Understanding the theory of Adversarial Attack and Defense Mechanisms in AI Systems

ABSTRACT

Within the realm of artificial intelligence systems, the specter of adversarial attacks poses significant challenges to information security. This paper delves into the intricate landscape of adversarial attacks and defense mechanisms within AI systems. Adversarial attacks, ranging from subtle manipulations to overt intrusions, undermine the integrity and reliability of AI models, necessitating robust defense strategies. This abstract elucidates key facets of adversarial attack methodologies and defense mechanisms, highlighting the critical imperative of safeguarding AI systems against evolving threats. Through an in-depth exploration of adversarial attack vectors, including adversarial examples, model inversion, model stealing, and data poisoning, coupled with comprehensive defense strategies such as adversarial training, model regularization, and input sanitization, this concentration underscores the pivotal role of proactive security measures in fortifying AI systems. Thereby, this paper aims to empower stakeholders with the necessary knowledge and tools to protect AI systems from malicious exploitation by delving into different forms of adversarial attacks, techniques for recognizing and preventing such attacks, a thorough investigation of defense tactics, an assessment of the consequences of these attacks, and insights drawn from recent studies.

KEYWORDS

1. **Adversarial Examples**
   Inputs to machine learning models that are intentionally crafted to cause misclassification or incorrect predictions, often by adding imperceptible perturbations to legitimate data.

2. **Model Inversion**
   A type of adversarial attack where an attacker attempts to reverse-engineer or extract sensitive information from a trained machine learning model, often by observing its outputs.

3. **Model Stealing**
   An attack where an adversary attempts to replicate or clone a target machine learning model by querying it with carefully crafted inputs and learning its internal structure or parameters.

4. **Data Poisoning**
   A method of attack where an adversary introduces malicious or misleading data into the training dataset of a machine learning model, compromising its performance or integrity.

5. **Adversarial Training**
   A defense strategy where machine learning models are trained using both legitimate and adversarial examples to improve their robustness against adversarial attacks.

6. **Model Regularization**
   A technique used to prevent overfitting in machine learning models by adding constraints or penalties to the model parameters during training.

7. **Input Sanitization**
   The process of filtering or cleansing input data to remove potentially malicious or adversarial components before it is processed by a machine learning model.

8. **Attack Vectors**
   Methods or pathways used by attackers to exploit vulnerabilities in AI systems, including adversarial examples, data poisoning, and model inversion.

9. **Transferability**
   The ability of adversarial examples crafted for one machine learning model to deceive other models, often with similar architectures or trained on similar datasets.

10. **Computer Vision**
    A field of artificial intelligence that enables computers to interpret and understand visual information from the real world, including images and videos.

11. **Malware**
    Malicious software is designed to disrupt, damage, or gain unauthorized access to computer systems, networks, or devices.

12. **Regularization**
    A technique used in machine learning to prevent overfitting by adding a penalty term to the model's loss function, encouraging simpler models and improving generalization.

13. **Distillation Technology**
    A method used to compress large-scale machine learning models into smaller, more efficient models while retaining their original accuracy and performance.

14. **Generative Adversarial Networks (GANs)**
    A class of machine learning algorithms that consist of two neural networks, the generator and the discriminator, which are trained simultaneously to generate realistic data samples.

15. **Seminal Works**
    Influential or groundbreaking research contributions that have significantly impacted or shaped a particular field of study, providing foundational knowledge or insights for further exploration and development.

INTRODUCTION

In the digital age, marked by ubiquitous internet access, exponential data growth, and the constant evolution of machine learning, artificial intelligence has become integral across industries worldwide (Gao, 2019). From image recognition to autonomous driving, AI technologies are driving transformative changes. Concurrently, machine learning advancements are reshaping traditional computer security research, with hackers increasingly leveraging ML for precise attacks. Recent studies have unveiled vulnerabilities in fields like computer vision and network security to adversarial threats (Esposito, 2021).

The seminal work of Szegedy et al (Szegedy, 2014). introduced adversarial samples, revealing neural network weaknesses and sparking widespread research interest. Following this, Liu et al. pioneered a method exploiting malware-based visual detectors for adversarial attacks, while Zhou et al. innovated an alternate model for training adversarial attacks without real data.

This paper aims to be a comprehensive resource for adversarial attack research, systematically categorizing and comparing algorithms across image, text, and malware domains. It reviews concepts, types, impacts, evolution, and defense mechanisms against adversarial attacks. Additionally, it discusses future research directions to support the growing field of adversarial attack security.

Key contributions include:

1. Presenting adversarial attacks in a structured format for rapid understanding.
2. Providing a roadmap for researchers entering adversarial attack research.
3. Offering recent, reputable references to ensure up-to-date coverage.
4. Analyzing methodological strengths and weaknesses to aid researchers in method selection.

5. Categorizing literature to facilitate navigation of adversarial attack research.

STRUCTURE OF THE PAPER

The paper’s structure is outlined as follows:

1. **Rationale for Studying Adversarial Attacks**
   This section advocates for the importance of studying adversarial attacks.

2. **Understanding Adversarial Attacks**
   Here, we delve into the concepts, classifications, risks, and methodologies related to adversarial attacks.

3. **Categorization and Comparison of Adversarial Attack Methods**
   This section meticulously categorizes and contrasts methods used in adversarial attacks across images, text, and malware.

4. **Exploring Defense Strategies**
   We examine various defense methods against adversarial attacks.

5. **Discussion Of Review Results**
   Comparative Analysis with Existing Reviews - This section highlights distinctions between our paper and similar reviews.

6. **Supplementary Considerations**
   Additional supplements necessary for effective research into adversarial attacks are presented.

7. **Practical implementation**
   Open-source library, repository, and toolboxes to measure the robustness, resilience and performance of the models. Examples: Adversarial Robustness Toolbox (ART), Cortex Certifai

8. **Conclusion**
   Discussion on Future Directions - We offer insights into potential avenues for future research, Overall Added Value, and Limitations of research. Finally, we summarize our findings and conclusions.

REVIEW PROCESS AND METHODOLOGY

For this paper, we undertook a rigorous process to ensure the quality and reliability of the content.

1. **Literature Review**
   Conducting a comprehensive review of existing literature and research findings to identify gaps, trends, and areas requiring further investigation.

2. **Experimental Analysis**
   Designing and conducting experiments to evaluate the efficacy of different security techniques and defense mechanisms in protecting AI systems against various types of attacks.

3. **Case Studies**
   Analyzing real-world case studies and scenarios to understand the practical implications of security vulnerabilities and the effectiveness of mitigation strategies.

4. **Collaboration and Knowledge Sharing**

5. Engaging with fellow researchers, industry experts, and practitioners to exchange ideas, collaborate on projects, and disseminate research findings through conferences, workshops, and publications.

6. **Expert Consultation**
   Input from subject matter experts in the fields of artificial intelligence, cybersecurity, and information security was sought to gain insights into the nuances of adversarial attacks and defense strategies. Expert consultations helped validate the relevance and accuracy of the information gathered and provided additional perspectives on the topic.

WHY STUDY ADVERSARIAL ATTACK

Studying adversarial attacks is imperative due to their profound threat to the efficacy of deep learning, particularly in practical applications such as computer vision. The surge in research contributions in this area underscores its critical importance. This paper presents a comprehensive survey of adversarial attacks in deep learning, encompassing the design of attacks, analysis of their existence, and proposals for defense mechanisms. Moreover, the exploration of adversarial attacks in real-world scenarios underscores their practical feasibility. Notably, the transferability of adversarial examples across neural networks, especially those with similar architectures, underscores the need for robust defenses against such attacks. The literature reveals diverse perspectives on the underlying reasons for deep neural network vulnerability to adversarial perturbations, emphasizing the necessity for systematic investigation in this realm (AKHTAR, 2018).

LITERATURE REVIEW

1. UNDERSTANDING ADVERSARIAL ATTACK

Focusing on adversarial examples in research not only keeps AI security at the forefront but also sustains researchers' motivation. Adversarial attacks involve subtly altering examples to mislead machine learning models while remaining imperceptible to human observers. This research direction ensures ongoing engagement and progress in AI security (Kong, 2021).

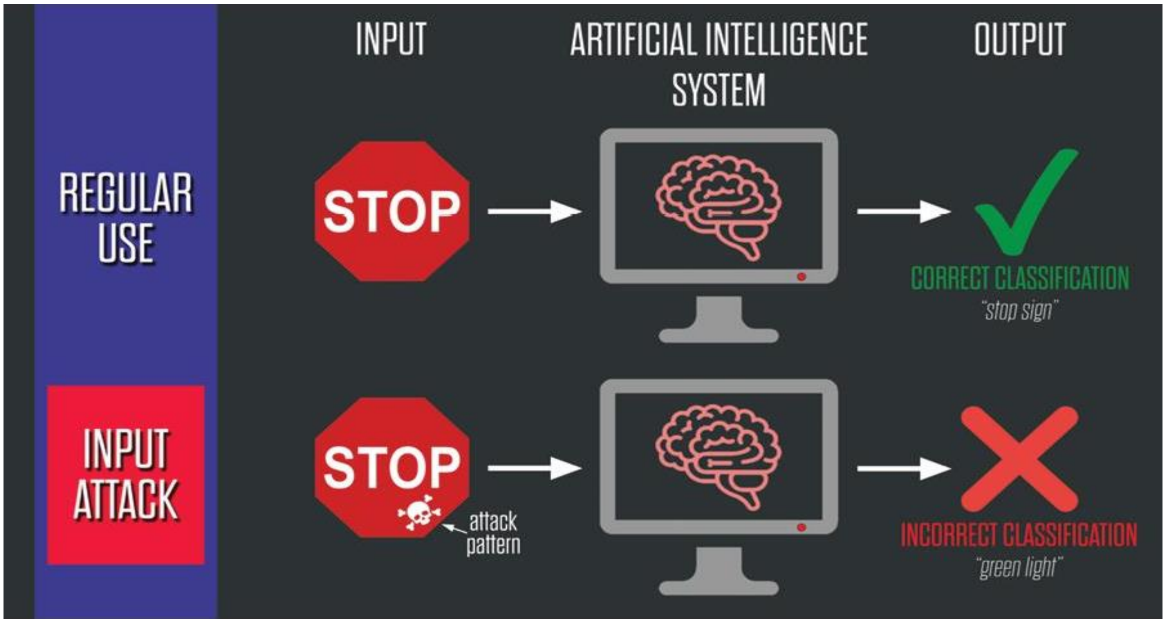


Figure: In regular use (top), the AI system takes a valid input, processes it with the model (brain), and returns an output. In an input attack (bottom), the input to the AI system is altered with an attack pattern, causing the AI system to return an incorrect output. Copy right and Adopted from: Marcus Comiter (2019) Attacking Artificial Intelligence: AI’s Security Vulnerability and What Policymakers Can Do About It.

2. CATEGORIZATION AND COMPARISON OF ADVERSARIAL ATTACK METHODS, (AKHTAR, 2018).

White-Box Attack

In a white-box attack, the attacker has full knowledge of the target model's architecture and parameters, allowing precise manipulation of inputs to exploit vulnerabilities and cause incorrect predictions or classifications.

Black-Box Attack

In a black-box attack, the attacker does not know the target model's internal structure. They can only manipulate inputs and observe outputs to craft attacks, making them less precise but potentially more versatile and applicable across various models.

Real-World Attack/Physical Attack

Real-world attacks, also known as physical attacks, operate without understanding the internal structure of the target model and typically involve manipulating inputs or environmental conditions to deceive the model. These attacks simulate real-world scenarios where adversaries may have limited knowledge or control over the model but can still exploit vulnerabilities to cause misclassifications or errors.

Targeted Attack

A targeted attack involves intentionally directing adversarial efforts towards a specific outcome, aiming to cause the model to incorrectly predict a predefined label or outcome for certain inputs. This type of attack requires prior knowledge of the target label or outcome and is designed to achieve a predetermined result, such as misclassification or system manipulation.

Untargeted Attack

An untargeted attack is not focused on achieving a specific outcome or label for the adversarial input. Instead, the goal is simply to induce the model to make a wrong prediction or classification, without specifying what that incorrect outcome should be. This type of attack aims to disrupt the model's decision-making process without predefining the desired misclassification.

Evade Attack

An evade attack involves introducing disturbances to test samples or modifying inputs during testing to deceive the model's detection mechanisms. The objective is to evade detection by the model, causing it to misclassify or fail to recognize adversarial inputs as malicious, thereby compromising its reliability and security.

Poisoning Attack

A poisoning attack involves injecting carefully crafted malicious examples into the model's training data. These examples are designed to subtly manipulate the model's learning process, introducing vulnerabilities or backdoors that can be exploited by attackers during inference or deployment. This type of attack aims to compromise the integrity and performance of the model by undermining its training process.

Backdoor Attack

A backdoor attack involves surreptitiously implanting specific patterns or features into the model during training to create a hidden vulnerability. This allows attackers to manipulate the model's behavior in specific ways when exposed to inputs containing the implanted trigger. Backdoor attacks enable unauthorized access to the model or its outputs, circumventing normal security measures and potentially causing significant harm or disruption.

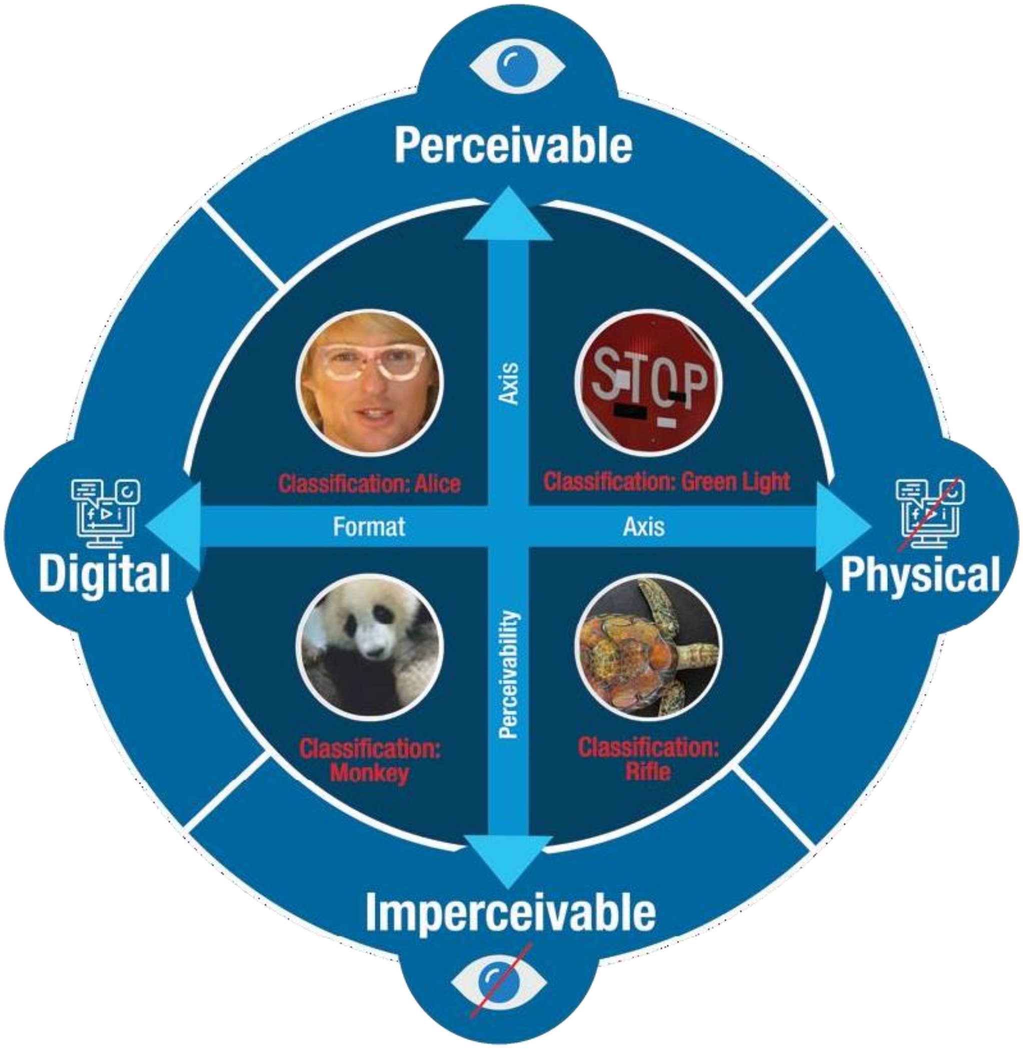


Figure: Taxonomy for categorizing input attacks. The horizontal axis characterizes the format of the attack, either in the physical world or digital. The vertical axis characterizes the perceivability of the attack, either perceivable to humans or imperceivable to humans. Copy right and Adopted from: Graphic by Marcus Comiter except for stop sign attack thumbnail from Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, panda attack thumbnail from Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples.” arXiv preprint arXiv:1412.6572 (2014), turtle attack thumbnail from Athalye, Anish, et al. “Synthesizing robust adversarial examples.” arXiv preprint arXiv:1707.07397 (2017), and celebrity attack thumbnail from Sharif, Mahmood, et al. “Adversarial generative nets: Neural network attacks on state-of-the-art face recognition.” arXiv preprint arXiv:1801.00349 (2017).

3. DEFENSE STRATEGIES FOR ADVERSARIAL ATTACKS

Defensive strategies against adversarial attacks encompass three main approaches: modifying data, modifying models, and utilizing auxiliary tools (Qiu, 2019).

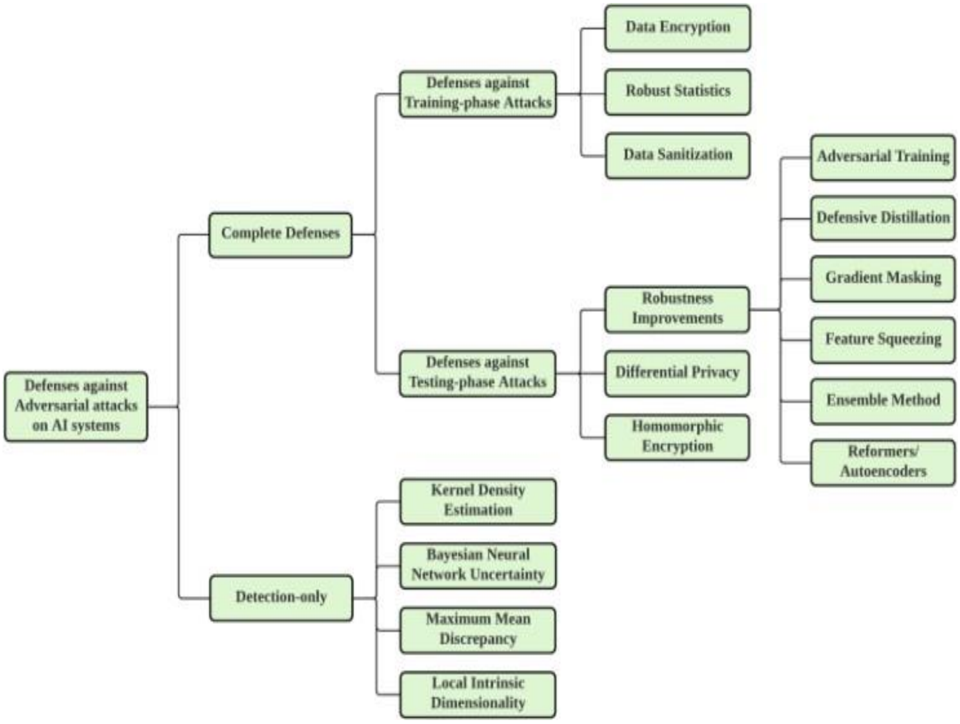


Figure: Taxonomy of defenses against AI system attacks. Copy right and adopted from: Ayodeji OseniNour MoustafaH. JanickePeng LiuZ. TariA. Vasilakos (2021). Security and Privacy for Artificial Intelligence: Opportunities and Challenges

1. Modifying Data

- Adversarial Training: Incorporates adversarial samples into the training dataset to enhance model robustness. This method involves training the model with modified labels to withstand adversarial inputs.
- Data Randomization: Introduces random resizing or texture additions to adversarial samples during training to reduce their effectiveness.

2. Modifying Models

- Regularization: Enhances model generalization by adding penalty terms to the cost function during training, limiting vulnerability to adversarial attacks.

- Defensive Distillation: Utilizes distillation technology to create a smoother output surface and reduce model sensitivity to disturbances. This approach involves training a distilled network on the probability vector predicted by the initial network.

3. **Using Auxiliary Tools:**
   - Defense-GAN: Utilizes generative adversarial networks to generate adversarial examples for model training, enhancing robustness.
   - MagNet: Implements a detection and correction mechanism to identify and mitigate adversarial perturbations in inputs.
   - High-Level Representation Guided Denoiser: Incorporates additional tools to assist the neural network model in detecting and filtering out adversarial inputs based on high-level representations.

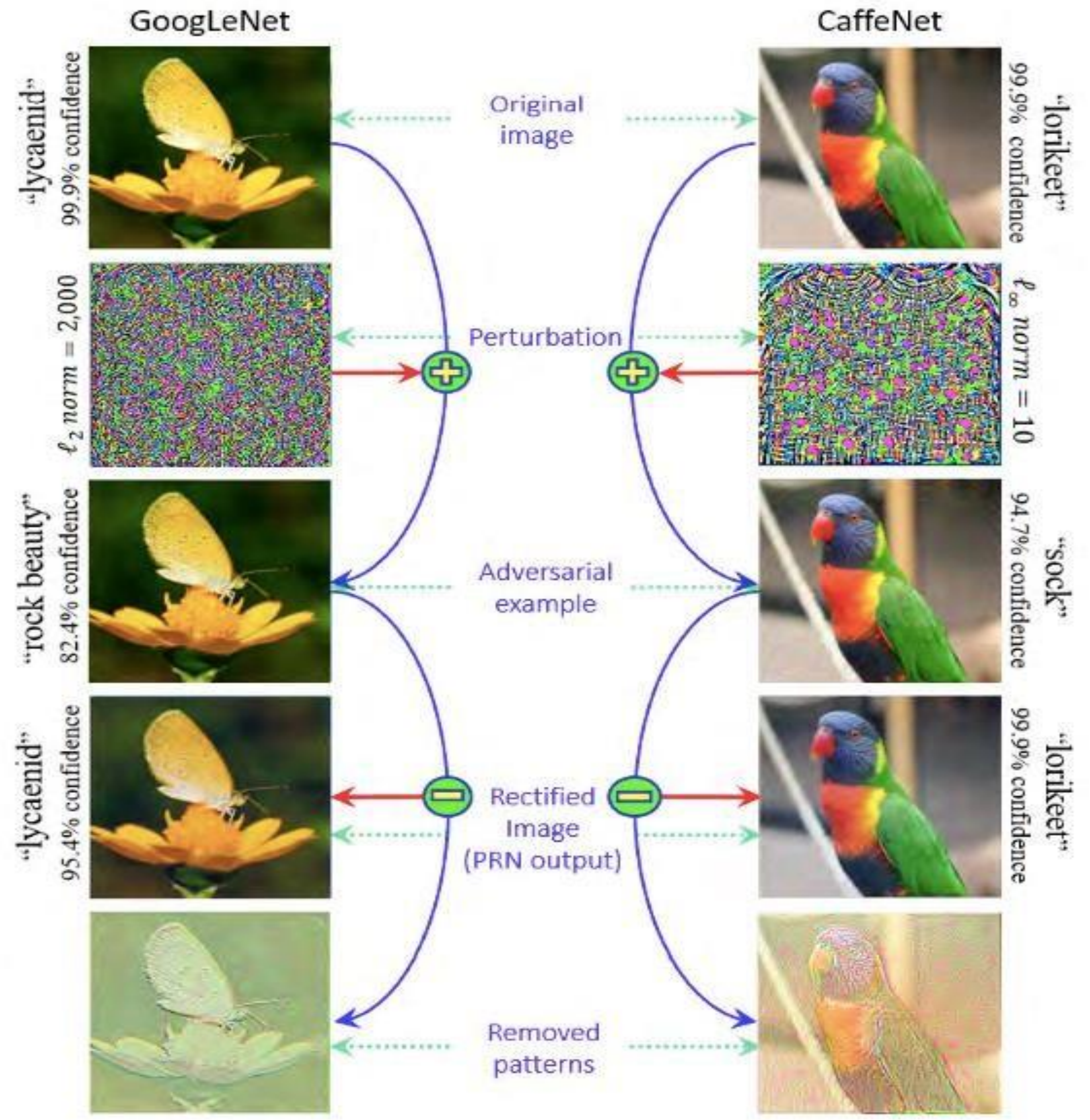


Figure: Illustration of defense against universal perturbations. The approach rectifies an image to restore the network prediction. The pattern removed by rectification is separately analyzed to detect perturbation. Copy right and adopted from: N. Akhtar, J. Liu, and A. Mian. (2017). ‘‘Defense against universal adversarial perturbations.’’ [Online]. Available: https://arxiv.org/abs/1711.05929

DISCUSSION OF REVIEW RESULTS

AUTHOR	MAIN CONTENT	REVIEW RESULTS
Naveed Akhtar, 2018	Deep learning, central to the current AI boom, is particularly dominant in computer vision applications like self-driving cars and surveillance. While deep neural networks excel in solving complex tasks, recent findings reveal vulnerability to adversarial attacks. These attacks involve subtle alterations to inputs, leading models to make incorrect predictions, even beyond human capacity. This threat has spurred a surge in research efforts. This paper offers the first comprehensive examination of adversarial attacks in computer vision, covering attack design, analysis, and defense strategies. Notably, it highlights real-world evaluations to underscore practical vulnerability. Additionally, it provides insights into the future trajectory of this research area.	Recent research highlights the vulnerability of deep neural networks to subtle input changes, prompting extensive study into adversarial attacks and defense mechanisms. This review focuses on seminal works in this domain, revealing the threat posed to deep learning's practical applications, especially in critical areas like safety and security. Deep learning models can be manipulated in both digital and physical settings. Despite these challenges, ongoing research suggests optimism for future advancements in building robust deep learning systems resilient to adversarial attacks.
Zixiao Kong, 2021	As AI applications expand with the internet's growth, so do adversarial attacks, prompting urgent research in security. This review outlines key theories and methods for researchers entering the adversarial attack field. It covers the significance, types, and risks of such attacks, focusing on image, text, and malicious code domains. Discussions on open issues and comparisons with similar reviews conclude the paper.	This review serves as a roadmap for researchers, drawing from recent high-quality articles to elucidate typical adversarial attacks in text, image, and malware domains. It also outlines defense technologies and highlights discussions on open issues. By providing this framework, the paper aims to empower new researchers to navigate the complexities of adversarial attack and defense effectively.
Shilin Qiu, 2019	In recent years, artificial intelligence (AI) technologies	This paper categorizes adversarial attacks based on their timing: during

	have found widespread application in various fields like computer vision and natural language processing. However, their susceptibility to adversarial attacks poses a significant challenge, particularly in security-critical domains. Enhancing AI system robustness against such attacks has become pivotal for further AI advancement. This paper aims to provide a comprehensive overview of recent research progress on adversarial attack and defense technologies in deep learning. It categorizes adversarial attacks based on the target model's stage of occurrence and explores their applications across different domains. Additionally, it discusses various defense methods, including data modification, model adjustment, and auxiliary tool utilization.	model training or testing. Training stage attacks modify datasets, labels, or inputs, while testing stage attacks employ white-box or black-box methods. White-box attacks use detailed model information for higher success, whereas black-box attacks demonstrate broader applicability. The paper also discusses attack applications and defense methods, stressing the need for comprehensive defense strategies to ensure AI security across diverse applications.

SUPPLEMENTARY CONSIDERATIONS

Research efforts to defend neural networks against adversarial attacks are met with countermeasures as attackers devise more potent methods. Notably, a Kaggle competition was hosted to develop defenses against such attacks, reflecting the growing interest in this area. This surge in research activity holds promise for enhancing the robustness of deep learning approaches, paving the way for their safe and secure deployment in real-world applications.

PRACTICAL IMPLEMENTATION – CORTEX CERTFAI TOOL

Certifai an open-source product, offers a platform tailored to evaluate AI models, focusing on their robustness, fairness, and explainability. This tool comes in two versions:

Certifai Toolkit

This edition equips users with the necessary resources to develop, execute, and review assessments locally.

Certifai Enterprise

Designed for multi-user server setups in Kubernetes, this version includes extra features like the Certifai AI Risk Assessment Questionnaire and Policy Select compliance toolkit.

The process involves several steps:

Scan Definitions

Data scientists define scans, specifying models and datasets for evaluation. A subset of the dataset may be utilized for explanations.

Evaluation Metrics

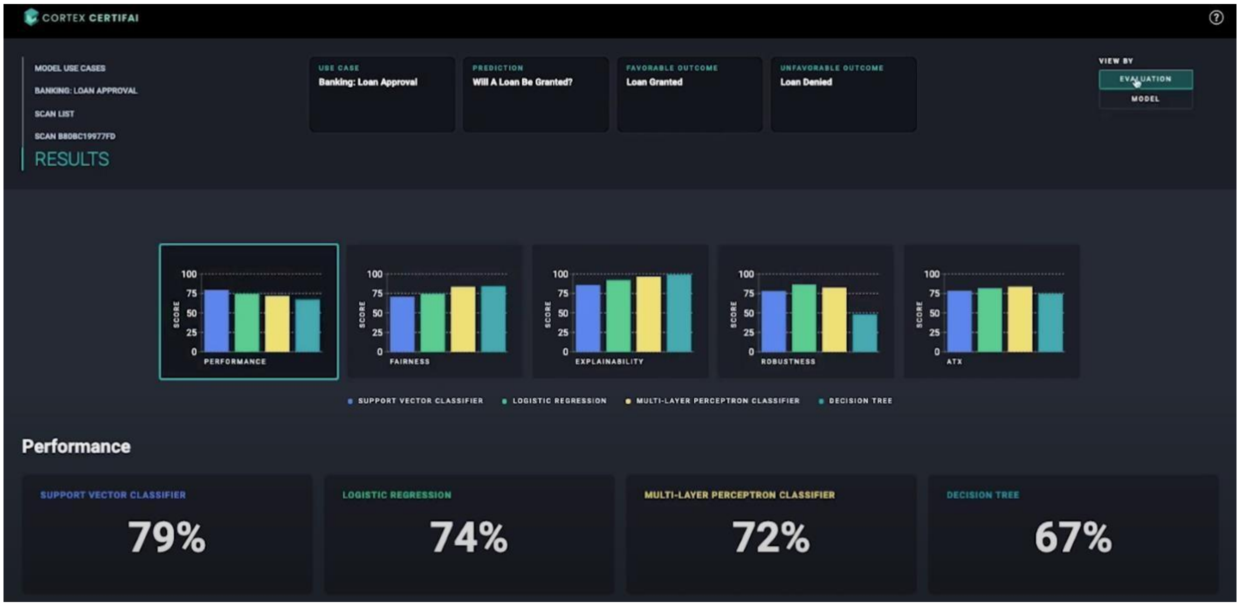
Models undergo assessment based on various metrics:

Performance Metric

Evaluates accuracy using test data or pre-calculated scores.

Robustness
Assesses how well models maintain outcomes with changing data feature values.
Fairness by Group
Measures the disparity in outcomes among different groups within features.
Explainability
Gauges the clarity of counterfactual explanations provided for each model.
Certifai generates counterfactual explanations demonstrating how altering dataset features can yield different outcomes. This aids in understanding which features influence results and to what extent.
Decision Making
Business decision-makers and Compliance Officers utilize visualizations and scores to compare evaluations and select models aligned with business objectives. They also ensure models meet predefined standards for robustness, fairness, and explainability.
Model Improvement
Data scientists leverage evaluation results to refine models and training processes, thereby enhancing the reliability and trustworthiness of AI models over time.

Certifai serves as a comprehensive toolset to validate AI models, fostering confidence in their performance, fairness, and transparency in various applications.



CORTEX CERTIFAI

MODEL USE CASES
BANKING: LOAN APPROVAL
SCAN LIST
SCAN B80BC19977FD
RESULTS

USE CASE: Banking: Loan Approval
PREDICTION: Will A Loan Be Granted?
FAVORABLE OUTCOME: Loan Granted
UNFAVORABLE OUTCOME: Loan Denied

VIEW BY: EVALUATION

Performance

SUPPORT VECTOR CLASSIFIER: 79%
LOGISTIC REGRESSION: 74%
MULTI-LAYER PERCEPTRON CLASSIFIER: 72%
DECISION TREE: 67%

Figure: Cognitive scale - Cortex Certifai console

CONCLUSION

Looking ahead, it is essential to sustain the momentum of research activity and collaboration in the field. By continuing to innovate and refine defense techniques, we can enhance the resilience of deep learning approaches, paving the way for their safe and secure deployment in safety-critical applications. Ultimately, addressing the multifaceted challenges posed by adversarial attacks requires a concerted effort from researchers across disciplines. Through collective action and interdisciplinary collaboration, we can advance towards a future where AI technologies are robust, reliable, and trustworthy in real-world settings.

Overall Added Value

This survey represents a significant contribution to the field of adversarial attacks on deep learning in Computer Vision, providing a comprehensive overview of recent research endeavors. It highlights the vulnerability of deep neural networks to subtle input perturbations and emphasizes the critical need for robust defense strategies. By reviewing influential works, the survey adds value by synthesizing key insights and trends, fostering a deeper understanding of adversarial attack dynamics.

Limitations of Research

While defense techniques have been proposed to mitigate known attacks, the ongoing evolution of adversarial methods presents a continuous challenge. Despite promising strides, current defense methods often target specific attack types and lack comprehensive coverage. Additionally, the efficacy of defenses may vary across different scenarios, highlighting the need for further research to address these limitations.

Suggestions for Future Direction

The high research activity in adversarial attack and defense underscores the importance of continued exploration in this area. Future research should focus on developing more robust defense strategies capable of mitigating a broader range of adversarial attacks. Additionally, there is a need for interdisciplinary collaboration to tackle adversarial threats across various domains, including safety-critical applications. By fostering innovation and collaboration, the field can advance toward ensuring the security of AI technology in diverse real-world scenarios.

Moving forward, we intend to investigate adversarial attack instances within the realm of malicious code. Given the structural resemblance between malicious code and textual data, we contemplate the adaptation of text-based adversarial attack algorithms to the domain of malicious code.

In conclusion, the landscape of adversarial attacks on deep learning in Computer Vision is evolving rapidly, fueled by the ongoing efforts to defend neural networks against increasingly sophisticated attacks. This survey has shed light on the vulnerabilities of deep neural networks to subtle perturbations, emphasizing the urgent need for robust defense mechanisms. While significant progress has been made in developing defense strategies, the dynamic nature of adversarial attacks poses ongoing challenges.

APPENDIX

https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack/

https://www.harvardmagazine.com/2018/12/ai-and-adversarial-attacks

https://www.researchgate.net/publication/374549251_An_Intelligent_Secure_Adversarial_Examples_Detection_Scheme_in_Heterogeneous_Complex_Environments

https://www.belfercenter.org/publication/AttackingAI

https://www.semanticscholar.org/paper/Security-and-Privacy-for-Artificial-Intelligence%3A-Oseni-Moustafa/04659ccb232c5ecbe42ef8a522bbb55f41f7a7aa

https://www.researchgate.net/publication/347019914_On_Adversarial_Examples_and_Stealth_Attacks_in_Artificial_Intelligence_Systems

https://www.linkedin.com/pulse/adversarial-attacks-aiml-models-everything-you-need-know-d-souza/

https://neurosciencenews.com/ai-vulnerabilities-neuroscience-25312/

https://research.ibm.com/topics/adversarial-robustness-and-privacy

https://cognitivescale.github.io/cortex-certifai/docs/about/

https://www.aitimejournal.com/common-types-of-attacks-on-ai-systems/46003/

https://ieeexplore.ieee.org/document/9827763

https://www.thesecuritybench.com/ai/adversarial-attacks-and-defense-mechanisms-in-ai-systems/

https://hls.harvard.edu/today/medical-ai-systems-could-be-vulnerable-to-adversarial-attacks/

https://www.infosys.com/iki/perspectives/securing-ai-adversarial-attacks.html

REFERENCE LIST

1. Ivan Y. Tyukin; Desmond J. Higham; Alexander N. Gorban (2020). On Adversarial Examples and Stealth Attacks in Artificial Intelligence Systems.
2. Shilin Qiu; ORCID,Qihe Liu; Shijie Zhou; Chunjiang Wu (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies.
3. Zixiao Kong, Jingfeng Xue, Yong Wang, Lu Huang, Zequn Niu, Feng Li (2021). A Survey on Adversarial Attack in the Age of Artificial Intelligence - School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
4. Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP (2019). A defence against Trojan attacks on deep neural networks," in Proceedings of the 35th Annual Computer Security Applications Conference, pp. 113–125, New York, NY, USA.
5. C. Esposito, M. Ficco, and B. B. Gupta (2021). "Blockchain-based authentication and authorization for smart city applications," Information Processing & Management, vol. 58, no. 2, p. 102468.
6. C. Szegedy, W. Zaremba, I. Sutskever et al.,(2014). "Intriguing properties of neural networks," in 2nd International Conference on Learning Representations, ICLR 2014, Banff, Canada.
7. Naveed Akhtar and Ajmal Mian (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey.
8. Shilin Qiu, ORCID,Qihe Liu, Shijie Zhou, andChunjiang Wu (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies.
9. Ayodeji OseniNour MoustafaH. JanickePeng LiuZ. TariA. Vasilakos (2021). Security and Privacy for Artificial Intelligence: Opportunities and Challenge.