iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IX **Month of publication:** September 2023

DOI: https://doi.org/10.22214/ijraset.2023.55716

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

Threats of Adversarial Attacks on Deep Learning

Sajja Regmi[1], Saksham Regmi[2]
[1, 2] *Department of Science, Kathmandu Model Secondary School, Kathmandu, Nepal*

Abstract: ***AI, accompanied with Deep Learning, is a growing technological innovation which has an extensible reach to various applications. The complex computational models i.e. Deep Neural Networks (DNNs) excel in natural language processing, image recognition, computer vision, and autonomous systems. Despite the merits, these intricate networks are vulnerable to adversarial attacks (such as black-box and white-box) which can be the primary threats to the AI-DL integration. For their mitigation, adversarial training, defense distillations and other defenses are introduced. This paper deals with the impediments posed by adversarial attacks on Deep Learning, and lists out the possible solutions to reduce their impacts.***
Keywords: Deep Learning, Deep Neural Networks, Adversarial Attacks, Adversarial Training, Defensive Distillation, Natural Language Processing, Image Recognition, Autonomous

I. INTRODUCTION

A. Artificial Intelligence (AI)

Artificial Intelligence, as defined by John McCarthy [1], is an innovation created with the goal of replicating human intelligence in machines including their comprehension, decisiveness and actions. Originated since the 1950s, AI imitates human cognitive phenomena like analytical and rational thinking [2]. It enlists various learning algorithms that create models from data for prediction and classification. Infact, core AI skills include reasoning, problem-solving, learning, perception and self correction. AI offers benefits like improved accuracy, swift decision-making, versatility, diligence and integration in fields like autonomous vehicles and facial recognition. In cybersecurity, AI excels in detecting malware beyond human capabilities [2]. It counts building networks of artificial neurons to emulate human neural processes. These neural networks extract patterns from the given data and serve human-like decisions. For instance, in 1997, Deep Blue, an IBM's program, defeated chess grandmaster Garry Kasparov [3] which was considered to be its milestone. AI's evolution merges human cognition and artificial neural networks, progressing towards human-like computational reasoning. This trajectory aligns with AI's essence of replicating human intelligence. However, besides all the advantages, its challenges include high setup, maintenance and upgrade cost along with limitations in creative tasks.

B. Deep Learning (DL)

Deep learning is a transformative subset of AI that harnesses neural networks for pattern recognition, image identification, speech processing, and natural language understanding. Tracing back to the 1970s, deep learning first gained prominence with its 2012 ImageNet success, almost halving error rates in object identification of natural images [4]. Also, the victory of Google's AlphaGo program against the world champion Lee Sedol in 2016 backs up the excellence of deep learning in pattern recognition [3]. Besides, it also finds applications in biometric security, facial recognition, and more.

Deep neural networks (DNNs) highlight deep learning's potential in diverse contexts. They enable systems with billions of parameters to predict accurate results- advancing computer vision and speech recognition. From Alexa to medical imaging, DNNs enhance accuracy and predictions beyond human capabilities, fueling AI progress. As deep learning reshapes AI, its potential to revolutionize industries and other firms becomes transparent.

C. Growing Intersection of AI and DL

The convergence of artificial intelligence (AI) and deep learning (DL) has been explicitly expanding in today's world. AI employs neural networks to enhance accuracy in analyzing large datasets through DL models. The classical AI deals with the symbol grounding problem [5], manually crafting representational elements, whereas DL excels in autonomously extracting intricate features from high-dimensional data with minimal human intervention. Interestingly, in contrast, DL's limitations also align with symbolic AI's strengths, suggesting an amicable integration of the two approaches. Symbolic representations offer reusable, data-efficient structures with high-level abstraction for better generalization, and their language-like feature aids human comprehension. Researchers are actively merging symbolic AI principles with DL, promising further progress. AI, driven by deep learning, efficiently analyzes extensive datasets via hidden neural layers, achieving heightened accuracy.

This accuracy improves with increased training data. The real-world instances like Siri and Google Photos showcase deep learning's accuracy enhancement. Additionally, AI-powered deep learning benefits medicine as well, potentially improving disease detection using MRI images. Hence, the union of AI and deep learning marks an era of transformative innovation across domains. [5]

D. Deep Neural Network

Deep neural networks (DNNs) are complex computational models that form a fundamental component of deep learning. DNNs have a multi-layer structure, which is why they're called "deep." These layers refine the way they understand information as data moves through them. It's a bit like training a computer to recognize different parts of an image one step at a time. For instance, the first layer may recognize edges and dots; the second may identify the iris or the tip of a nose; the third may distinguish the whole eyes and nose and subsequently, the layers proceed till it recognizes the whole face. The connections between these layers are similar to the pathways in a computer's circuitry. As the network processes more data, it gets better at understanding and recognizing patterns [6]. Deep neural networks (DNNs) are now standard in applications like image classification, object detection, and natural language processing. They excel in speech recognition and natural language understanding, and are close to achieving human-level accuracy in image recognition.

II. THREATS ON NEURAL NETWORKS

While DNNs excel in various tasks, recent criticism has highlighted shortcomings like data inefficiency, poor generalization, and lack of interpretability. Their black-box (an adversarial attack) nature and reliance on large datasets raise questions about their practical applicability, which further doubts their real world viability[5]. Deep neural networks (DNNs) are highly liable to adversarial attacks [7]. Szegedy et al. first noticed adversarial examples in image classification- capable of altering image classifications through minor transformations, often undetectably small [8]. According to Goodfellow et al., the design of modern DNNs forces them to behave linearly in high dimensional spaces which makes them more vulnerable to adversarial attacks [9]. Consequently, such attacks restrict the potential domains for neural network usage, as seen with their application in self-driving cars, where adversaries could manipulate the vehicle's actions. This vulnerability also raises concerns for applications of medical diagnosis and beyond. Moreover, DNN vulnerabilities extend to deep learning platforms like graph neural networks (GNNs). Attacks on GNNs involve corrupting graph topology where the attacker carefully selects a small number of edges and manipulates them through perturbation and rewiring, resulting in incorrect predictions [10]. These concerns hold significance in domains related to public trust, human decision making, and human well-being. Even if the accumulation of neighboring nodes' information is a powerful principle of representation learning, the process of GNNs exchanging the information along the nodes makes them more prone to adversarial attacks [11].

A. Adversarial Attacks

Adversarial examples- specialized forms of noise- were first discovered by Szegedy et al in the image classification domain. This noise is meticulously crafted to mislead classifiers without human detection, effectively altering classifications to any desired class. This potency is amplified by transferability, where adversarial images designed for one network can deceive others with different architectures or training sets [7]. The existence of adversarial examples highlights a lack of robustness in deep learning. The minor input modifications can effectively mislead even state-of-the-art deep neural networks into making incorrect predictions [7],[9]. The real-world impact of adversarial examples is broad; for instance, they could manipulate automatic indexing to change document categories and bypass spam filters in text analysis [12].

Adversarial attacks are the malicious attacks that trick machine learning programmes by deceitful inputs i.e. adversarial examples. These attacks exploit model vulnerabilities, manipulating inputs which are unperceivable by humans, causing models to confidently make incorrect predictions. Adversarial attacks, based on the threat model, are classified as white-box, gray-box, and black-box. Some of the techniques like the Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), DeepFool, and Projected Gradient Descent (PGD) are often employed for such attacks [13].

1) *White Box:* In adversarial contexts, white-box attacks assume adversaries comprehend the target model entirely, granting them full access to its intricacies like its architecture, parameters, design, training process and potentially its training data [14]. The full access to target model parameters facilitates the creation of precisely targeted adversarial samples. For instance, using knowledge about a Deep Neural Network's (DNN) architecture and parameters, adversaries can subtly manipulate input images, causing the DNN classifier to confidently misclassify the changed input [13], despite the changes being almost imperceptible to humans.

2) *Black Box:* Black-box attacks exploit transferable adversarial examples, posing threats without specific target model knowledge such as its parameters and architectures. They focus on transferability, allowing adversarial examples from local surrogate models to target models with no direct access [15]. Real-world situations often involve adversaries with black-box access to models having no direct but query access to predictions as seen in commercial AI offerings like IBM and Google [15]. The adversaries may send the input samples to the model and receive the model's predictions in response, including class probabilities [15]. While training data enables attacking surrogate models, using target model outputs for gradient estimation is another approach to formulate adversarial examples [15].

B. *Effects of Adversarial Attacks*

The real-life threats posed by adversarial attacks on neural networks are apparently increasing.

1) In "One Pixel Attack for Fooling Deep Neural Networks", Su et al. described that a single pixel's change can deceive the classifier in One Pixel Attack [16].
2) Szegedy et al.'s "Box-Constrained L-BFGS" has demonstrated that small perturbations could trick deep learning models, giving rise to adversarial images- intentionally perturbed to trick machine learning techniques [7].
3) Carlini and Wagner introduced "CARLINI and WAGNER ATTACKS (C&W)" to render perturbations which are quasi-imperceptible, leading to defensive distillation failures [17].
4) "Universal Adversarial Perturbations' introduced by Moosavi Dezfooli et al. can deceive the network on 'any' image with high probability, which is almost imperceptible to human vision [18].
5) "Houdini" by Cisse et al. tricks the gradient-based learning machines by designing adversarial examples which are specifically customized to task losses [19].
6) Nguyen et al.'s work was an initiation of fooling state-of-the-art neural networks misclassifying images to specific classes with high confidence, unrecognized by humans [20].
7) As studied by Jiajun et al, adversarial attacks extend beyond pixel perturbations transitioning them to the physical world- such as affecting road signs [21]. In one experiment, the approach succeeded in causing the meaning of a 'STOP' roadside sign to be transformed into a '30mph' speed limit sign as shown in Fig. 1 [22].
8) In 2014, few researchers from google and NYU showed that an image of a panda can be subtly altered to be misclassified as a gibbon due to an optical illusion as illustrated in Fig. 2 [9]. However, for the neural network, it showcased the potential consequences of adversarial attacks in the AI landscape. The implications of these threats underscore the need for safeguards, as trivial perturbations can lead to unforeseen hazardous outcomes.

Fig. 1. Perturbations on a sign, created by shining crafted light on it, distorts how it is interpreted in a machine learning system [22].

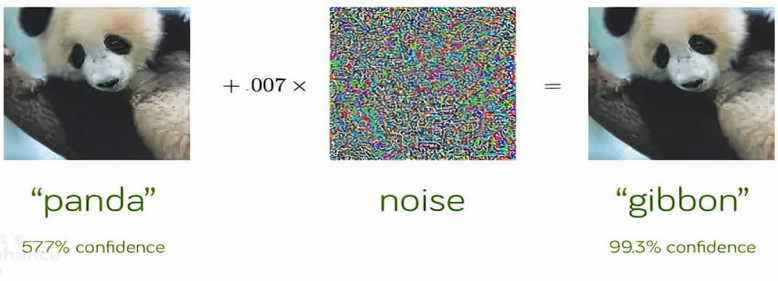


Fig. 2. An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon [9].

a) *Natural Language Processing*: In Natural Language Processing (NLP), deep learning models like Recurrent Neural Networks (RNNs) and transformers enable more accurate language models, enhancing human-machine communication, automating customer service, and advancing search engines. NLP empowers machines to comprehend language data. It also teaches computers to analyze text- as seen in tools like Google Translate, Google Assistant, Microsoft's Cortana and Amazon Alexa. Besides, in finance, NLP is used in Prudential's chat bot, Bank of America's Erica and more. In enterprise domains, it is used for spam, malware, intrusion detection, and beyond [23]. However, in smart speakers applications, an adversary exploits NLP's vulnerabilities by using slightly altered inputs- via voice squatting- to trigger false and malicious device activation [24]. Also, intrusion detection systems can misclassify malware; and NLP's role in critical decisions- such as finance- can be manipulated. Real-world deployment has laid out NLP's lack of robustness- as seen in Amazon's gender-biased recruitment tool [25]. Additionally, according to some German researchers at Ruhr University, the orders- imperceptible to the human ear- can be inserted into audio files for hacking voice assistants [26].

b) *Image Recognition and Computer Vision:* Image recognition, a prime aspect of deep learning, trains models to identify objects in images. This field also exposes machine learning's vulnerability to adversarial attacks extending to adversarial photographs challenging model robustness, even via a camera. For instance, in convolutional neural networks trained for image classification, slight alterations to input images can lead to misclassification of adversarial images, while these two images remain nearly imperceptible to observers [7]. Given image classification's significance in various domains, countering adversarial attacks is vital for system reliability and resilience. In Computer Vision, deep learning gained prominence in 2012 with the groundbreaking performance of Krizhevsky et al. 's Convolutional Neural Network (CNN) model- rooted on a large-scale visual recognition task. But, a new landmark was set in another 2012 competition when Hinton's team won with a 16% error rate, surpassing the prior standard of 25%. This marked the rise of deep learning- especially in 2017 leading to <5% errors in the competition, corresponding to human performance [27]. Practical applications like facial recognition ATMs and mobile phone Face ID security highlight the impending impact of deep learning in Computer Vision on daily life. While deep learning significantly advanced Computer Vision, it is still susceptible to adversarial attacks- undermining its performance. Adversarial attacks have showcased the potential to severely impair the performance of deep learning techniques across multiple computer vision tasks [27].

c) *Autonomous Systems:* Deep Learning's success in various fields like robot navigation and reinforcement learning highlights its potential for autonomous systems. Drones and self-driving cars offer transportation efficiency and safety improvements. The integration of AI-based systems, exemplified by Tesla's driverless cars, emphasizes the significance of AI and DL in enhancing driving performance through features like braking, lane changing, and obstacle navigation, powered by computer vision, image recognition and DL [2]. Deep Neural Networks (DNNs) have made significant progress, especially in vision-based autonomous driving and utilizing cost-effective cameras. However, building reliable autonomous systems capable of making correct decisions under diverse adversarial conditions remains challenging due to hardware uncertainties including camera distortion and noises. Despite their success, DNN-based vision models are vulnerable to minor image corruptions, lighting shifts, and data distribution changes [28]. These concerns have prompted adversarial attacks, revealing model vulnerabilities. The attacks such as FGSM attack and PGD attack tend to mislead the model to misclassify the object orientation of the vehicle or even could incorrectly perceive class 'Car' as class 'Background' potentially leading to an accident [29] .

III. DEFENSES

The vulnerability of deep neural networks to adversarial perturbations has prompted extreme research into both attack methods and defense strategies. The convergence of deep learning and adversarial attacks has prompted systematic investigations to counter these issues:

A. *Adversarial Training (AT)*

Adversarial Training is a prominent defense strategy, known for its empirical robustness, scalability across deep networks, and effectiveness against various threat models. During adversarial training, mini arrays of training samples are introduced to adversarial perturbations, then used to upgrade network architectures and parameters till resistance against such attacks is learnt by the model. This method fosters models to acquire robust features, enhance generative abilities, and improve performance on clean data [30]. In 2015, Goodfellow et al proposed a simple and fast adversarial training method which involved training models on both clean and adversarial images- challenging deep neural networks with deliberately perturbed images [9].

Brute-force adversarial training was first seen to regularize the network to lessen over-fitting, eventually improving robustness of the networks against the adversarial attacks[31]. As inspired by that, Miyato et al. introduced a *'Virtual Adversarial Training'* method in 2016 to even out the distributions of the neural networks' outputs [32]. In 2017, under NLP's context, he applied virtual adversarial training to text classification tasks by perturbing word vectors of input sentences [33]. Later, Zheng et al. proposed a similar *'Stability Training'* approach to enhance the robustness of neural networks against minute distortions to input images [34]. In 2018, Sato et al., pivoted on enhancing the interpretability of adversarial examples in the context of NLP [35]; Wang et al. studied AT's effects on different sets of variables in MRC tasks [36]; and a recent work in 2019 by Sato et al. investigates the effects of AT on neural machine translation [37].

Despite its benefits, adversarial training can reduce accuracy on larger datasets. Also, it may not fully counter iterative attacks; and robustness might rely on the range of adversarial training examples. While a potent defense, adversarial training's efficacy is still being refined. However, combining adversarial training with defensive distillation can offer potential solutions.

B. Defensive Distillation

Defensive distillation, a method proposed by Papernot et al., has emerged as a robust defense strategy against adversarial attacks, aiming to harden neural networks by reducing their vulnerability to input perturbations. One of its remarkable features is its compatibility with various feed-forward neural networks (artificial neural networks where information flows in forward direction only), requiring only a single re-training step and offering strong security guarantees against adversarial examples [38]. Defensive distillation stands effective against known attacks such as the Fast Gradient Sign Method (FGSM) and the Jacobian-based Saliency Map Approach (JSMA) [39]. It achieves this by destroying gradients essential to attack heuristics, enhancing robustness in the proximity of test data without significantly compromising accuracy [40]. The prospect of combining adversarial training with defensive distillation to establish a mapping between adversarial examples, further showcases its compliance and potential for enhancing security. However, a limitation of defensive distillation could be its applicability restriction to only DNN models which produce an energy based probability distribution [38].

C. Other Defenses

The pursuit of enhancing defense strategies encompasses three primary avenues: 1) usage of modified training or testing procedures, 2) network modifications involving increased layers or alternate loss/activation functions, and 3) integration of external models as add-ons during classification of new examples [31]. Certifiably robust defense methods like random word masking have been proposed to counter word substitution and character-level attacks [41]. GNN GUARD serves as a protective solution dedicated to enhancing the resilience of Graph Neural Networks (GNNs) against poisoning attacks [42]. Gradient masking, a novel approach that nullifies gradients central to attack heuristics rather than solely minimizing model error, has emerged as another promising defense strategy [40]. Metzen et al. introduced a subnetwork augmentation named "Perturbation detection method" that detects adversarial perturbations by connecting to each layer of the primary deep neural network [43]. Although successful to a certain extent, this subnetwork remains vulnerable to adversarial examples generated by unseen attack methods. Researchers prioritize NLP robustness through pre-deployment testing, such as Rebeiro et al.'s CheckList method [44], which uncovers vulnerabilities even after internal testing.

On a different note, SafetyNet proposed by Lu et al. involves a Support Vector Machine to discern between natural and perturbed data using quantified features from a DNN. While potentially providing more robust results, SafetyNet still grapples with the task of handling unforeseen adversarial patterns [45]. Data augmentation methods, as demonstrated by Goodfellow et al. [9], maintain robustness without accuracy trade-offs. Li et al. in 2020 use backtranslation and self-learning to generate augmented training data [46]. Lu et al. in 2020 formally propose Counterfactual Data Augmentation (CDA) for gender bias mitigation, which involves causal interventions that break associations between gendered and gender-neutral words [47]. These advancements highlight the continuous evolution and intricate nature of strategies in mitigating adversarial attacks and deploying effective defense mechanisms.

IV. CONCLUSION

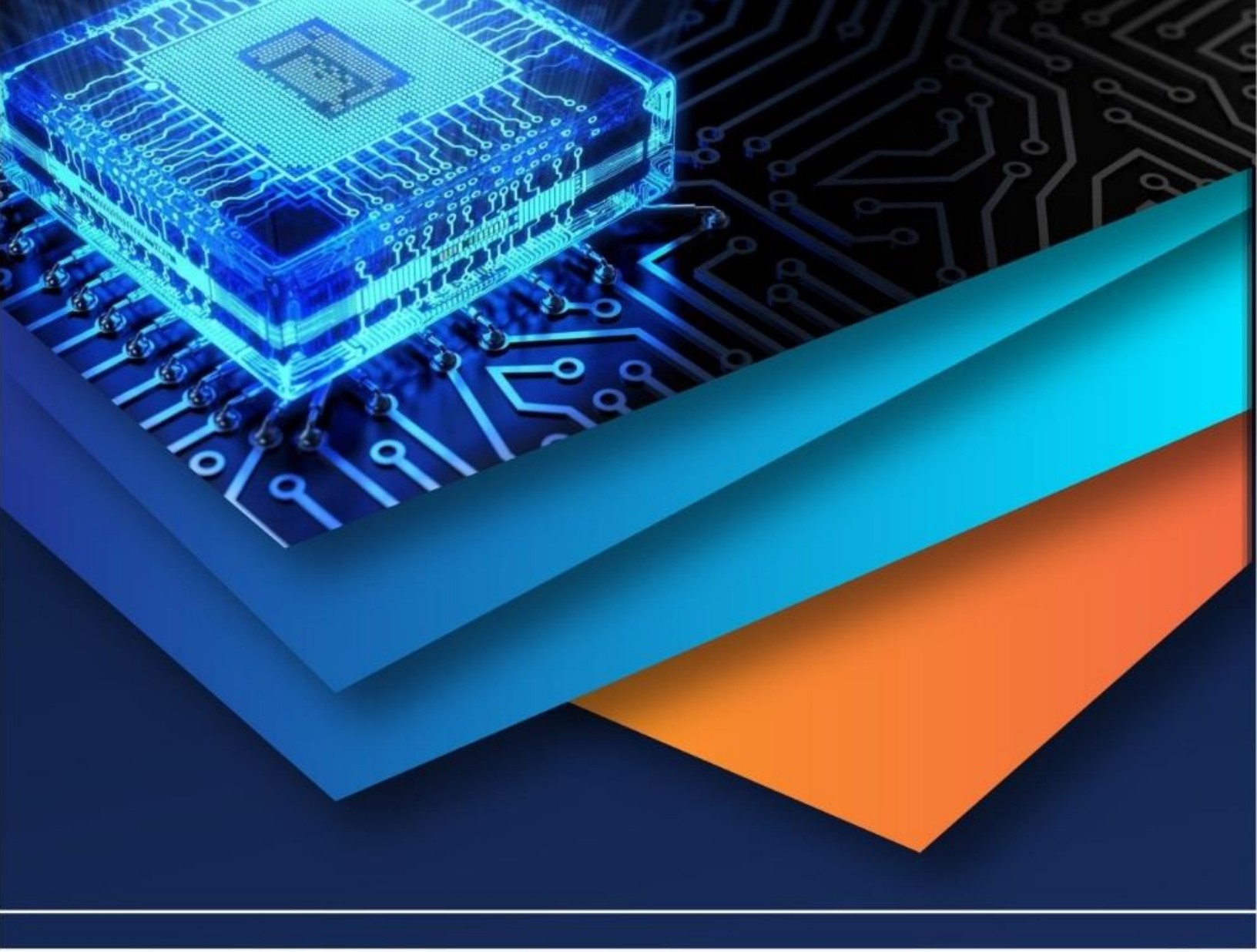
Deep learning has unquestionably revolutionized artificial intelligence (AI), propelling it to acquire unprecedented capability and application. Techniques like deep neural networks, and convolutional neural networks have transformed domains such as natural language processing, image recognition, and autonomous systems, leading to breakthroughs that enhance human-machine interaction, decision-making, and innovation across sectors.

This evolution, however, has brought forth a formidable challenge in the form of adversarial attacks, exposing vulnerabilities in AI systems and raising concerns about their security, reliability, and ethical implications. Addressing this challenge becomes not just critical but essential to ensure the continued integration and trustworthiness of AI technologies. This requires an integrated approach that combines explainable AI, interdisciplinary collaboration, and advanced defense strategies including adversarial training and defensive distillation, to fortify deep learning models against adversarial vulnerabilities. Eventually, it leads to promising a future where the partnership between AI and DL thrives amidst emerging challenges. The potential of deep learning within AI encompasses robust and interpretable model development, insights from diverse fields such as medicine and agriculture. Anchored in addressing adversarial vulnerabilities, AI empowered by deep learning stands poised to drive industries and transform the world through safety, reliability, and innovation- showcasing the immense potential of human ingenuity harnessed through this evolving partnership.

REFERENCES

[1] J. McCarthy, M.L. Minsky, N. Rochester and C.E. Shannon, “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”, August 31, 1955, AI Magazine, 27(4), 12, https://doi.org/10.1609/aimag.v27i4.1904
[2] V. R. Jokanović, Artificial Intelligence, 1st ed., 2022.
[3] H. Sheikh, C. Prins and E. Schrijvers, "Artificial Intelligence: Definition and Background," in Mission AI: The New System Technology, Springer International Publishing, 2023, [Online].
[4] A. Krizhevsky, I. Sutskever and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
[5] M. Garnelo and M. Shanahan, "Reconciling deep learning with symbolic artificial intelligence: representing objects and relations," Current Opinion in Behavioral Sciences, vol. 29, pp. 17-23, 2019. ISSN 2352-1546. doi: 10.1016/j.cobeha.2018.12.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352154618301943
[6] R. R. Nadikattu, “The Supremacy of Artificial Intelligence and Neural Networks.” 5. 2320-2882. 10.1729/Journal.24071, (2017).
[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks.” In ICLR, 2014.
[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.Goodfellow, “Intriguing properties of neural networks”, arXiv:1312.6199, 2013.
[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples”, In ICLR, 2015.
[10] X. Tang, Y Li, Y Sun, H Yao, P Mitra, and S Wang, “Transferring robustness for graph neural network against poisoning attacks”, In WSDM, 2020.
[11] D. Zügner, A. Akbarnejad, and S. Günnemann, “Adversarial attacks on neural networks for graph data”, In KDD, 2018.
[12] M. Soll, T. Hinz, S. Magg, S. Wermter, “Evaluating Defensive Distillation for Defending Text Processing Neural Networks Against Adversarial Examples”, (2019), 10.1007/978-3-030-30508-6_54.
[13] Z. Zheng and P. Hong, "Robust Detection of Adversarial Attacks by Modeling the Intrinsic Properties of Deep Neural Networks," in Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), Montréal, Canada, 2018, pp. 7924-7933, Curran Associates Inc.
[14] M. Bohara, K. Patel, A. Saiyed, A.Ganatra, “Adversarial Artificial Intelligence Assistance for Secure 5G-Enabled IoT”, (2021). 10.1007/978-3-030-67490-8_13.(https://www.researchgate.net/publication/350757204_Adversarial_Artificial_Intelligence_Assistance_for_Secure_5G-Enabled_IoT)
[15] A.N. Bhagoji, W. He, B. Li, D. Song, "Exploring the Space of Black-box Attacks on Deep Neural Networks", (2017), ArXiv preprint arXiv:1712.09491,
[16] J. Su, D.V. Vargas, K. Sakurai, “One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation”, (2019), DOI: 10.1109/TEVC.2019.2890858.
[17] N. Carlini and D. Wagner, ‘‘Towards evaluating the robustness of neural networks.’’ (2016), [Online], arXiv preprint arXiv:1608.04644.
[18] S-Mohsen M-Dezfooli, AFawzi, O Fawzi, and P Frossard, “Universal adversarial perturbations”, 2016 ,arXiv preprint arXiv:1610.08401.
[19] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, (2017), ‘‘Houdini: Fooling deep structured prediction models’’, [Online], arXiv preprint arXiv:1707.05373 .
[20] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 427–436, 2015.
[21] J. Lu, H. Sibai, E. Fabry, and D. Forsyth, “Standard detectors aren’t (currently) fooled by physical adversarial stop signs”, arXiv preprint arXiv:1710.03337, 2017.
[22] A. Gnanasambandam, A.M. Sherman, & S.H. Chan, “Optical Adversarial Attack”, arXiv preprint, arXiv:2108.06247 (2021).
[23] A. Abusnaina, A. Khormali, D. Nyang, M. Yuksel, and A. Mohaisen, ‘‘Examining the robustness of learning-based DDoS detection in software defined networks,’’ in Proc. IEEE Conf. Dependable Secure Comput. (DSC), Nov. 2019, pp. 1–8.
[24] S. S. Alchekov , M. A. Al-Absi , A. A. Al-Absi , “Lee HJ. Inaudible Attack on AI Speakers. Electronics”, 2023; 12(8):1928,
[25] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, ‘‘SummEval: Re-evaluating summarization evaluation,’’ Trans. Assoc. Comput. Linguistics, vol. 9, pp. 391–409, Apr. 2021.
[26] A. J. Dellinger, Research finds the sound of chirping birds can be used to hack Alexa. Digitaltrends, Available online: https://www.digitaltrends.com/home/voice-assistants-hacked-adverserial-attack-birds-chirping-alexa/?utm_source=sendgrid&utm_medium=email&utm_campaign=daily-brief (accessed on 2 October 2018).
[27] A. Krizhevsky, “Learning multiple layers of features from tiny images”, 2009.
[28] M. Shu, Y. Shen, M. C. Lin and T. Goldstein, "Adversarial Differentiable Data Augmentation for Autonomous Systems", 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021, pp. 14069-14075, doi: 10.1109/ICRA48506.2021.9561205.
[29] Q. Sun, A. A. Rao, X. Yao, B. Yu and S. Hu, "Counteracting Adversarial Attacks in Autonomous Driving", 2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD), San Diego, CA, USA, 2020, pp. 1-7.
[30] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, “Adversarial examples improve image recognition”, In CVPR, 2020.

[31] N. Akhtar, A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey", ArXiv preprint arXiv:1801.00553 (2018).
[32] T. Miyato, A. M. Dai, and I. Goodfellow, (2016), ''Adversarial training methods for semi-supervised text classification'' [Online], arXiv preprint arXiv:1605.07725 .
[33] T. Miyato, A. M. Dai, and I. J. Goodfellow, 2017, "Adversarial training methods for semi-supervised text classification", In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.
[34] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, ''Improving the robustness of deep neural networks via stability training,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun, 2016, pp. 4480–4488.
[35] M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto, 2018, "Interpretable adversarial perturbation in input embedding space for text", In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, pages 4323–4330.
[36] J. Wang, X. Fu, G. Xu, Y. Wu, Z. Chen, Y. Wei, and L. Jin. 2018, "A3net: Adversarial-and-attention network for machine reading comprehension", In NLPCC.
[37] M. Sato, J. Suzuki, and S. Kiyono, 2019, "Effective adversarial regularization for neural machine translation", In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 204–210.
[38] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks", In Security and Privacy (SP), 2016 IEEE Symposium on, pages 582–597. IEEE, 2016
[39] N. Papernot et al., "The limitations of deep learning in adversarial settings," in Security and Privacy (EuroS&P), 2016 IEEE European Symposium on IEEE, 2016, pp. 372–387.
[40] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ACM, 2017, pp. 506–519.
[41] J. Zeng, X. Zheng, J. Xu, L. Li, L. Yuan, and X. Huang, ''Certified robustness to text adversarial attacks by randomized [MASK],'' arXiv:2105.03743, 2021
[42] X. Zhang & M. Zitnik, " GNNGuard: Defending Graph Neural Networks against Adversarial Attacks", 2020.
[43] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations", ArXiv preprint, 2017, arXiv:1702.04267.
[44] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, ''Beyond accuracy: Behavioral testing of NLP models with CheckList", 2020, arXiv:2005.04118
[45] J. Lu, T. Issaranon, and D. Forsyth, "Safetynet: Detecting and rejecting adversarial examples robustly", CoRR, abs/1704.00103, 2017.
[46] Y. Li, X. Li, Y. Yang, and R. Dong,, "A diverse data augmentation strategy for low-resource neural machine translation", Information, 11(5) , 2020b.
[47] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, "Gender Bias in Neural Natural Language Processing", Springer International Publishing, Cham, pages 189–202, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY