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

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. The potential repercussions of such misclassification are severe, with the risk of causing traffic accidents and posing a notable safety threat. 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; AI; Smart Systems; Complex Problems

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 come significant vulnerability: the susceptibility to adversarial attacks. These attacks, which exploit weakness in AI systems, aim to manipulate decision-making processes, potentially leading to incorrect predictionsor 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, development of defensive strategies, and consideration of real-world impacts. By fostering adeeper understanding of the dynamics between AI and cybersecurity, stakeholders can empower themselves to navigate the future of AI responsibly and ethically. In summary, the intersection of artificial intelligence and cybersecurityraises complex questions about the security and robustness of AI systems. By acknowledging these challenges and investing in proactive measures, we can strive to develop AI technologies that are secure, resilient, and aligned with ethical principles.

Doi : https://doi.org/10.54216/JCHCI.070105
Received: August 21, 2023 Revised: October 14, 2023 Accepted: January 27, 2024

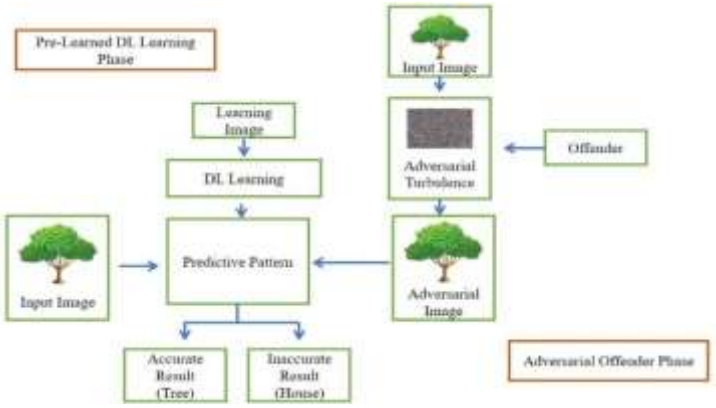

Figure 1: Neural Intrusions

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, utilizing a secure key sequence generated by SHA-256. Y. Chen et al [2], approached machine learning models against fast gradient search method (FGSM). The focusis on designing a robust architecture for CNN-based image classification models to counter such attacks.

K. Khullar et al [3], evaluated the effectiveness of two cutting-edge model architectures against adversarial attacks on the CIFAR- 10 dataset. Using both L∞ metric and Wasserstein distance strategies to generate adversarial examples, the study applies defenses such as input preprocessing and adversarial training. D. Pedraza A et al [4], introduced a chaos theory-based method to discern adversarial examples from regular images in deep learning. Experimental findings indicate the limited robustness of Lyapunov exponents, while combining them with entropy significantly enhances discriminatory power. M.A. Pandya et al [5], proposed that Deep neural networks excel in computer vision but present challenges in interpretability. The author employs an iterative targeted attack on base architectures with image classes, evaluating the impact on explanation algorithms pre- and post-attack. C. Xiao et al [6], explored an innovative defense strategy against adversarial examples in image classification, focusing on deep neural network systems. The technique demonstrates enhanced robustness on CIFAR-10and Tiny ImageNet datasets compared to existing methods, offering lower training costs than conventional adversarial training. Y. Wang et al [7], aimed to improve face recognition (FR) performance by incorporating Gabor face representations into DCNN-based systems and proposed the & quot Gabor DCNN (GDCNN)& quot; ensemble method, strategically using various Gabor face representations during both training and testing. Y. Liu et al. [8], presented a novel algorithm designed to enhance the security of widely- used deep models by addressing vulnerabilities to adversarial attacks. The algorithm autonomously locates target objects, utilizes semantic segmentation to generate a mask matrix, and effectively deceives deep detection models. Zhang YA et al [9], proposed a preprocessing defense framework utilizing image compression reconstruction. The framework, based on adversarial sensitivity, compresses pixel depth to eliminate perturbations and employs super-resolution image reconstruction for quality restoration. Xuejun Tian et al [10], investigated the limitations of AI in image classification through an exploration of deep learning drawbacks and presented an interactive game where users influence deep neural networks to misclassify images, allowing them to experience the technology, leading the model to inaccurately classify other categories. A. Agarwal et al.[11], proposed a non-deep learning approach using image transforms like Discrete Wavelet Transform and Discrete Sine Transform. The algorithm neutralizes adversarial perturbation impact through a wavelet decomposition-based denoising filtering method, emphasizing effectiveness across diverse perturbation methods on multiple image databases. D. Vyas et al. [12], demonstrated a study that looks into how cyber attackers can trick Deep Learning-based Network Intrusion Detection Systems (NIDS) using techniques like FGSM, JSMA, PGD, and C&W. The research evaluated these methods on the CICIDS-2017 dataset in three stages: before an attack, after an attack, and after using a defense technique. S. Niu et al. [13], comprehensively explores TL, shedding light on its current advancements, emerging trends, applications, and the ongoing challenges within diverse domains such as image processing, speech recognition, and natural language processing. S. Rezaei et al. [14], introduced a security vulnerability as the models are publicly accessible, enabling attackers to launch efficient brute force attacks without additional target-specific information. The author demonstrated a target-agnostic attack, rendering previous defenses impractical, and highlighted a fundamental security challenge in deep neural network transfer learning. A. Abdelkader et

Doi : https://doi.org/10.54216/JCHCI.070105
Received: August 21, 2023 Revised: October 14, 2023 Accepted: January 27, 2024

al. [15], incorporated an adversarial perturbation-based regularizes into the classification objective. By integrating resistance to attacks during training, the method outperforms existing approaches on diverse datasets and neural network architectures.

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 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 aims to develop a methodology to analyze and mitigate such attacks using a pretrained model. The approach entails several steps. Firstly, it delves into Adversarial Image Attacks, employing techniques like the Fast Gradient Sign Method (FGSM) to generate perturbed images inducing misclassification by deep learning models. Secondly, it employs Training Using a Pretrained Model, leveraging the established AlexNet architecture. By retaining the lower layers trained on extensive datasets like ImageNet and fine-tuning the upper layers, the model can efficiently adapt to new datasets, thus optimizing time and resource utilization. To combat adversarial example attacks, adversarial training emerges as a prevalent defense mechanism. This method involves retraining the classification network using a blend of clean and adversarial examples, with the aim of fortifying the network's resilience. In this study, we employ adversarial training to assess the efficacy of the retrained classification network in thwarting adversarial attacks. Notably, throughout the training process, the labels of adversarial examples in the dataset remain unchanged, preserving their true labels rather than reflecting any erroneous predictions made by the classification network. For instance, an adversarial example resembling a cat retains its original "cat" label, rather than being mislabeled as "dog" due to the network's misclassification. This strategy is designed to enhance the robustness of the classification network, ensuring accurate predictions—a crucial aspect for real-world applications.

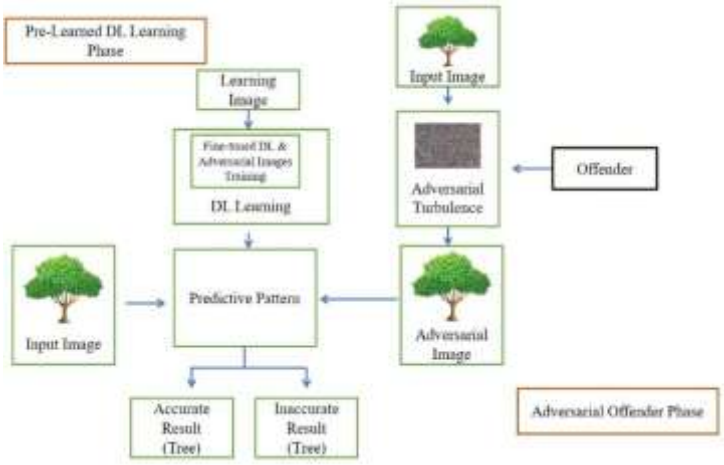

Figure 2: Strategies for Resilience Against Neural Intrusions

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 the 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 their prowess in 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, leveraging their ability to accurately analyze and interpret visual data. Convolution involves sliding small filters over the input image, allowing the network to detect patterns and features at different spatial locations. This approach enables CNNs to learn hierarchical representations of the input, progressively extracting more abstract features as information flows through the network. As a result, CNNs excel at tasks such as image

Doi : https://doi.org/10.54216/JCHCI.070105
Received: August 21, 2023 Revised: October 14, 2023 Accepted: January 27, 2024

classification, object detection, and image segmentation, thanks to their ability to capture intricate patterns and structures within visual data.

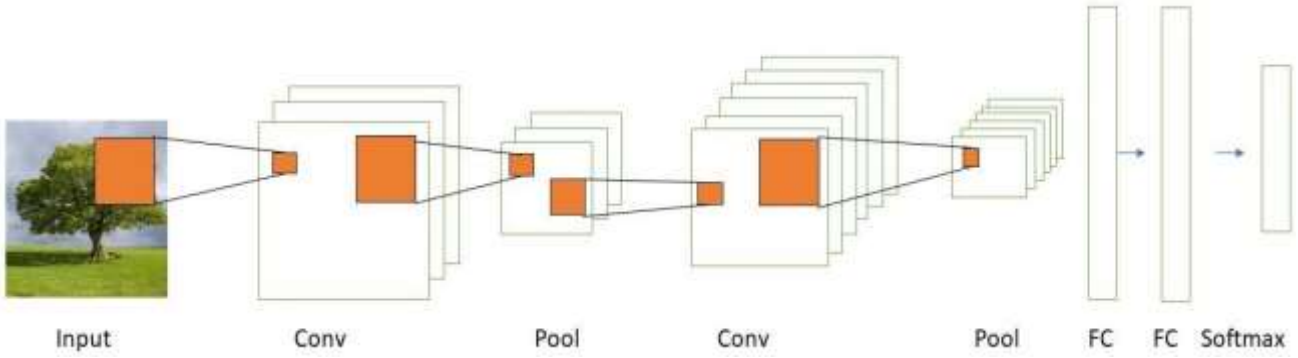

Figure 3: CNN Architecture

5. Results and Discussions:

The Alex Net model, renowned for its image recognition capabilities, has been found vulnerable to Adversarial Image Attacks. These attacks, utilizing techniques like the Fast Gradient Sign Method (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. 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 innovative 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, fostering a safer AI ecosystem.

Figure 4: Pretrained DL Model Without Attack Phase

Doi : https://doi.org/10.54216/JCHCI.070105
Received: August 21, 2023 Revised: October 14, 2023 Accepted: January 27, 2024

Figure 5: Pretrained DL Model With Adversarial Attack Phase

6. Conclusion

Artificial intelligence (AI) has propelled remarkable advancements in technology, transforming our interactions with the world. However, it brings challenges. Exploring adversarial attacks on deep learning models reveals vulnerabilities in AI systems, necessitating robust security measures and ethical guidelines. Techniques like the Fast Gradient Sign Method (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 aims to explore universal adversarial example attacks targeting image classification models, aiming to uncover the reasons behind the susceptibility of these models to such attacks. Understanding these vulnerabilities is crucial for devising effective strategies to counter adversarial examples. Through a series of experiments, we investigated this phenomenon. Our evaluations revealed that while there is minimal discrepancy in the pixel value distribution between clean and adversarial examples, the primary difference lies in the salient regions of the images. Based on our findings, we deduce that adversarial training can enhance both the accuracy and resilience of the original classification model. Additionally, we observed that training a discriminator to discern adversarial examples shows promise in mitigating such attacks, albeit with challenges, particularly 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," in IEEE Transactions onNeural Networks and Learning Systems, vol. 33, no. 9, pp. 4915-4929, doi: 10.1109/TNNLS.2021.3062754, Sept. 2022.

[2] Y. Chen, M. Zhang, J. Li and X. Kuang, "Adversarial Attacks and Defenses in ImageClassification: A Practical Perspective," 2022 7th International Conference on Image, Vision andComputing (ICIVC), Xi'an, China, pp. 424-430,doi: 10.1109/ICIVC55077.2022.9886997,2022.

[3] Sweeta Bansal,Karan Kohli,K. K. Vishwakarma,Kush Gupta. "Graph Algo Visualizer." Journal of Cognitive Human-Computer Interaction, Vol. 3, No. 2, 2022 ,PP. 36-41.

[4]. Pedraza A, Deniz O, Bueno G. Approaching Adversarial Example Classification withChaos Theory. Entropy (Basel). Oct 24;22(11):1201. doi: 10.3390/e22111201. PMID:33286969; PMCID: PMC7712112,2020.

[5]. Y. Wang, L. Xie, X. Liu, J. -L. Yin and T. Zheng, "Model-Agnostic Adversarial Example Detection Through Logit Distribution Learning," 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, pp. 3617-3621, doi: 10.1109/ICIP42928.2021.9506292,2021.

Doi : https://doi.org/10.54216/JCHCI.070105
Received: August 21, 2023 Revised: October 14, 2023 Accepted: January 27, 2024

[6]. M. A. Pandya, P. Siddalingaswamy and S. Singh, "Explainability of Image Classifiers for Targeted Adversarial Attack," 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India,pp.1-6,doi: 10.1109/INDICON56171.2022.10039871,2022.

[7]. C. Xiao and C. Zheng, "One Man's Trash Is Another Man's Treasure: Resisting Adversarial Examples by Adversarial Examples," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Seattle, WA, USA, 2020pp.409-418. doi: 10.1109/CVPR42600.2020.00049,2020.

[8].Y. Wang, T. Li, S. Li, X. Yuan and W. Ni,"New Adversarial Image Detection Based on Sentiment Analysis," in IEEE Transactions on NeuralNetworks and Learning Systems, doi: 10.1109/TNNLS.2023.3274538.

[9].Y. Liu, W. Zhang and N. Yu, "Query-Free Embedding Attack Against Deep Learning," 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, pp. 380-386, doi: 10.1109/ICME.2019.00073,2020.

[10]. Zhang YA, Xu H, Pei C, Yang G.Adversarial example defense based on image reconstruction. PeerJ Comput Sci. Dec24;7: e811. doi: 10.7717/peerj-cs.811. PMID: 35036533; PMCID: PMC8725667,2021.

[11]. R. Venkatesan,M.Sumithra,B. Buvaneswari,R. Selvalingeshwaran. (2022). Food Ordering Systems' Newness. Journal of Cognitive Human-Computer Interaction, 4 (1), 15-20.

[11].Tian, Xuejun, Tian, Xinyuan, and Pan, Bingqin. 'Similarity Attack: An Adversarial Attack Game for Image Classification Based on Deep Learning'. 1 Jan. 2023 : 1467 – 1478

[12]. A.Agarwal, R. Singh, M. Vatsa and N. Ratha, "Image Transformation-Based DefenseAgainst Adversarial Perturbation on Deep Learning Models," in IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 5, pp. 2106-2121, 1 Sept.-Oct. 2021, doi: 10.1109/TDSC.2020.3027183

[13]. J. Ji, B. Zhong and K. -K. Ma, "Multi-ScaleDefense of Adversarial Images," 2019 IEEEInternational Conference on Image Processing (ICIP),Taipei, Taiwan, 2019, pp. 4070-4074, doi: 10.1109/ICIP.2019.8803408.

[14]. C. -Y. Lin, F. -J. Chen, H. -F. Ng and W. -Y. Lin, "Invisible Adversarial Attacks on Deep Learning-Based Face Recognition Models," in IEEE Access, vol. 11, pp. 51567-51577, 2023, doi: 10.1109/ACCESS.2023.3279488.

[15] M. Ashok Kumar,M.Sumithra,B.Buvaneswari,R. Selvalingeshwaran. "Online Cafeteria E." Journal of Cognitive Human-Computer Interaction, Vol. 4, No. 1, 2022 ,PP. 21-28.

[16].D. Vyas and V. V. Kapadia, "Evaluation ofAdversarial Attacks and Detection on Transfer Learning Model," 2023 7th International Conference on Computing Methodologies and Communication(ICCMC), Erode, India, 2023, pp. 1116-1124, doi: 10.1109/ICCMC56507.2023.10084164.

[17].S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010-2020)," IEEE Transactions on Artificial Intelligence, 2021.

[18]. S. Rezaei and X. Liu, "A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning," arXiv preprint arXiv:1904.04334, 2019.

[19]. Ajith R. ,Mercy Beullah. "Automated System for Management of Exam Cell." Journal of Cognitive Human-Computer Interaction, Vol. 4, No. 1, 2022 ,PP. 29-38.

Doi : https://doi.org/10.54216/JCHCI.070105
Received: August 21, 2023 Revised: October 14, 2023 Accepted: January 27, 2024

[20].Meenakshi and G. Maragatham, “A reviewon security attacks and protective strategies of machine learning,” in Int. Conf. on Emerging CurrentTrends in Computing and Expert Technology, Cham, Springer, pp. 1076–1087, 2019.

[21].B.Pal and S. Tople, “To transfer or not to transfer: Misclassification attacks against transfer learned text classifiers,” arXiv preprint arXiv:2001.02438, 2020.

[22].Z. Yan, Y. Guo and C. Zhang, “Deep defe nse: Training DNNs with improved adversarial robustness,” arXiv preprint arXiv: 1803.00404, 2018.

[23]. P. Yang, J. Chen, C. J. Hsieh, J. L. Wang and M. Jordan, “Ml-loo: Detecting adversarial exampl es with feature attribution,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 4, pp. 6639–6647, 2020.

[24].Athalye, N. Carlini and D. Wagner, “Obfu scated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in Int. Conf. on Machine Learning, PMLR, Stockholm, Sweden,pp. 274–283, 2018.

[25].Y. Ji, X. Zhang, S. Ji, X. Luo, and T. Wang,“Model-reuse attacks on deep learning systems,” in Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018, pp.349– 363.

Doi : https://doi.org/10.54216/JCHCI.070105
Received: August 21, 2023 Revised: October 14, 2023 Accepted: January 27, 2024