Baghdad Science Journal

Volume 22 | Issue 8 Article 27

8-27-2025

Emerging Challenges in Adversarial Deep Learning for Computer Vision and Cybersecurity

Younis Al-Arbo
Department of computer Science, College of Education for Pure Science, University of Mosul, Mosul, Nineveh, Iraq

Asmaa Alqassab
Department of computer Science, College of Education for Pure Science, University of Mosul, Mosul, Nineveh, Iraq

How to Cite this Article

Al-Arbo, Younis and Alqassab, Asmaa (2025) "Emerging Challenges in Adversarial Deep Learning for Computer Vision and Cybersecurity," *Baghdad Science Journal*: Vol. 22: Iss. 8, Article 27.
DOI: https://doi.org/10.21123/2411-7986.5038

BAGHDAD SCIENCE JOURNAL 2025;22(8):2791–2799

RESEARCH ARTICLE

Emerging Challenges in Adversarial Deep Learning for Computer Vision and Cybersecurity

Younis Al-Arbo *, **Asmaa Alqassab**

Department of computer Science, College of Education for Pure Science, University of Mosul, Mosul, Nineveh, Iraq.

ABSTRACT

The emergence of "Deep Learning" DL has revolutionized the scope of cybersecurity and computer vision. However, this technology is not immune to emerging challenges that can affect its performance and security. One major challenge is the availability of large datasets for training DL algorithms. Furthermore, there is a need for improved algorithms and architectures that can effectively process such datasets. Another challenge is the constant evolution of cyber threats, which require the development of new DL models to defend against them. Additionally, the interpretability and explain ability of DL models in cybersecurity pose a significant challenge, as their black-box nature can make them difficult to understand and mitigate against. Therefore, the emerging challenges in DL for computer vision and cybersecurity require a coordinated effort from researchers and practitioners in the felid of neural network specially with generative adversarial network to overcome handicaps and effectively leverage the technology to enhance security and surveillance in various domains.

Keywords: Adversarial attacks, Deep learning (DL), Generative adversarial network (GAN), Deep neural network (DNN), Computer vision, Cybersecurity

Introduction

The widespread use of "Deep Learning" DL techniques faces a significant danger from research in adversarial DL: these methods are susceptible to malignant adversaries' meticulously planned attacks. "Deep Neural Networks" DNN For example, struggle to accurately categorize hostile images through tiny perturbations that are added to clean photos. first go over the three primary categories of attacks that can be made against DL technologies: poisoning, evasion, and privacy threats. [1,2]

In Cybersecurity, DL offers the opportunity to detect and prevent security breaches, but the challenge lies in detecting increasingly sophisticated and elusive attacks. Although deep learning has the ability to process data of large amounts for the purpose of identifying objects, but also, it demands vast amount of memory and processing power, particularly in real-time situations. [3]

In computer vision, one of DL considerable difficulties is the processing of vast amount of data created with cameras and sensors, furthermore, computer vision algorithms need to be very powerful and able to be adapted to different environments, weather conditions for example, which can affect forecast accuracy. While in the field of cybersecurity, DL has the problem of handling security breaches which are continuously increased. [4]

Attackers are continuously coming up with new ways to evade being identified, and that's why DL algorithms have to have the ability to stay up to date with such sophisticated threat scene. Likewise, DL algorithms have to run in an extremely safe environment to prevent hostile actors from manipulating them. Having such issues in mind, researchers still

Received 11 September 2023; revised 17 May 2024; accepted 19 May 2024.
Available online 27 August 2025

* Corresponding author.
E-mail addresses: younis.bayati@uomosul.edu.iq (Y. Al-Arbo), asmaa_mow@uomosul.edu.iq (A. Alqassab).

https://doi.org/10.21123/2411-7986.5038

attempting to come up with enhanced version of DL algorithms which are capable of adapting to the varying of circumstances and handling real time data, besides, fresh methods like “adversarial training” as well as employing “detection of anomalies” in order to identify assaults that may not be obvious. Ultimately, while deep learning has demonstrated immense potential in both of “cybersecurity” and “computer vision”, there stills more work needs to be accomplished to cope with the obstacles which come up with the evolving of the technology. [5,6]

Adversarial deep learning for cybersecurity

In computer vision, adversarial DL can be used to improve the accuracy and robustness of image recognition models by identifying and mitigating the effects of adversarial perturbations. [7] As cyber threats become increasingly sophisticated and prevalent, adversarial DL is likely to play an increasingly important role in protecting against such threats. The security flaw in DL algorithms when compared to hostile samples has recently gained widespread recognition. While they are perceived as harmless by humans, fabricated samples can cause various undesirable behaviors in DL models. Adversarial attacks are successfully used in situations seen in the real world to show how practical they are. As a result, adversarial attack and defensive solutions have gained more and more interest in recent years from both the ML and security fields. [8]

DNNs’ outstanding performance has made DL useful in a wide range of industries. However, the possible risks presented by adversarial demos have slowed down the general adoption of DL. In some cases, the model’s final performance suffers noticeably from hostile disturbances that are invisible to the human sight. In the area of DL, numerous studies on adversarial assaults and their defenses have been published. [9] Instead of poisoning attacks, which involve feeding contaminated data into the training data, the majority of them concentrate on evasion attacks, where hostile cases are discovered during testing. Furthermore, because there are no accepted evaluation techniques, it is challenging to determine the actual threat posed by hostile attacks or the potency of a DL model. [10]

Since DL is implemented for deployment in many vital systems, it is crucial to take its dependability into account because these algorithms are susceptible to hostile attacks. Similar to how they attack a firewall, hackers use hostile samples to expose weaknesses in the DL framework’s operations. Prior to implementation, it is crucial to take into account the DL framework’s deficiencies by carrying out stress testing in hostile environments to find any weak points. This study is referred to as the Adversarial DL. DL’s constraints are made worse by the fact that many DL frameworks operate in crucial systems like black boxes. Since the framework’s choice is left unexplained in such circumstances, it is exceedingly difficult for clients and professionals to perceptive the model’s findings. Since there is no assurance regarding the reliability of DL frameworks, their usage in a secure critical system is unlikely. [11,12]

Domain creation algorithms (DCAs)

Several forms of malware have taken use of Domain Creation Algorithms (DCAs), these are employed to create Command and Control (C & C) linkages that lead to “Distributed Denial-of-Service” DDoS assaults. Recently, DL-based architectures based on “generative adversarial networks” (GANs) have been trained to generate competing domains to avoid detection by DL. An anti-black box offensive strategy was provided by another study to circumvent DGA without taking the beforehand classification architecture into consideration. Utilizing the DMD-2018 dataset, the approach was able to reduce the A1 rating from 0.981 to 0.402. [13]

Anti-malware software

The novel Malware Reconfiguration Variation (MRV) method creates copies of antagonistic malware based on semantic analysis of current malware in order to go around the malware detector and increase the detector’s efficacy. [14] This method employs three protection measures. [15] Another study used adversarial instance generation methods to produce malware while retaining the invasive virus’ capabilities. A malware detection technique based on transmitted GAN (“tGAN”) was created to recognize zero-data attacks. A 97.22% accuracy rate and good learning stability were demonstrated by this approach. In order to make the DL model more resistant to adversarial attack, an infection identification model is also suggested to eliminate random elements of the data. In an effort to strengthen adversarial attacks, six heuristic principles have been employed, comprising API call-based visualization and evasion classifiers such recurrent neural network (RNN), DNN, and ML classifiers [16]

Intrusion detection systems (IDS)

A PCA-based detection system’s accuracy was reduced by 15% when compared to an adaptive AE-based IDS system for assessing its robustness against hostile cases. According to several research,

black box attacks against IDS using GANs have a high success rate because they generate hostile network traffic to evade detection, including smart vehicle networks. It was suggested to use antagonistic anomalies to train the Artificial neural network (ANN) model to create a host-based identity system (HIDS) based on GAN. Similar to this, in three aggressive black-box assaults against DNN-based NIDS that were examined, two data supplementation modules were employed to address the problem of inadequate information in Network Identification Systems (NIDS).[17,18]

Fig. 1. An unnoticeable picture manipulation attack on a deep visual model results in a highly confidently inaccurate prediction (a) Manipulated Gibbon with 98% confidence (b) Original Panda with 52% confidence.

The adversarial attacks on computer vision and deep neural networks

Due to the widespread usage of DL as well as computer vision in vital applications like facial recognition and autonomous driving, adversarial assaults on these advancements are becoming a serious problem. Adversarial attacks are malicious entries designed to trick these systems, resulting in incorrect or unwanted results. These attacks can take various forms, including modifying input data, exploiting vulnerabilities in the learning process itself, or using complex generation models to generate hostile examples. Mitigating hostile attacks is a major challenge, requiring robust and flexible deep learning models that can effectively detect and defend against such attacks.[19] This is an active area of machine learning research, with ongoing efforts to develop new defense mechanisms and techniques to make deep learning systems more secure and trustworthy. Given the superhuman skills of DL, it is thought that computer vision-based Artificial Intelligence (AI) has matured to the point where it can be implemented in crucial safety and security systems. The use of facial recognition technology in mobile devices, ATMs, and vehicles are some of the most notable instances of how modern cultures are coming to trust computer vision technologies in the real world. With ongoing DL-based vision research for self-driving cars, facial recognition, robotics, monitoring systems, etc., anticipate deep learning's ubiquitous presence in the security-critical industry. applications for computer vision. But now that the adversarial vulnerability of DL has been found, there are legitimate worries about this potential.[20]

DNN forecasts can be evaluated with little to no input interruption, according to Szegedy et al..[21] In the case of images, even when these disturbances are small-scale and imperceptible to the human visual system, the resulting predictions of a deep vision model can alter as showing in Fig. 1.[22,23] The picture categorization problem led to the initial discovery of these perplexing signals. But for a variety of typical computer vision problems, like semantic segmentation, which can disclose things, and object tracking, the existence of it is already well known.[24,25] Deep learning's viability as a practical technology is gravely threatened by a variety of hostile perturbation characteristics that are highlighted in the literature. For example, it has been noted time and time again that the models that have been challenged typically exhibit high confidence in their inaccurate predictions for the modified images.[26,27] Additionally, it has been demonstrated that the same illness frequently deceives several models. Besides, universal perturbations—precomputed perturbations that may be introduced to "any image" with high chance to deceive a specific model—have been seen in the literature. These discoveries have significant ramifications for crucial security applications, particularly given that DL solutions are frequently thought to possess predictive abilities that are on par with or even superior to those of humans.[28]

The research community has paid a great deal of attention to adversarial attacks (and their defenses) over the past five years due to the topic's importance. provides an overview of the developments in this area up to the beginning of 2018.[29]

The majority of these research can be viewed as first-generation technologies that examine algorithms and fundamental strategies to trick or protect DL from adversarial attacks. A few of these algorithms have sparked a steady stream of advancements in assault and defense strategies that have been further developed and adapted.[27,28] Other vision tasks, rather than only a classification problem, which was the main focus of early contributions in this approach, have been discovered to be more the focus of these second-generation systems. In recent years, this trend in computer vision has significantly developed. By expanding on the concepts of and later literature, gaining the ability to produce more precise explanations of technical terms for this rapidly evolving study path. This resulted in the literature examined in this article having a more logical structure, about which gives brief explanations based on how the research community now understands the terminology.[29,30] Additionally, the focus on peer-reviewed

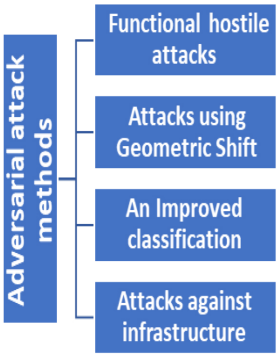

Fig. 2. Methods of dealing with Adversarial attacks.

works that are disseminated in the top ML and computer vision journals. By focusing on ground-breaking contributions, provide scientists in computer vision and machine learning with a clearer perspective of this approach. Furthermore, this article provides the most thorough analysis of this tendency to date by reviewing the most recent contributions to this quickly developing topic.[31]

Adversarial attack methods

This part, goes over several common adversary attack techniques and strategies. As showing in Fig. 2 These methods try to attack and apply to different DL paradigms.

1. Functional hostile attacks

In contrast to conventional lp-ball attacks, functional adversarial assaults allow the perturbations function to be employed to conceal input attributes in order to construct an adversarial instance. Operational adversarial attacks are further limited in a number of ways since features cannot be selectively concealed.

ReColorAdv and other authors suggested a practical adversarial approach on pixel colors. By uniformly altering the colors of the input image, ReColorAdv generates anti-examples that deceive image classifiers. In an aggressive instance, ReColorAdv maps every single pixel color c in the source file to a replacement pixel color *f(c)* using a soft-parameter algorithm f. The attack potential may be greatly increased by integrating functionally adversarial assaults with existing attacks that employ the l_p standard. As a consequence, the model may modify input locally, repeatedly, and globally. The most potent attack at the moment can be created, according to experiments, by combining ReColorAdv with other attacks.[32,33]

2. Attacks using geometric shift

In order to create adversarial test models, geometry-based attacks rotate or zoom targets in photos. According to the geometry transformation's invariance, the class of image classification jobs must yield the same results regardless of how significant the geometry transformation of the input image is. Algorithms based on gradients and those based on geometric transformation invariants frequently coexist. Only straightforward transformations, such as rotation and translation, are sufficient to deceive DNN, as demonstrated by Engstrom et al..[34] Mani-Fool is a method for locating the smallest, worst-case geometric modifications for images that Kanpak et al.[35] suggested. According to Xiao et al.,[36] it is possible to change the scene's geometry while maintaining the image's original appearance.

Translation-Invariant (TI) attack technique was suggested by Dong et al.[37] to produce more adversarial convertible cases for defense models. The antagonistic samples are strengthened by TI using a set of localized pictures, which increases their mobility and reduces sensitivity to the specific parts of the white-box models being attacked. TI can be utilized to improve the attacks by coupling the gradient of the uncorrected image to a preset kernel.

Weak Shi et al.[38] Utilize the Diversified Input Method (DIM) to increase the portability of hostile examples. DIM inputs the updated images into a class for scale computation after applying a set of label-preserving modifications to training images at random. The data enrichment strategy served as inspiration for this technique. DIM can be combined with momentum-based approaches to further boost portability. When compared to the top opposition systems and official starting locations from the 2018 NIPS competitive competition, the M-DI^2-FGSM (Fast Gradient Sign Method) Improved Offensive outperforms the initial assault submission in the NIPS competition by an amount of 5.8%.[39]

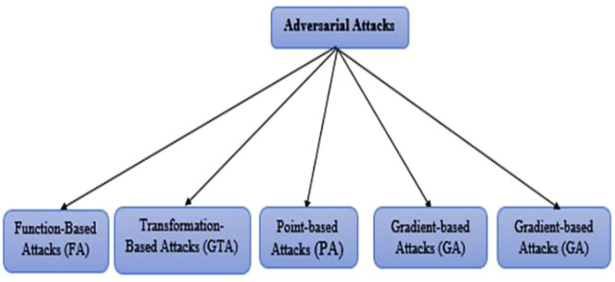

Fig. 3. Improved classification of adversary attacks.

3. An improved classification

A better classification of hostile assaults, As appears in Fig. 3 which is depicted in Fig. 1, based on the previous categories. First specifically divide hostile attacks into five groups: Gradient-based attacks (GA), Point-based attacks (PA), Geometry Transformation-Based Attacks (GTA), Function-Based Attacks (FA), and Transportation-Based Attacks (TA) are the five types of attacks. In addition, categorize attacks based on optimization and assaults based on delicate features from as gradient-based attacks since they are nearly identical to one another but have different objectives, and categorize attacks based on generative models from as transformation-based attacks. Because of the fact that class procedures are always performed in accordance with hits, decision-based attacks are handled by as a particular case of point-based attacks. Finally, both of the classifications in and are included in our categorization. [40,41]

4. Attacks against infrastructure

Transfer-based attacks need a grasp of the training data rather than model knowledge. This gives you the option to select between white box and black box assaults. In 2017, Szegedy et al. [42] presented the idea of proactive testing models and noted that accidental examples created for one model may be effectively transferred to another, regardless of the model's design. This was known as the paradigm's portable. and then Papernot et al. [43] go into great detail about this.

In 2018, Papernot et al. [44] presented a transfer-based technique that uses inputs that were fabricated by the adversary and classified by the target DNN to train a local model to replace it. In this study, this approach is mentioned as an alternative. In 2017, Liu et al. published a revolutionary strategy that, for the first time, enables a significant number of target hostile instances to switch across several contexts by combining a variety of different paradigms. [45]

TREMBA. A technique dubbed "TRansferable Embedding" (TREMBA) based Black Box Attack, which combines "transform-based attack" with "point-based attack", was proposed by Huang et al.. [46] While point-based assault increases success rate, transform-based attack increases query efficacy. In contrast to earlier assault strategies, this strategy trains information-enhanced surrogate models to behave like the target model. TREMBA operates in two stages. Using low-dimensional integration space, the decoder is trained to generate inverse issues of the original network in the first stage. In the second stage, the low-dimensional embedded space of the generator is searched for examples of an anti-target mechanism using the Natural Evolution Strategy (NES). TREMBA's success rate is around 10% greater than that of earlier black box attacks, although it receives more than 50% smaller requests.

Analyzed recent assaults in computer vision in terms of their enhanced classification comparison with proposed work

Through our study of the topic .it has been shown previously an unnoticeable picture manipulation attack on a deep visual model results in a highly confidently inaccurate prediction Fig. 4. gives (a) Manipulated Gibbon with 98% Confidence (b) Original Panda with 52% Confidence of computer vision and deep neural networks.

A technique dubbed TRansferable EMbedding (TREMBA) based Black Box Attack). TREMBA's success rate is around 10% greater than that of earlier black box attacks, although it receives more than 50% smaller requests.

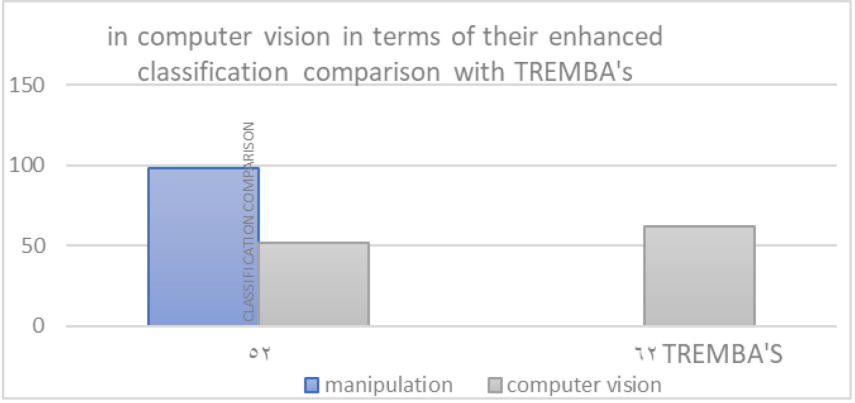

Fig. 4. Analyzed recent assaults.

Results

Results show the mentioned Adversarial Deep Learning for Cybersecurity in effective attack detection Domain Creation Algorithms (DCAs). the approach was able to reduce the A1 rating from 0.981 to 0.402. Anti-malware software.

A malware detection technique based on transmitted GAN (tGAN) was created to recognize zero-data attacks. A 97.22% accuracy rate and good learning stability were demonstrated by this approach.

Intrusion Detection Systems (IDS) A PCA-based detection system's accuracy was reduced by 15% when compared to an adaptive AE-based IDS system for assessing its robustness against hostile cases.

Conclusion

"Adversarial Deep Learning" offers a big issue to "Cybersecurity" and "Computer Vision". Adversarial assaults on deep neural networks may produce erroneous or malignant outcomes, impacting computer vision systems' reliability and accuracy. Likewise, cybersecurity assaults can take advantage of deep learning algorithms' flaws, leaving valuable information at hazard. To address this issue, researchers are investigating strategies incorporate detection methods, defensive distillation and enhanced adversarial, that aim to render deep learning algorithms safe and robust. Overall, the problem of adversarial DL is a crucial area of study that will only become more significant as DL technology is applied in more and more fields.

In this study looked into the concepts and processes behind the suggested algorithms and procedures. Based on the most recent research, also examined the efficacy of various hostile defenses. In the past two years, new adversary assaults and defenses have been created. Investigations were also conducted into some fundamental problems, like the reason for contradictory samples and the presence of all-encompassing strong boundaries. Thoroughly examined and analyzed the most recent assaults in computer vision in terms of their enhanced classification, as well as performing an upgraded adversarial attack classification that incorporates current classifications.

Author's declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Mosul.

Author's contribution

Y.AL. presented the idea and concern with the design and responsible for methodology. A. Alqassab concern with writing, proof reading, editing the manuscript.

References

1. Yiyun Z, Meng H, Liyuan L, Jing H, Xi G. The adversarial attacks threats on computer vision: A survey. IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW). 2019;105–108. https://doi.org/10.1109/MASSW.2019.00012.
2. Sholihin M, Fudzee M F, Ismail M N. AlexNet-Based Feature Extraction for Cassava Classification: A Machine Learning Approach. Baghdad Sci J. Dec 5 2023;20(6 (Suppl.)):2624. https://doi.org/10.21123/bsj.2023.20.6(Suppl.).
3. Zhou S, Liu C, Ye D, Zhu T, Zhou W, Yu PS. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity. ACM Comput Surv. 2022;55(8):1–39, https://doi.org/10.1145/3547330.
4. Jassim OA, Abed MJ, Saied ZH. Indoor/Outdoor Deep Learning Based Image Classification for Object Recognition Applications. Baghdad Sci J. Dec 5 2023;20(6 (Suppl.)):2540. https://doi.org/10.21123/bsj.2023.20.6(Suppl.).
5. Hyrum SA, Jonathan W, Bobby F. DeepDGA:Adversarially-tuned domain generation and detection. Proceedings of the 2017 ACM Workshop on Artificial Intelligence and Security. arXiv:1610.01969. 2017:13–21. ACM. https://doi.org/10.1145/2996758.2996767.
6. Sidi L, Nadler A, Shabtai A. MaskDGA: A Black-box Evasion Technique Against DGA Classifiers and Adversarial Defenses. arXiv preprint arXiv:1902.08909. 2019;12P. https://doi.org/10.48550/arXiv.1902.08909.
7. Yang W, Kong D, Xie T, Gunter CA. Malware detection in adversarial settings: Exploiting feature evolutions and confusions in android apps. Proceedings of the 33rd Annual Computer Security Applications Conference. 2018. https://doi.org/10.1145/3134600.3134642.
8. Kim JY, Bu SJ, Cho SB. Malware detection using deep transferred generative adversarial networks. In International Conference on Neural Information Processing, November 2017;556–564. https://doi.org/10.1007/978-3-319-70087-8_58.
9. Ghassemi N, Shoeibi A, Rouhani M. Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images. Biomedical Signal Processing and Control. Mar 1 2020;57:101678. https://doi.org/10.1016/j.bspc.2019.101678.
10. Li D, Li Q, Ye Y, Xu S. Enhancing Robustness of Deep Neural Networks Against Adversarial Malware Samples: Principles, Framework, and AICS'2019 Challenge. arXiv

preprint arXiv:1812.08108, 2019. https://doi.org/10.48550/arXiv.1812.08108.
11. Madani P, Vlajic N. Robustness of deep auto-encoder in intrusion detection under adversarial contamination. Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security. 2019;1. https://doi.org/10.1145/3190619.3190637.
12. Seo E, Song HM, Kim HK. GIDS: GAN based Intrusion Detection System for In-Vehicle Network. In 2018 16th Annual Conference on Privacy, Security and Trust (PST). 2018. https://doi.org/10.1109/PST.2018.8514157.
13. Salem M, Taheri S, Yuan JS. Anomaly Generation using Generative Adversarial Networks in Host Based Intrusion Detection. arXiv preprint arXiv: 1812.04697, 2019. https://doi.org/10.48550/arXiv.1812.04697.
14. Zhang H, Yu X, Ren P, Luo C, Min G. Deep Adversarial Learning in Intrusion Detection: A Data Augmentation Enhanced Framework. arXiv preprint arXiv:1901.07949, 2019. https://doi.org/10.48550/arXiv.1901.07949.
15. David S, Julian S, Karen S, Ioannis A, Aja H, Arthur G, *et al.* Mastering the game of go without human knowledge. Nature. 2018;550. https://doi.org/10.1038/nature24270.
16. Charlotte M. China unveils world's first facial recognition atm 2020.
17. Apple. About face id advanced technology. 2020. https://support.apple.com/en-au/HT208108.
18. Sorin G, Bogdan T, Tiberiu C, Gigel M. A survey of deep learning techniques for autonomous driving. J Field Robot. 2020;37(3):362–386. https://doi.org/10.1002/rob.21918.
19. Niko S, Oliver B, Walter S, Raia H, Dieter F, Jurgen L, *et al.*, The limits and potentials of deep learning for robotics, Int J Robot Res 2019;37(4–5):403–404. https://doi.org/10.1177/0278364918770733.
20. Maryam MN, Flavio V, Taghi MK, Naeem S, Randall W, Edin M. Deep learning applications and chal-lenges in big data analytics. J Big Data. Dec 2015;2:1–21. https://doi.org/10.1186/s40537-014-0007-7.
21. Christian S, Wojciech Z, Iiya S, Joan B, Dumitru E, Ian G, etal Intriguing properties of neural networks. arXiv preprint arXiv: 1312.6199, 2015. https://doi.org/10.48550/arXiv.1312.6199.
22. Wang S, Zhang J, Liu M, Liu B, Wang J, Yang S. Large-signal behavior modeling of GaN P-HEMT based on GA-ELM neural network. Circ Syst Signal L Pr. Apr 1 2022:1–4. https://doi.org/10.1007/s00034-021-01891-7.
23. Luo H, Yan X, Zhang J, Guo Y. A neural network-based hybrid physical model for GaN HEMTs. IEEE Trans Microw Theory Techn. Sep 26 2022;70(11):4816–26. https://doi.org/10.1109/TMTT.2022.3206442.
24. Jia YJ, Lu Y, Shen J, Chen QA, Chen H, Zhong Z, *et al.* Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In International Conference on Learning Representations (ICLR'20). Jan. 2020 https://doi.org/10.48550/arXiv.1905.11026.
25. Chen X, Yan X, Zheng F, Jiang Y, Xia ST, Zhao Y, *et al.* One-shot adversarial attacks on visual tracking with dual attention. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2020:10176–10185.
26. Naveed A Ajmal M. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access. 2018;6:14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385.
27. Zheng H, Zhang Z, Gu J, Lee H, Prakash A. Efficient adversarial training with transferable adversarial examples. Proc IEEE Comput Soc Conf. Comput Vis Pattern Recognit. 2020:1181–1190. https://doi.org/10.48550/arXiv.1912.11969.
28. Mishra S, Stoller D, Benetos E, Sturm BL, Dixon S. GAN-based generation and automatic selection of explanations for neural networks. arXiv preprint arXiv:1904.09533. Apr 21 2019. https://doi.org/10.48550/arXiv.1904.09533.
29. Naveed A, Mohammed AJ, Mohammed B, Ajmal M. Label universal targeted attack. arXiv preprint. arXiv:1905.11544. 2019;78–89. https://doi.org/10.48550/arXiv.1905.11544.
30. Oriol V, Igor B, Wojciech MC, Michael M, Aandrew D, Junyoung C, *et al.* Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature. 2019;575:350–354. https://doi.org/10.1038/s41586-019-1724-z.
31. Xu B, Zhou D, Li W. Image enhancement algorithm based on GAN neural network. IEEE Access. Mar 29 2022;10:36766–77. https://doi/10.1109/ACCESS.2022.3163241.
32. Laidlaw C, Feizi S. Functional adversarial attacks. Adv Neural Inf Process Syst. 2019. https://doi.org/10.48550/arXiv.1906.00001.
33. Yeo YJ, Shin YG, Park S, Ko SJ. Simple yet effective way for improving the performance of GAN. IEEE Trans Neural Netw Learn Syst. Jan 1 2021;33(4):1811–8. https://doi.org/10.1109/TNNLS.2020.3045000.
34. Kanbak C, Moosavi-Dezfooli SM, Frossard P. Geometric robustness of deep networks: analysis and improvement. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2018;4441–4449. https://doi.org/10.48550/arXiv.1711.09115.
35. Xiao C, Zhu JY, Li B, He W, Liu M, Song D. Spatially transformed adversarial examples. arXiv preprint arXiv: 1801.02612. Jan 8 2018. https://doi.org/10.48550/arXiv.1801.02612.
36. Dong Y, Pang T, Su H, Zhu J. Evading defenses to transferable adversarial examples by translation-invariant attacks. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2019;4312–4321. https://doi.org/10.48550/arXiv.1904.02884.
37. Xie C, Zhang Z, Zhou Y, Bai S, Wang J, Ren Z, *et al.* mproving transferability of adversarial examples with input diversity. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2019;2730–2739. https://doi.org/10.48550/arXiv.1803.06978.
38. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands. Oct 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing. 2016;630–645. https://doi.org/10.48550/arXiv.1603.05027.
39. Serban AC, Poll E, Visser J. Adversarial examples-a complete characterization of the phenomenon. arXiv preprint arXiv: 1810. 2018. https://doi.org/10.48550/arXiv.1810.01185.
40. Zhou Y, Han M, Liu L, He J, Gao X. The adversarial attacks threats on computer vision: A survey. In: 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW). 2019. https://doi.org/0.1109/MASSW.2019.00012.
41. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, *et al.* Ntriguing properties of neural networks. arXiv preprint arXiv:1312.6199. Dec 21 2013. https://doi.org/10.48550/arXiv.1312.6199.
42. Papernot N, McDaniel P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277. May 24 2016. https://doi.org/10.48550/arXiv.1605.07277.
43. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. Proc ACM Conf Comput. Commun Secur. Apr 2 2017;506–519. https://doi.org/10.48550/arXiv.1602.02697.

44. Liu Y, Chen X, Liu C, Song D. Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770. Nov 8 2016. https://doi.org/10.48550/arXiv.1611.02770.
45. Huang Z, Zhang T. Black-box adversarial attack with transferable model-based embedding. arXiv preprint arXiv: 1911.07140. Nov 17 2019. https://doi.org/10.48550/arXiv.1911.07140.
46. Wierstra D, Schaul T, Glasmachers T, Sun Y, Peters J, Schmidhuber J. Natural evolution strategies. J Mach. Learn Res. 2015;15(1):215–255. https://doi.org/10.48550/arXiv.1106.4487.

التحديات المنبثقة في التعلم العميق العدائي فيما يخص الرؤية الحاسوبية والامن السيبراني

يونس ال-عربو، أسماء القصاب

علوم الحاسوب، كلية التربية للعلوم الصرفة، جامعة الموصل، موصل، العراق.

المستخلص

أحدث ظهور التعلم العميق والذي يرمز له بـDL , ثورة في مجال الأمن السيبراني و رؤية الحاسوب. ومع ذلك، فإن هذه التكنولوجيا ليست محمية ضد التحديات الناشئة التي يمكن أن تؤثر على أدائها وأمنها. يتمثل أحد التحديات الرئيسية في توافر مجموعات كبيرة من البيانات لتدريب خوارزميات التعلم العميق. علاوة على ذلك، هناك حاجة إلى تحسين وتطوير الخوارزميات والبنى التي يمكنها معالجة مجموعات البيانات هذه بشكل فعال. التحدي الآخر هو التطور المستمر للتهديدات السيبرانية، الأمر الذي يتطلب تطوير نماذج DL جديدة للدفاع ضدها. بالإضافة إلى ذلك، فإن قابلية شرح وترجمة نماذج التعلم العميق في الأمن السيبراني تشكل عائقا" لايستهان به، كون طبيعتها المبهمة من الممكن ان تجعلها صعبة الفهم . بالتالي،فأن مثل هذة العوائق في التعلم العميق فيما يخص الرؤية الحاسوبية والامن السبراني تتطلب جهدًا منسقًا من الباحثين والممارسين في مجال الشبكات العصبية وخاصة مع شبكة الخصومة التوليدية للتغلب على المعوقات والاستفادة بشكل فعال من التكنولوجيا لتعزيز الأمن والمتابعة في مختلف المجالات.

الكلمات المفتاحية: الهجمات العدائية، التعلم العميق، الشبكة العدائية التوليدية، الشبكة العصبية العميقة، الرؤية الحاسوبية، الامن السيبراني.