AG Publishing House

Handbook On Fundamentals And Methods Of Machine And Deep Learning(VOLUME-1)

Volume 1
Year: 2024

AG PH Books

Adversarial Attacks and Defenses in Deep Learning Systems

Dr. Endluri Venkata Naga Jyothi[1*]

[1]*Associate professor, CMR College of Engineering & Technology, Hyderabad.*

Abstract

The rapidly expanding Internet applications of deep learning and related situations have led to a major increase in interest in adversarial attacks and defences in machine learning and deep neural networks. In this discipline, more and more researchers are employed. Give a thorough analysis of the ideas and approaches that make it possible for scholars to study adversarial attacks and defences. Furthermore, since there are no established assessment techniques, it is difficult to assess the true danger posed by adversarial assaults or the resilience of a deep learning model. Moreover, make an effort to provide the first analytical framework for a methodical comprehension of adversarial assaults. The framework is designed with cybersecurity in mind, including a lifecycle for hostile assaults and countermeasures. It was noted that no defence method now in use defeats hostile samples with both efficiency and efficacy.

Keywords: Adversarial attacks and defenses, Machine learning, Deep neural network, Artificial intelligence, Algorithms.

1 Introduction

The term "deep learning" describes a group of machine learning techniques that are based on deep neural networks (DNNs) and are often used for tasks like classification and prediction. Multiple layers, a huge number of computational neurones, and nonlinear activation functions are features of DNNs, a kind of mathematical model. A typical deep learning system's workflow consists of two stages: the training phase and the inference phase [1].

Machine learning methods have been successfully used in many different contexts. Deep learning in

* ISBN No. - 978-81-974433-9-8

particular is quickly emerging as a vital tool for a variety of jobs. But in many cases, a deep learning or machine learning model's failure might result in major safety issues. For instance, with driverless cars, mistaking a traffic sign for another might result in a serious collision. Therefore, before a model is widely used, it is essential to train it to be accurate and stable. Regretfully, a depressing occurrence in model security has been shown by several research in recent years: deep learning models may be susceptible to hostile instances, or samples that have been purposefully disturbed by an opponent. Even if these manipulated models may demonstrate great accuracy with innocuous samples, there is a good chance that they may provide incorrect predictions. The term "adversarial attack" may then be used to refer to a large class of assaults that try to trick a machine learning model by introducing adversarial samples into the inference phase (also termed an evasion attack) or the training phase (also known as a poisoning attack). Both types of attacks will raise concerns about model security and dramatically reduce the resilience of deep learning models. Furthermore, the real-world discovery of these model security issue vulnerabilities in deep learning solutions has raised doubts about the degree of trustworthiness of deep learning technology [2].

2 Literature Review

(Sarala & Gangappa, 2024) [3] The Internet is developing at a fast pace, which means that artificial intelligence areas are finding more and more applications. The age of AI has here. Adversarial assaults are also common in the area of AI at the same time. Consequently, there is a pressing need for research on adversarial assault security. We begin by outlining the importance of an adversarial assault. Next, we present the ideas, forms, and risks associated with adversarial attacks. Lastly, we go over the standard assault and defence strategies for each application area. This article addresses the adversarial attack classifications and techniques of three data types—image, text, and malicious code—in response to the increasingly complicated neural network model. This will enable researchers to promptly identify the appropriate research topic. We also discussed several outstanding concerns and had a comparison with other evaluations that were comparable at the conclusion of this review.

(Begum et al., 2024) [4] The complicated relationships that result in adversarial flaws in deep learning systems are examined in this research. In order to assess how effectively different adversarial attack techniques, such as FGSM and PGD, may compromise model integrity, this evaluates them. These findings underscore the continuous game of cat and mouse between attackers and defenders using deep learning techniques in security. Even if model resilience has increased significantly, the absence of a globally defined approach emphasises the need for a diverse security policy. This research demonstrates the necessity for ongoing innovation as well as the ongoing challenge of safeguarding deep learning models against malicious attacks.

(Vizcarra et al., 2024) [5] Unexpectedly, deep neural networks have been applied to optical character recognition (OCR) to improve the performance of OCR systems to the point where they are now an essential preprocessing step in text analysis pipelines for critical applications where OCR accuracy is critical, like license plate recognition (LPR) systems. Investigating substitute defence mechanisms, such

image denoising and in painting, offers a strong strategy for enhancing the resistance of LPR systems against hostile assaults. The operational needs of real-world LPR systems are met by giving realistic implementation and integration of image denoising and inpainting methods top priority. These techniques provide practical and approachable ways to improve security without adding a large amount of computational overhead, and they can be easily included into processes that are already in place. Through the use of a multipronged strategy that leverages the advantages of conventional image processing methods, the study aims to create all-encompassing and adaptable defence plans that are customised to the unique weaknesses and needs of LPR systems. By strengthening LPR systems against hostile attacks, this all-encompassing strategy hopes to promote greater confidence and dependability in the implementation of OCR and LPR technologies across a range of fields and applications.

(Wang et al., 2023) [6] This review offers a thorough summary of the most recent developments in adversarial attack and defence strategies, emphasising the use of deep neural networks in classification models. In particular, we do a thorough categorisation of contemporary adversarial attack strategies and cutting-edge adversarial defence tactics according to assault principles, and we display the results in eye-catching tables and tree diagrams. This is predicated on a thorough assessment of the previous works, which includes a breakdown of their advantages and disadvantages. Additionally, we group the techniques into two categories: robustness augmentation and counter-attack detection, with an emphasis on regularization-based techniques for the latter. A hierarchical classification of the most recent defence strategies is provided, and new attack avenues such as search-based, decision-based, drop-based, and physical-world attacks are also investigated. The challenges of ensuring method transferability, overcoming the impact of gradient masking, maintaining clean accuracy, and balancing training costs with performance are highlighted. Finally, a summary of the problems and lessons learnt is provided, along with recommendations for further study.

(Ma et al., 2023) [7] A new method for the low-cost authentication of wireless Internet of things (IoT) devices is radio frequency fingerprint identification, or RFFI. Utilising distinct hardware limitations as device identifiers, RFFI makes extensive use of deep learning as a feature extractor and classifier. Deep learning is susceptible to adversarial assaults, however, in which clean data is perturbed in order to provide hostile instances that lead the classifier to predict incorrectly. It has been shown that deep learning-based RFFI is susceptible to these kinds of assaults; yet, no investigation has yet been conducted into viable adversarial strategies against a variety of RFFI classifiers. In this work, we provide our research on two methods—the fast gradient sign technique (FGSM) and projected gradient descent (PGD)—for investigating white-box assaults, both targeted and non-targeted. Real datasets were gathered and a LoRa testbed constructed. It has been empirically shown that these adversarial instances operate well against gated recurrent units (GRU), long short-term memory (LSTM) networks, and convolutional neural networks (CNNs).

(Zhou et al., 2022) [8] Deep neural networks' exceptional performance has fuelled the development of deep learning applications across a wide range of industries. However, the large-scale use of deep learning has been impeded by the possible threats posed by adversarial samples. In many cases, the

model's ultimate performance is greatly reduced by hostile perturbations that are undetectable to the human eye. In the field of deep learning, several publications on adversarial assaults and their defences have been published. As contrast to poisoning attacks, which involve inserting poisoned data into the training set, most concentrate on evasion attacks, in which the adversarial cases are discovered during testing. Furthermore, since there are no established assessment techniques, it is difficult to assess the true danger posed by adversarial assaults or the resilience of a deep learning model. Therefore, we examine the existing literature in this paper. We also try to provide the first analytical framework for a methodical comprehension of adversarial assaults. The framework is designed with cybersecurity in mind, including a lifecycle for hostile assaults and countermeasures.

(Domingo & Borondo, 2021) [9] In order to create the ground and excited wave functions of various Hamiltonians appropriate for the study of molecular system vibrations, we construct and implement two Deep Learning models in this research. In order to train the created neural networks, analytically solved Hamiltonians are used, and the network is asked to generalise these answers to increasingly complicated Hamiltonian functions. The excited vibrational wave functions of various chemical potentials may be replicated using this method. Since every methodology used here is data-driven, it makes no assumptions on the details of the system's underlying physical model. Because of this, this method is flexible and may be used to the study of many quantum chemical systems.

(Ren et al., 2020) [10] The rapid advancements in deep learning (DL) and artificial intelligence (AI) have made it imperative to guarantee the stability and security of the implemented algorithms. It has been well known recently that DL algorithms are vulnerable to adversarial samples in terms of security. Although the synthetic samples seem harmless to humans, they may cause the DL models to behave in a variety of ways. Adversarial assaults have been successfully used in actual physical-world situations, proving its applicability. Because of this, adversarial attack and defence strategies have gained popularity as a study area in recent years and are receiving more attention from the machine learning and security sectors. The theoretical underpinnings, algorithms, and practical uses of adversarial attack strategies are originally presented in this study. Next, we outline some research endeavours pertaining to defence strategies, including the wide spectrum of this topic. We next go over a number of unresolved issues and obstacles in the hopes of sparking further investigation into this important field.

(Zhang et al., 2020) [11] It is crucial to defend against these hostile assaults, yet doing so presents difficult obstacles. We herein provide TIKI-TAKA, a generic framework that can be used to: (i) evaluate the resilience of cutting-edge deep learning-based network intrusion detection systems (NIDS) against adversarial manipulations; and (ii) integrate defence mechanisms that we suggest strengthen resistance against assaults that use these evasion strategies. In particular, we use five state-of-the-art adversarial attack types to undermine three widely used NN-based malicious traffic detectors. We conduct experiments using datasets that are made accessible to the public. We take into account both one-to-all and one-to-one classification situations, which include differentiating between criminal and benign traffic and detecting particular forms of anomalous traffic among many observed instances. According to the data, attackers may bypass network intrusion detection systems (NIDS) with success rates as high

as 35.7% by simply changing the traffic's time-based characteristics. We suggest three defence mechanisms: query detection, ensembling adversarial training, and model voting ensembling to mitigate these shortcomings. We show that these techniques may block assaults with potentially disastrous outcomes (such botnets) and restore intrusion detection rates to almost 100% against the majority of malicious traffic. This demonstrates the efficacy of our methods and supports their use in the development of durable and trustworthy deep anomaly detectors.

3 Conclusion

An overview of current representative adversarial attack and defence strategies is provided in this study. looked at the concepts and procedures of the suggested algorithms and approaches. The usefulness of these adversarial defences in light of the most recent developments was also covered. furthermore, offered a five-stage analytic approach akin to this one for antagonistic defences. The goals of defensive tactics at various phases line up with the adversarial assault lifecycle. To reduce the risks to the target models, several forms of defences might be combined at different phases according to the suggested framework. To put it briefly, the use of contemporary RFFI for physical layer security is seriously threatened by the achievements of both targeted and non-targeted assaults. In a white-box attack, the attacker is fully aware of the victim and can determine the gradient with accuracy, which poses a serious security risk. Adversarial training, the most successful defence mechanism, is too computationally costly for widespread use, and other good heuristic defences have been shown to be susceptible to adaptive white-box adversaries. Additionally, a number of unresolved issues and difficulties in this important field were covered, offering helpful research guidelines to further this field's progress.

References

[1] M. Ozdag, “Adversarial Attacks and Defenses Against Deep Neural Networks : A Survey,” Procedia Comput. Sci., vol. 140, pp. 152–161, 2018, doi: 10.1016/j.procs.2018.10.315.

[2] Y. Li and Y. Wang, “Defense Against Adversarial Attacks in Deep Learning,” 2019, doi: 10.3390/app9010076.

[3] D. V Sarala and D. T. Gangappa, “Adversarial Attacks and Defense Strategy in Deep Learning,” vol. 24, no. 1, pp. 127–132, 2024.

[4] K. S. Begum et al., “Adversarial Attacks and Defenses in Deep Learning Models,” vol. 12, pp. 857–865, 2024.

[5] C. Vizcarra et al., “Deep learning adversarial attacks and defenses on license plate recognition system,” Cluster Comput., vol. 27, no. 8, pp. 11627–11644, 2024, doi: 10.1007/s10586-024-04513-4.

[6] Y. Wang et al., “Adversarial Attacks and Defenses in Machine Learning-Powered Networks : A Contemporary Survey,” pp. 1–46, 2023.

[7] J. Ma, J. Zhang, G. Shen, A. Marshall, and C. Chang, “White-Box Adversarial Attacks on Deep Learning-Based Radio Frequency Fingerprint Identification,” 2023.

[8] S. Zhou, C. H. I. Liu, D. Ye, and T. Zhu, “Adversarial Attacks and Defenses in Deep Learning : From a Perspective of Cybersecurity,” vol. 55, no. 8, 2022, doi: 10.1145/3547330.

[9] L. Domingo and F. Borondo, “Deep learning methods for the computation of vibrational wavefunctions,” 2021.

[10] K. Ren, T. Zheng, Z. Qin, and X. Liu, “Adversarial Attacks and Defenses in Deep Learning,” Engineering, vol. 6, no. 3, pp. 346–360, 2020, doi: 10.1016/j.eng.2019.12.012.

[11] C. Zhang, X. Costa-p, S. Member, P. Patras, and S. Member, “Adversarial Attacks Against Deep Learning-based Network Intrusion Detection Systems and Defense Mechanisms,” 2020.