International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at :www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2361054

A Comprehensive Review on Adversarial Attack Detection Analysis in Deep Learning

Soni Kumari*[1], Sheshang Degadwala[2]

*[1]Research Scholar, Dept. of Computer Engineering, Sigma Institute of Engineering, Gujarat, India

soni.kumari177@gmail.com[1]

[2]Associate Professor & Head of Department, Dept. of Computer Engineering, Sigma University, Gujarat, India

sheshang13@gmail.com [2]

ARTICLEINFO

Article History:

Accepted: 10 Oct 2023

Published: 20 Nov 2023

Publication Issue

Volume 9, Issue 10

September-October-2023

Page Number

319-325

ABSTRACT

This comprehensive review investigates the escalating concern of adversarial attacks on deep learning models, offering an extensive analysis of state-of-the-art detection techniques. Encompassing traditional machine learning methods and contemporary deep learning approaches, the review categorizes and evaluates various detection mechanisms while addressing challenges such as the need for benchmark datasets and interpretability. Emphasizing the crucial role of explaining ability and trustworthiness, the paper also explores emerging trends, including the integration of technologies like explainable artificial intelligence (XAI) and reinforcement learning. By synthesizing existing knowledge and outlining future research directions, this review serves as a valuable resource for researchers, practitioners, and stakeholders seeking a nuanced understanding of adversarial attack detection in deep learning.

Keywords: Adversarial Attacks, Deep Learning, Detection Techniques, Machine Learning, Interpretability, Explainable Artificial Intelligence (XAI), Reinforcement Learning.

I. INTRODUCTION

The pervasive integration of deep learning models across diverse sectors has undeniably propelled artificial intelligence into new frontiers of innovation. However, the heightened complexity of these models has also exposed a vulnerability—adversarial attacks. These attacks, characterized by subtle manipulations of input data, pose a significant threat to the reliability and security of deep learning systems. As the stakes continue to rise with the increasing reliance on these models for critical tasks, understanding and mitigating the risks associated with adversarial attacks have become imperative [1,4].

In this context, this review paper embarks on a comprehensive examination of the current state of adversarial attack detection in deep learning [6,8]. The exploration begins by elucidating the

fundamental concepts of adversarial attacks, shedding light on the motivations and techniques employed by adversaries to exploit vulnerabilities in neural networks. Subsequently, we delve into an in-depth analysis of existing detection mechanisms, ranging from traditional machine learning approaches to cutting-edge deep learning-based solutions. By synthesizing this knowledge, we aim to provide a holistic view of the challenges and advancements in the field, laying the foundation for future research directions.

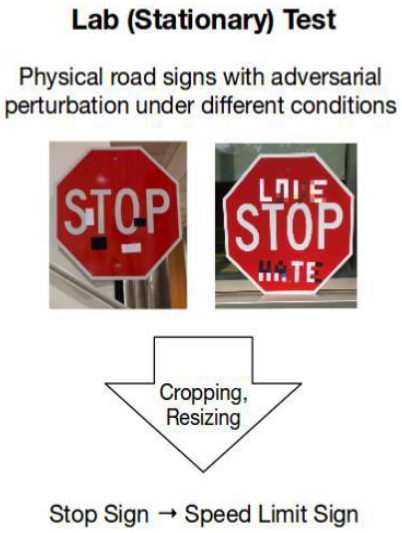

Figure 1. Example of Adversarial Attack [10]

As artificial intelligence technologies become integral to our daily lives, the insights derived from this review are pivotal for researchers, practitioners, and stakeholders alike, offering a roadmap to fortify deep learning models against the evolving landscape of adversarial threats [11,15]. Through a nuanced understanding of the challenges posed by adversarial attacks, we strive to contribute to the ongoing efforts to enhance the robustness and trustworthiness of deep learning systems in the face of adversarial manipulation.

II. LITERATURE STUDY

Certainly! Here are more detailed summaries for each paper without specific page references:

In [1], G. Ryu and D. Choi proposed a novel methodology for detecting adversarial attacks based on image entropy. Their study involves a meticulous analysis of various methods, exploring their applications and implications in the context of adversarial attacks. By delving into the intricacies of different techniques, the authors provide a comprehensive understanding of the proposed image entropy approach and its potential applications in bolstering the security of deep learning models against adversarial threats.

In [2], X. Cui delved into the intricate world of targeting image-classification models. The paper provides an in-depth exploration of this specific adversarial attack strategy. Cui meticulously analyzes the nuances of targeting image-classification models, shedding light on the methods employed, potential vulnerabilities exploited, and broader implications. This extensive examination contributes valuable insights into the evolving landscape of adversarial attacks in the context of image classification.

In [3], M. Kim and J. Yun introduced AEGuard, an innovative image feature-based independent adversarial example detection model. This model contributes to the field of adversarial attack detection by relying on distinctive image features. The paper not only outlines the technical details of AEGuard but also provides a broader context for understanding the significance of image feature-based approaches in enhancing the robustness of deep learning models against adversarial examples.

In [4] P. Lorenz, M. Keuper, and J. Keuper presented a notable contribution by unfolding local growth rate estimates for near-perfect adversarial detection. Their work advances the understanding of detection mechanisms by addressing challenges in achieving nearly perfect detection rates. The paper discusses the methodology's strengths and limitations, providing valuable insights for researchers and practitioners working on improving the efficacy of adversarial attack detection.

In [5], L. Shi, T. Liao, and J. He proposed a noise-fusion method to defend deep neural network image classification models against adversarial attacks. The paper not only introduces the novel defense strategy but also meticulously discusses its implementation, effectiveness, and potential impact on the broader landscape of adversarial attack defense mechanisms. This contribution adds to the arsenal of defense strategies aimed at safeguarding image classification models.

In [6] A. S. Almuflih et al. made a significant stride in adversarial attack detection by introducing a novel exploit feature-map-based method. The paper provides an in-depth exploration of this detection mechanism, detailing its theoretical foundations and practical applications. This work contributes to the ongoing efforts in developing sophisticated and effective detection strategies against adversarial attacks.

In [7], M. Khan and collaborators conducted an in-depth analysis of Alpha Fusion Adversarial Attack using deep learning. The paper not only explores the methodologies employed in this particular adversarial attack but also delves into the implications and potential countermeasures. This comprehensive examination contributes valuable insights to the understanding of adversarial attacks using alpha fusion and adds to the broader knowledge base in adversarial attack analysis.

In [8] N. Ghaffari Laleh et al. delved into the intricate realm of adversarial attacks and robustness in computational pathology. The paper provides a comprehensive study, exploring the challenges and potential solutions in the context of computational pathology. By analyzing the vulnerabilities and proposing robustness measures, this contribution is valuable for researchers working on securing machine learning models in medical image analysis.

In [9], Y. Wang and collaborators conducted a contemporary survey on adversarial attacks and defenses in machine learning-powered networks. The survey spans various aspects, offering a thorough review of the current state of adversarial attacks and defenses. The paper not only summarizes existing knowledge but also identifies gaps and future research directions, serving as a comprehensive resource for researchers, practitioners, and stakeholders in the field of adversarial attacks in machine learning.

In [10] H. Hirano, A. Minagi, and K. Takemoto investigated universal adversarial attacks on deep neural networks for medical image classification. This work explores the vulnerabilities of medical image classification models to universal adversarial attacks, providing insights into potential risks and countermeasures. The study is particularly relevant in the medical imaging domain, where the security and reliability of deep learning models are of paramount importance.

In [11], A. Talk, F. Wikipedia, A. Wikipedia, and C. Wikipedia discussed the University of Science and Technology of China, providing information on an institution potentially relevant to adversarial attack research. While this paper primarily serves as a reference for the mentioned institution, it highlights the interdisciplinary nature of research in the field of adversarial attacks.

In [12] Y. Zheng and S. Velipasalar proposed part-based feature squeezing for detecting adversarial examples in person re-identification networks. The paper explores the application of part-based feature squeezing in the context of person re-identification networks, contributing to the understanding of adversarial attacks in this specific domain. The study provides insights into the effectiveness of feature squeezing as a defense strategy in person re-identification applications.

In [13], B. Liang et al. presented an adaptive noise reduction approach for detecting adversarial image examples in deep neural networks. The paper meticulously details the methodology of adaptive noise reduction, offering insights into its implementation, strengths, and potential limitations. This work contributes to the ongoing research on detecting adversarial examples through noise

reduction, enhancing our understanding of effective defense mechanisms.

In [14] M. A. Ahmadi, R. Dianat, and H. Amirkhani introduced an adversarial attack detection method based on a re-attacking approach. The paper provides a thorough examination of the re-attacking methodology, elucidating its theoretical foundations and practical applications. This contribution adds to the repertoire of adversarial attack detection methods, shedding light on the complexities of re-attack strategies and potential countermeasures.

In [15] K. Ren, T. Zheng, Z. Qin, and X. Liu conducted a comprehensive study on adversarial attacks and defenses in deep learning. The paper delves into various aspects of adversarial attacks, providing a comprehensive overview of existing defenses and their effectiveness. The study not only synthesizes current knowledge but also identifies challenges and future research directions, contributing to the evolving landscape of adversarial attacks in deep learning.

III.METHODOLOGY

A. Types of Adversarial Attacks

Black Box Attacks [1,3,5]: In this category, adversaries do not have access to the internal parameters or architecture of the target model. Black box attacks rely on the external observation of model outputs and seek to craft adversarial examples without detailed knowledge of the model's structure, training data, or weights. This methodology involves exploring vulnerabilities in the model's decision boundary through input-output interactions.

White Box Attacks [1,3,5]: Conversely, white box attacks assume complete knowledge of the target model, including its architecture, parameters, and training data. Adversaries exploit this information to craft highly tailored adversarial examples, leveraging a deep understanding of the model's internal workings. White box attacks often pose a more significant threat due to the comprehensive knowledge available to attackers.

B. Pixel-Based Adversarial Attacks Detection Methods

Pixel-based adversarial attacks involve manipulating individual pixel values in input images to deceive the model. To detect such attacks, various methodologies are employed:

Statistical Analysis [2,5,9]: Detection methods may leverage statistical metrics such as image entropy, pixel intensity distributions, or spatial correlations to identify anomalies introduced by adversarial manipulations.

Gradient-Based Approaches [6,11]: By analyzing the gradients of the loss function concerning input pixels, detection methods can identify regions where gradients deviate significantly from the norm, signalling potential adversarial perturbations.

Image Restoration Techniques [4,8]: Employing image restoration methods, such as denoising or inpainting, can help detect pixel-based adversarial attacks by highlighting inconsistencies or artifacts introduced during the attack process.

C. Model-Based Adversarial Attacks Detection Methods

Model-based adversarial attacks focus on exploiting vulnerabilities in the model's architecture and decision-making processes. Detecting these attacks involves methodologies that scrutinize model behavior:

Anomaly Detection [1,3,12]: Anomaly detection techniques, including monitoring model outputs for unexpected deviations from normal behavior, can identify instances where adversarial examples lead to abnormal predictions.

Expandability and Interpretability [7,10,14]: Leveraging interpretability tools such as saliency maps or attention mechanisms helps visualize the model's

decision process, enabling the identification of unexpected patterns indicative of adversarial inputs.

Ensemble Methods [13,15]: Constructing ensemble models with diverse architectures or incorporating multiple classifiers enhances robustness against adversarial attacks. Inconsistencies in predictions across ensemble members can signal potential adversarial inputs.

TABLE I

COMPARATIVE ANALYSIS

Methodology	Pros	Cons
Black Box Attacks [1,3,5]	- Less information available to attackers.	- Crafting effective adversarial examples may be challenging. - Limited understanding of the model's internal structure.
White Box Attacks [1,3,5]	- More potent in crafting tailored adversarial examples. - Comprehensive understanding of model internals.	- Higher threat level due to extensive knowledge. - Greater potential for successful attacks.
Statistical Analysis [2,5,9]	- Utilizes statistical metrics for anomaly detection. - Broad applicability to various models.	- May struggle with subtle attacks that avoid statistical abnormalities. - Limited effectiveness if the attack is carefully crafted.
Gradient-Based Approaches [6,11]	- Identifies deviations in gradient norms for pixel values. - Can be effective in detecting localized adversarial perturbations.	- Sensitive to noise and may yield false positives. - May struggle with attacks designed to evade gradient analysis.
Image Restoration Techniques [4,8]	- Uses restoration methods to highlight inconsistencies. - Provides additional insight into attack artifacts.	- Limited effectiveness against certain attack strategies. - Computational overhead for image restoration methods.
Anomaly Detection 1,3,12]	- Monitors model outputs for unexpected deviations. - Can identify abnormal predictions caused by adversarial examples.	- May have false positives in cases of legitimate model behavior changes. - Limited to detecting anomalies post-prediction.
Expandability and Interpretability [7,10,14]	- Visualizes model decision processes for pattern identification.	- Interpretability tools may have limitations in complex models.
Ensemble Methods [13,15]	- Helps understand unexpected patterns indicative of adversarial inputs.	- Requires additional computational resources for visualization.

IV.CONCLUSION

In conclusion, the analysis of adversarial attack methodologies and detection techniques underscores the intricate landscape of securing machine learning models. While black box and white box attacks represent contrasting approaches, each with its own set of challenges, a comprehensive defense strategy requires a multi-faceted approach. The evaluation of pixel-based and model-based detection methods reveals the importance of combining these techniques for improved performance. Pixel-based approaches leverage statistical and gradient-based analyses to scrutinize image-level perturbations, while model-based methods focus on understanding and fortifying the internal workings of the model itself. The synergy between these two approaches provides a more holistic defense mechanism, addressing vulnerabilities at both the input data and model architecture levels.

Further research and development should concentrate on refining and optimizing the integration of pixel-based and model-based detection methods. Investigating advanced ensemble techniques that seamlessly blend pixel-level anomaly detection with model interpretability and expandability could yield even more robust defenses against evolving adversarial threats. Additionally, exploring the incorporation of machine learning algorithms capable of adapting to emerging attack strategies can enhance the adaptability of the defense mechanisms. Future work should also consider the scalability and efficiency of these combined approaches to ensure practical implementation in real-world, resource-constrained environments. As the field progresses, a concerted effort toward developing standardized evaluation metrics and benchmark datasets will be crucial for objectively comparing the effectiveness of different methodologies and promoting advancements in adversarial defense strategies.

V. REFERENCES

[1] G. Ryu and D. Choi, “Detection of adversarial attacks based on differences in image entropy,” International Journal of Information Security, 2023, doi: 10.1007/s10207-023-00735-6.

[2] X. Cui, “Targeting Image-Classification Model,” pp. 1–13, 2023.

[3] M. Kim and J. Yun, “AEGuard: Image Feature-Based Independent Adversarial Example Detection Model,” Security and Communication Networks, vol. 2022, 2022, doi: 10.1155/2022/3440123.

[4] P. Lorenz, M. Keuper, and J. Keuper, “Unfolding Local Growth Rate Estimates for (Almost) Perfect Adversarial Detection,” pp. 27–38, 2023, doi: 10.5220/0011586500003417.

[5] L. Shi, T. Liao, and J. He, “Defending Adversarial Attacks against DNN Image Classification Models by a Noise-Fusion Method,” Electronics (Switzerland), vol. 11, no. 12, 2022, doi: 10.3390/electronics11121814.

[6] A. S. Almuflih, D. Vyas, V. V Kapdia, M. R. N. M. Qureshi, K. M. R. Qureshi, and E. A. Makkawi, “Novel exploit feature-map-based detection of adversarial attacks,” Applied Sciences, vol. 12, no. 10, p. 5161, 2022.

[7] M. Khan et al., “Alpha Fusion Adversarial Attack Analysis Using Deep Learning,” Computer Systems Science and Engineering, vol. 46, no. 1, pp. 461–473, 2023, doi: 10.32604/csse.2023.029642.

[8] N. Ghaffari Laleh et al., “Adversarial attacks and adversarial robustness in computational pathology,” Nature Communications, vol. 13, no. 1, pp. 1–10, 2022, doi: 10.1038/s41467-022-33266-0.

[9] Y. Wang et al., “Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey,” pp. 1–46, 2023, [Online]. Available: http://arxiv.org/abs/2303.06302

[10] H. Hirano, A. Minagi, and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," BMC Medical Imaging, vol. 21, no. 1, pp. 1–13, 2021, doi: 10.1186/s12880-020-00530-y.

[11] A. Talk, F. Wikipedia, A. Wikipedia, and C. Wikipedia, "University of Science and Technology of China," no. 6, p. 29201, 2001.

[12] Y. Zheng and S. Velipasalar, "Part-Based Feature Squeezing To Detect Adversarial Examples in Person Re-Identification Networks," Proceedings - International Conference on Image Processing, ICIP, vol. 2021-September, pp. 844–848, 2021, doi: 10.1109/ICIP42928.2021.9506511.

[13] B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang, "Detecting Adversarial Image Examples in Deep Neural Networks with Adaptive Noise Reduction," IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 1, pp. 72–85, 2021, doi: 10.1109/TDSC.2018.2874243.

[14] M. A. Ahmadi, R. Dianat, and H. Amirkhani, "An adversarial attack detection method in deep neural networks based on re-attacking approach," pp. 10985–11014, 2021.

[15] 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.

Cite this article as :

Soni Kumari, Sheshang Degadwala, "A Comprehensive Review on Adversarial Attack Detection Analysis in Deep Learning ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 10, pp.319-325, September-October-2023. Available at doi : https://doi.org/10.32628/CSEIT2361054

Journal URL : https://ijsrcseit.com/CSEIT2361054