perturbations [10]. By reducing the input space, feature squeezing can make it harder for adversarial perturbations to cause significant

impact.

Gradient masking or obfuscation: Gradient masking is a defense technique that involves intentionally obscuring or hiding gradient information during adversarial attacks. Adversarial attacks often rely on computing gradients of the loss function with respect to the input data to generate perturbations. By introducing random noise or perturbations to the gradients, the defense strategy can make it more difficult for attackers to craft effective adversarial perturbations [14].

Adversarial detection: This approach involves incorporating additional modules or classifiers within the CNN model to detect adversarial examples. These modules analyze the input data and make predictions on whether the input is clean or adversarial. By identifying and rejecting adversarial inputs, the defense strategy can protect the model from making incorrect predictions based on such inputs.

In summary, CNN model-based adversarial attack defense strategies employ specific design principles or techniques within the CNN model to enhance its robustness against adversarial attacks. These strategies involve approaches such as adversarial training, defensive distillation, feature squeezing, gradient masking, and adversarial detection. By incorporating these techniques, the defense strategies aim to mitigate the impact of adversarial perturbations and improve the resilience of CNN models against attacks.

2.3.4. Generative Adversarial Network (GAN)

GAN-based adversarial attack defense strategies refer to defense mechanisms that utilize the concept of Generative Adversarial Networks (GANs) to protect machine learning models from adversarial attacks [8,9]. These strategies leverage the adversarial training framework of GANs to generate robust and resilient models that are less susceptible to adversarial perturbations.

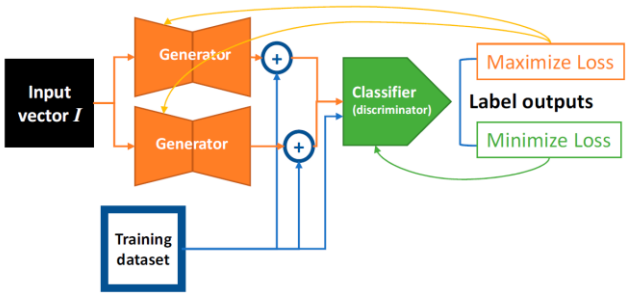


Fig 5. Generative Adversarial Network [26]

As shown in fig 5. GAN-based defense strategy typically involves the following steps:

Generation of adversarial examples: GANs are trained in an adversarial manner, consisting of a generator and a discriminator [12] The generator aims to generate synthetic examples that resemble the training data, while the discriminator tries to distinguish between the real and synthetic examples. Adversarial examples can be generated by perturbing the input data with the gradients obtained from the discriminator.

Adversarial training with GANs: The machine learning model is trained on a combination of clean examples and the adversarial examples generated by the GAN. The model is exposed to both clean and adversarial inputs during training, enabling it to learn robust representations and decision boundaries that are less susceptible to adversarial perturbations [19.22].

Adversarial regularization: In this approach, an additional term is added to the loss function during training to encourage the model to make more robust predictions. This regularization term is typically based on the divergence or discrepancy between the predictions made on clean examples and the predictions made on their adversarial counterparts. By minimizing this discrepancy, the model is incentivized to be more resilient against adversarial attacks [25].

Adversarial transformation or augmentation: GANs can be used to generate transformed or augmented versions of the input data, which can help expose the model to a wider range of potential adversarial perturbations. By training on these transformed examples, the model can learn to generalize and adapt to various types of adversarial attacks, improving its robustness.

Defense module using GANs: GANs can also be utilized as a separate defense module to identify and filter out adversarial examples [26]. The GAN discriminator is employed to distinguish between clean and adversarial inputs, allowing the defense module to reject or handle the adversarial examples appropriately.

It is important to note that GAN-based defense strategies are not foolproof and have their limitations. Adversarial attacks continue to evolve, and attackers can adapt their techniques to bypass GAN-based defenses [27]. Therefore, it is crucial to combine GAN-based strategies with other defense techniques and regularly evaluate their effectiveness against emerging attack methods.

In summary, GAN-based adversarial attack defense strategies leverage the adversarial training framework of GANs to generate robust models and improve their resilience against adversarial perturbations. These strategies involve steps such as generation of adversarial examples, adversarial training with GANs, adversarial regularization, adversarial transformation, or augmentation, and using GANs as a defense module. By incorporating GANs, the defense strategies aim to enhance the robustness of machine learning models against adversarial attacks.

3. Comparative Study

Below Table 1. Shows the different defense strategy for adversarial attacks among them most suitable method is GAN based modeling.

Table 1. Comparative Study

Defense Strategy	Strengths	Limitations
Filtering [4,13,15]	- Simple and easy to implement	- Limited effectiveness against sophisticated attacks
	- Can effectively remove noise and perturbations	- May introduce false positives or false negatives
Encoder-Decoder [3,7,6,24]	- Low computational overhead	- Vulnerable to adaptive attacks
	- Able to reconstruct original input from adversarial input	- Limited effectiveness against targeted or sophisticated attacks
	- Reduces impact of adversarial perturbations	- Performance depends on the quality of encoder-decoder model
Deep Learning [1,16,17,10,14]	- Can be used in combination with other defense strategies	- Additional computational overhead due to reconstruction
	- Robustness against simple attacks	- Vulnerable to adaptive attacks
	- Generalization to unseen adversarial examples	- Limited effectiveness against sophisticated attacks
	- Can handle complex data types (images, text, etc.)	- High computational requirements for training large models
GAN Model [8,9,12,19,20,21,22,25,26]	- Can be enhanced with additional defense techniques	- May suffer from performance degradation due to defenses
	- Can generate robust models less susceptible to attacks	- Attacks can still be effective if the substitute model is not sufficiently like the target model
	- Resilient against certain transfer-based attacks	- Computational overhead due to training and using GANs
	- Can be combined with other defense strategies	- Vulnerable to advanced and adaptive attacks

4. Results Analysis

Below are commonly used evaluation metrics in the field of image processing and computer vision. They are used to measure different aspects of image quality and similarity.

MAE (Mean Absolute Error) is calculated as the average absolute difference between corresponding pixels in two images:

$$MAE = (1/N) \sum |I1(i,j) - I2(i,j)| \quad (1)$$

PSNR (Peak Signal-to-Noise Ratio) is computed as the logarithm of the ratio between the maximum pixel value and the mean squared error (MSE) between two images:

$$PSNR = 20 * log10\,(MAX) - 10 * log10\,(MSE) \quad (2)$$

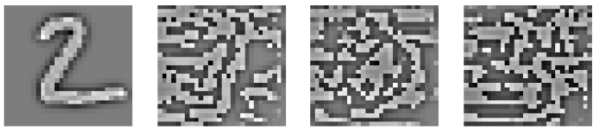
SSIM (Structural Similarity Index) measures the structural similarity by comparing local image patches and computing luminance, contrast, and structural components:

$$SSIM = (2\mu1\mu2 + C1) * (2\sigma12 + C2) / (\mu1\text{^}2 + \mu2\text{^}2 + C1) * (\sigma1\text{^}2 + \sigma2\text{^}2 + C2) \quad (3)$$

Correctly classified: 184
Incorrectly classified: 316
MAE: 0.19804340600967407
SSIM: 0.0017039392841979861
ACC: 0.368

Correctly classified: 132
Incorrectly classified: 368
MAE: 0.17982342839241028
SSIM: 0.002825278090313077
ACC: 0.264

(a) FGSM **(b) PGD**



Correctly classified: 202
Incorrectly classified: 298
MAE: 0.279354065656662
SSIM: 0.007923208177089691
ACC: 0.404

(c) DeepFool

Fig 6. Filtering Defense on (a) FGSM (b) PGD and (c) DeepFool Attacks

As shown in Fig 6. Filtering Defense shows when DeepFool Attack is applied the defense process does not work to denoise attacks pixels.

Correctly classified: 467
Incorrectly classified: 33
MAE: 0.14346034824848175
SSIM: 0.6606622338294983
ACC: 0.934

Correctly classified: 469
Incorrectly classified: 31
MAE: 0.1220720037817955
SSIM: 0.6971783638000488
ACC: 0.938

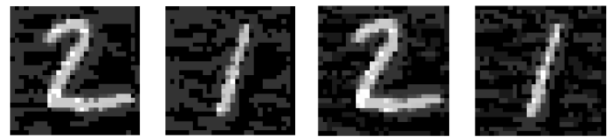
(a) FGSM **(b) PGD**

Correctly classified: 181
Incorrectly classified: 319
MAE: 0.06408143788576126
SSIM: 0.9069164395332336
ACC: 0.362

(c) DeepFool

Fig 7. Encoder-Decoder Defense on (a) FGSM (b) PGD and (c) DeepFool Attacks

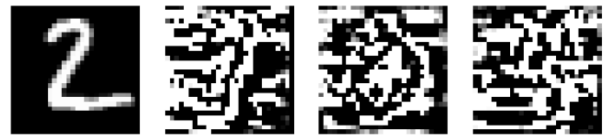
As shown in Fig 7. Encoder-Decoder Defense shows when DeepFool Attack is applied the defense process does not work to denoise attacks pixels.



Correctly classified: 191
Incorrectly classified: 309
MAE: 3.8526343226763515e-10
SSIM: 1.0
ACC: 0.382

Correctly classified: 84
Incorrectly classified: 416
MAE: 0.0003430326178204268
SSIM: 0.9999883770942688
ACC: 0.168

(a) FGSM **(b) PGD**

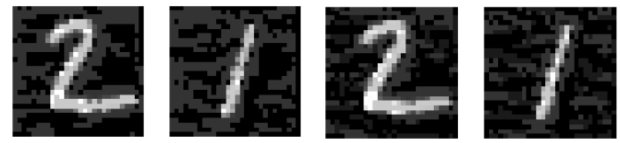


Correctly classified: 167
Incorrectly classified: 333
MAE: 0.0002752728760242462
SSIM: 0.999998927116394
ACC: 0.334

(c) DeepFool

Fig 8. Deep Learning Defense on (a) FGSM (b) PGD and (c) DeepFool Attacks

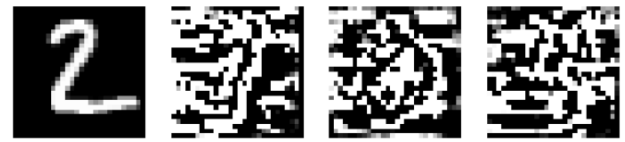
As shown in Fig 8. Deep Learning Defense shows when DeepFool Attack is applied the defense process does not work to denoise attacks pixels.



Correctly classified: 497
Incorrectly classified: 3
MAE: 3.8526343226763515e-10
SSIM: 1.0
ACC: 0.994

Correctly classified: 497
Incorrectly classified: 3
MAE: 0.0003430326178204268
SSIM: 0.9999883770942688
ACC: 0.994

(a) FGSM **(b) PGD**



Correctly classified: 167
Incorrectly classified: 333
MAE: 0.0002752728760242462
SSIM: 0.999998927116394
ACC: 0.334

(c) DeepFool

Fig 9. GAN Defense on (a) FGSM (b) PGD and (c) DeepFool Attacks

As shown in Fig 9. GAN Defense shows when DeepFool Attack is applied the defense process does not work to denoise attacks pixels.

Below Table 2. Gives numerical analysis of different preventive methods for Adversarial Attacks among them GAN based approach gives best performance than others.

Table 2. Comparative Analysis of Adversarial Attacks Defense

NO	Preventive	Dataset	FGSM			PGD			DeepFool		
			MAE	SSIM	ACC	MAE	SSIM	ACC	MAE	SSIM	ACC
1	Filtering	MINIST [1]	0.19	0.001	0.36	0.17	0.002	0.26	0.27	0.007	0.40
		CelebA [11]	0.17	0.002	0.33	0.12	0.012	0.23	0.21	0.012	0.46
		Fashion [16]	0.15	0.006	0.32	0.11	0.027	0.32	0.22	0.034	0.48
2	Encoder-Decoder	MINIST [1]	0.14	0.66	0.93	0.12	0.69	0.93	0.06	0.90	0.36
		CelebA [11]	0.18	0.72	0.89	0.11	0.79	0.73	0.07	0.92	0.37
		Fashion [16]	0.16	0.82	0.89	0.11	0.79	0.73	0.09	0.92	0.39
3	Deep Learning	MINIST [1]	0.003	1.0	0.38	0.00	0.99	0.16	0.01	0.99	0.33
		CelebA [11]	0.01	1.0	0.38	0.00	0.99	0.16	0.12	0.99	0.32
		Fashion [16]	0.02	1.0	0.38	0.00	0.99	0.16	0.23	0.99	0.31
4	GAN Model	MINIST [1]	0.003	1.0	0.99	0.01	0.99	0.99	0.1	0.99	0.39
		CelebA [11]	0.02	1.0	0.99	0.01	0.99	0.99	0.1	0.99	0.33
		Fashion [16]	0.03	1.0	0.99	0.01	0.99	0.99	0.1	0.99	0.39

5. Conclusion

In conclusion, our comprehensive study focused on exploring the impact of magnitude perturbation in adversarial attacks and defense strategies. We investigated various attack techniques and defense mechanisms to gain insights into the effectiveness and limitations of different approaches.

Through our research, we observed that adversarial attacks leveraging magnitude perturbation can significantly compromise the performance and reliability of machine learning models. Attacks such as FGSM, PGD, and Deep Fool demonstrated their ability to generate adversarial examples that deceive models and cause misclassifications. These attacks highlighted the vulnerability of models to small perturbations, emphasizing the need for robust defense mechanisms.

On the defense front, we examined different strategies, including filtering-based, encoder-decoder-based, deep learning/CNN model-based, and GAN model-based defenses. Each approach exhibited its strengths and limitations in mitigating adversarial attacks. Filtering-based defenses demonstrated simplicity and effectiveness in removing noise and perturbations but had limited effectiveness against sophisticated attacks. Encoder-decoder-based defenses showed promise in reconstructing original inputs and reducing the impact of adversarial perturbations, but they may struggle against targeted or sophisticated attacks. Deep learning/CNN model-based defenses exhibited robustness against simple attacks and generalization to unseen adversarial examples but were vulnerable to adaptive attacks. GAN model-based defenses provided the potential to generate more robust models, but their effectiveness relied on the similarity between the substitute and target models.

Overall, our study underscores the dynamic nature of adversarial attacks and defense strategies. It highlights the ongoing need for continuous research and development in this field to stay ahead of evolving attack techniques. Additionally, combining multiple defense strategies and exploring hybrid approaches may offer more comprehensive protection against adversarial attacks.

As the arms race between attackers and defenders persists, our comprehensive study contributes to the growing body of knowledge in adversarial attack and defense research. By shedding light on the impact of magnitude perturbation and analyzing various defense strategies, we hope to inspire further advancements in developing robust and resilient machine learning models that can withstand adversarial attacks in real-world scenarios.

6. References and Footnotes

Author contributions

Conceptualization—Bhasha Anjaria and Dr. Jaimeel Shah; Methodology—Bhasha Anjaria; Software—Bhasha Anjaria; Validation—Dr. Jaimeel Shah; Formal Analysis—Bhasha Anjaria and Dr. Jaimeel Shah; Investigation—Dr. Jaimeel Shah; Resources—Bhasha Anjaria; data curation—Bhasha Anjaria; Writing—review and editing—Bhasha

Anjaria and Dr. Jaimeel Shah; Visualization—Bhasha Anjaria; supervision—Dr. Jaimeel Shah; Project administration—Bhasha Anjaria.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Akhtar, N., Mian, A., Kardan, N., Shah, M.: Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. IEEE Access. 9, 155161–155196 (2021). https://doi.org/10.1109/ACCESS.2021.3127960.

[2] Almuflih, A.S., Vyas, D., Kapdia, V. V, Qureshi, M.R.N.M., Qureshi, K.M.R., Makkawi, E.A.: Novel exploit feature-map-based detection of adversarial attacks. Applied Sciences. 12, 5161 (2022).

[3] Bakhti, Y., Fezza, S.A., Hamidouche, W., Deforges, O.: DDSA: A Defense against Adversarial Attacks Using Deep Denoising Sparse Autoencoder. IEEE Access. 7, 160397–160407 (2019). https://doi.org/10.1109/ACCESS.2019.2951526.

[4] Dasgupta, D., Gupta, K.D.: Dual-filtering (DF) schemes for learning systems to prevent adversarial attacks. Complex and Intelligent Systems. (2022). https://doi.org/10.1007/s40747-022-00649-1.

[5] Deldjoo, Y., Noia, T. Di, Merra, F.A.: A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks. ACM Comput. Surv. 54, (2021). https://doi.org/10.1145/3439729.

[6] Despiegel, V., Despiegel, V.: ScienceDirect Dynamic Dynamic Autoencoders Autoencoders Against Against Adversarial Adversarial Attacks Attacks. 00, (2023). https://doi.org/10.1016/j.procs.2023.03.104.

[7] Folz, J., Palacio, S., Hees, J., Dengel, A.: Adversarial defense based on structure-to-signal autoencoders. Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020. 3568–3577 (2020). https://doi.org/10.1109/WACV45572.2020.9093310.

[8] Han, S., Lin, C., Shen, C., Wang, Q., Guan, X.: Interpreting Adversarial Examples in Deep Learning: A Review. (2023). https://doi.org/10.1145/3594869.

[9] Hu, W., Tan, Y.: Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN BT - Data Mining and Big Data. Presented at the (2022).

[10] Jin, L., Tan, F., Jiang, S.: Generative Adversarial Network Technologies and Applications in Computer Vision. Computational Intelligence and Neuroscience. 2020, (2020). https://doi.org/10.1155/2020/1459107.

[11] Kuribayashi, M.: Defense Against Adversarial Attacks BT - Frontiers in Fake Media Generation and Detection. Presented at the (2022). https://doi.org/10.1007/978-981-19-1524-6_6.

[12] Laykaviriyakul, P., Phaisangittisagul, E.: Collaborative Defense-GAN for protecting adversarial attacks on classification system. Expert Systems with Applications. 214, 118957 (2023). https://doi.org/https://doi.org/10.1016/j.eswa.2022.118957.

[13] Li, F., Du, X., Zhang, L.: Adversarial Attacks Defense Method Based on Multiple Filtering and Image Rotation. Discrete Dynamics in Nature and Society. 2022, (2022). https://doi.org/10.1155/2022/6124895.

[14] Liang, H., He, E., Zhao, Y., Jia, Z., Li, H.: Adversarial Attack and Defense: A Survey. Electronics (Switzerland). 11, 1–19 (2022). https://doi.org/10.3390/electronics11081283.

[15] Liu, S., Zhuang, Y., Ma, X., Wang, H., Cao, D.: An Adversarial Sample Defense Method Based on Saliency Information BT - Ubiquitous Security. Presented at the (2023).

[16] Ren, K., Zheng, T., Qin, Z., Liu, X.: Adversarial Attacks and Defenses in Deep Learning. Engineering. 6, 346–360 (2020). https://doi.org/10.1016/j.eng.2019.12.012.

[17] Ryu, G., Choi, D.: A hybrid adversarial training for deep learning model and denoising network resistant to adversarial examples. Applied Intelligence. 9174–9187 (2022). https://doi.org/10.1007/s10489-022-03991-6.

[18] Samangouei, P., Kabkab, M., Chellappa, R.: Defense-Gan: Protecting classifiers against adversarial attacks using generative models. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. (2018).

[19] Singh, A.B., Awasthi, L.K., Urvashi: Defense Against Adversarial Attacks Using Chained Dual-GAN Approach BT - Smart Data Intelligence. Presented at the (2022).

[20] Taheri, S., Khormali, A., Salem, M., Yuan, J.S.: Developing a robust defensive system against adversarial examples using generative adversarial networks. Big Data and Cognitive Computing. 4, 1–15 (2020). https://doi.org/10.3390/bdcc4020011.

[21] Wang, J., Wu, H., Wang, H., Zhang, J., Luo, X., Ma, B.: Immune Defense: A Novel Adversarial Defense Mechanism for Preventing the Generation of Adversarial Examples. (2023).

[22] Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., Hossain, E., Poor, H.V.: Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey. 1–46 (2023).

[23] Yadav, A., Upadhyay, A., Sharanya, S.: An integrated Auto Encoder-Block Switching defense approach to prevent adversarial attacks. (2022).

[24] Yang, J., Shao, M., Liu, H., Zhuang, X.: Generating adversarial samples by manipulating image features with auto-encoder. International Journal of Machine Learning and Cybernetics. 14, 2499–2509 (2023). https://doi.org/10.1007/s13042-023-01778-w.

[25] Yu, F., Wang, L., Fang, X., Zhang, Y.: The defense of adversarial example with conditional generative adversarial networks. Security and Communication Networks. 2020, (2020). https://doi.org/10.1155/2020/3932584.

[26] Zhao, W., Mahmoud, Q.H., Alwidian, S.: Evaluation of GAN-Based Model for Adversarial Training. Sensors. 23, (2023). https://doi.org/10.3390/s23052697.

[27] Zhou, M., Niu, Z., Wang, L., Zhang, Q., Hua, G.: Adversarial Ranking Attack and Defense. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 12359 LNCS, 781–799 (2020). https://doi.org/10.1007/978-3-030-58568-6_46.