A COMPREHENSIVE REVIEW OF ADVERSARIAL ATTACKS IN MACHINE LEARNING

Abdirashid Abukar Ahmed[1], Dr. Nirvair Neeru[2],

1 Scholar at Department of Computer Science & Engineering, Punjabi University, Patiala, Punjab - 147002, India
Abdirashidit14@gmail.com

2 Assistant Professor at Department of Computer Science & Engineering, Punjabi University, Patiala, Punjab - 147002,
Nirvair.ce@pbi.ac.in

----------***----------

ABSTRACT

Adversarial attacks pose a substantial risk to the security and dependability of machine learning (ML) models. These attacks entail creating harmful inputs, known as adversarial instances, which can lead models to provide inaccurate predictions. This article offers a thorough examination of adversarial assaults in machine learning, including their many forms, techniques of generation, current research, and potential future research areas. We analyze well-known attack techniques such as FGSM, DeepFool, Carlini & Wagner (C&W), and ZOO, emphasizing their advantages and constraints. The research limitations we have identified relate to the lack of comprehensive comparative analysis and the absence of a structured decision-making framework for offensive technique selection. In addition, we investigate the research obstacles related to adversarial variety, dynamic assault environments, the capacity to transfer knowledge across different domains, and the assessment of resilience in real-world scenarios. The paper highlights the need for investigating adversarial assaults to improve the resilience of models, enhance security measures, guide decision-making, stimulate innovation, and encourage responsible development of AI. In conclusion, we suggest potential areas for future study, such as the creation of improved defensive mechanisms, robust modeling tools, and the incorporation of multidisciplinary approaches.

***Key Words*:** adversarial Attacks, Machine Learning, Adversarial Examples, robustness, Fast Gradient Sign Method, DeepFool, Carlini & Wagner (C&W), Zoo-Adversarial Instance Optimization

INTRODUCTION AND BACKGROUND CONTEXT

In an era of the Internet, with a vast accumulation. The amounts of data and the development of computer power, Continuous innovation and development of machine learning methods and frameworks, artificial intelligence (AI) technologies It includes image recognition, machine translation and Autonomous vehicles have been widely used and widespread applied all over the world [1]. For mankind, artificial intelligence has advanced to historical moments. At the same The research on the ancient field of computer security is also influenced by machine learning algorithms [2]. Besides using machine learning ML to develop various malicious detections and attack identification systems, hackers can also use it in order to make more accurate attacks. Recent studies have revealed the vulnerability of a wide range of applications, ranging from computer vision to network security. To the threat of an adversarial attack [3],[4].

The concept of hostile samples, which is a very interesting weakness in neural networks, was originally suggested. The paper stimulated a great deal of interest from researchers in adversarial attacks and, as economic benefits become more apparent, the number of attacks will continue to rise [5]. In the context of image recognition, an adversarial attack consists of modifying an original image so that the changes

are almost undetectable by a human [6]. The modified image is called an adversarial image, which will be misclassified by the neural network, while the original one is correctly classified. One of the most famous real life attacks is to change the image of the road sign so that it's misinterpreted by an autonomous vehicle [7]. The modification of illegal content to make it undetectable by automatic moderation algorithms is another application.[8]. Gradient based methods, where the attacker changes the image in the direction of the loss function of the input image, thus increasing the misclassification rate, are the most common attacks [8],[9],[10].

ADVERSARIAL ATTACKS: UNSEEN VULNERABILITIES IN AI

The integrity and dependability of AI models are greatly at risk from a class of risks known as adversarial attacks. These exploits trick AI systems into making bad or potentially disastrous conclusions by using cunningly prepared inputs. Contrary to their original status as a theoretical curiosity, adversarial assaults have moved beyond the confines of theoretical study to become actual dangers in far-reaching effects [9]. Artificial intelligence models' dependability and integrity are seriously threatened by adversarial attacks. Attackers can fool the model into making bad or inaccurate judgments by providing it with meticulously designed inputs. These attacks now pose a threat in the real world and have moved beyond theoretical investigation.

EXAMPLES OF ADVERSARIAL ATTACKS

Consider an autonomous vehicle reliant on computer vision for navigation. Adversarial attackers could potentially manipulate subtle markings on road signs or lanes, causing the vehicle's AI system to misinterpret them, leading to dangerous situations [11]. Similarly, healthcare applications that utilize medical imaging classification tools could be susceptible to adversarial manipulation, leading to misinterpretations of critical scans [12].

ADVERSARIAL ATTACK MODELS

The algorithm or method used to generate adversarial samples is referred to as an adversarial attack model. Given a model F(x) and a natural input image x, an adversarial sample x' is created by introducing a carefully designed perturbation δ to the original input x. This perturbation aims to steer the model's prediction towards a desired misclassification, while remaining imperceptible to the human eye. Mathematically, this can be represented as:

$x' = x + \delta$ where $||\delta|| \leq$ epsilon

Here, $||\ ||$ represents a distance metric (e.g., L0, L2, L∞) that restricts the magnitude of the perturbation (ϵ) to ensure the adversarial sample remains visually indistinguishable from the original one. Different distance metrics capture the notion of "imperceptibility" in various ways [13].

LITERATURE REVIEW

INTRODUCTION TO ADVERSARIAL ATTACKS IN MACHINE LEARNING

In the field of machine learning, adversarial assaults have become a significant obstacle to maintaining the security and dependability of models used in a variety of applications. With the growing prevalence of machine learning algorithms, attackers have developed advanced methods to manipulate these models by exploiting their weaknesses [5].

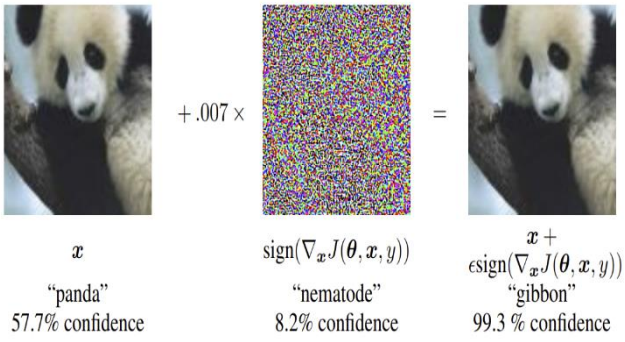

Figure 1: an adversarial example

Adversarial attacks refer to the creation of harmful inputs, called adversarial instances, with the purpose of tricking machine learning models into producing inaccurate predictions or classifications [9].

The impacts of adversarial assaults are significant, affecting several fields like image recognition, natural language processing, and autonomous systems [14]. These assaults have the potential to cause substantial disruptions in vital systems, putting their integrity at risk and eroding user confidence. It is crucial to comprehend the characteristics of adversarial assaults and create strong defensive mechanisms in order to protect machine learning systems from possible threats ([15].

This literature review examines the complex field of adversarial assaults in machine learning, covering important studies, current progress, comparative evaluations, theoretical frameworks, and areas where further study is needed. Through a thorough analysis of the current literature, our objective is to clarify the fundamental principles of adversarial assaults, evaluate the effectiveness of defense techniques, and pinpoint opportunities for future research and innovation ([16].

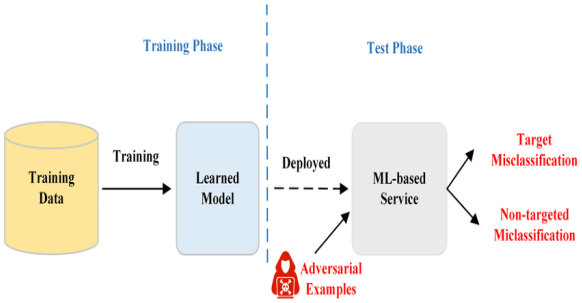

Figure 2: an adversarial example

This study aims to offer a thorough comprehension of adversarial assaults in machine learning and contribute to the advancement of robust and reliable machine learning systems that can survive adversarial threats.

OVERVIEW OF EXISTING ADVERSARIAL ATTACK METHODS

The concept of adversarial attacks in machine learning has received considerable interest after influential studies that established the basis for comprehending the susceptibility of models to adversarial manipulation. These groundbreaking papers presented essential concepts and approaches, significantly influencing the field of adversarial machine learning research.

[5] initially showcased the vulnerability of neural networks to adversarial perturbations, uncovering the presence of undetectable alterations to input data that result in misclassification. Their study pioneered the investigation of adversarial attacks and ignited further research in the topic.

[17] first presented the notion of adversarial examples and put out an approach for creating them. They showed that even minor, meticulously designed changes to the input data can lead neural networks to generate incorrect outputs with a strong level of certainty. This study yielded significant insights into the susceptibility of machine learning models and established the foundation for further investigations on adversarial assaults.

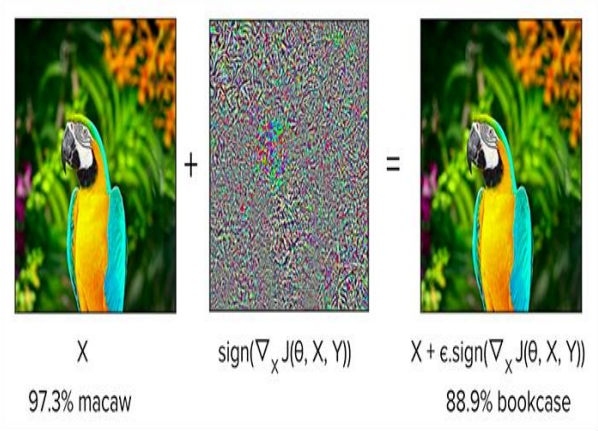

Figure 3: Fast Gradient Sign Method (FGSM) for Adversarial Image Generation

[14] introduced DeepFool, a technique for creating adversarial perturbations that take use of the linear properties of deep neural networks. Their methodology systematically calculates the smallest possible alteration needed to incorrectly categorize an input, offering a rapid and efficient method for generating adversarial instances.

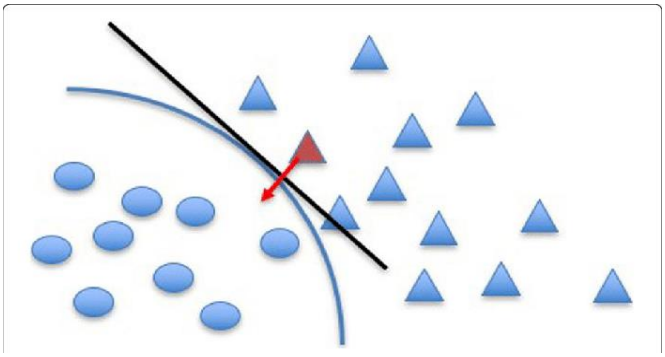

Figure 4: DeepFool-The-DeepFool-attack-approximates-the-decision-boundary-close-to-a-targeted

Carlini and Wagner proposed the Carlini & Wagner (C&W) approach [15], which presents the creation of hostile cases as an optimization issue. Their approach may effectively overcome different protection systems and achieve high success rates by making subtle changes, emphasizing the need to assess the resilience of neural networks against advanced attackers.

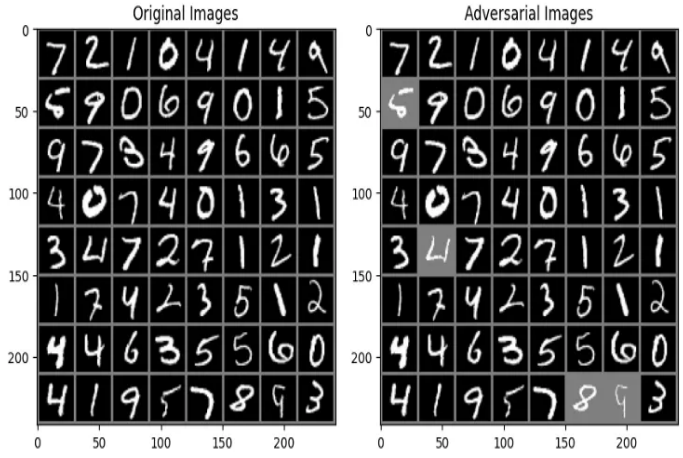

Figure 5: Comparing the same digits subtly perturbed by the Carlini & Wagner attack

ZOO (Zeroth Order Optimization) attack proposed in [18]. Similar to the techniques of Carlini & Wagner (C&W) [15], ZOO formulates the task of creating adversarial cases as an optimization issue. But ZOO does this without requiring you to know the inner workings of the model. Because of this opaqueness, ZOO assaults are especially risky. They can get beyond several defenses put in place to fend off "white-box" attacks—attacks carried out with complete awareness of the model. Furthermore, ZOO maintains low input changes while achieving high success rates, simulating real-world situations and making identification even more challenging.

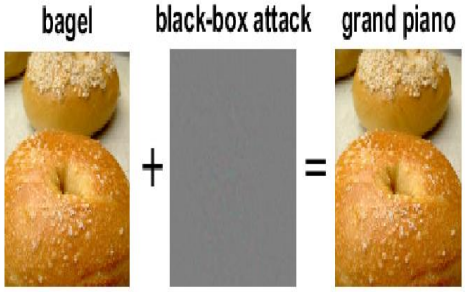

Figure 6: ZOO blackbox targeted attack example

These influential publications have made significant contributions to the area of adversarial machine learning by offering fundamental ideas and approaches for comprehending and reducing the susceptibilities of machine learning models to adversarial manipulation.

RESEARCH GAPS & CHALLENGES:

1 RESEARCH GAPS

- Limited Comparative Analysis: The existing literature lacks a comprehensive comparative analysis of prevalent adversarial attack methods, such as FGSM, DF, C&W, and ZOO. There is a need to systematically compare the effectiveness, efficiency, and practical considerations of these methods to understand their relative strengths and weaknesses.
- Decision-Making Framework: The absence of a clear framework for selecting appropriate adversarial attack strategies poses a challenge for practitioners seeking to enhance model robustness. Without guidance on how to evaluate and compare different attack methods, practitioners may struggle to make informed decisions.

2 RESEARCH CHALLENGES

The extensive study of adversarial methods in machine learning has revealed a spectrum of attack techniques, their applications, and the dynamic evolution of adversarial threats. Despite this progress, distinct research challenges have emerged, paving the way for future investigations:

- Adversarial Diversity: Developing defenses against diverse attacks, exploring variations in strengths, data, and model structures.
- Dynamic Adversarial Landscapes: Constructing dynamic defenses that anticipate and respond to novel adversarial methods.
- Transferability Across Domains: Investigating how attacks transfer across domains, impacting different applications and datasets.
- Real-World Robustness Assessment: Developing practical methodologies for effective real-world Défense scenarios.

THE SIGNIFICANCE OF THE STUDY

This research has the following significance

1. Enhancing Model Robustness: Our research compares adversarial attack methods to strengthen machine learning models.
2. Advancing Security Measures: Insights aid in developing robust defense mechanisms for critical applications.
3. Informing Decision-Making: Clear insights guide stakeholders in model development and risk mitigation.
4. Fostering Innovation: Collaboration drives the development of novel techniques in machine learning security.
5. Ethical Implications: We advocate for responsible research practices to promote societal well-being.

CONCLUSION

Adversarial attacks provide a substantial obstacle to the security and dependability of machine learning models. This study conducted a thorough examination of the characteristics of these assaults, investigated several methodologies used in these attacks, and deliberated on the existing research deficiencies and obstacles. We emphasized the significance of investigating adversarial assaults to improve the resilience of models, progress security measures, guide decision-making, stimulate innovation, and encourage responsible development of AI. In conclusion, we have suggested potential areas for future study, such as the creation of more advanced protection mechanisms, robust modeling tools, and the incorporation of multidisciplinary approaches. Continued study on adversarial assaults is essential in the field of machine learning to develop AI systems that are secure and reliable in real-world situations.

REFERENCES

[1] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in Proceedings of the 35th Annual Computer Security Applications Conference, 2019, pp. 113–125.

[2] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018, pp. 2154–2156.

[3] C. Esposito, M. Ficco, and B. B. Gupta, "Blockchain-based authentication and authorization for smart city applications," Inf. Process. Manag., vol. 58, no. 2, p. 102468, 2021.

[4] D. Li, L. Deng, B. B. Gupta, H. Wang, and C. Choi, "A novel CNN based security guaranteed image watermarking generation scenario for smart city applications," Inf. Sci. (Ny)., vol. 479, pp. 432–447, 2019.

[5] C. Szegedy et al., "Intriguing properties of neural networks," arXiv Prepr. arXiv1312.6199, 2013.

[6] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," IEEE Trans. neural networks Learn. Syst., vol. 30, no. 9, pp. 2805–2824, 2019, doi: 10.1109/TNNLS.2018.2886017.

[7] K. Eykholt et al., "Robust physical-world attacks on deep learning visual classification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1625–1634.

[8] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," IEEE Trans. neural networks Learn. Syst., vol. 30, no. 9, pp. 2805–2824, 2019.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv Prepr. arXiv1412.6572, 2014.

[10] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in Artificial intelligence safety and security, Chapman and Hall/CRC, 2018, pp. 99–112.

[11] A. Kurakin et al., "Adversarial attacks and defences competition," in The NIPS'17 Competition: Building Intelligent Systems, 2018, pp. 195–231.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., pp. 1–28, 2018.

[13] Z. Wang, M. Song, S. Zheng, Z. Zhang, Y. Song, and Q. Wang, "Invisible adversarial attack against deep neural networks: An adaptive penalization approach," IEEE Trans. Dependable Secur. Comput., vol. 18, no. 3, pp. 1474–1488, 2019.

[14] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 2574–2582, 2016, doi: 10.1109/CVPR.2016.282.

[15] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," Proc. - IEEE Symp. Secur. Priv., pp. 39–57, 2017, doi: 10.1109/SP.2017.49.

[16] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., pp. 1–22, 2018.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–11, 2015.

[18] P. Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C. J. Hsieh, "ZOO: Zeroth order optimization based black-box atacks to deep neural networks without training substitute models," AISec 2017 - Proc. 10th ACM Work. Artif. Intell. Secur. co-located with CCS 2017, pp. 15–26, 2017, doi: 10.1145/3128572.3140448.

[19] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing : Detecting Adversarial Examples in Deep Neural Networks," no. February, 2018.

[20] et al Shokri, Reza, "Membership Inference Attacks Against Machine Learning Models".