Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 07, No. 01, 2024 ♦ PP. 20–23

Article DOI: https://doi.org/10.54216/JCHCI.070105

FULL LENGTH ARTICLE

Defense Against Adversarial AI

Bhavani G.[1,*] **Soundarya S.**[1] **Tejashwini V.**[1] **Sumitha S.**[1]

[1] Panimalar Engineering College, Chennai, Tamil Nadu, India

Emails: bhavanigovar@gmail.com · soundisweetie.kolathur@gmail.com · tejasreedevi6@gmail.com · sumithasuresh07@gmail.com

Received: August 21, 2023 Revised: October 14, 2023 Accepted: January 27, 2024 ⋆ Corresponding author

ABSTRACT

The increasing prevalence of deep learning technology has paved the way for a new era of AI-powered capabilities, promising revolutionary advancements across various societal domains such as healthcare and autonomous vehicles. Despite offering potent solutions to complex problems, the formidable power of these AI systems is accompanied by a susceptibility that malicious actors could exploit. Adversarial attacks, particularly targeting deep learning models, involve the crafting of altered inputs, often imperceptible changes to images, to deceive or undermine the functionality of the AI system. Within the domain of autonomous driving systems, adversarial attacks pose a severe risk. Envision a situation where a precisely manipulated adversarial attack targets a red traffic light sign, causing the AI system to misclassify it as an entirely unrelated object, perhaps identifying it as a bird. The potential consequences of such misclassifications underscore the serious impact that adversarial attacks can exert on the safety and dependability of autonomous vehicles. Ensuring the resilience and security of AI technologies against adversarial threats is of utmost importance as AI continues to play a pivotal role in critical applications such as healthcare, finance, and autonomous systems. It necessitates a holistic strategy that melds advanced research, meticulous testing, and the deployment of robust security measures. This comprehensive approach is essential for fostering trust and mitigating potential harm in an ever-growing, AI-driven world.

Keywords: Deep Learning ▪ Artificial Intelligence ▪ Smart Systems ▪ Complex Problems ▪ Adversarial Attacks

1. INTRODUCTION

In today's dynamic landscape of artificial intelligence (AI), deep learning models represent the pinnacle of computational achievement, showcasing remarkable capabilities in pattern recognition, prediction, and data processing. However, alongside their power comes a significant vulnerability: susceptibility to adversarial attacks. These attacks exploit weaknesses in AI systems and aim to manipulate decision-making processes, potentially leading to incorrect predictions or classifications.

The concept of hacking deep learning models delves into this realm of subversion, where sophisticated techniques are employed to deceive and compromise AI systems. Adversarial attacks involve crafting subtle alterations to input data, known as adversarial examples, which can mislead the model while appearing normal to human observers. Motivations for such attacks vary, ranging from academic research to more malicious intentions such as digital impersonation or misinformation campaigns.

The implications of hacked deep learning models extend beyond technical concerns to ethical and societal considerations. Erroneous medical diagnoses, flawed autonomous vehicle decisions, and compromised communication systems are among the potential consequences, highlighting the critical need to understand and address vulnerabilities in AI systems. Addressing these challenges requires a multifaceted approach involving the exploration of attack vectors, the development of defensive strategies, and consideration of real-world impacts. By fostering a deeper understanding of the dynamics

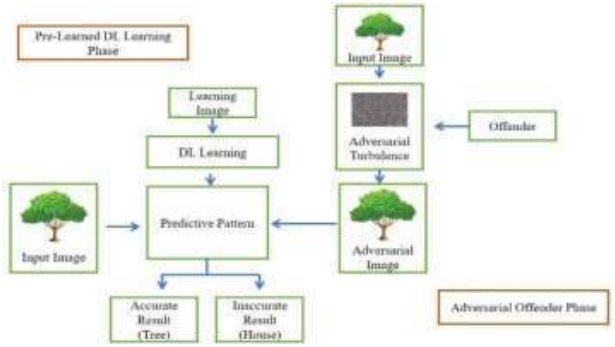

Figure 1. Neural intrusions.

between AI and cybersecurity, stakeholders can navigate the future of AI responsibly and ethically.

2. RELATED WORK

Y. Ding et al. [1] suggested a lightweight encryption approach for diverse-sized medical images to ensure privacy in smart medical technology. It efficiently encrypts multiple images at a cost comparable to single-image encryption, utilizing a secure key sequence generated by SHA-256. Y. Chen et al. [2] approached machine learning models against the Fast Gradient Sign Method (FGSM), focusing on a robust CNN-based image classification architecture.

K. Khullar et al. [3] evaluated the effectiveness of model architectures against adversarial attacks on the CIFAR-10 dataset, using both L_∞ metrics and Wasserstein distance strategies to generate adversarial examples. Pedraza et al. [4] introduced a chaos-theory-based method to discern adversarial examples from regular images in deep learning. Their findings indicate that combining Lyapunov exponents with entropy significantly enhances discriminatory power.

Pandya et al. [6] examined explainability of image classifiers under targeted adversarial attack. Xiao and Zheng [7] explored a defense strategy against adversarial examples in image classification, demonstrating enhanced robustness on CIFAR-10 and Tiny ImageNet datasets. Wang et al. [8] proposed adversarial image detection based on sentiment analysis, while Liu et al. [9] presented a query-free embedding attack against deep learning systems.

Zhang et al. [10] proposed a preprocessing defense framework based on image compression reconstruction. Tian et al. [12] investigated limitations of AI in image classification through an interactive adversarial attack game. Agarwal et al. [13] proposed a non-deep-learning defense using image transforms such as Discrete Wavelet Transform and Discrete Sine Transform. Additional studies addressed adversarial attacks in intrusion detection, transfer learning, face recognition, model reuse, and feature-attribution-based detection [14, 15, 17, 18, 19, 24, 25, 23].

3. PROPOSED METHODOLOGY

Artificial intelligence plays a pivotal role in autonomous systems by learning from data and facilitating decision-making processes. However, it is susceptible to adversarial attacks, which aim to deceive or compromise model performance. These attacks are categorized into white-box attacks, assuming full access to the model, and black-box attacks, relying solely on input-output interactions. Of particular concern

Table 1. Summary of adversarial attack and defense directions discussed in related work.

Direction	Representati\ Idea	Security Role
Image encryption	Deep key generation and secure image streams	Protects sensitive visual data before model use
Adversarial training	Retraining with clean and perturbed examples	Improves model robustness against evasion attacks
Image reconstruction	Compression and super-resolution restoration	Removes or weakens perturbation noise
Transform-based denoising	Wavelet and sine transform filtering	Neutralizes adversarial perturbations without retraining
Detection methods	Logit distribution, entropy, and feature attribution	Identifies suspicious adversarial inputs

are adversarial image black-box attacks, which can produce images indistinguishable from regular ones to humans but cause misinterpretation by deep learning models.

This project develops a methodology to analyze and mitigate such attacks using a pretrained model. First, it studies adversarial image attacks using techniques such as FGSM to generate perturbed images that induce misclassification by deep learning models. Given an input image x, label y, model parameters θ, and loss function J, FGSM can be expressed as:

$$x_{\text{adv}} = x + \varepsilon \operatorname{sign}\left(\nabla_x J(\theta, x, y)\right), \tag{1}$$

where ε controls perturbation strength.

Second, the method uses training with a pretrained model by leveraging the established AlexNet architecture. By retaining lower layers trained on extensive datasets such as ImageNet and fine-tuning upper layers, the model can efficiently adapt to new datasets while optimizing time and resource utilization.

To combat adversarial example attacks, adversarial training emerges as a prevalent defense mechanism. This method retrains the classification network using a blend of clean and adversarial examples, with the aim of fortifying the network's resilience. During training, labels of adversarial examples remain unchanged, preserving their true labels rather than reflecting erroneous predictions. For instance, an adversarial example resembling a cat retains its original "cat" label instead of being mislabeled as "dog" due to network misclassification.

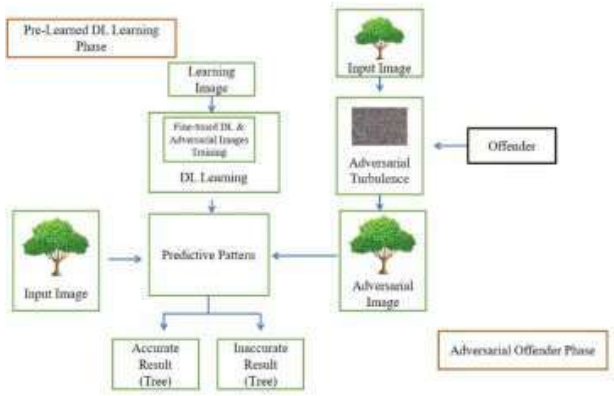

Figure 2. Strategies for resilience against neural intrusions.

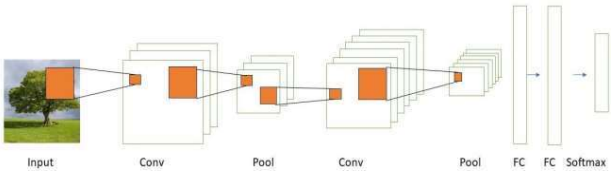

Figure 3. CNN architecture.

4. ALGORITHM DESCRIPTION

Convolutional Neural Networks (CNNs) utilize a series of layers, each designed to identify different aspects of an input image. The network's complexity depends on its specific purpose, with layers ranging from a few to potentially thousands. These layers iteratively enhance their comprehension of the input by incorporating insights from preceding stages, enabling the CNN to discern nuanced patterns and details within the data.

CNNs take cues from connectivity patterns found in the human visual cortex, optimizing their structure for efficient interpretation of visual information. Their artificial neurons are strategically arranged to process entire images effectively, resembling the brain's ability to perceive visuals. Renowned for object recognition, CNNs are extensively deployed in computer vision applications such as image classification and object detection. They find diverse applications across industries, including autonomous driving, facial recognition technology, and medical image analysis.

Convolution involves sliding small filters over the input image, allowing the network to detect patterns and features at different spatial locations. For an input feature map X and kernel K, a convolutional response can be represented as:

$$Y(i,j) = \sum_m \sum_n X(i+m, j+n)K(m,n). \qquad (2)$$

This approach enables CNNs to learn hierarchical representations of the input, progressively extracting more abstract features as information flows through the network.

5. RESULTS AND DISCUSSION

The AlexNet model, renowned for its image recognition capabilities, has been found vulnerable to adversarial image attacks. These attacks, utilizing techniques like FGSM, subtly alter pixels in images, making them imperceptible to humans but misleading to the model, resulting in misclassifications. This investigation delves into the escalating trend of AI model hacking, which poses significant risks across various domains.

Figure 4. Pretrained DL model without attack phase.

Figure 5. Pretrained DL model with adversarial attack phase.

While AI technologies offer numerous benefits, they also entail substantial risks if exploited maliciously. The study examines diverse adversarial techniques, including data poisoning, model inversion, and evasion attacks, elucidating their exploitation of AI system weaknesses. Additionally, it proposes defense mechanisms aimed at fortifying AI systems against such attacks, informed by a thorough understanding of adversarial methodologies and constraints.

By offering effective solutions and guidance, this study aims to inform future research and development in AI security, addressing ethical and responsible challenges in AI deployment. Its contributions are pivotal in navigating the complexities of the AI security landscape and fostering a safer AI ecosystem.

6. CONCLUSION

Artificial intelligence has propelled remarkable advancements in technology, transforming our interactions with the world. However, exploring adversarial attacks on deep learning models reveals vulnerabilities in AI systems, necessitating robust security measures and ethical guidelines. Techniques such as FGSM underscore the importance of proactive defense mechanisms.

We propose a defense framework focusing on adversarial training with pretrained models to counter adversarial image attacks, enhancing AI system resilience against evolving threats. This paper explores universal adversarial example attacks targeting image classification models and aims to uncover the reasons behind their susceptibility. Understanding these vulnerabilities is crucial for devising effective strategies to counter adversarial examples.

Through experiments, the study observed minimal discrepancy in pixel-value distribution between clean and adversarial examples; the primary difference lies in salient image regions. Based on these findings, adversarial training can enhance both the accuracy and resilience of the original classification model. Additionally, training a discriminator to discern adversarial examples shows promise in mitigating attacks, although challenges remain when handling large datasets.

REFERENCES

[1] Y. Ding, F. Tan, Z. Qin, M. Cao, K.-K. R. Choo, and Z. Qin, "Deep Keygen: A Deep Learning-Based Stream Cipher Generator for Medical Image Encryption and Decryption," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4915–4929, 2022.

[2] Y. Chen, M. Zhang, J. Li, and X. Kuang, "Adversarial Attacks and Defenses in Image Classification: A Practical Perspective," in *ICIVC*, Xi'an, China, pp. 424–430, 2022.

[3] S. Bansal, K. Kohli, K. K. Vishwakarma, and K. Gupta, "Graph Algo Visualizer," *Journal of Cognitive Human-Computer Interaction*, vol. 3, no. 2, pp. 36–41, 2022.

[4] A. Pedraza, O. Deniz, and G. Bueno, "Approaching Adversarial Example Classification with Chaos Theory," *Entropy*, vol. 22, no. 11, 1201, 2020.

[5] Y. Wang, L. Xie, X. Liu, J.-L. Yin, and T. Zheng, "Model-Agnostic Adversarial Example Detection Through Logit Distribution Learning," in *IEEE ICIP*, pp. 3617–3621, 2021.

[6] M. A. Pandya, P. Siddalingaswamy, and S. Singh, "Explainability of Image Classifiers for Targeted Adversarial Attack," in *INDICON*, Kochi, India, pp. 1–6, 2022.

[7] C. Xiao and C. Zheng, "One Man's Trash Is Another Man's Treasure: Resisting Adversarial Examples by Adversarial Examples," in *CVPR*, pp. 409–418, 2020.

[8] Y. Wang, T. Li, S. Li, X. Yuan, and W. Ni, "New Adversarial Image Detection Based on Sentiment Analysis," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[9] Y. Liu, W. Zhang, and N. Yu, "Query-Free Embedding Attack Against Deep Learning," in *IEEE ICME*, pp. 380–386, 2020.

[10] Y. A. Zhang, H. Xu, C. Pei, and G. Yang, "Adversarial Example Defense Based on Image Reconstruction," *PeerJ Computer Science*, e811, 2021.

[11] R. Venkatesan, M. Sumithra, B. Buvaneswari, and R. Selvalingeshwaran, "Food Ordering Systems' Newness," *Journal of Cognitive Human-Computer Interaction*, vol. 4, no. 1, pp. 15–20, 2022.

[12] X. Tian, X. Tian, and B. Pan, "Similarity Attack: An Adversarial Attack Game for Image Classification Based on Deep Learning," pp. 1467–1478, 2023.

[13] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Image Transformation-Based Defense Against Adversarial Perturbation on Deep Learning Models," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2106–2121, 2021.

[14] J. Ji, B. Zhong, and K.-K. Ma, "Multi-Scale Defense of Adversarial Images," in *IEEE ICIP*, pp. 4070–4074, 2019.

[15] C.-Y. Lin, F.-J. Chen, H.-F. Ng, and W.-Y. Lin, "Invisible Adversarial Attacks on Deep Learning-Based Face Recognition Models," *IEEE Access*, vol. 11, pp. 51567–51577, 2023.

[16] M. Ashok Kumar, M. Sumithra, B. Buvaneswari, and R. Selvalingeshwaran, "Online Cafeteria E," *Journal of Cognitive Human-Computer Interaction*, vol. 4, no. 1, pp. 21–28, 2022.

[17] D. Vyas and V. V. Kapadia, "Evaluation of Adversarial Attacks and Detection on Transfer Learning Model," in *ICCMC*, pp. 1116–1124, 2023.

[18] S. Niu, Y. Liu, J. Wang, and H. Song, "A Decade Survey of Transfer Learning (2010–2020)," *IEEE Transactions on Artificial Intelligence*, 2021.

[19] S. Rezaei and X. Liu, "A Target-Agnostic Attack on Deep Models: Exploiting Security Vulnerabilities of Transfer Learning," arXiv:1904.04334, 2019.

[20] Ajith R. and M. Beullah, "Automated System for Management of Exam Cell," *Journal of Cognitive Human-Computer Interaction*, vol. 4, no. 1, pp. 29–38, 2022.

[21] Meenakshi and G. Maragatham, "A Review on Security Attacks and Protective Strategies of Machine Learning," in *Emerging Current Trends in Computing and Expert Technology*, pp. 1076–1087, 2019.

[22] B. Pal and S. Tople, "To Transfer or Not to Transfer: Misclassification Attacks Against Transfer Learned Text Classifiers," arXiv:2001.02438, 2020.

[23] Z. Yan, Y. Guo, and C. Zhang, "Deep Defense: Training DNNs with Improved Adversarial Robustness," arXiv:1803.00404, 2018.

[24] P. Yang, J. Chen, C. J. Hsieh, J. L. Wang, and M. Jordan, "ML-LOO: Detecting Adversarial Examples with Feature Attribution," *AAAI*, vol. 34, no. 4, pp. 6639–6647, 2020.

[25] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," in *ICML*, pp. 274–283, 2018.

[26] Y. Ji, X. Zhang, S. Ji, X. Luo, and T. Wang, "Model-Reuse Attacks on Deep Learning Systems," in *ACM CCS*, pp. 349–363, 2018.