AG Publishing House

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

Volume 1
Year: 2024

AG PH Books

Overview on Adversarial Attacks and Defenses in Deep Learning Systems

Prof. Nilima D.Bobade[1], Dr. Swati S. Sherekar[2*]

[1]*Assistant Professor AT Prof. Ram Meghe Institute of Technology & Research Badnera*.

[2]*Professor AT S.G.B.A.U/ Computer Department, Amravti, India*.

Abstract

In "machine learning and deep neural networks", there is a significant surge in interest in adversarial assaults and answers due to the quickly growing Internet usage of deep learning and related scenarios. The use of deep neural networks in models for classification is highlighted in this article, which provides an extensive overview of the most recent advancements in "adversarial attack and defense tactics". People who manipulate the deep learning model are prone to making mistakes that lead to incorrect predictions and inappropriate actions. Learning the specified approach is crucial to keeping the machine in good working order and preventing it from being disrupted by hostile attack users. Through meticulous testing and data analysis, the study has produced verifiable evidence to support the ongoing discussion in the area of deep learning security.

Keywords: Adversarial attacks and defenses, Deep learning, Machine learning.

1 Introduction

Thanks to a trillion-fold growth in computer capability, deep learning (DL) is currently extensively employed for various machine learning (ML) tasks such as natural language processing, game theory, and picture classification. Nevertheless, researchers have identified a serious security risk to the current DL algorithms: By interfering with innocuous samples without being detected by humans, adversaries may easily trick DL models. Tiny perturbations that are invisible to the human sight or ear might cause

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

the model to provide an extremely confident false prediction. It is believed that the adversarial sample phenomenon will provide a significant obstacle to the mainstream adoption of DL algorithms in industries. Considerable research has been done on this unsolved issue. The threat model divides current adversarial assaults into three categories: “white-box, gray-box, and black-box threats”. The adversary' level of knowledge determines how the three models vary from one another. Attackers are assumed to be completely knowledgeable of the target model's architecture and features in white-box assault threat models. Hence, they can generate adversarial samples on an intended model in any configuration. In the gray-box hazard model, the adversaries only know the structure of the target model. In the black-box hazard concept, query access is the only way for adversaries to create hostile samples [1].

Simultaneously, many certificated and heuristic defenses have been recently proposed for adversarial sample classification and detection. A defensive mechanism known as a heuristic defense is one that effectively repels particular attackers while not guaranteeing theoretical correctness. The most efficient heuristic defense available at the moment is adversarial training, which tries to strengthen the DL model's resistance by introducing hostile sample into the training stage. Other heuristic defenses generally rely on eliminating and input/feature adjustments to reduce the adversarial effects in the data/feature domains. However, under a certain class of adversarial assaults, certified defenders may always provide certificates for their lowest accuracy. Using convex relaxations to establish the upper limit of an adversarial polytope is a lately popular method for network certification. For trained deep learning models, the relaxed upper bound is a certification that ensures no attack, subject to certain restrictions, may exceed the certificated attack success rate, which is roughly represented by the upper bound. Nevertheless, these certified defenses' real performance still falls well short of the adversarial training's [2]

2 Literature Review

(Khaleel et al., 2024) [3] The literature review of adversarial assaults and responses presented in this work is based on highly referenced conference proceedings and publications that have been published in the Scopus database. This research classifies and evaluates the literature from many disciplines, such as Graph Neural Networks, Deep Learning Models for IoT Systems, and others, via the categorization and evaluation of 128 systematic articles: 80 original papers and 48 review papers until May 15, 2024. The study makes recommendations for further research and development in the field for adversarial robustness and protection mechanisms, based on results on identified metrics, citation analysis, and contributions from these works. Presenting the fundamentals of adversarial assaults and countermeasures as well as the significance of preserving machine learning platforms' flexibility is the work's stated goal. The goal here is to support the development of effective and long-lasting defense systems for AI applications across a range of sectors.

(Begum et al., 2024) [2] 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 analyzes 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 emphasizes 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.

(Wang et al., 2023) [4] In particular, using attack principles, we do a thorough categorization of modern adversarial attack strategies and cutting-edge adversarial defensive tactics, presenting the results in eye-catching tables and tree diagrams. This has been developed based on a detailed analysis of the prior studies, which includes a summary of their benefits and drawbacks. We also classify the techniques into robustness improvement and counter-attack detection categories, emphasizing regularization-based techniques for the latter. In addition to a hierarchical categorization of the latest security techniques, novel attack paths including drop-based, decision-based, search-based, and physical-world assaults are also explored. These problems include mitigating the impact of gradient masking, maintaining clean accuracy, ensuring technique transferability, and striking a balance between training costs and performance. Finally, a summary of the issues raised and the lessons learned are provided, along with suggestions for further study.

(Muoka et al., 2023) [5] A thorough conceptual analysis is also included, along with a number of hostile assaults and countermeasures intended for the understanding of medical imagery. This survey, which combines qualitative and quantitative data, concludes with a thorough discussion and new research directions by addressing the problems with adversarial attack and defensive mechanisms unique to medical image processing systems. "Adaptability, transferability, labeling, dataset, robustness against target attacks, computational resources, real-time detection, interpretability, explainability, response, and adversarial attacks" in multi-modal fusion are the main problems we identified with adversarial attack and defense in medical imaging. By addressing these knowledge gaps and working toward these goals, the area of medical visualization adversarial attack and defensive mechanisms may be able to develop deep learning systems that are safer, more dependable, and more beneficial therapeutically.

(Zhou et al., 2022) [6] 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 defenses 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.

(Chen et al., 2021) [7] These days, voice processing systems, or VPSes, are extensively used and have ingrained themselves into people's everyday lives. They assist with tasks like driving a vehicle, unlocking smartphones, making online purchases, and more. Unfortunately, recent research has shown that systems

built on deep neural networks are vulnerable to adversarial instances, a finding that highlights the need of VPS security. This page provides a thorough review of the background material on adversarial attacks, including how adversarial instances, psychoacoustic models, and evaluation indicators are created. Next, we provide a succinct overview of techniques for defending against hostile assaults. In order to help newcomers to this discipline better comprehend the structure and categorization, we conclude by proposing a systematic classification of adversarial assaults and response strategies.

(Ren et al., 2020) [8] 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 attacks have been effectively used in real-world situations, demonstrating its applicability. Because of this, the domains of "machine learning and security" have been paying greater attention to adversarial attack and defense tactics as a research topic in recent years. This paper initially addressed the theoretical foundations, algorithms, and real-world applications of adversarial attack techniques. Next, we outline some research endeavors pertaining to defensive 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.

(Yuan et al., 2019) [1] Deep learning has made considerable strides and been successfully used in a broad range of safety-critical applications. However, it has recently been shown that adversarial examples—well-designed input samples—can weaken deep neural networks. When deep neural networks are being tested or deployed, adversarial instances may readily trick them while remaining undetectable to humans. When using deep neural networks in safety-critical situations, one of the main dangers is their susceptibility to adversarial instances. As such, there is much interest in the attacks and defenses on adversarial instances. In this study, we discuss the approaches for producing adversarial instances, provide a taxonomy of these approaches, and analyze recent results on adversarial examples versus deep neural networks. Applications for adversarial instances are examined under the taxonomy. We discuss countermeasures for hostile instances in more detail, as well as the difficulties and possible solutions.

3 Conclusion

We have provided a broad summary of current representative adversarial attack and defensive strategies in this work. We have looked at the concepts and procedures of the suggested algorithms and approaches. On the basis of the most recent developments, we have also spoken about how successful these hostile defenses are. The most effective defense mechanism, adversarial training, is too computationally expensive to be used widely, and other effective heuristic defenses have been demonstrated to be vulnerable to "adaptive white-box adversaries". We have also addressed some unresolved difficulties and concerns in this area to give a useful research guideline to support the progress of this important topic. Aggressive assaults in a methodical manner and, last but not least, examined current defensive strategies from various defense angles. This demonstrates that the models' vulnerability to adversarial attacks renders them unsuitable for use in security-critical applications.

References

[1] 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.

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

[3] Y. L. Khaleel, M. A. Habeeb, and H. Alnabulsi, “Adversarial Attacks in Machine Learning: Key Insights and Defense Approaches,” Appl. Data Sci. Anal., pp. 121–147, 2024, doi: 10.58496/adsa/2024/011.

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

[5] G. W. Muoka et al., “A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense,” Mathematics, vol. 11, no. 20, 2023, doi: 10.3390/math11204272.

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

[7] X. Chen, S. Li, and H. Huang, “Adversarial Attack and Defense on Deep Neural Network-Based Voice Processing Systems: An Overview,” Appl. Sci., vol. 11, no. 18, 2021, doi: 10.3390/app11188450.

[8] 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.