International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

A REVIEW OF ADVERSARIAL ATTACK METHODS ON DEEP NEURAL NETWORKS

Ashish Sagar*

Research Scholar

Department of Computer Science

NIT, Jalandhar

ABSTRACT

Because Deep Neural Networks (DNNs) can simulate complicated data representations, they have shown impressive results in a variety of applications, including autonomous systems, natural language processing, and picture classification. Despite their exceptional performance, DNNs are extremely susceptible to adversarial assaults, in which models provide inaccurate predictions due to carefully planned, undetectable modifications to input data. Particularly in security-sensitive applications like financial systems, healthcare diagnostics, and autonomous driving, this vulnerability presents serious issues. This review paper offers a thorough examination of protection mechanisms intended to lessen the threat of various adversarial attack techniques, such as backdoor, data poisoning, black-box, and white-box attacks. This study highlights important trends, difficulties, and possible avenues for further research in adversarial machine learning by reviewing the body of existing literature. The study's narrative literature review methodology provides insights into new trends while combining important research findings.

Keywords: *Adversarial Attack, Deep Neural Networks, White-box Attack, Black-box Attack, Data Poisoning, Backdoor Attack, Adversarial Defense, Machine Learning Security*

I. INTRODUCTION

Deep learning has transformed artificial intelligence in the past ten years by making it possible for machines to do tasks like speech recognition, image recognition, and natural language understanding on par with humans. With their numerous layers of nonlinear processing units, Deep Neural Networks (DNNs) are able to extract intricate patterns and representations from vast amounts of data, leading to breakthroughs in domains such as computer vision, autonomous driving, and medical diagnostics.

However, DNNs' vulnerability to adversarial assaults is a serious flaw. The technique of slightly altering input data in a way that is nearly undetectable to humans yet causes the model to provide inaccurate results is known as an adversarial attack. Szegedy et al. (2014) carried out the first significant study pointing out this issue, demonstrating that even little changes in input photos could result in a significant misdiagnosis by DNNs.

When DNNs are used in practical applications, the presence of such flaws presents significant security problems. In medical diagnostics, an adversarial attack could result in erroneous disease identification, putting patient health at risk, while an adversarial attack on an autonomous car could cause catastrophic accidents.

The state of adversarial assault techniques and defense tactics is reviewed in this work. It seeks to provide researchers and practitioners with a comprehensive grasp of the state of affairs, obstacles, and potential paths forward in protecting deep learning systems from hostile attacks.

1.1 Deep Neural Networks

Multiple layers of artificial neurons make up Deep Neural Networks (DNNs), a form of machine learning model. These layers enable DNNs to model intricate relationships in data without the need for manual feature engineering by hierarchically extracting more complicated features from input data (LeCun, Bengio, & Hinton, 2015).

DNNs have achieved outstanding success across a variety of applications, such as:

- **Image Classification**: Recognizing objects and scenes from images (Krizhevsky, Sutskever, & Hinton, 2012).
- **Natural Language Processing (NLP)**: Machine translation, sentiment analysis, and text generation (Vaswani et al., 2017).
- **Autonomous Vehicles**: Real-time decision-making and object detection for safe driving (Bojarski et al., 2016).
- **Medical Diagnostics**: Automatic disease detection using medical images such as MRIs and X-rays (Esteva et al., 2017).
- **Financial Forecasting**: Stock market prediction and credit risk assessment (Heaton, Polson, & Witte, 2017).

However, the power of DNNs comes with major limitations—most notably their vulnerability to adversarial attacks, where small perturbations in input data can lead to incorrect model outputs.

1.2 Emergence of Adversarial Attacks

Szegedy et al. (2014) were the first to reveal that DNNs are susceptible to adversarial assaults. They demonstrated that subtle changes made to input images could cause significant misclassification. The Fast Gradient Sign Method (FGSM), later proposed by Goodfellow et al. (2015), allows attackers to quickly create adversarial samples by figuring out the gradient of the loss function with respect to the input.

Different types of adversarial attacks include:

- **White-box Attacks**: The attacker has full knowledge of the model architecture, parameters, and gradients.
- **Black-box Attacks**: The attacker can only query the model and observe outputs without access to internal structures [Papernot et al., 2017].
- **Data Poisoning Attacks**: Malicious samples are inserted into the training data, causing the model to learn incorrect patterns [Biggio et al., 2012].

- **Backdoor Attacks**: Hidden triggers are embedded during training, causing the model to misclassify when the trigger is present [Gu et al., 2017].

These attacks are especially harmful in real-world applications like financial systems (false fraud detection), healthcare systems (inaccurate medical diagnosis), and autonomous driving (misclassification of traffic signs).

1.3 Research Objectives

The primary research objectives of this review paper are:

1. To classify and analyze various adversarial attack methods targeting deep neural networks.
2. To review and evaluate defense mechanisms proposed in the literature to counter adversarial attacks.
3. To identify existing research gaps and recommend potential future research directions that enhance the robustness and security of DNNs.

II. LITERATURE REVIEW

According to Stanly, Shalinie, and Paul (2023), millions of photos being added to repositories every millisecond in this rapidly evolving digital age. Deep neural networks extract and analyze these context-rich images with a wealth of underlying data for countless research applications. Major advances in deep neural networks are yielding multiple advantages including increasing prediction and accuracy rates. Even while deep learning models are improving vertically, security threats always exist and undermine the power of deep neural networks. An opponent manipulating archived photos will still appear to the user to be identical. These days, there is an increase in these manipulations, known as adversarial assaults, which fool deep learning networks into producing incorrect predictions. This study examines the various forms of adversarial attacks, providing insight into their history, multifaceted development, uses, and difficulties.

Artificial intelligence technologies have been widely applied in computer vision, natural language processing, automatic driving, and other sectors in recent years, according to a study by Qiu et al. (2019). Nevertheless, adversarial assaults can target artificial intelligence systems, which restricts the use of AI technology in critical security domains. Therefore, strengthening AI systems' resistance to hostile attacks has become more and more crucial to the advancement of AI. The goal of this paper is to provide a thorough overview of the most recent developments in deep learning research on adversarial attack and defense systems. This paper explains the adversarial attack techniques in the training and testing stages of the target model, respectively, based on the various stages where the adversarial assault took place. Next, we categorize the ways in which adversarial attack technologies are used in the physical world, computer vision, natural language processing, and cyberspace security. Lastly, we outline the three primary categories of current adversarial defense techniques: data modification, model modification, and auxiliary tool use.

According to Zhou et al. (2022), deep learning applications have been encouraged across a wide range of disciplines due to the exceptional performance of deep neural networks. However, the widespread adoption of deep learning has been impeded by the possible hazards posed by adversarial samples. In these situations, the model's final performance is greatly reduced by hostile perturbations that are invisible to the human eye. Adversarial attacks and their defenses in the field of deep learning have been the subject of numerous published articles. In contrast to poisoning attacks, which involve inserting tainted data into the training set, most concentrate on evasion attacks, in which the adversarial samples are discovered during testing. Furthermore, because there are no established evaluation techniques, it is challenging to assess the resilience of a deep learning model or the actual threat of adversarial attacks. Therefore, we examine the existing literature in this work. We also try to provide the first framework for analyzing adversarial attacks in a methodical way. The framework is designed to offer a lifecycle for hostile assaults and defenses from a cybersecurity standpoint.

According to Abomakhelb et al. (2025), research on artificial intelligence (AI) security is extremely significant and promising in the present decade. More focus is being placed on deep neural network (DNN) security in particular. DNNs are the most commonly used and have a substantial share in both industry and research, despite their recent rise to prominence as a prominent tool for handling complex challenges across a variety of machine learning (ML) tasks. However, DNNs are susceptible to adversarial attacks, in which small but deliberate perturbations can fool DNN models. As a result, a number of studies have suggested that DNNs are vulnerable to novel assaults. Researchers must investigate defenses that lessen the hazards involved and improve the dependability of modifying DNNs for a range of crucial applications in light of the growing frequency of these attacks. Consequently, a number of defense techniques have been used to safeguard DNNs from adversarial attacks. As a fundamental technique for all ML activities, DNN is our main emphasis. In this work, we thoroughly review and present the most recent DNN security research based on different machine learning tasks, emphasizing the adversarial attacks that lead to DNN failure and the defense tactics that keep the DNNs safe. We examine, discuss, and clarify the workings of common adversarial assaults and response strategies that are relevant to all DNN-based ML tasks. The majority of the most recent state-of-the-art attacks and defenses are thoroughly and robustly reviewed in our evaluation, which offers a thorough taxonomy for attacker and defender concerns. We also take a close look at the latest systematic assessment of the metrics used to assess the effectiveness of attack or defense strategies. We conclude by discussing the field's present difficulties, unresolved problems, and potential future study avenues.

As AI has become the industry's focus and GenAI applications have gained popularity globally, adversarial machine learning (AML) assaults have become a significant problem for enterprises in recent years, according to Malik et al. (2023). Although businesses are keen to invest in GenAI applications and create their own huge language models, they must contend with a number of security and privacy concerns, especially those related to AML assaults. Many large-scale machine learning models have been compromised by AML assaults. AML attacks have the potential to drastically lower machine learning models' accuracy and efficiency if they are executed effectively. In the context of autonomous transportation and vital healthcare, they have far-reaching detrimental effects. This study uses adversarial strategies and methodologies to identify, assess, and classify AML attacks. Open-source tools for evaluating AI and ML models against AML assaults are also suggested by this study. Additionally, this study recommends certain defenses against every attack. It seeks to provide companies with guidelines on how to protect themselves from AML threats and ensure that ML models are secure.

According to Li et al. (2021), recent research has shown that many classification techniques, particularly Deep Neural Networks (DNNs), are susceptible to adversarial examples, which are deliberately constructed to trick a well-trained classification model while being indistinguishable from natural data to humans. This is true even though machine learning systems are efficient and scalable. Because of this, using DNNs or similar techniques in security-critical areas may be dangerous. Many efforts have been made in this area since Biggio et al. and Szegedy et al. initially recognized this problem, including creating attack strategies to produce adversarial cases and defense strategies to prevent such examples. With an emphasis on the creation and protection of hostile examples, this article seeks to familiarize the statistical community with this subject and its most recent advancements. Readers can examine the examined methods by using the publicly available computing codes (in Python and R) utilized in the numerical trials. The authors believe that this paper will inspire additional statisticians to pursue this vital and fascinating area of creating and defending against hostile cases.

According to a study by Barr (2025), adversarial assaults are putting the security and reliability of neural network designs at greater risk. These attacks can interrupt applications, generate false positives, and reduce performance, especially on resource-constrained Internet of Things (IoT) devices. A two-step methodology is used in this study: first, a resilient Convolutional Neural Network (CNN) that performs well on the MNIST dataset is designed; second, its robustness against sophisticated adversarial techniques like Deepfool and L-BFGS is assessed and improved. According to preliminary tests, the suggested CNN is susceptible to adversarial attacks even though it does well on common classification tasks. The suggested CNN was re-trained using APE-GAN, a cutting-edge adversarial training technique, to address this issue. This greatly increased the CNN's resistance to adversarial attacks while maximizing performance for embedded systems with constrained computational power. By surpassing traditional techniques and proving to be a trailblazing solution in adversarial machine learning, systematic experimentation shows how well APE-GAN works to improve the accuracy and robustness of the suggested CNN. This research represents a major advancement in tackling the difficulties presented by adversarial attacks by incorporating APE-GAN into the training process, which guarantees the safe and effective functioning of the suggested CNN in actual IoT applications.

Goodfellow et al. (2015), FGSM is a gradienta) s(-based method where the adversarial example x′x'x′ is computed as:

x′=x+ϵ⋅sign(∇xJ(θ,x,y))x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))x′=x+ϵ⋅sign(∇x J(θ,x,y))

Where ϵ\epsilonϵ is a small perturbation factor, JJJ is the loss function, and θ\thetaθ represents model parameters. FGSM is fast and efficient but often less powerful compared to iterative attacks.

Madry et al. (2018) extended FGSM into an iterative method where multiple small steps are taken toward the adversarial direction, projecting back into the valid input space after each iteration. PGD is one of the most widely used benchmarks for robustness evaluation.

The attack was formulated by Carlini and Wagner (2017) as an optimization problem that achieves high confidence in misclassification while minimizing the perturbation. It avoids a lot of detecting techniques and is more computationally costly yet quite successful.

By taking advantage of the fact that various models share similar decision boundaries, Papernot et al. (2016) showed that adversarial examples created on a substitute model could transfer to the target model.

In order to create adversarial examples without internal knowledge, Ilyas et al. (2018) presented techniques in which the attacker repeatedly queries the target model with different inputs and watches outputs.

According to Biggio et al. (2012), well-designed poisoning sites might cause the model to misclassify particular inputs or reduce its accuracy. This is especially risky when training data is gathered from unreliable sources or crowdsourced.

According to Gu et al. (2017), a hidden pattern (trigger) is injected during training, causing the model to function correctly when the trigger is absent but to produce a predetermined wrong label when it is present. Since the model seems to function effectively under typical conditions, these attacks are covert and challenging to identify.

III. DEFENSE STRATEGIES

3.1 Adversarial Training

One of the most studied and successful defenses is adversarial training. The model may learn to accurately identify both clean and adversarial samples using this strategy, which incorporates adversarial data into the training process. Using Projected Gradient Descent (PGD) examples, Madry et al. (2018) showed that adversarial training greatly increases robustness. On clean data, however, this method frequently results in a minor reduction in accuracy and requires significant processing resources.

3.2 Input Transformation

Before supplying data to the model, input preprocessing methods like feature squeezing, picture denoising, and JPEG compression seek to eliminate adversarial perturbations. Although these techniques are simple to use and lightweight, they do not ensure complete robustness, particularly when facing adaptive adversaries.

3.3 Model Regularization

By lessening the model's sensitivity to slight input changes, regularization strategies like weight decay and dropout enhance generalization and lessen overfitting. Regularization is useful, but it is not enough to fend off sophisticated adversarial attacks.

3.4 Certified Defenses

Provable assurances that model predictions stay constant within a given perturbation range are offered by certified defenses. One prominent technique for classifying inputs with several noisy versions is randomized smoothing (Cohen et al., 2019). Although certified defenses are theoretically sound, they can reduce accuracy on clean data and are typically computationally costly.

IV. RESEARCH METHODOLOGY

The narrative literature review approach serves as the foundation for this review work. The study entails a thorough examination of scholarly journals, conference proceedings, peer-reviewed articles, and other reputable publications that concentrate on adversarial attack techniques and defense measures in deep neural networks (DNNs).

From 2014 to 2023, pertinent publications were chosen from scholarly databases like IEEE Xplore, Springer, arXiv, and MDPI. Without limiting the study to a tight, methodical process, the book selection concentrated on offering insights into important adversarial assault tactics and accompanying mitigation mechanisms.

Through the synthesis of important research findings, critical discussion, and insights into new trends and upcoming difficulties in the field of adversarial machine learning, this approach seeks to present a thorough knowledge.

V. CONCLUSION

The security and dependability of deep neural networks are still seriously threatened by adversarial assaults. While backdoor and data poisoning attacks focus on the training process, white-box and black-box attacks take use of flaws in the model architecture. Although they are not always reliable, current protection techniques like input preprocessing and adversarial training show promise. Although certified defenses come with computational expenses, they provide theoretical assurances.

The absence of established evaluation criteria, difficulties in practically deploying defenses, and a lack of adaptive defense tactics that can be applied to many types of attacks are important research gaps. Enhancing model interpretability, creating provable resilient architectures, and improving real-time adversarial input detection and mitigation should be the main goals of future research.

REFERENCES

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6572
2. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1706.06083
3. Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. *IEEE Symposium on Security and Privacy*. https://arxiv.org/abs/1608.04644
4. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIACCS)*. https://arxiv.org/abs/1602.02697
5. Ilyas, A., Engstrom, L., Athalye, A., & Lin, J. (2018). Black-box Adversarial Attacks with Limited Queries and Information. *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/1804.08598

6. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning Attacks against Support Vector Machines. *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/1206.6389
7. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. https://arxiv.org/abs/1708.06733
8. Hamil Stanly, Mercy Shalinie S., Riji Paul, A review of generative and non-generative adversarial attack on context-rich images, Engineering Applications of Artificial Intelligence, Volume 124, 2023, 106595, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2023.106595.m/science/article/abs/pii/S0952197623007790
9. Qiu, Shilin & Liu, Qihe & Zhou, Shijie & Wu, Chunjiang. (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies. Applied Sciences. 9. 909. 10.3390/app9050909..
10. Zhou, Shuai & Liu, Chi & Ye, Dayong & Zhu, Tianqing & Zhou, Wanlei & Yu, Philip. (2022). Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity. ACM Computing Surveys. 55. 10.1145/3547330.
11. Abomakhelb, Abdulruhman & Jalil, Kamarularifin & Buja, Alya & Alhammadi, Abdulraqeb & Alenezi, Abdulmajeed. (2025). A Comprehensive Review of Adversarial Attacks and Defense Strategies in Deep Neural Networks. Technologies. 13. 202. 10.3390/technologies13050202.
12. JASMITA MALIK , RAJA MUTHALAGU AND PRANAV M. PAWAR (2024), A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies; Systematic Review of AML Attacks, Defensive Controls, and Technologies
13. Li, Y., Cheng, M., Hsieh, C. J., & Lee, T. C. M. (2022). A Review of Adversarial Attack and Defense for Classification Methods. The American Statistician, 76(4), 329–345. https://doi.org/10.1080/00031305.2021.2006781
14. M. Barr, “A Robust Neural Network against Adversarial Attacks”, Eng. Technol. Appl. Sci. Res., vol. 15, no. 2, pp. 20609–20615, Apr. 2025.