ISSN 0361-7688, Programming and Computer Software, 2023, Vol. 49, Suppl. 2, pp. S115–S121. © Pleiades Publishing, Ltd., 2023.
Russian Text © The Author(s), 2023, published in Journal of Information Technologies and Computing Systems, 2023, No. 4.

Analysis of Vulnerabilities of Neural Network Image Recognition Technologies

A. V. Trusov[a,b,*], **E. E. Limonova**[a,b,**], **V. V. Arlazarov**[a,b,***], **and A. A. Zatsarinnyy**[a,****]

[a] *Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia*

[b] *Smart Engines Service LCC, Moscow, Russia*

**e-mail: trusov.anton.02@gmail.com*

***e-mail: elena.e.limonova@gmail.com*

****e-mail: vva777@gmail.com*

*****e-mail: azatsarinny@ipiran.ru*

Received October 9, 2023; revised October 15, 2023; accepted November 3, 2023

Abstract—The problem of vulnerability of artificial intelligence technologies based on neural networks is considered. It is shown that the use of neural networks generates a lot of vulnerabilities. Examples of such vulnerabilities are demonstrated, such as incorrect classification of images containing adversarial noise or patches, failure of recognition systems in the presence of special patterns in the image, including those applied to objects in the real world, training data poisoning, etc. Based on the analysis, the need to improve the security of artificial intelligence technologies is shown, and some considerations that contribute to this improvement are discussed.

Keywords: neural networks, adversarial attacks on neural networks, adversarial images, neural network security

DOI: 10.1134/S0361768823100079

1. INTRODUCTION

Over the last decade, the use of artificial intelligence (AI) systems based on deep neural networks (DNN) in various fields of activity has become commonplace both in many countries around the world and in Russia. This is due to the fact that DNNs are quite effective tools for modeling complex objects and processes that cannot be described using classical mathematical models and methods. Thus, DNNs have become an indispensable tool for solving many computer vision problems (e.g., classification and comparison of images, semantic segmentation, object detection, and many others), and are widely used in complex software systems, including but not limited to, person identification systems [1], recognition of documents [2], and autonomous driving [3].

However, despite all the advantages, DNNs also have a number of fundamental drawbacks that can lead to disruptions in the operation of AI-based information systems. Thus, among the potential vulnerabilities of DNNs, there are the possibility of poisoning the training dataset (adding specially prepared images to it) [4] to create backdoors, i.e., vulnerabilities that an attacker can take advantage of during the execution of the DNN; the ability to invert the model to extract data about the training set or the architecture of model [5]; all sorts of ways to distort the input signal in order to force the neural network to produce an incorrect answer, even in those examples in which a person does not notice the distortions [6]. And this is not a complete list of DNN vulnerabilities. Essentially, the use of DNNs in software systems leads to the emergence of new vulnerabilities that are not always obvious from the standpoint of the classical paradigm of ensuring and controlling information security. It is important that systems using DNNs often process private and personal data and therefore their incorrect operation can have a significant negative impact on the safety of specific people.

At the same time, at present, according to our opinion, both users of systems with DNNs and developers of such systems have little understanding of the range of potential threats generated by the use of these technologies. And one of the reasons is the lack of scientific research in this area.

This paper is devoted to the analysis of some aspects of the use of DNNs from the standpoint of their vulnerability, and it makes it possible, to a certain extent, to form a more realistic view of this important problem.

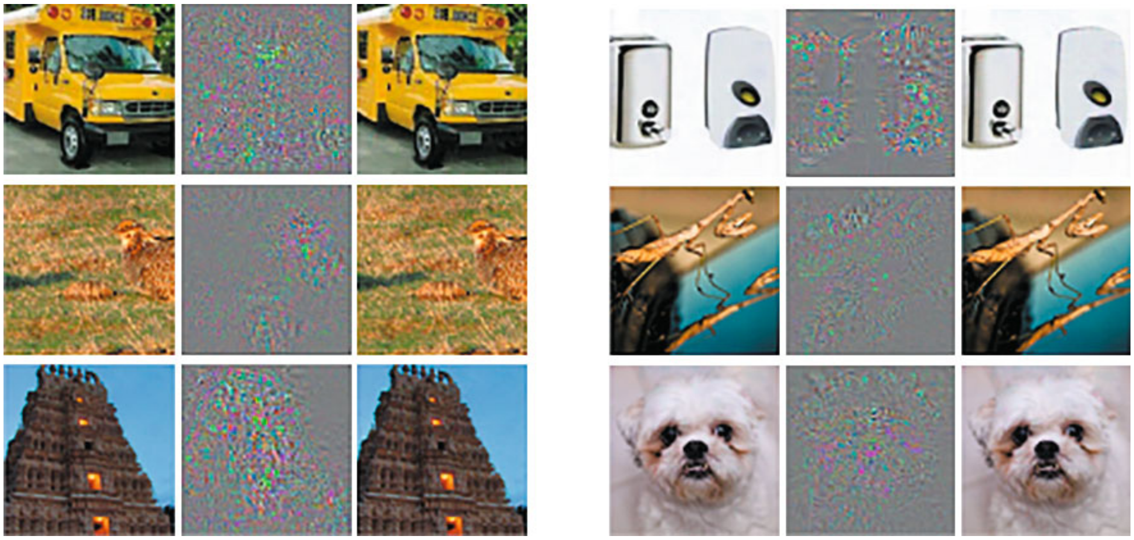

Fig. 1. Deceiving the AlexNet network using special noise.

2. EXAMPLES OF VULNERABILITY OF NEURAL NETWORKS

In the opinion of many researchers, DNNs are a very convenient, available, and effective tool for approximating complex objects and processes. Unfortunately, neural networks, like any approximation, are not perfect. They have a number of limitations that attackers can use to influence the results produced by a network. Let us look at some examples based on an analysis of the literature.

Adding small specially generated noise can lead to incorrect image recognition, even if this noise is invisible to the human eye. For example, the left columns in Fig. 1 show images from the ImageNet dataset (1000-class classification problem) that are correctly recognized by the AlexNet neural network, but after adding specially generated "noise" from the central column (in the figure, the noise is amplified 10 times for clarity), all images are recognized as belonging to the same class *Ostrich,* although visually they have nothing in common with this bird [6].

A specially prepared sticker can fool neural network classifiers even if such a sticker is printed and photographed in the real world. For example, Figure 2 shows a sticker that "turns" a banana into a toaster, according to the VGG16 neural network [7].

To fool facial recognition systems, you can use special makeup [8] or photo transformations that are visually similar to applying makeup [9], as well as various types of masks [10]. Figure 3 shows such an attack: a person without a mask (left picture) and in a regular medical mask (central) is recognized correctly, but in a mask with a special pattern (right picture) the person is not recognized, although the part of the face overlapped by the mask is the same as is the case with a medical mask. Specific infrared lighting devices [11] or clothing with special patterns [12] and patches [13] can also be used to fool automatic person recognition in CCTV systems. For example, Figure 4 shows a sweater with a special coloring thanks to which the person wearing it is not detected by the neural network unlike most other people in the same image [12]. Recent research suggests that distortion does not always have to make noticeable changes to the image or be similar to high-frequency noise, which is usually obtained as result of gradient optimization of the input image with respect to the adversarial objective function. Distortion can affect only individual high-level features of the image [14]. Figure 5 shows an example of such distortion (right) created using a diffusion DNN.

Along with the above examples, we note that information about the architecture and parameters of the neural network, which has significant commercial value, can be recovered and stolen by an attacker if he has physical access to the equipment on which the neural network is executed [15–17]. Moreover, if an attacker has access to the training set (even a small part of it), he can poison this sample by adding special examples to it [4]. Training on poisoned data is dangerous because the resulting neural network may contain backdoors [18] or work incorrectly on certain input images [19]. Finally, there is a threat to data privacy since training data can be extracted either from an already trained neural network [5] or during its distributed training [20].

3. ATTACKS AGAINST NEURAL NETWORKS

The actions described above (using some examples) to find and exploit neural network vulnerabilities are usually called *attacks against neural networks,* and the images on which networks give a wrong answer are

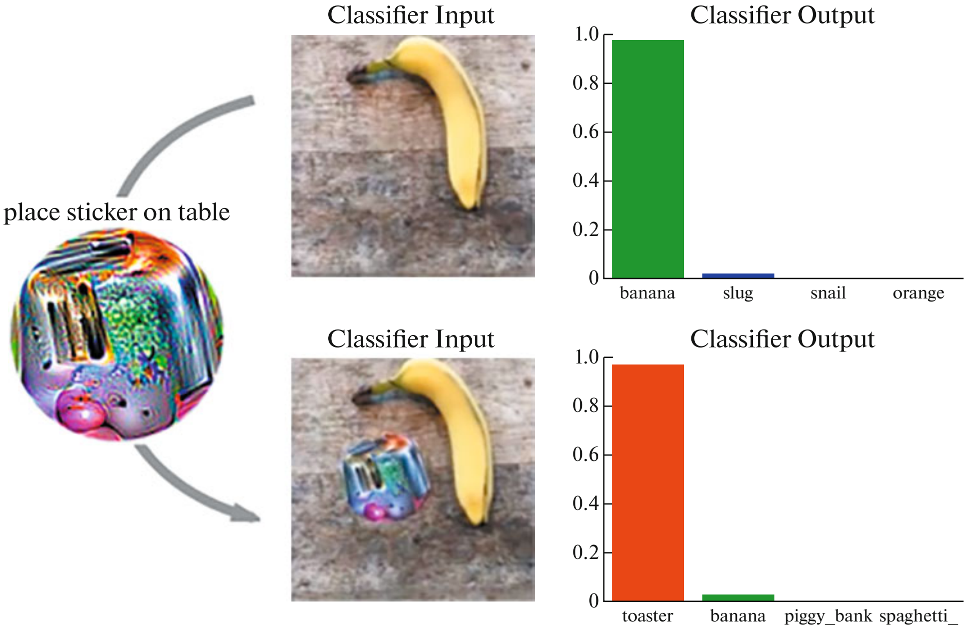

Fig. 2. An attack against the VGG16 model using a sticker (left).

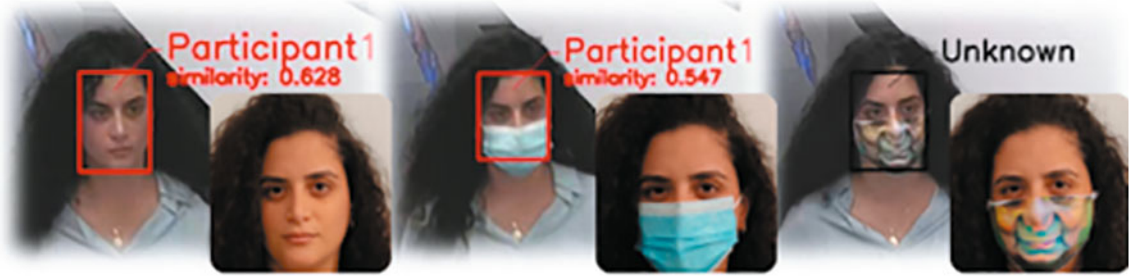

Fig. 3. A mask that prevents automatic facial recognition (right).

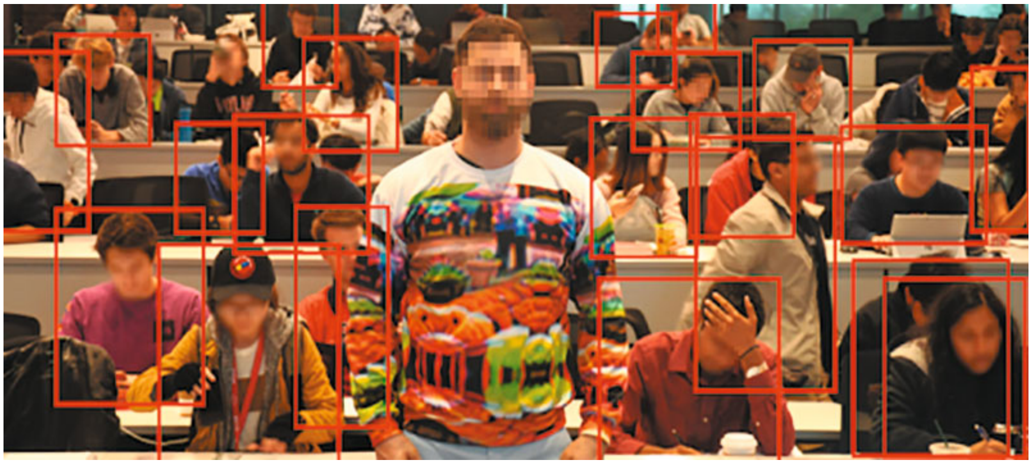

Fig. 4. An invisibility sweater (in the center) that deceives the human detection DNN.

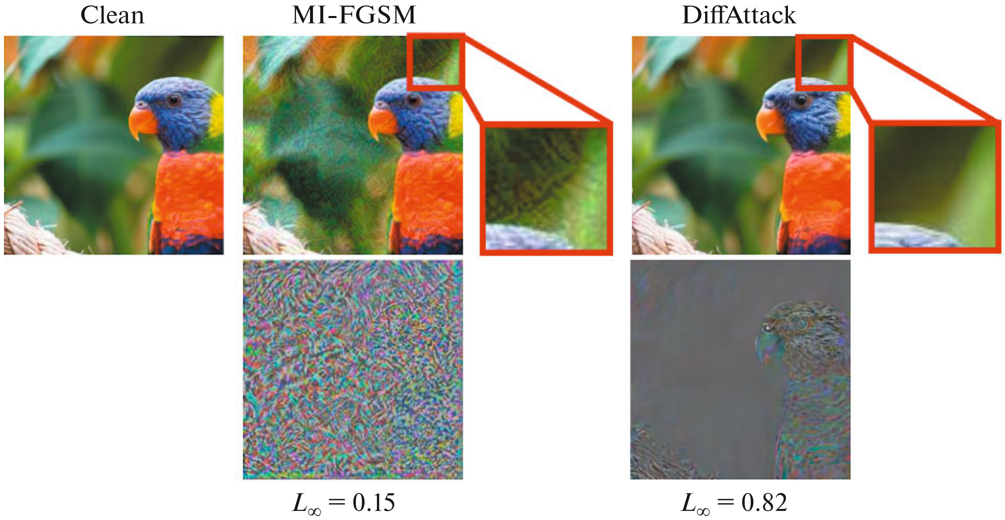

Fig. 5. Attacks against parrot images (clean left), gradient optimization (middle), and diffuse DNN (right).

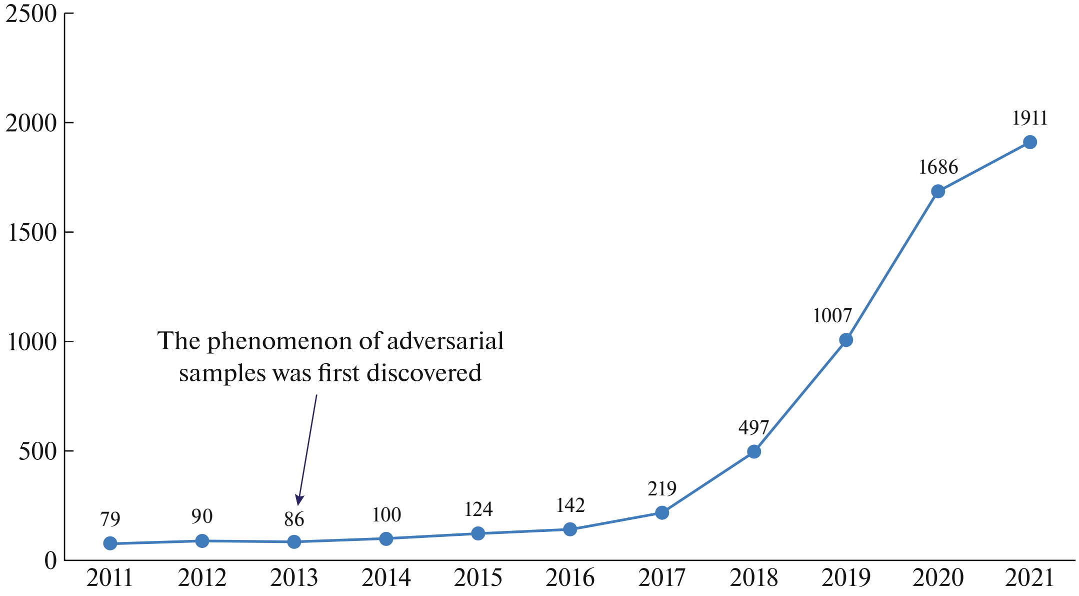

Fig. 6. Number of publications devoted to attacks against machine learning algorithms.

called *adversarial images* [21]. There are many different methods and algorithms that are used to prepare attacks against neural networks, and their number is growing every day. This is evidenced by publications in the Scopus database devoted to attacks against machine learning algorithms. The number of such publications is constantly growing. This number depending on the year of publication according to the study [24] is presented in Fig. 6 (the arrow in the figure indicates the time of detection of the existence of adversarial images that fool neural networks). For example, a very detailed review of attacks against neural networks in computer vision published in 2018 [22] mentions 12 different attacks; another review from 2021 [23] mentions 33 attacks; and the paper [24] published in 2022 lists more than 50 attacks organized in accordance with their taxonomy. This taxonomy itself is quite complex and is evolving rapidly.

As new methods of attacking neural networks develop, means of protection against them are also being developed. A 2021 study mentions more than 60 different security algorithms [25]. Currently, the race between attacks and protection tools is far from over. Thus, any heuristic protection (protection based on applying an attack to data during training to improve the stability of the DNN) resists a given attack well,

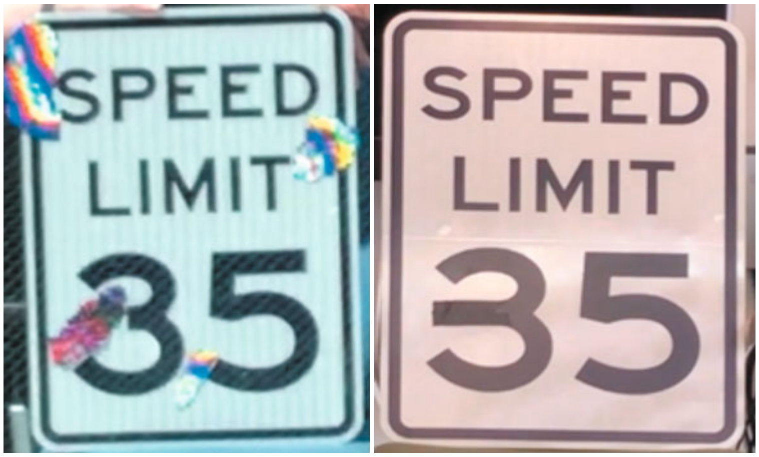

Fig. 7. Stickers that "turn" 35 mph into 85.

but is easily circumvented using other attacks, and guaranteed (certified) protection tools have three significant drawbacks:

- they reduce the quality of neural networks;
- they significantly complicate training (in the sense of computational complexity and convergence rate);
- they ensure protection only against small image distortions in a certain metric space, while attackers can use various metrics [25].

Attacks against neural networks can be carried out not only in computer vision systems but also in speech recognition [26], text analysis [27], reinforcement learning [28], instruction recognition in Internet of Things (IoT) systems [29], and any other areas where neural networks are used.

Unfortunately, the current state of the industry demonstrates insufficient understanding of the threat posed by attacks on machine learning algorithms (and neural networks, in particular), and ignorance of ways to protect against them. For example, 22 out of 25 European and American organizations that use various machine learning algorithms in their products, which in recent years have become fashionable to call AI algorithms, do not have the necessary tools to protect their machine learning systems and are in need of guidance [30].

4. DISCUSSION OF RESULTS AND RECOMMENDATIONS

Thus, attacks against machine learning algorithms and especially neural networks have remained a relevant and widely discussed topic in the scientific community over the past decade. However, this topic has not yet received due attention in the development of practical applications. Despite the fact that the threat to AI systems is mentioned in the review of the problems of deploying software systems based on machine learning [31], it does not contain specific recommendations for protecting against these threats. Some high-level recommendations can be found in [32, 33]. Here are some of them:

- Analyze possible vulnerabilities of AI systems and the corresponding attack techniques.
- Test the system for resistance to attacks.
- Engage specialists for controlled hacking of the system and identifying its weak points.
- Inform users of AI systems about potential threats to their security due attacks.
- Ensure the security of private data if it was used for training neural networks.
- Improve communication between researchers working on neural network attacks, neural network resilience, and those implementing AI-based systems.

Even though these recommendations may be useful, they do not seem practical because they do not provide a clear and straightforward answer to the question of how to secure an AI application system against malicious attacks. Many industry representatives remain unaware of the existence of neural network vulnerabilities and the threats they pose. Any regulations regarding the security of neural networks against possible attacks have not yet been formed or have not received widespread use.

However, interest from society and industry in this topic is also beginning to grow. There are examples of clothing camouflage to protect against automated identification systems (e.g., the Italian brand Cap_able [34]) and image distortion to prevent automated image theft (e.g., the Russian classified ads site Avito [35]). Large cyber security companies are beginning to examine AI-powered systems for potential vulnerabilities. For example, McAffe spent 18 months hacking and analyzing Tesla and MobileEye systems for autonomous vehicles [37]. They showed how various stickers attached to road signs could fool these systems. An example of such stickers is shown in Fig. 7. On the left are those that fooled MobileEye, on the right are the stickers that fooled Tesla. Both systems recognized these signs as the speed limit being 85 mph instead of 35.

Despite all this, practical attacks on neural networks in computer vision are not yet very common.

5. CONCLUSIONS

Results of an analysis of a new class of vulnerabilities in systems using deep neural networks that can arise at different stages from training to practical application are presented. Examples of various attacks are described that significantly affect recognition results both in digital systems and in the real world. It is shown that the issues of information security and stability of AI systems are extremely relevant today. The paper presents for discussion general considerations and recommendations that can improve the security of operation of the systems based on deep neural networks.

FUNDING

This work was supported by ongoing institutional funding. No additional grants to carry out or direct this particular research were obtained.

CONFLICT OF INTEREST

The author declares that he have no conflicts of interest.

REFERENCES

1. Ye, M. et al., Deep learning for person re-identification: A survey and outlook, *IEEE Trans. Pattern Anal. Mach. Intell.,* 2021, vol. 44, no. 6, pp. 2872–2893.
2. Arlazarov, V.V., Andreeva, E.I., Bulatov, K.B., Nikolaev, D.P., Petrova, O.O., Savelev B.I., and Slavin, O.A., Document image analysis and recognition: A survey, *Komput. Optika,* 2022, vol. 46, no. 4, pp. 567–589.
3. Yang, B. et al., Edge intelligence for autonomous driving in 6G wireless system: Design challenges and solutions, *IEEE Wireless Commun.,* 2021, vol. 28, no. 2, pp. 40–47.
4. Gu, T., Dolan-Gavitt, B., and Garg, S., Badnets: Identifying vulnerabilities in the machine learning model supply chain, arXiv:1708.06733, 2017.
5. Fredrikson, M., Jha, S., and Ristenpart, T., Model inversion attacks that exploit confidence information and basic countermeasures, *Proc. of the 22nd ACM SIGSAC Conference on Computer and Communications Security,* 2015, pp. 1322–1333.
6. Szegedy C. et al., Intriguing properties of neural networks, arXiv:1312.6199, 2013.
7. Brown, T.B. et al., Adversarial patch, arXiv:1712.09665, 2017.
8. Lin, C.S. et al., Real-world adversarial examples via makeup, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, 2022, pp. 2854–2858.
9. Hu, S. et al. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2022, pp. 15014–15023.
10. Zolfi, A. et al., Adversarial Mask: Real-World Universal Adversarial Attack on Face Recognition Models, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* Cham: Springer Nature Switzerland, 2022, pp. 304–320.
11. Zhou, Z. et al., Invisible mask: Practical attacks on face recognition with infrared, arXiv:1803.04683, 2018.
12. Wu, Z., Lim, S.N., Davis, L.S., and Goldstein, T., Making an invisibility cloak: Real world adversarial attacks on object detectors, *Proc. of the Computer Vision—ECCV 2020: 16th European Conference,* Glasgow, UK, 2020, Part 4, pp. 1–17.
13. Thys, S., Van Ranst, W., and Goedemé, T., Fooling automated surveillance cameras: adversarial patches to attack person detection, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,* 2019. https://openaccess.thecvf.com/content_CVPRW_2019/html/CV-COPS/Thys_Fooling_Automated_Surveillance_Cameras_Adversarial_Patches_to_Attack_Person_Detection_CVPRW_2019_paper.html.
14. Chen, J. et al., Diffusion Models for Imperceptible and Transferable Adversarial Attack, arXiv:2305.08192, 2023.
15. Hong, S. et al., Security analysis of deep neural networks operating in the presence of cache side-channel attacks, arXiv:1810.03487, 2018.
16. Oh, S.J., Schiele, B., and Fritz, M., Towards reverse-engineering black-box neural networks, in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,* 2019 pp. 121–144.
17. Chmielewski, Ł., and Weissbart, L., On reverse engineering neural network implementation on GPU, *Proc. of the Applied Cryptography and Network Security Workshops: ACNS 2021 Satellite Workshops, AIBlock, AIHWS, AIoTS, CIMSS, Cloud S&P, SCI, SecMT, and SiMLA,* Kamakura, Japan, 2021, Springer, 2021, pp. 96–113.
18. Goldblum, M. et al., Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses, *IEEE Trans. Pattern Anal. Mach, Intell.,* 2022, vol. 45, no. 2, pp. 1563–1580.
19. Shafahi, A. et al., Poison frogs! Targeted clean-label poisoning attacks on neural networks, *Advances in neural information processing systems,* 2018, vol. 31.
20. Wang, Y. et al. Sapag: A self-adaptive privacy attack from gradients, arXiv:2009.06228, 2020.
21. Warr, K., *Strengthening Deep Neural Networks: Making AI Less Susceptible to Adversarial Trickery,* O'Reilly, 2019.
22. Akhtar, N. and Mian, A., Threat of adversarial attacks on deep learning in computer vision: A survey, *IEEE Access,* vol. 2018, no. 6, pp. 14410–14430
23. Machado, G.R., Silva, E., and Goldschmidt, R.R., Adversarial machine learning in image classification: A

survey toward the defender's perspective, *ACM Comput. Surveys (CSUR),* 2021, vol. 55, no. 1, pp. 1–38.

24. Long, T. et al., A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions, *Comput. & Security,* 2022, p. 102847.
25. Ren, K. et al., Adversarial attacks and defenses in deep learning, *Engineering,* 2020, vol. 6, no. 3, pp. 346–360.
26. Zhang, X. et al., Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition, *Complex & Intell. Syst.,* 2023, vo. 9, no. 1, pp. 65–79.
27. Kwon, H. and Lee, S., Ensemble transfer attack targeting text classification systems, *Comput. & Security,* 2022, vol. 117, p. 102695.
28. Mo, K. et al. Attacking deep reinforcement learning with decoupled adversarial policy, *IEEE Trans. Dependable Secure Comput.,* 2022, vol. 20, no. 1, pp. 758–768.
29. Zhou, X. et al., Hierarchical adversarial attacks against graph-neural-network-based IoT network intrusion detection system, *IEEE IoT J.,* 2021, vol. 9, no. 12, pp. 9310–9319.
30. Kumar, R.S.S. et al., Adversarial machine learning-industry perspectives, *IEEE Security and Privacy Workshops (SPW),* IEEE, 2020, pp. 69–75.
31. Paleyes, A., Urma, R.G., and Lawrence, N.D., Challenges in deploying machine learning: A survey of case studies, *ACM Comput. Surveys,* 2022, vol. 55, no. 6, pp. 1–29.
32. Ala-Pietilä, P. et al., *The Assessment List for Trustworthy Artificial Intelligence (ALTAI),* European Commission, 2020.
33. Musser, M. et al., Adversarial machine learning and cybersecurity: Risks, challenges, and legal implications, arXiv:2305.14553, 2023.
34. Facial recognition's latest foe: Italian knitwear. https://therecord.media/facial-recognitions-latest-foe-italian-knitwear. Accessed July 20, 2023.
35. How we fight content copying, or the first adversarial attack in production. https://habr.com/ru/companies/avito/articles/452142. Accessed July 20, 2023.
36. Povolny, S. and Trivedi, S., Model hacking ADAS to pave safer roads for autonomous vehicles. https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/. Accessed July 20, 2023.

Translated by A. Klimontovich

Publisher's Note. Pleiades Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.