AI vulnerabilities

Andrei Viureanu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
andrei.vizureanu@gmail.com

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—The paper is intended as a guideline for the research regarding past, recent, and future AI threats and vulnerabilities. We start by exploring today's context, implementation, and vulnerabilities of AI systems. We then look at the current perspective on AI threats and taxonomy. Further, we focus on the latest types of AI-related attacks and elaborate on the adversarial attacks. The approach to adversarial attacks is both offensive and defensive as we present current techniques for misleading and improving the robustness of the AI models. Finally, we propose some perspectives on further work and contributions to research on AI vulnerabilities.**

***Index Terms*—AI threats, vulnerabilities, ML-DL, adversarial attacks, model enhancement**

I. Introduction

We've embraced AI in our everyday lives and leveraged its advantages in industry, healthcare, fraud detection, customer service, natural language processing, computer vision, and autonomous vehicles. However, the expanding use of AI, and in particular deep learning (DL), in real-life applications and critical systems obliges us to focus beyond the benefits and more on the advert threats and vulnerabilities.

Shaping the digital future, legislation and frameworks have already been developed by authorities and international forums and organizations. Addressing risks involving AI applications the EU, in the AI ACT [1] determines a risk-based level of applications and tries to set clear requirements for each type of risk-level application and the system that integrates them.

"High-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities. The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks. The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws."[1]

Considering the above, the work outlines the vulnerabilities of AI, with a focus on the ML models and their security concerns facing adversarial attacks.

1st Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 8 November, 2024

II. AI incidents

AI incidents, their behavior, and their occurrence are a big concern for lots of organizations that categorize, keep track of, and build databases for further analysis (e.g. AI Incident Database [2],[3]).

From the harms of relevance to the public policy community of AI incidents perspective, the CSET (Center for Security and Emerging Technology) AI Harm Taxonomy characterizes the harms, entities, and technologies involved in AI incidents and the circumstances of their occurrence. According to their database, the fields most affected are information and communication, transportation and storage, arts, entertainment and recreation, law enforcement, wholesale and retail trade, public administration, human health and social work, and administrative and support services.

Another taxonomy, described in [4], the Goals, Methods, and Failures (GMF) taxonomy is a failure cause analysis taxonomy for AI systems in the real world. The most affected technologies highlighted in the incident database are autonomous driving, chatbot, face recognition, automatic skill assessment, content search, deepfake video generation, robotic manipulation, autonomous drones, and voice generation.

In digital security, when it comes to specially designed attacks we can consider these three concepts assets, vulnerabilities, and threats that can be addressed based on their layer (e.g. data, software, storage, system, and network).

Some public databases refer to TTPs (Tactics, Techniques, and Procedures) such as Adversarial Tactics, Techniques, and Common Knowledge (ATT and CK) [5]. Traditional attack phases are divided into pre-attack phases and attack phases.

III. Adversarial attacks

The findings of Szegedy [6] ,[7] along with other research [8], [9], [10] made it clear that neural networks, by design, are vulnerable to attacks in the form of adversarial noise, also known as adversarial perturbations or adversarial examples.

This type of attack refers to applying subtle perturbation to input data to mislead or cause errors in machine learning models. No matter the domain or the target data, the amount of adversarial noise crafted over the original input data is designed to be undetected or hard to detect by human observers while significantly impacting the AI model's predictive outcome. In terms of security concerns, aside from the method's

stealthiness and transferability, another big threat lies in the fact that this can be performed even if the adversary lacks access to the underlying model.

A. Status quo

Organizations such as Mitre study the adversarial threats and built a documented platform, ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)[11] for adversary tactics and techniques against Al-enabled systems. Machine learning attacks follow the same phases but with techniques and procedures adapted to the ML context [11]: Pre-attack phases, that involve reconnaissance and resource development. In this stage, the hostile actors gather information and capabilities to support further attacks, obtain relevant ML artifacts, and target ML capabilities used by the victim. The attack phase starts with the Initial Access when hostile actors try to approach the local, or cloud-enabled ML system (e.g. the network, mobile device, or a sensor platform). After that, adversaries can interact with the ML model through an available API or indirectly via a product or service that integrates the ML model as part of its processes. In the execution phase, one can run malicious code in ML artifacts or software. To ensure the persistence of the attack, maintain some artifacts on the victim system, or keep access to the system, adversaries may introduce a backdoor into the ML model. Inserting a backdoor trigger ensures that the vulnerability can be activated at a later time by specific data samples. The privilege escalation tactics are most suitable in LLM-enabled systems, where hostile actors can gain higher-level permissions on a system performing techniques like LLM Prompt (Direct, Indirect), LLM Plugin Compromise, and LLM jailbreak. Special actions are performed to evade the defense and avoid attack detection at the ML level (e.g. against malware detection). Keylogging or credential dumping techniques are used to steal credentials, such as account names and passwords, in the Credential Access phase. The attack on the ML model can include proxy models, poisoning the target model, or generating adversarial data to feed the target model. From ML's perspective, in Exfiltration the adversary is trying to obtain machine learning artifacts or other information on the ML-enabled System. Finally, by disrupting availability and integrity and interfering with business and operational processes, unfriendly entities act by manipulating, interrupting, misleading, or destroying machine learning systems and data.

B. Adversarial attacks taxonomy

Adversarial attacks are responsible for many of the AI incidents above and are to be categorized in several ways, considering the attacker's information and goal, the place in the ML system, and the type of model being targeted. Considering the attacker's knowledge there are white-box, black-box, and grey-box attacks. The White-box attacks scenario [12][13] is where hostile actors have information and access to the AI/ML model and know the configuration, parameters, and data used to train it.

In the Black-box attacks[14][15][16][17][18], hostile actors don't have prior information on the AI/ML model and don't know the configuration, parameters, and data used to train it. In this scenario, the attackers rely only on the inputs and outputs behaviors of the system (e.g. through the model's API). In transferability Attacks, also known as Grey-box attacks, the attacker leverages the transferability of adversarial examples, exploiting shared vulnerabilities across different ML models. This involves using adversarial examples created to deceive one specific machine learning (ML) model and mislead other ML models.

From the attacker's goal perspective, we can talk about targeted and untargeted attacks. While the untargeted actions aim for example just to misclassify the data and maximize the classification error, the targeted ones try to mislead the ML prediction into a specific way or outcome. The location in which the attack in the ML application pipeline is being performed can also be a way to classify focus on manipulating the training data, through poisoning or training attacks, on creating a backdoor in the ML model, or targeting the deployed models and applying inference or evasion attacks. The goal of poisoning [19][20] is to infect the training data to compromise the model's integrity early in the learning phase. The backdoor approach [21][22] is to implant a specific trigger pattern that can be activated at need. Taking advantage of the model's vulnerabilities, the inference and evasion attacks generate adversarial examples that are hard to detect but force the model to produce incorrect results.

If we were to take into consideration the targeted AI model we could split the ML attacks into DNN-based and other types of models. There are several studies and incidents based on algorithms and models that use Naive Bayes, Logistic Regretion, Decision Tree, SVM, PCA/LASSO, clustering, Graph Neural Networks [23], Binarized Neural Networks (BNNs) [24], Spiking Neural Networks [25][26], cellular neural networks [27], Diffractive Deep Neural Network [28].

The popularity of the DNNs makes these models more prone to adversarial attacks. The DNN domain is mostly represented by the Convolutional Neural Networks - CNN (Yolo [29], stochastic CNNs [30]) and the GAI (transformer-based neural networks [31], generative adversarial network (GAN) [32], LLM[33][34]) subdomains.

Attacks don't occur solely in the digital space as adversaries may also act directly, to manipulate the physical environment, for ML systems that capture and use real-world data. As digital attacks imply the adversary has direct access to the data fed into the model, physical ones don't have knowledge about the digital representation of the data, and the model is directly fed with sensor inputs (e.g. images from video cameras and microphones).

C. Generating adversarial samples

Several methods for generating adversarial examples have been proposed in different studies:

- BFGS or L-BFGS is one of the first types of attack. It was exemplified by Szegedy et al. [7] and uses the

Limited Memory Broyden–Fletcher–Goldfarb–Shanno L-BFGS optimization algorithm. Though efficient, it uses an expensive linear search method to find the optimal value, which is time and computational-consuming. Further, Zhang, Jiebao, et al. [35] propose incorporating the perturbation pixel selection strategy into Limited Pixel - BFGS (LP-BFGS);

- Fast Gradient Sign Attack technique (FGSM)[6][36] was introduced by Ian Goodfellow et. all, and is a simple yet efficient technique for generating adversarial examples in white-box attacks. The method implies computing the loss, the gradient of the image, and based on that, slightly modifying the image pixels in the gradient's direction to maximize the loss. There are lots of optimization based on FGSM: Complex-FGSM [37], Basic Iterative Method (BIM)- Iterative FGSM (IFGSM), Momentum Iterative-FGSM (MI-FGSM [38]), Adaptive FGSM (Ada-FGSM[39], Adam-FGSM [40]), and F-MIFGSM [41].
- DeepFool[42], based on an iterative linearization of the classifier, can also be used for computing adversarial examples and generating minimal perturbations that are sufficient to change the classification labels of state-of-the-art classifiers.
- Projected Gradient Descent (PGD)[43][44][45] generates adversarial examples by applying small but iteratively adjusted perturbations to the input data, to maximize the model's prediction error. Compared to similar tactics, this method is more effective, more complex has a more fine-grained control, is more flexible and it's able to bypass simple defense mechanisms.
- CW2(Carlini and Wagner)[46][43] attack refines the adversarial example through multiple iterations. The method manages to generate small perturbations with high impact, making it suitable for inconspicuous real-world scenarios
- Jacobian-based saliency map (JSMA)[47] is a targeted attack that iteratively saturates a few pixels in an image to their maximum/minimum values. Some improved versions of the JSMA are Weighted JSMA (WJSMA)[47], Taylor JSMA (TJSMA)[48], Euclidean Jacobian-based Saliency Maps Attack(EJSMA) and Rapid Jacobian-based Saliency Maps Attacks(RJSMA) [49];
- SparseFool[50] a geometry-inspired sparse attack that exploits the low mean curvature of the decision boundary that is fast and can scale to high dimensional data.
- Edge Manipulation method via Hierarchical Deep Reinforcement Learning (EMHDRL)[23] is a method used against graph neural networks (GNNs) that has strong transferability, improves the imperceptibility of GNNs attacks by assuring the whole graph remains unchanged, and reduces the performance of GNNs in classifying.
- Like noise-based attacks, Generative Adversarial Networks (GANs)[51][52][53][54] create adversarial samples by generating perturbation to input images to force the model to misclassify to a specific target class.
- Zeroth-Order Optimisation Attack (ZOO)[55] is a black-box attack inspired by the CW. The attack uses the zeroth-order optimization for the input image and the confidence score vector output of the black-box DNN.
- One pixel [56] attack is a method for generating one-pixel adversarial perturbations. This method is suitable for black-box attacks since it requires modifying just one pixel in the input image.
- Adaptive Decay Attack (ADA) is a method proposed in speaker recognition attack [57] which is stealthiness is very close to the CW2(Carlini and Wagner), with much less computation time than CW2.
- The quantum adversarial sample generation algorithm (QASGA) [58] encodes real samples into quantum states and superimposes them with the generated perturbations, using the qGAN [59][60], to build the adversarial samples.
- Generating multiple transformed images by randomly rotating the original images enables the rotation model enhancement algorithm [61] to craft adversarial examples with a single model, and boosts attacks on multiple models, increasing the transferability and success rate for black-box attacks.
- Adversarial Transformation Networks (ATNs) [62], [63] are a type of neural network that generates adversarial example against a target network or set of networks that can be trained in a black-box or white-box manner.

D. Mitigation

Defense mechanisms: Akhtar and Mian [64] proposed three groups for categorizing defenses against adversarial attacks:

- methods that modify the target models for robustness,
- methods that alter inputs to remove perturbations (Defensive Distillation, GAN, Autoencoders) and
- methods that integrate external modules into the model.

Among some of the most popular methods for enforcing the robustness of the AI models against adversarial attacks are:

- Defense-GAN is a framework trained to model the distribution of unperturbed images and can find a close output to a given image that does not contain adversarial changes.
- DRAGAN [65] (Deep Regret Analytic Generative Adversarial Networks) uses GANs to resist attacks in image classification tasks. Using an improved version of the Defense-GAN, the method focuses on training the GAN on unperturbed images. It then uses the GAN to reconstruct the input images before sending them to the classifier.
- Using Convolutional Auto-Encoders, that effectively counter adversarial perturbations introduced to the input images is possible to enhance the robustness of targeted classifier models against adversarial attacks [66].
- Purifying Variational Autoencoder (PuVAE) [67] is a method that is being proposed to clean adversarial examples. The performance and robustness of PuVAE have been tested against various attack methods. The test

results indicate that the method is competitive with state-of-the-art solutions and 100 times faster than the Defense-GAN.

- Enhancing the model's robustness with adversarial training by using adversarial examples among the training dataset is a common technique employed in different applications [68], [69], [70].
- Using approximate computing based on pseudo-random bit-streams, stochastic computing (SC) [71], [72] is presented as a novel mechanism to build (SCNNs) and fortify NN models against adversarial attacks.
- The generative framework, Evasion Vaccination (EVAX) [73] is an accurate detector that is not fooled by evasive attacks and can generalize to novel zero-day attacks.

IV. Next steps

Based on the actual state, there are several ways to contribute to research in the AI vulnerabilities field.

The next step could be a state-of-the-art survey on adversarial attacks, both offensive and defensive with digital and psychical applications. The research may also contain a comparative benchmarking on the latest adversarial attack methods based on their success rate, transferability, or stealthiness on the latest AI open-source computer vision object detection models.

Alternatively we could be researching the impact of adversarial attacks on trend or promising technologies and applications. The repercussions that adversarial attacks may have on continuous learning models are startling. Studding and enforcing these models, especially in real life application scenarios would have clear advantages.

Further research of impact of adversarial attacks on multimodal large language models or vision language models(VLM) would also be of great use as application based on these technologies are on growing trend.

All course of action above can have a focus on psychical world scenarios with real-live application implications.

References

[1] https://artificialintelligenceact.eu/ai-act-explorer/

[2] AI Incident Database," AI Incidents. [Online]. Available: https://incidentdatabase.ai/

[3] S. McGregor, "Preventing repeated real world ai failures by cataloging incidents: The ai incident database," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 17, 2021, pp. 15 458–15 463.

[4] Pittaras, Nikiforos, and Sean McGregor. "A taxonomic system for failure cause analysis of open source AI incidents." arXiv preprint arXiv:2211.07280 (2022).

[5] "AI Incident Database," AI Incidents. [Online]. Available: https://incidentdatabase.ai/

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, arXiv:1412.6572.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proceedings of the 2014 International Conference on Learning Representations. Computational and Biological Learning Society, 2014.

[8] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in Proceedings of the 1st IEEE European Symposium on Security and Privacy. IEEE, 2016.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proceedings of the 2015 International Conference on Learning Representations.

[10] A. Kurakin, I. J. Goodfellow, S. Bengio, Adversarial examples in the physical worls, workshop track - ICLR 2017.

[11] "Altas framework," MITRE Corporation. [Online]. Available: https://atlas.mitre.org/matrix

[12] M. Azadmanesh, B. S. Ghahfarokhi and M. A. Talouki, "A White-Box Generator Membership Inference Attack Against Generative Models," 2021 18th International ISC Conference on Information Security and Cryptology (ISCISC), Isfahan, Iran, Islamic Republic of, 2021, pp. 13-17, doi: 10.1109/ISCISC53448.2021.9720436.

[13] I. Bankole-Hameed, A. Parikh and J. Harguess, "Exploring the Effect of Adversarial Attacks on Deep Learning Architectures for X-Ray Data," 2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), DC, USA, 2022, pp. 1-9, doi: 10.1109/AIPR57179.2022.10092220.

[14] H. Zhang, Z. Li, H. Liu, B. Yang, C. Li and J. Wang, "Rotation Model Enhancement for Adversarial Attack," 2022 International Conference on Networking and Network Applications (NaNA), Urumqi, China, 2022, pp. 431-436, doi: 10.1109/NaNA56854.2022.00080.

[15] Adv-Makeup: A New Imperceptible and Transferable Attack on Face Recognition Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, Cong Liu, 2021, International Joint Conference on Artificial Intelligence

[16] K. Agrawal and C. Bhatnagar, "A Black-box based Attack Generation Approach to Create the Transferable Patch Attack," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 1376-1380, doi: 10.1109/ICICCS56967.2023.10142656

[17] J. a. Yu and L. Peng, "Black-box Attacks on DNN Classifier Based on Fuzzy Adversarial Examples," 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2020, pp. 965-969, doi: 10.1109/ICSIP49896.2020.9339329.

[18] K. Mahmood, P. H. Nguyen, L. M. Nguyen, T. Nguyen and M. Van Dijk, "Besting the Black-Box: Barrier Zones for Adversarial Example Defense," in IEEE Access, vol. 10, pp. 1451-1474, 2022, doi: 10.1109/ACCESS.2021.3138966.

[19] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice,V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in Proc. 10th ACM Workshop Artif. Intell. Secur., B. M. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, and A. Sinha, Eds. Dallas, TX, USA, Nov. 2017, pp. 27–38.

[20] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! Targeted clean-label poisoning attacks on neural networks," in Proc. Annu. Conf. Neural Inf. Process. Syst., S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Dec. 2018, pp. 6106–6116.

[21] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, arXiv:1712.05526.

[22] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015.

[23] G. Sun, J. Ding and X. Li, "EMHDRL: Adversarial Attacks on Graph Neural Networks via Hierarchical Reinforcement Learning," 2023 42nd Chinese Control Conference (CCC), Tianjin, China, 2023, pp. 8745-8750, doi: 10.23919/CCC58697.2023.10239808.

[24] V. -N. Dinh, N. -M. Bui, V. -T. Nguyen, K. -S. Nguyen, Q. -M. Duong and Q. -K. Trinh, "A Study on Adversarial Attacks and Defense Method on Binarized Neural Network," 2022 International Conference on Advanced Technologies for Communications (ATC), Ha Noi, Vietnam, 2022, pp. 304-309, doi: 10.1109/ATC55345.2022.9943040.

[25] S. Sharmin, P. Panda, S. S. Sarwar, C. Lee, W. Ponghiran and K. Roy, "A Comprehensive Analysis on Adversarial Robustness of Spiking Neural Networks," 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8851732.

[26] M. Leontev, D. Antonov and S. Sukhov, "Robustness of spiking neural networks against adversarial attacks," 2021 International Conference on Information Technology and Nanotechnology (ITNT), Samara, Russian Federation, 2021, pp. 1-6, doi: 10.1109/ITNT52450.2021.9649179.

[27] A. Horváth, "On The Resilience of Cellular Neural Networks to Low-intensity Adversarial Attacks," 2021 17th International Workshop on

Cellular Nanoscale Networks and their Applications (CNNA), Catania, Italy, 2021, pp. 1-4, doi: 10.1109/CNNA49188.2021.9610769.

[28] Y. Li and C. Yu, "Late Breaking Results: Physical Adversarial Attacks of Diffractive Deep Neural Networks," 2021 58th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2021, pp. 1374-1375, doi: 10.1109/DAC18074.2021.9586204.

[29] Nganyewou Tidjon, Lionel and Khomh, Foutse. (2022). Threat Assessment in Machine Learning based Systems. 10.48550/arXiv.2207.00091.

[30] F. Neugebauer, V. Vekariya, I. Polian and J. P. Hayes, "Stochastic Computing as a Defence Against Adversarial Attacks," 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), Porto, Portugal, 2023, pp. 191-194, doi: 10.1109/DSN-W58399.2023.00053.

[31] L. Zhang, S. Lambotharan and G. Zheng, "Adversarial Learning in Transformer Based Neural Network in Radio Signal Classification," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 9032-9036, doi: 10.1109/ICASSP43922.2022.9747508.

[32] M. Azadmanesh, B. S. Ghahfarokhi and M. A. Talouki, "A White-Box Generator Membership Inference Attack Against Generative Models," 2021 18th International ISC Conference on Information Security and Cryptology (ISCISC), Isfahan, Iran, Islamic Republic of, 2021, pp. 13-17, doi: 10.1109/ISCISC53448.2021.9720436.

[33] Su, Jingtong, Julia Kempe, and Karen Ullrich. "Mission impossible: A statistical perspective on jailbreaking llms." arXiv preprint arXiv:2408.01420 (2024).

[34] Geisler S, Wollschläger T, Abdalla MH, Gasteiger J, Günnemann S. Attacking large language models with projected gradient descent. arXiv preprint arXiv:2402.09154. 2024 Feb 14.

[35] Zhang, Jiebao, et al. "LP-BFGS attack: An adversarial attack based on the Hessian with limited pixels." Computers and Security 140 (2024): 103746.

[36] J. Li, "Analyse of Influence of Adversarial Samples on Neural Network Attacks with Different Complexities," 2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM), Montreal, ON, Canada, 2022, pp. 329-333, doi: 10.1109/ISPCEM57418.2022.00072.

[37] Y. Li and C. Yu, "Late Breaking Results: Physical Adversarial Attacks of Diffractive Deep Neural Networks," 2021 58th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2021, pp. 1374-1375, doi: 10.1109/DAC18074.2021.9586204.

[38] Y. Dong et al., "Boosting Adversarial Attacks with Momentum," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 9185-9193, doi: 10.1109/CVPR.2018.00957. keywords: Iterative methods;Robustness;Training;Data models;Adaptation models;Security,

[39] Shi, Yucheng and Han, Yahong and Zhang, Quanxin and Kuang, Xiaohui. (2020). Adaptive Iterative Attack towards Explainable Adversarial Robustness. Pattern Recognition. 105. 107309. 10.1016/j.patcog.2020.107309.

[40] Zhang, Jiebao, Wenhua Qian, Rencan Nie, Jinde Cao, and Dan Xu. "Generate adversarial examples by adaptive moment iterative fast gradient sign method." Applied Intelligence 53.1 (2023): 1101-1114.

[41] S. Liu, Z. Zhang, X. Zhang and H. Feng, "F-MIFGSM: adversarial attack algorithm for the feature region," 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2020, pp. 2164-2170, doi: 10.1109/ITAIC49862.2020.9338937.

[42] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2574–2582.

[43] Arindam Jati, Chin-Cheng Hsu, Monisankha Pal, Raghuveer Peri, Wael AbdAlmageed, Shrikanth Narayanan, "Adversarial attack and defense strategies for deep speaker recognition systems", Computer Speech and Language, Volume 68, 2021, 101199, ISSN 0885-2308,

[44] Geisler S, Wollschläger T, Abdalla MH, Gasteiger J, Günnemann S. Attacking large language models with projected gradient descent. arXiv preprint arXiv:2402.09154. 2024 Feb 14.

[45] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.Towards deep learning models resistant to adversarial attacks.International Conference on Learning Representations, ICLR, pp. 1–28, 2018.

[46] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." 2017 ieee symposium on security and privacy (sp). Ieee, 2017.

[47] Wiyatno, Rey, and Anqi Xu. "Maximal jacobian-based saliency map attack." arXiv preprint arXiv:1808.07945 (2018).

[48] Combey, Théo, et al. "Probabilistic jacobian-based saliency maps attacks." Machine learning and knowledge extraction 2.4 (2020): 558-578

[49] Y. Ling, Z. Yong and W. Pengfei, "Euclidean and Rapid Jacobian-based Saliency Maps Attacks," 2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Chengdu, China, 2021, pp. 355-361, doi: 10.1109/ISKE54062.2021.9755384.

[50] Modas, Apostolos, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. "Sparsefool: a few pixels make a big difference." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[51] Chenhan Zhang, Shui Yu, Zhiyi Tian, and James J. Q. Yu. 2023. Generative Adversarial Networks: A Survey on Attack and Defense Perspective. ACM Comput. Surv. 56, 4, Article 91 (April 2024), 35 pages. https://doi.org/10.1145/3615336

[52] Xiao, Chaowei, et al. "Generating adversarial examples with adversarial networks." arXiv preprint arXiv:1801.02610 (2018)

[53] Q. Zhang, J. Yang, X. Zhang and T. Cao, "Generating Adversarial Examples in Audio Classification with Generative Adversarial Network," 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China, 2022, pp. 848-853, doi: 10.1109/ICIVC55077.2022.9886154.

[54] C. -S. Shieh et al., "Synthesis of Adversarial DDoS Attacks Using Wasserstein Generative Adversarial Networks with Gradient Penalty," 2021 6th International Conference on Computational Intelligence and Applications (ICCIA), Xiamen, China, 2021, pp. 118-122, doi: 10.1109/ICCIA52886.2021.00030.

[55] Chen, Pin-Yu, et al. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models." Proceedings of the 10th ACM workshop on artificial intelligence and security. 2017

[56] Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." IEEE Transactions on Evolutionary Computation 23.5 (2019): 828-841.

[57] X. Zhang, Y. Xu, S. Zhang and X. Li, "A Highly Stealthy Adaptive Decay Attack Against Speaker Recognition," in IEEE Access, vol. 10, pp. 118789-118805, 2022, doi: 10.1109/ACCESS.2022.3220639.

[58] W. Cheng, S. Zhang and Y. Lin, "Study on the Adversarial Sample Generation Algorithm Based on Adversarial Quantum Generation Adversarial Network," 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS), Chengdu, China, 2023, pp. 238-243, doi: 10.1109/ISCTIS58954.2023.10213103.

[59] Lloyd, S., and Weedbrook, C. (2018). Quantum Generative Adversarial Learning. Physical review letters, 121 4, 040502 .

[60] Dallaire-Demers, P.,and Killoran, N. (2018). Quantum generative adversarial networks. ArXiv, abs/1804.08641

[61] H. Zhang, Z. Li, H. Liu, B. Yang, C. Li and J. Wang, "Rotation Model Enhancement for Adversarial Attack," 2022 International Conference on Networking and Network Applications (NaNA), Urumqi, China, 2022, pp. 431-436, doi: 10.1109/NaNA56854.2022.00080.

[62] Baluja, Shumeet, and Ian Fischer. "Adversarial transformation networks: Learning to generate adversarial examples." arXiv preprint arXiv:1703.09387 (2017).

[63] Baluja, Shumeet, and Ian Fischer. "Learning to attack: Adversarial transformation networks." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[64] Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." Ieee Access 6 (2018): 14410-14430.

[65] A. ArjomandBigdeli, M. Amirmazlaghani and M. Khalooei, "Defense against adversarial attacks using DRAGAN," 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), Mashhad, Iran, 2020, pp. 1-5, doi: 10.1109/ICSPIS51611.2020.9349536.

[66] Mandal, Shreyasi. "Defense Against Adversarial Attacks using Convolutional Auto-Encoders." arXiv preprint arXiv:2312.03520 (2023).

[67] Hwang, Uiwon, et al. "Puvae: A variational autoencoder to purify adversarial examples." IEEE Access 7 (2019): 126582-126593.

[68] K. Roshan, A. Zafar and S. B. Ul Haque, "A Novel Deep Learning based Model to Defend Network Intrusion Detection System against Adversarial Attacks," 2023 10th International Conference on Computing

for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 386-391

[69] Xhonneux, Sophie, et al. "Efficient adversarial training in llms with continuous attacks." arXiv preprint arXiv:2405.15589 (2024).

[70] lex Lamb, Vikas Verma, Kenji Kawaguchi, Alexander Matyasko, Savya Khosla, Juho Kannala, Yoshua Bengio, "Interpolated Adversarial Training: Achieving robust neural networks without sacrificing too much accuracy", Neural Networks, Volume 154, 2022, Pages 218-233, ISSN 0893-6080, https://doi.org/10.1016/j.neunet.2022.07.012.

[71] F. Neugebauer, V. Vekariya, I. Polian and J. P. Hayes, "Stochastic Computing as a Defence Against Adversarial Attacks," 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), Porto, Portugal, 2023, pp. 191-194, doi: 10.1109/DSN-W58399.2023.00053.

[72] Banitaba, Faeze S., Sercan Aygun, and M. Hassan Najafi. "Late Breaking Results: Fortifying Neural Networks: Safeguarding Against Adversarial Attacks with Stochastic Computing." arXiv preprint arXiv:2407.04861 (2024).

[73] Samira Mirbagher Ajorpaz, Daniel Moghimi, Jeffrey Neal Collins, Gilles Pokam, Nael Abu-Ghazaleh, and Dean Tullsen. 2023. Evax: Towards a Practical, Pro-Active; Adaptive Architecture for High Performance ; Security. In Proceedings of the 55th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '22). IEEE Press, 1218–1236. https://doi.org/10.1109/MICRO56248.2022.00085.