Анализ уязвимостей нейросетевых технологий распознавания образов

А. В. Трусов[I, II], Е. Е. Лимонова[I, II], В. В. Арлазаров[I, II], А. А. Зацаринный[I]

[I] Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия
[II] ООО «Смарт Энджинс Сервис», Москва, Россия

Аннотация. В статье рассмотрена актуальная проблема уязвимости технологий искусственного интеллекта на основе нейронных сетей в задаче распознавания образов. Показано, что применение нейронных сетей порождает множество уязвимостей. Приведены конкретные примеры таких уязвимостей: некорректная классификация изображений, содержащих вредоносный шум или заплатки, отказ распознающих систем при наличии на изображении особых узоров, в том числе нанесенных на объекты реального мира, отравление обучающей выборки и др. На основе проведенного анализа показана необходимость улучшения безопасности технологий искусственного интеллекта и даны предложения, способствующие этому улучшению.

Ключевые слова: нейронные сети, атаки на нейронные сети, вредоносные изображения, нейросетевая безопасность.

DOI 10.14357/20718632230405 **EDN** BVJZWS

Введение

За последнее десятилетие использование систем искусственного интеллекта (ИИ) на основе глубоких нейронных сетей (ГНС) в различных сферах деятельности стало обычным явлением как во многих странах мира, так и в России. Это обусловлено тем, что ГНС являются достаточно эффективным инструментарием для моделирования сложных объектов и процессов, которые невозможно описать с использованием классических математических моделей и методов. Так, ГНС стали незаменимым инструментом для решения многих задач компьютерного зрения (например, классификации и сопоставления изображений, семантической сегментации, детекции объектов и многих других), широко используются в сложных программных комплексах, включая, помимо прочего, системы идентификации людей [1], распознавания документов [2] и автономного вождения [3].

Однако при всех достоинствах ГНС обладают и рядом принципиальных недостатков, которые могут приводить к нарушениям работы информационных систем на основе ИИ. Так, среди потенциальных уязвимостей ГНС можно назвать возможность отравления обучающей выборки (добавления в нее специально подготовленных изображений) [4] для создания закладок – уязвимостей, которыми сможет воспользоваться злоумышленник во время исполнения ГНС; возможность инверсии модели для извлечения данных об обучающей выборке или архитектуре модели [5]; всевозможные способы искажения входного сигнала с целью заставить нейронную сеть выдать неправильный ответ даже на тех примерах, на которых человек не замечает искажений [6]. И это далеко не полный перечень уязвимостей ГНС. По существу,

использование ГНС в программных комплексах приводит к появлению новых уязвимостей, которые не всегда очевидны с позиций классической парадигмы обеспечения и контроля информационной безопасности. При этом важно, что в системах, использующих ГНС, зачастую обрабатывается частная и персональная информация и поэтому их некорректное функционирование может оказать существенное негативное влияние на безопасность конкретных людей.

Вместе с тем, в настоящее время, по мнению авторов, и пользователи систем с ГНС, и разработчики таких систем слабо представляют спектр потенциальных угроз, порождаемых применением этих технологий. И одна из причин - недостаточность научных исследований в этой области.

Настоящая статья, посвященная анализу некоторых аспектов применения ГНС с позиций их уязвимости, позволяет в определенной степени сформировать более реальный взгляд на эту актуальную проблему.

1. Примеры уязвимости нейронных сетей

По мнению многих исследователей, ГНС являются весьма удобным, доступным и эффективным инструментом для аппроксимации сложных объектов и процессов. К сожалению, нейронные сети, как и любая аппроксимация, не идеальны. Они имеют ряд ограничений, которые могут использовать злоумышленники для влияния на результаты работы сети. Рассмотрим несколько примеров на основе анализа известных публикаций.

Добавление небольшого, особым образом сгенерированного, шума может привести к неправильному распознаванию изображения, даже если этот шум незаметен для человеческого глаза. Так, в левых столбцах Рис. 1 показаны изображения из набора данных ImageNet (задача 1000-классовой классификации), которые корректно распознает нейронная сеть AlexNet, но после добавления специально сгенерированного “шума” из центрального столбца (на рисунке шум усилен в 10 раз для наглядности), все изображения распознаются, как принадлежащие к одному классу “Страус”, хотя визуально не имеют ничего общего с этой птицей [6].

Специально подготовленная заплатка может обмануть нейросетевые классификаторы, даже если такая наклейка распечатана и сфотографирована в реальном мире. Например, на Рис. 2 показана наклейка, "превращающая” банан в тостер, по версии нейронной сети VGG16 [7].

Для обхода систем распознавания лиц можно использовать специальный макияж [8] или преобразования фотографий, визуально похожие на нанесение макияжа [9], а также различные типы масок [10]. На Рис. 3 показана такая атака: человек без маски (левый рисунок), так и в обычной

Рис. 1. Обман сети AlexNet с помощью специального шума

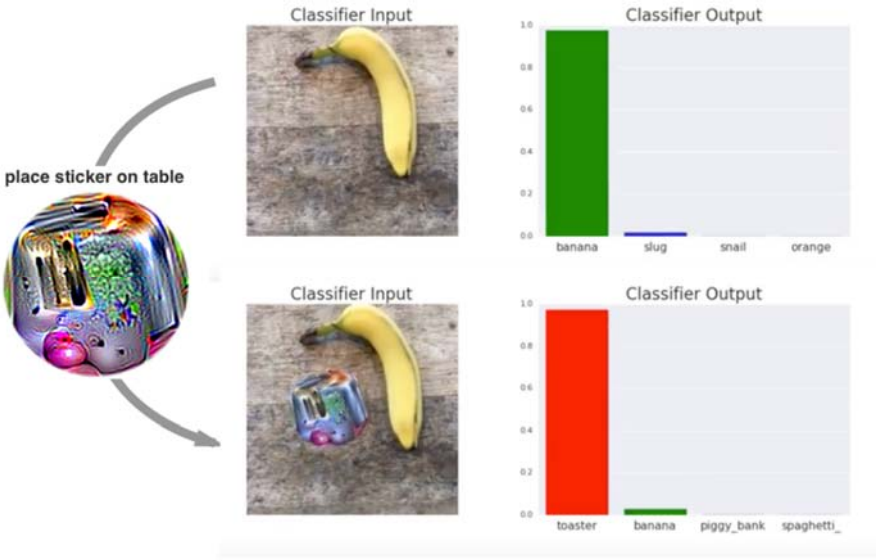

Рис. 2. Атака на модель VGG16 с помощью наклейки (показана слева)

Рис. 3. Маска, предотвращающая автоматическое распознавание лица (справа)

медицинской маске (центральный), распознается корректно, однако в маске с особым узором (правый рисунок) человек не распознается, хотя доля перекрытия лица такая же, как и в случае с медицинской маской.

Специальные инфракрасные осветительные приборы [11], одежда с особыми узорами [12] и накладками [13] также могут быть использованы для обмана систем видеонаблюдения. Например, на Рис. 4 показан свитер со специальной раскраской, благодаря которой человек, носящий его не детектируется нейронной сетью, в отличие от большинства остальных людей на том же изображении [12]. Последние исследования говорят о том, что искажение не всегда должно вносить заметные изменения в изображение или быть похожим на высокочастотный шум, который обычно получается при градиентной оптимизации входного изображения по функционалу атаки. Искажение может затрагивать лишь отдельные высокоуровневые признаки изображения [14]. На Рис. 5 показан пример такого искажения (справа), созданный с помощью диффузной ГНС.

Наряду с приведенными примерами отметим, что информация об архитектуре и параметрах ГНС, имеющая значительную коммерческую ценность, может быть восстановлена и украдена злоумышленником, если он имеет физический доступ к оборудованию, на котором исполняется нейронная сеть [15-17]. Более того, если злоумышленник имеет доступ к обучающей выборке (даже к небольшой ее части), он может "отравить" эту выборку, добавив в нее специальные примеры [4]. Обучение на отравленных данных опасно, поскольку полученная таким образом нейронная

Рис. 4. Свитер-невидимка (в центре), обманывающий ГНС детекции человека

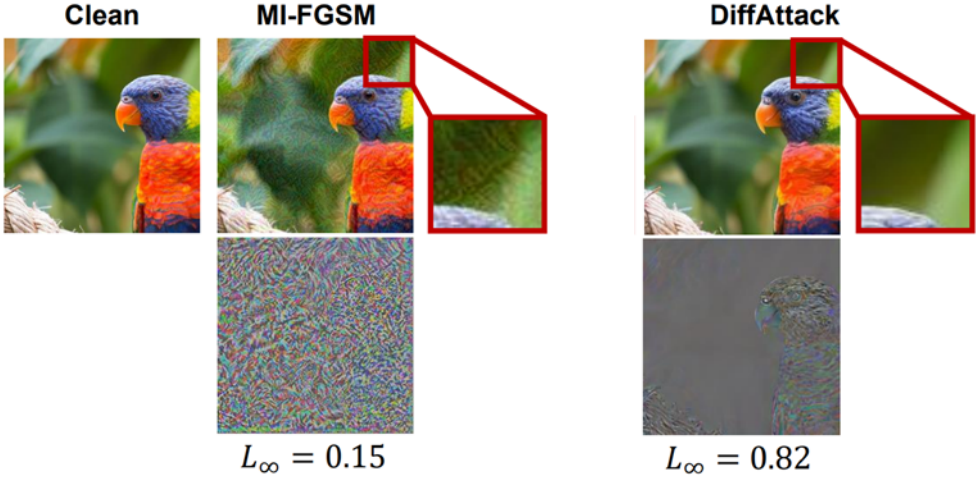

Рис. 5. Атаки на изображения попугая (чистое слева), с помощью градиентной оптимизации (по центру) и диффузной ГНС (справа)

сеть может содержать закладки [18] или работать некорректно на определенных входных изображениях [19]. Наконец, существует угроза конфиденциальности данных, поскольку обучающие данные могут быть извлечены как из уже обученной нейронной сети [5], так и в процессе ее распределенного обучения [20].

2. Атаки на нейронные сети

Описанные выше (на некоторых примерах) действия по поиску и использованию уязвимостей нейронных сетей обычно называют *атаками на нейронные сети*, а изображения, на которых сети дают неправильный ответ, – *вредоносными изображениями* [21]. Существует множество различных методов и алгоритмов, которые используются для формирования атак на нейронные сети, и их число растет с каждым днем. Об этом свидетельствуют публикации в базе данных Scopus, посвященные атакам на алгоритмы машинного обучения. Число таких публикаций постоянно растет. Данные в зависимости от года публикации согласно исследованию [22] представлены на Рис. 6 (стрелка на рис. указывает на время обнаружения существования вредоносных изображений, обманывающих нейронные сети). Например, в очень подробном обзоре атак, применимых к нейронным сетям в компьютерном зрении, опубликованном в 2018 году [23], упоминается 12 различных атак; в другом обзоре от 2021 года [24] – 33 атаки, в статье, опубликованной в 2022 году [22], – более 50 атак, организованных в соответствии с их таксономией. Сама эта таксономия довольно сложная и быстро эволюционирует.

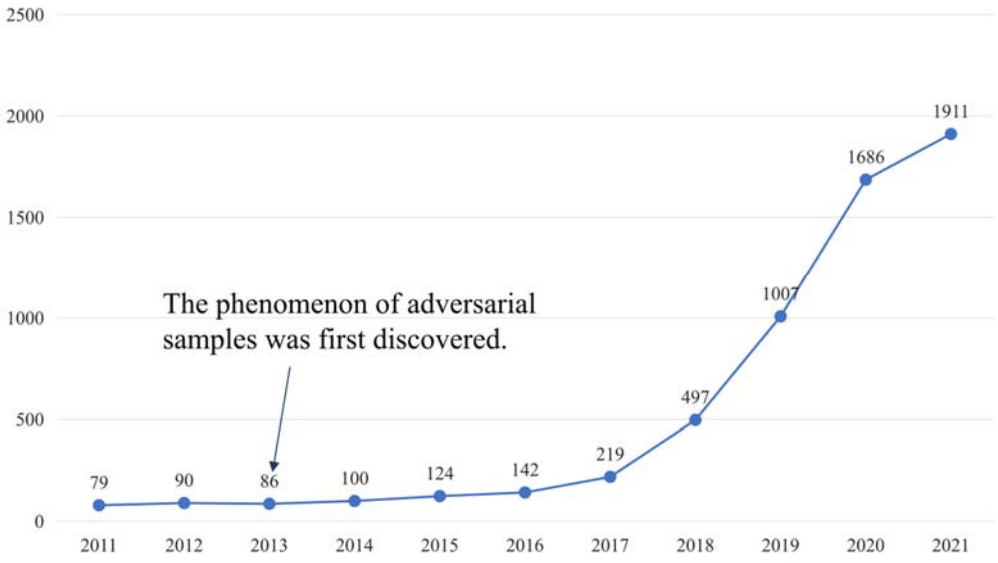

Рис. 6. Число публикаций, посвященных атакам на алгоритмы машинного обучения

По мере развития новых методов атак на нейронные сети в противовес им разрабатываются и средства защиты. В исследовании 2021 года упоминается более 60 различных алгоритмов защиты [25]. В настоящее время гонка между атаками и защитами еще далека от завершения. Так, любая эвристическая защита (основанная на применении атаки к данным во время обучения, для улучшения устойчивости ГНС) хорошо противостоит заданной атаке, но легко обходится с помощью других атак, а гарантируемые (сертифицированные) защиты имеют три существенных недостатка:

- приводят к снижению качества нейронных сетей;
- существенно усложняют (в смысле вычислительной сложности и скорости сходимости) процесс обучения;
- обеспечивают защиту только от небольших искажений изображения в некотором метрическом пространстве, в то время как атаки могут использовать различные метрики [25].

Атаки на нейронные сети могут производиться не только в системах компьютерного зрения, но и распознавания речи [26], анализа текста [27], обучения с подкреплением [28], распознавания инструкций в системах интернета вещей (IoT) [29] и в любых других областях, где применяются нейронные сети.

К сожалению, современное состояние индустрии демонстрирует недостаточное понимание угрозы, исходящей от атак на алгоритмы машинного обучения (и нейронные сети, в частности), и незнание способов защиты от них. Например, 22 из 25 европейских и американских организаций, использующих в своих продуктах различные алгоритмы машинного обучения, которые в последние годы стало модно называть алгоритмами ИИ, "не имеют необходимых инструментов для защиты своих систем машинного обучения и нуждаются в руководстве" [30].

3. Обсуждение результатов и рекомендации

Таким образом, атаки на алгоритмы машинного обучения и особенно нейронные сети остаются актуальной и широко обсуждаемой темой в научной среде на протяжении последнего десятилетия. Однако данная тема еще не получила должного внимания при разработке практических приложений. Несмотря на то, что угроза атак на ИИ упоминается в обзоре проблем развертывания программных комплексов на основе машинного обучения [31], в нем нет определенных рекомендаций по борьбе с ней. Некоторые высокоуровневые рекомендации можно найти в [32 и 33]. Приведем некоторые из них:

Рис. 7. Стикеры “превращающие” 35 миль в час в 85

- Анализировать возможные уязвимости систем ИИ и соответствующие способы атак.
- Тестировать систему на устойчивость к атакам.
- Привлекать специалистов для контролируемого взлома системы и выявления ее слабых мест.
- Информировать пользователей систем ИИ о потенциальных угрозах их безопасности, связанной с возможностью атак.
- Обеспечивать безопасность приватных данных, если они использовались при обучении нейронных сетей.
- Улучшать коммуникацию между исследователями, работающими над атаками на нейронные сети, устойчивостью нейронных сетей и теми, кто внедряет системы на основе ИИ.

Хотя эти рекомендации могут быть полезны, они кажутся не практичными, поскольку не дают четкого и прямого ответа на вопрос, как обеспечить безопасность прикладной системы ИИ от атак злоумышленников. Многие представители индустрии остаются в неведении относительно существования нейросетевых уязвимостей и угроз, которые они несут. Какие-либо регламенты, касающиеся обеспечения безопасности нейронных сетей от возможных атак, еще не сформированы или не получили массового распространения.

Тем не менее, интерес со стороны общества и индустрии к данной теме также начинает расти. Существуют примеры камуфляжа одежды для защиты от систем распознавания (например, итальянский бренд Cap_able [34]) и искажений изображений для предотвращения автоматизированного воровства изображений (например, российский сайт объявлений Avito [35]). Крупные компании, занимающиеся кибербезопасностью, начинают исследовать системы на базе ИИ в поисках потенциальных уязвимостей. Например, McAffe потратила 18 месяцев на взлом и анализ систем Tesla и MobileEye для автономных транспортных средств [36]. Они показали, как различные стикеры, прикрепленные к дорожным знакам, могут обмануть эти системы. Пример таких стикеров показан на Рис. 7. Слева те, что обманывают MobileEye, справа – Tesla. Обе системы распознали такие знаки как ограничение скорости в 85 миль в час вместо 35.

Несмотря на все это практические атаки на нейронные сети в компьютерном зрении пока еще не очень распространены.

Заключение

В работе приведены результаты анализа нового класса уязвимостей систем, использующих глубокие нейронные сети, которые могут возникать на разных этапах от обучения до непосредственного применения. Приведены примеры различных атак, значительно влияющих на результаты распознавания, как в цифровых системах, так и в реальном мире. Показано, что вопросы информационной безопасности и стабильности ИИ на сегодняшний день крайне актуальны. В работе приведены для обсуждения общие соображения и рекомендации, позволяющие повысить безопасность функционирования систем на основе глубоких нейронных сетей.

Литература

1. Ye M. et al. Deep learning for person re-identification: A survey and outlook //IEEE transactions on pattern analysis and machine intelligence. – 2021. – Т. 44. – №. 6. – С. 2872-2893.
2. Arlazarov V. V., Andreeva E. I., Bulatov K. B., Nikolaev D. P., Petrova O. O., Savelev B. I., Slavin O. A. Document image analysis and recognition: A survey // Компьютерная оптика. — 2022. — Т. 46. — № 4. — С. 567-589
3. Yang B. et al. Edge intelligence for autonomous driving in 6G wireless system: Design challenges and solutions //IEEE Wireless Communications. – 2021. – Т. 28. – №. 2. – С. 40-47.
4. Gu T., Dolan-Gavitt B., Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain //arXiv preprint arXiv:1708.06733. – 2017.
5. Fredrikson M., Jha S., Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures //Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. – 2015. – С. 1322-1333.
6. Szegedy C. et al. Intriguing properties of neural networks //arXiv preprint arXiv:1312.6199. – 2013.
7. Brown T. B. et al. Adversarial patch //arXiv preprint arXiv:1712.09665. – 2017.
8. Lin C. S. et al. Real-world adversarial examples via makeup //ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2022. – С. 2854-2858.
9. Hu S. et al. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2022. – С. 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. – С. 304-320.
11. Zhou Z. et al. Invisible mask: Practical attacks on face recognition with infrared //arXiv preprint arXiv:1803.04683. – 2018.
12. Wu Z., Lim S.N., Davis L.S., Goldstein T. Making an invisibility cloak: Real world adversarial attacks on object detectors. InComputer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16 2020 (pp. 1-17). Springer International Publishing.
13. Thys S., Van Ranst W., Goedemé T. Fooling automated surveillance cameras: adversarial patches to attack person detection //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. – 2019. – С. 0-0.
14. Chen J. et al. Diffusion Models for Imperceptible and Transferable Adversarial Attack //arXiv preprint 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 preprint arXiv:1810.03487. – 2018.
16. Oh S. J., Schiele B., Fritz M. Towards reverse-engineering black-box neural networks //Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. – 2019. – С. 121-144.
17. Chmielewski Ł., Weissbart L. On reverse engineering neural network implementation on gpu //Applied Cryptography and Network Security Workshops: ACNS 2021 Satellite Workshops, AIBlock, AIHWS, AIoTS, CIMSS, Cloud S&P, SCI, SecMT, and SiMLA, Kamakura, Japan, June 21–24, 2021, Proceedings. – Springer International Publishing, 2021. – С. 96-113.
18. Goldblum M. et al. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses //IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2022. – Т. 45. – №. 2. – С. 1563-1580.
19. Shafahi A. et al. Poison frogs! targeted clean-label poisoning attacks on neural networks //Advances in neural information processing systems. – 2018. – Т. 31.
20. Wang Y. et al. Sapag: A self-adaptive privacy attack from gradients //arXiv preprint arXiv:2009.06228. – 2020.
21. Warr K. Strengthening deep neural networks: Making AI less susceptible to adversarial trickery. – O'Reilly Media, 2019.
22. Long T. et al. A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions //Computers & Security. – 2022. – С. 102847.
23. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey //Ieee Access. – 2018. – Т. 6. – С. 14410-14430
24. Machado G. R., Silva E., Goldschmidt R. R. Adversarial machine learning in image classification: A survey toward the defender's perspective //ACM Computing Surveys (CSUR). – 2021. – Т. 55. – №. 1. – С. 1-38.
25. Ren K. et al. Adversarial attacks and defenses in deep learning //Engineering. – 2020. – Т. 6. – №. 3. – С. 346-360.
26. Zhang X. et al. Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition //Complex & Intelligent Systems. – 2023. – Т. 9. – №. 1. – С. 65-79.
27. Kwon H., Lee S. Ensemble transfer attack targeting text classification systems //Computers & Security. – 2022. – Т. 117. – С. 102695.
28. Mo K. et al. Attacking deep reinforcement learning with decoupled adversarial policy //IEEE Transactions on Dependable and Secure Computing. – 2022. – Т. 20. – №. 1. – С. 758-768.
29. Zhou X. et al. Hierarchical adversarial attacks against graph-neural-network-based IoT network intrusion detection system //IEEE Internet of Things Journal. – 2021. – Т. 9. – №. 12. – С. 9310-9319.
30. Kumar R. S. S. et al. Adversarial machine learning-industry perspectives //2020 IEEE Security and Privacy Workshops (SPW). – IEEE, 2020. – С. 69-75.
31. Paleyes A., Urma R. G., Lawrence N. D. Challenges in deploying machine learning: a survey of case studies //ACM Computing Surveys. – 2022. – Т. 55. – №. 6. – С. 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 preprint arXiv:2305.14553. – 2023.
34. Facial recognition's latest foe: Italian knitwear [Электронный ресурс] // The Record. URL: https://therecord.media/facial-

recognitions-latest-foe-italian-knitwear (дата обращения: 20.07.2023).
35. Как мы боремся с копированием контента, или первая adversarial attack в проде. [Электронный ресурс] // Хабр. URL: https://habr.com/ru/companies/avito/articles/452142 (дата обращения: 20.07.2023).
36. Povolny S., Trivedi S. Model hacking ADAS to pave safer roads for autonomous vehicles. [Электронный ресурс] // McAfee Blogs. URL: https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/ (дата обращения: 20.07.2023).

Трусов Антон Всеволодович. Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия, программист первой категории. Область научных интересов: нейронные сети, обработка изображений, высокопроизводительные вычисления. E-mail: trusov.anton.02@gmail.com

Лимонова Елена Евгеньевна. Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия. Научный сотрудник, кандидат технических наук. Область научных интересов: нейронные сети, обработка изображений, распознавания образов на мобильных устройствах. E-mail: elena.e.limonova@gmail.com

Арлазаров Владимир Викторович Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия. Заведующий отделом, доктор технических наук. Область научных интересов: искусственный интеллект, машинное обучение, системы распознавания, информационные технологии. E-mail: vva777@gmail.com

Зацаринный Александр Алексеевич. Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия. Главный научный сотрудник, доктор технических наук, профессор. Область научных интересов: интегрированные информационно-телекоммуникационные системы и сети, цифровая трансформация общества. E-mail: AZatsarinny@ipiran.ru

Vulnerability Analysis of Neural Networks in Computer Vision

A. V. Trusov[I, II], E. E. Limonova[I, II], V. V. Arlazarov[I, II], A. A. Zatsarinnyy[I]

[I] Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia
[II] Smart Engines Service LCC, Moscow, Russia

Abstract. The work considers the actual problem of the vulnerability of artificial intelligence technologies based on neural networks. We show that the use of neural networks generates many vulnerabilities. We demonstrate specific examples of such vulnerabilities: incorrect classification of images containing adversarial noise or patches, failure of recognition systems in the presence of special patterns on the image, including those applied to objects in the real world, training data poisoning, etc. Based on the analysis, we show the need to improve the security of artificial intelligence technologies and suggest some considerations that contribute to this improvement.

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

DOI 10.14357/20718632230405 **EDN** BVJZWS

References

1. Ye M., Shen J., Lin G., Xiang T., Shao L., Hoi S.C. Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence. 2021 Jan 26;44(6):2872-93. doi: 10.1109/TPAMI.2021.3054775.
2. Arlazarov V.V., Andreeva E.I., Bulatov K.B., Nikolaev D.P., Petrova O.O., Savelev B.I., Slavin O.A. Document image analysis and recognition: A survey. Computer Optics. 2022; 46(4):567–589. doi: 10.18287/2412-6179-CO-1020.
3. Yang B., Cao X., Xiong K., Yuen C., Guan Y.L., Leng S., Qian L., Han Z. Edge intelligence for autonomous driving in 6G wireless system: Design challenges and solutions. IEEE Wireless Communications. 2021 Apr;28(2):40-7. doi:10.1109/MWC.001.2000292.
4. Gu T., Dolan-Gavitt B., Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733. 2017 Aug 22.
5. Fredrikson M., Jha S., Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. InProceedings of the 22nd ACM SIGSAC conference on computer and communications security 2015 Oct 12 (pp. 1322-1333).
6. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199. 2013 Dec 21.

7. Brown TB, Mané D, Roy A, Abadi M, Gilmer J. Adversarial patch. arXiv preprint arXiv:1712.09665. 2017 Dec 27.
8. Lin C.S., Hsu C.Y., Chen P.Y., Yu C.M. Real-world adversarial examples via makeup. InICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022 May 23 (pp. 2854-2858). IEEE. doi:10.1109/ICASSP43922.2022.9747469.
9. Hu S., Liu X., Zhang Y., Li M., Zhang L.Y., Jin H., Wu L. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 (pp. 15014-15023).
10. Zolfi A., Avidan S., Elovici Y., Shabtai A. Adversarial Mask: Real-World Universal Adversarial Attack on Face Recognition Models. InJoint European Conference on Machine Learning and Knowledge Discovery in Databases 2022 Sep 19 (pp. 304-320). Cham: Springer Nature Switzerland.
11. Zhou Z, Tang D, Wang X, Han W, Liu X, Zhang K. Invisible mask: Practical attacks on face recognition with infrared. arXiv preprint arXiv:1803.04683. 2018 Mar 13.
12. Wu Z., Lim S.N., Davis L.S., Goldstein T. Making an invisibility cloak: Real world adversarial attacks on object detectors. InComputer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16 2020 (pp. 1-17). Springer International Publishing.
13. Thys S., Van Ranst W., Goedemé T. Fooling automated surveillance cameras: adversarial patches to attack person detection. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops 2019 (pp. 0-0).
14. Chen J., Chen H., Chen K., Zhang Y., Zou Z., Shi Z. Diffusion Models for Imperceptible and Transferable Adversarial Attack. arXiv preprint arXiv:2305.08192. 2023 May 14.
15. Hong S., Davinroy M., Kaya Y., Locke S.N., Rackow I., Kulda K., Dachman-Soled D., Dumitraş T. Security analysis of deep neural networks operating in the presence of cache side-channel attacks. arXiv preprint arXiv:1810.03487. 2018 Oct 8.
16. Oh S.J., Schiele B., Fritz M. Towards reverse-engineering black-box neural networks. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. 2019:121-44.
17. Chmielewski Ł., Weissbart L. On reverse engineering neural network implementation on gpu. InApplied Cryptography and Network Security Workshops: ACNS 2021 Satellite Workshops, AIBlock, AIHWS, AIoTS, CIMSS, Cloud S&P, SCI, SecMT, and SiMLA, Kamakura, Japan, June 21–24, 2021, Proceedings 2021 (pp. 96-113). Springer International Publishing.
18. Goldblum M., Tsipras D., Xie C., Chen X., Schwarzschild A., Song D., Mądry A., Li B., Goldstein T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022 Mar 25;45(2):1563-80. doi:10.1109/TPAMI.2022.3162397.
19. Shafahi A., Huang W.R., Najibi M., Suciu O., Studer C., Dumitras T., Goldstein T. Poison frogs! targeted clean-label poisoning attacks on neural networks. Advances in neural information processing systems. 2018;31.
20. Wang Y., Deng J., Guo D., Wang C., Meng X., Liu H., Ding C., Rajasekaran S. Sapag: A self-adaptive privacy attack from gradients. arXiv preprint arXiv:2009.06228. 2020 Sep 14.
21. Warr K. Strengthening deep neural networks: Making AI less susceptible to adversarial trickery. O'Reilly Media; 2019 Jul 3.
22. Long T., Gao Q., Xu L., Zhou Z. A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions. Computers & Security. 2022 Jul 22:102847.
23. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. Ieee Access. 2018 Feb 19;6:14410-30. doi:10.1109/ACCESS.2018.2807385.
24. Machado G.R., Silva E., Goldschmidt R.R. Adversarial machine learning in image classification: A survey toward the defender's perspective. ACM Computing Surveys (CSUR). 2021 Nov 23;55(1):1-38.
25. Ren K., Zheng T., Qin Z., Liu X. Adversarial attacks and defenses in deep learning. Engineering. 2020 Mar 1;6(3):346-60.
26. Zhang X., Zhang X., Sun M., Zou X., Chen K., Yu N. Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition. Complex & Intelligent Systems. 2023 Feb;9(1):65-79.
27. Kwon H., Lee S. Ensemble transfer attack targeting text classification systems. Computers & Security. 2022 Jun 1;117:102695.
28. Mo K., Tang W., Li J., Yuan X. Attacking deep reinforcement learning with decoupled adversarial policy. IEEE Transactions on Dependable and Secure Computing. 2022 Jan 18;20(1):758-68.
29. Zhou X., Liang W., Li W., Yan K., Shimizu S., Kevin I., Wang K. Hierarchical adversarial attacks against graph-neural-network-based IoT network intrusion detection system. IEEE Internet of Things Journal. 2021 Nov 24;9(12):9310-9. doi:10.1109/JIOT.2021.3130434.
30. Kumar R.S., Nyström M., Lambert J., Marshall A., Goertzel M., Comissoneru A., Swann M., Xia S. Adversarial machine learning-industry perspectives. In2020 IEEE Security and Privacy Workshops (SPW) 2020 May 21 (pp. 69-75). IEEE. doi:10.1109/SPW50608.2020.00028.
31. Paleyes A., Urma R.G., Lawrence N.D. Challenges in deploying machine learning: a survey of case studies. ACM Computing Surveys. 2022 Dec 7;55(6):1-29.
32. Ala-Pietilä P., Bonnet Y., Bergmann U., Bielikova M., Bonefeld-Dahl C., Bauer W., Bouarfa L., Chatila R., Coeckelbergh M., Dignum V., Gagné J.F. The assessment list for trustworthy artificial intelligence (ALTAI). European Commission; 2020 Jul 17.
33. Musser M, Lohn A, Dempsey JX, Spring J, Kumar RS, Leong B, Liaghati C, Martinez C, Grant CD, Rohrer D, Frase H. Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications. arXiv preprint arXiv:2305.14553. 2023 May 23.
34. The Record. Facial recognition's latest foe: Italian knitwear. Available from: https://therecord.media/facial-recognitions-latest-foe-italian-knitwear [Accessed 20 July 2023]
35. Habr. How we fight content copying, or the first adversarial attack in prod. Available from:

https://habr.com/ru/companies/avito/articles/452142 [Accessed 20 July 2023]
36. Povolny S., Trivedi S. Model hacking ADAS to pave safer roads for autonomous vehicles. McAfee Blogs. Available from: https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/ [Accessed 20 July 2023]

Trusov Anton V. Federal Research Center «Computer Science and Control» of Russian Academy of Sciences, Moscow Russia, 1-st grade programmer. Research interests: neural networks, image processing, high performance computing.
E-mail: trusov.anton.02@gmail.com

Limonova Elena E. Federal Research Center «Computer Science and Control» of Russian Academy of Sciences, Moscow, Russia, PhD, researcher. Research interests: neural networks, image processing, pattern recognition on mobile devices.
E-mail: elena.e.limonova@gmail.com

Arlazarov Vladimir V. Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences, Moscow, Russia. Doctor of Science in technology, Head of the Department. Research interests: artificial intelligence, machine learning, recognition systems, information technology, queuing theory. Email address: vva777@gmail.com

Zatsarinnyy Alexander A. Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences, Moscow, Russia. Doctor of Science in technology, professor, principal scientist. E-mail: azatsarinny@ipiran.ru