конференция
РусКрипто

Ежегодная международная научно-практическая конференция

«РусКрипто'2024»

Повышение защищенности интеллектуальных систем в условиях деструктивного воздействия на основе генеративно-состязательных сетей

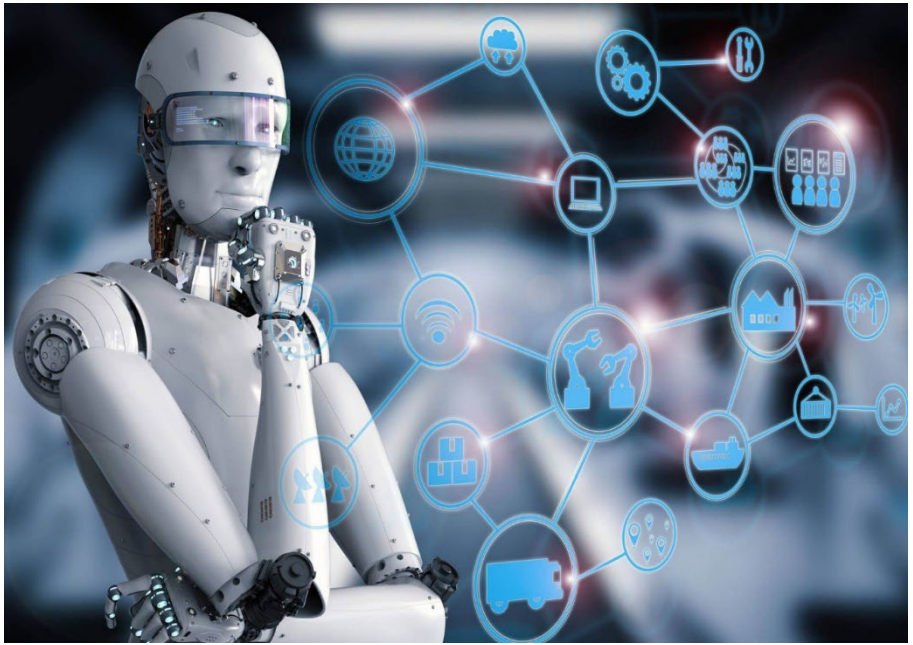
Васильев Никита Алексеевич, с.н.с., Военная академия связи
Лаута Олег Сергеевич, д.т.н., доцент, ГУМРФ С.О. Макарова
Котенко Игорь Витальевич, д.т.н., профессор, заслуженный деятель науки РФ, главный научный сотрудник и руководитель научно-исследовательской лаборатории проблем компьютерной безопасности, СПб ФИЦ РАН
Делова Мария Алексеевна, к.м.н, Военная академия связи

Содержание (1)

- **Введение**
- Релевантные работы
- Классификация атак
- Наиболее распространённые угрозы
- Методы защиты от атак
- Дискуссия и заключение

Интеллектуальные системы (1)

Интеллектуальные системы - это комплексные программные и аппаратные системы, способные анализировать информацию, принимать решения и выполнять задачи, которые обычно требуют человеческого интеллекта. Они используются для автоматизации различных процессов, улучшения производительности, предоставления рекомендаций и помощи в принятии решений.



Интеллектуальные системы (2)

Интеллектуальные системы играют все более значительную роль в современном обществе и оказывают влияние на различные аспекты жизни людей: автоматизация и оптимизация процессов, улучшение предсказаний и принятия решений, развитие новых технологий, улучшение медицинской помощи, улучшение безопасности и т.д.

Содержание (2)

- Введение
- **Релевантные работы**
- Классификация атак
- Наиболее распространённые угрозы
- Методы защиты от атак
- Дискуссия и заключение

Релевантные работы (1)

P_1	P_2	P_3	P_4	P_5	P_6	P_7
[1]	2020	Q1	203	11	0.05	Deep Learning
[2]	2020	Q1	72	18	0.25	Malicious Code
[3]	2020	Q2	139	25	0.18	Machine Learning
[4]	2020	Q2	136	12	0.09	Images, Graphics, Text
[5]	2020	Q3	115	4	0.03	Deep Learning
[6]	2020	Q3	13	4	0.31	Electricity, Smart Grids
[7]	2021	Q1	450	339	0.75	Computer Vision
[8]	2021	Q1	152	34	0.22	Electricity, Smart Grids
[9]	2021	Q1	78	1	0.01	Deep Learning
[10]	2021	Q2	163	38	0.23	Cybersecurity
[11]	2021	Q2	65	42	0.65	Electricity, Smart Grids
[12]	2021	Q3	132	75	0.57	Images, Text, Malicious Code
[13]	2022	Q1	185	121	0.65	Deep Learning

- P_1 – научный труд
- P_2 – год опубликования
- P_3 – квартиль журнала
- P_4 – количество ссылок на источники
- P_5 – количество ссылок на источники новее 2019 года
- P_6 – доля ссылок на источники новее 2019 года
- P_7 – предметная область

Релевантные работы (2)

P_1	P_2	P_3	P_4	P_5	P_6	P_7
[14]	2022	Q1	128	45	0.35	Deep Learning
[15]	2022	Q1	52	21	0.40	Deep Learning
[16]	2022	Q1	46	26	0.57	Digital Signals
[17]	2022	Q1	34	27	0.79	Deep Learning
[18]	2022	Q2	103	48	0.47	Deep Learning
[19]	2022	Q2	49	21	0.43	Electricity, Smart Grids
[20]	2022	Q2	46	7	0.15	Images
[21]	2023	Q1	246	140	0.57	Autonomous Vehicles
[22]	2023	Q1	176	119	0.68	Text
[23]	2023	Q1	53	23	0.43	Cybersecurity
[24]	2023	Q1	166	64	0.39	Cybersecurity
[25]	2023	Q1	254	134	0.53	Autonomous Vehicles
[26]	2023	Q2	179	48	0.27	Graphics

- P_1 – научный труд
- P_2 – год опубликования
- P_3 – квартиль журнала
- P_4 – количество ссылок на источники
- P_5 – количество ссылок на источники новее 2019 года
- P_6 – доля ссылок на источники новее 2019 года
- P_7 – предметная область

Содержание (3)

- Введение
- Релевантные работы
- **Классификация атак**
- Наиболее распространённые угрозы
- Методы защиты от атак
- Дискуссия и заключение

Классификация атак (1)

Значение признака	Название атаки	Сокращение	Область
Белый ящик	Fast Gradient Sign Method	FGSM	Изображение
	Iterative Gradient Sign Method	IGSM	Изображение
	Jacobian Saliency Map Attack	JSMA	Изображение
	Block Input Manipulation	BIM	Изображение
	Undetectable Perturbation	UP	Изображение
	Feature Adversary	FA	Изображение
	Carlini and Wagner's Attack	C&W	Изображение
	Iterative Least-Likely Class Method	ILLCM	Изображение
	One-Step Target Class Method	OSTCM	Изображение
	Deep Fool	DF	Изображение
	Hot/Cold method	HCM	Изображение
	Ground-Truth Attack	GTA	Изображение
	Targeted Audio Adversarial Examples	TAAE	Аудио

Классификация атак (2)

Значение признака	Название атаки	Сокращение	Область
Черный ящик	Boundary Attack	BA	Изображение
	Zero-Query Attacks	ZQA	Изображение
	Generative Adversarial Network	GAN	Изображение
	One Pixel Attack	OPA	Изображение
	Zeroth Order Optimization	ZOO	Изображение
	Natural Evolution Strategies	NES	Изображение
	Genetic Algorithms	GA	Аудио
	Improved Genetic Algorithm	IGA	Аудио
	Real-World Noise	RWN	Аудио
	Probability Weighted Word Saliency	PWWS	Текст
	Greedy Search Algorithm	GSA	Текст
	Insertion and Removal of Words	IRW	Текст

Классификация атак (3)

Значение признака	Название атаки	Сокращение	Область
Серый ящик	Cross-Site Scripting	CSS	Текст
	Password Guessing	PG	Текст
	Cross-Site Request Forgery	CSRF	Текст
	SQL Injection	SQLI	Текст
	Buffer Overflow Attack	BOA	Общий
	Weak Authentication Attack	WAA	Общий

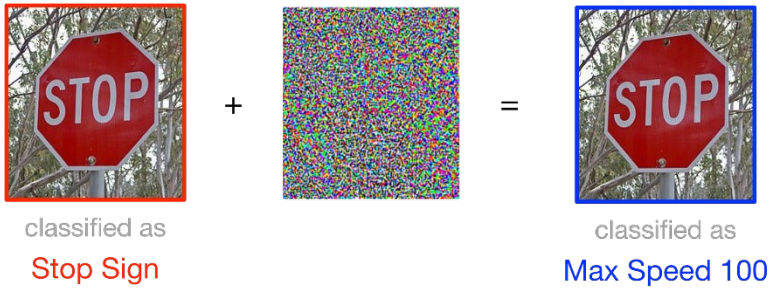
Содержание (4)

- Введение
- Релевантные работы
- Классификация атак
- **Наиболее распространённые угрозы**
- Методы защиты от атак
- Дискуссия и заключение

Fast Gradient Sign Method

FGSM – это метод атаки «белого ящика» на нейронные сети, который используется для обмана моделей, обученных для распознавания изображений. Метод FGSM заключается в том, чтобы изменить незначительно изображение таким образом, чтобы обученная модель ошибочно идентифицировала его другим классом.

$$Z^* = Z + \epsilon \cdot \text{sign}\left(\nabla_Z J(\theta, Z, W)\right)$$



Boundary Attack

Алгоритм граничной атаки – это типовой метод атаки «черного ящика», основанный на принятии решений. Начиная с исходного состязательного изображения, в нем используется бинарный поиск для нахождения точки выборки, которая находится вблизи границы классификации. Производится случайное блуждание по границе между двумя противоположными областями, чем уменьшается расстояние от целевого изображения. В соответствии с этим шагом продолжается итерация и постепенно уменьшается расстояние от исходного изображения.

Iterative Gradient Sign Method

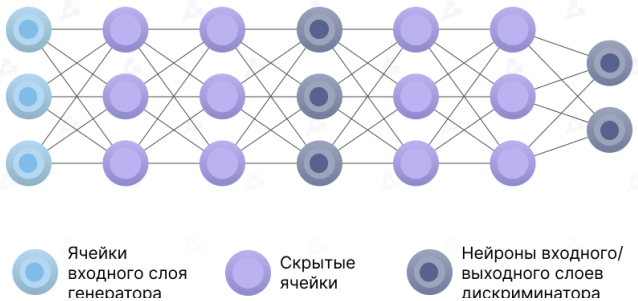
$$Z'_0 = Z;\ Z'_{n+1} = Clip_{Z,\epsilon}\{Z'_n + \alpha \cdot \text{sign}(\nabla_Z J(\theta, Z'_n, W))\}$$

IGSM является алгоритмом оптимизации, который начинается с исходного изображения и продолжает обновлять его через серию итераций с использованием FGSM. В каждой итерации значения пикселей изменяются в направлении увеличения потерь целевой функции. В отличие от FGSM, который использует только одну итерацию для создания поддельных изображений, IGSM повторяет процедуру атаки на каждой итерации, что дает лучший эффект, но требует больших вычислительных ресурсов.

Generative Adversarial Network

На первом этапе генератор создает поддельные примеры данных, которые пере-даются дискриминатору вместе с настоящими примерами из обучающего набора. Дискриминатор обучается отличать настоящие данные от поддельных, и генератор учится создавать такие данные, чтобы их было сложно отличить от реальных. На втором этапе генератор использует полученные знания о структуре данных, чтобы создать злоумышленные атаки на модель машинного обучения. Эти атаки могут быть различными в зависимости от типа модели и задачи, которую она решает.

Генеративно-состязательная нейросеть (GAN)



One Pixel Attack

Атака OPA относится к атакам «черного ящика» и основывается на алгоритмах МО. Она использует уязвимости в работе нейронных сетей, которые определяют изображения на основе цветовых значений каждого пикселя. Основной принцип работы этой атаки состоит в том, чтобы изменить значение всего лишь одного пикселя на изображении таким образом, чтобы нейронная сеть неправильно классифицировала это изображение.

Атака OPA использует эволюционные алгоритмы, позволяющие определить оптимальные пиксели и изменить их значения таким образом, чтобы обмануть нейронную сеть. Использование таких алгоритмов позволяет достичь максимальной эффективности атаки при минимальном числе изменений на изображении.

Teapot(24.99%)
Joystick(37.39%)

Содержание (5)

- Введение
- Релевантные работы
- Классификация атак
- Наиболее распространённые угрозы
- **Методы защиты от атак**
- Дискуссия и заключение

Методы защиты от состязательных атак (1)

Метод защиты	Описание подхода к защите	Атаки
Competitive Training	Ансамблевое состязательное обучение. Методология обучения, которая включает в себя возмущенные входные данные, переданные из других предварительно обученных моделей	FGSM, IGSM, ILLCM
	Расширенное состязательное и виртуальное состязательное обучение как средство упорядочивания текстового классификатора путем стабилизации функции классификации	IRW
	Обучение современному распознаванию речевых эмоций на смеси чистых и состязательных примеров, чтобы помочь упорядочению	RWN

Методы защиты от состязательных атак (2)

Метод защиты	Описание подхода к защите	Атаки
Defensive Distillation	Основная используемая идея заключается в двукратном обучении модели. Первоначально с использованием меток истинности с одним горячим основанием, но в конечном итоге с использованием вероятности исходной модели в качестве выходных данных для повышения надежности	FGSM, IGSM, RWN
Input Data Reconstruction	Состязательные примеры преобразуются в чистые данные с помощью реконструкции. Аппроксимируется многообразие нормальных примеров. Состязательные примеры перемещаются к многообразию нормальных примеров, чтобы правильно классифицировать состязательные примеры с небольшим возмущением	FGSM, IGSM, DF, C&W
Defense-GAN	Фреймворк, использующий возможности генеративных моделей для защиты глубоких нейронных сетей от враждебных атак	FGSM, IGSM, JSMA, DF, C&W

Методы защиты от состязательных атак (3)

Метод защиты	Описание подхода к защите	Атаки
Model Reinforcement	Аналог метода кодирования, который вставляет кодировщик перед входным слоем модели, а затем обучает модель для устранения конфликтных возмущений	GSA, GA, IGA, PWWS
	Архитектура, использующая байесовские классификаторы для построения более надежных нейронных сетей	FGSM, C&W
	Использование совокупности классификаторов с взвешенным или средневзвешенным значением их прогноза для повышения устойчивости к атакам	FGSM, BIM

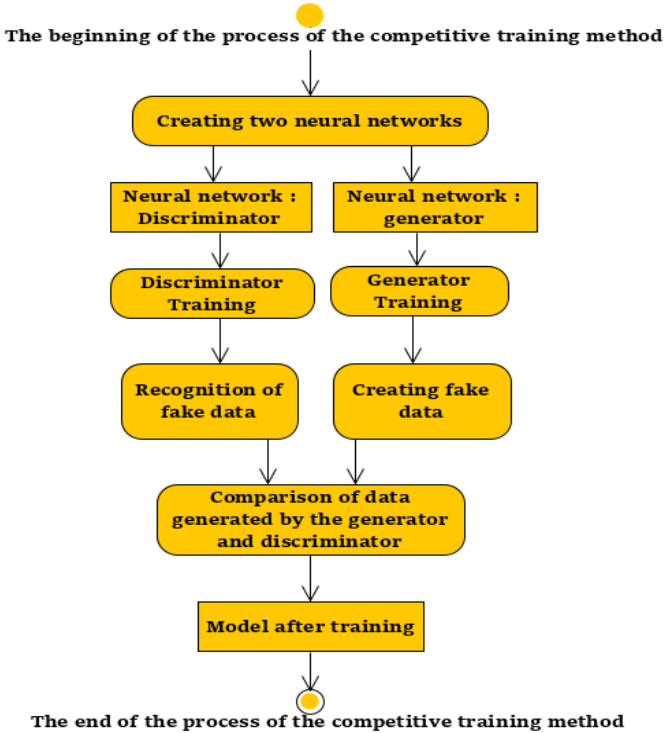
Методы защиты от состязательных атак (4)

Метод защиты	Описание подхода к защите	Атаки
Detection of Adversarial Examples	Сначала объекты сжимаются либо путем уменьшения разрядности цвета каждого пикселя, либо путем сглаживания выборки с помощью пространственного фильтра. Затем создается двоичный классификатор, который использует в качестве признаков результаты предсказаний целевой модели до и после сжатия входной выборки	FGSM, JSMA, BIM, C&W
	Фреймворк, который использует десять ненавязчивых функций качества изображения для различения образцов легитимных и состязательных атак	FGSM, IGSM, JSMA, DF
	Многоверсионное программирование основано на подходе к обнаружению аудио adversarial examples, который использует несколько готовых систем автоматического распознавания речи для определения того, является ли аудиовход adversarial example	TAAE, GA&GE

Методы защиты от состязательных атак (5)

Метод защиты	Описание подхода к защите	Атаки
Protection from Preprocessing	Использование PCA, фильтрации нижних частот, сжатия JPEG, программных методов определения пороговых значений в качестве метода предварительной обработки для повышения надежности	FGSM, IGSM, C&W
	Использование двух операций рандомизации: случайное изменение размера входных изображений и случайное заполнение нулями вокруг входных изображений	FGSM, DF, C&W

Схема метода состязательной тренировки



Процесс работы фреймворка Defense-GAN

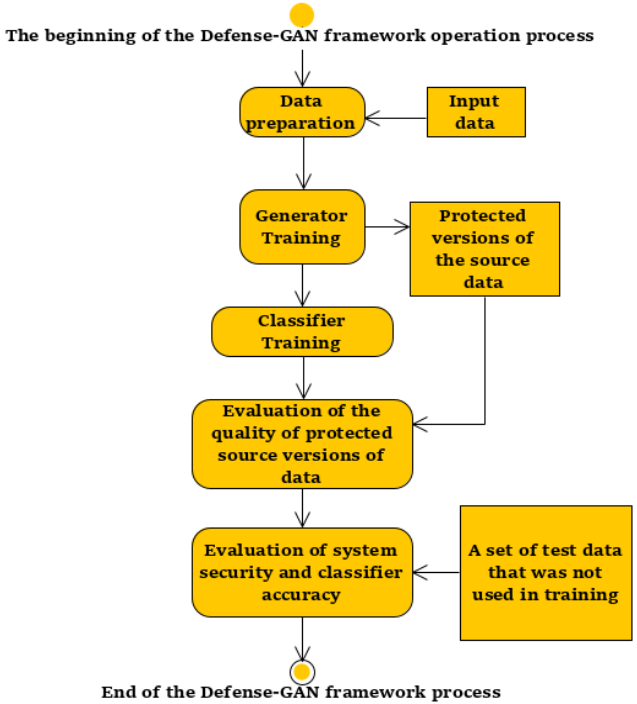


Схема метода защиты от предварительной обработки

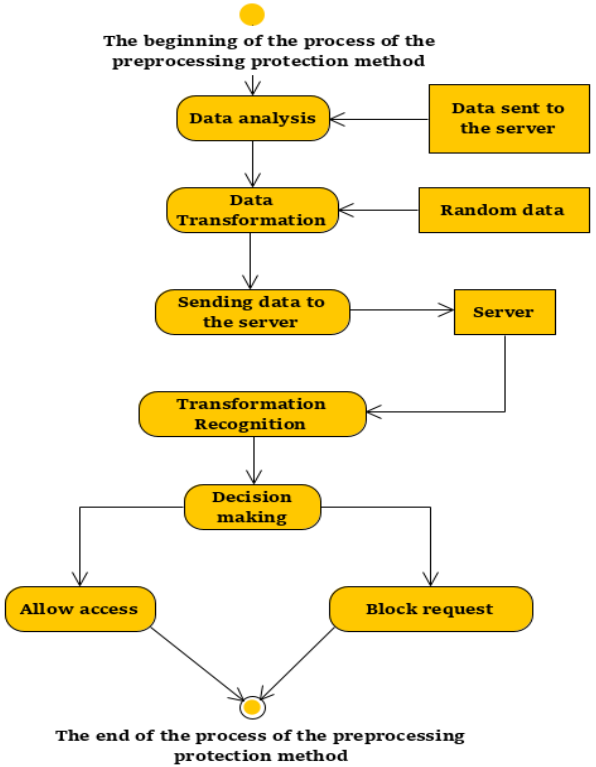


Схема анализа данных на выявление изменений в распределении

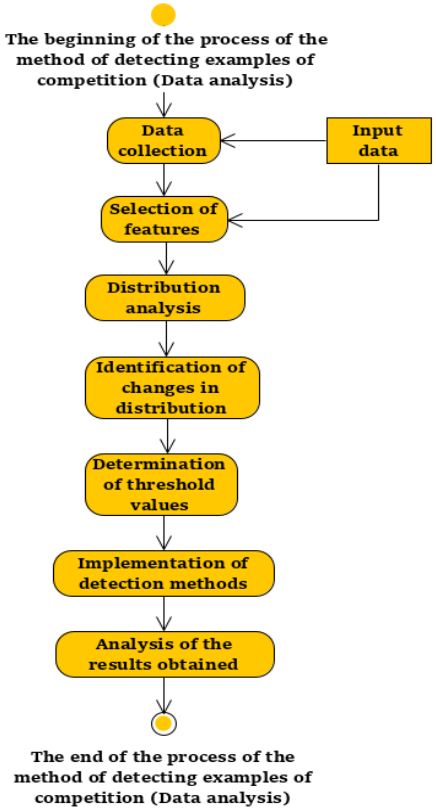
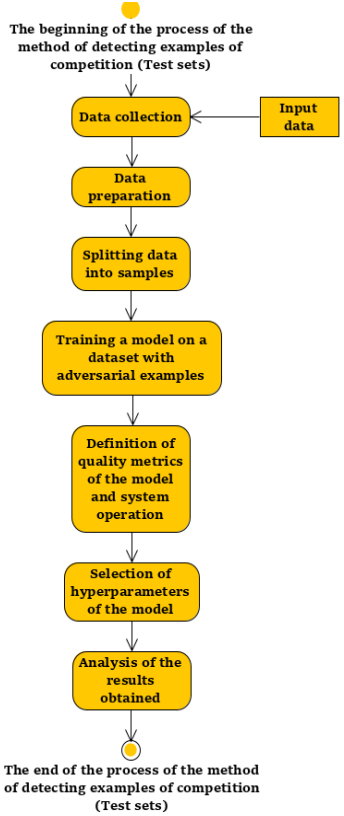


Схема использования тестовых наборов данных с состязательными примерами



Содержание (6)

- Введение
- Релевантные работы
- Классификация атак
- Наиболее распространённые угрозы
- Методы защиты от атак
- **Дискуссия и заключение**

Дискуссия (1)

Следует отметить, что несмотря на все достоинства фреймворка Defense GAN, у него существует один очень значительный недостаток: отсутствие зависимости от точки инициализации нейронной сети в прикладных задачах защиты информации влечет за собой то, что оптимальный дискриминатор будет присваивать более высокое значение для функции потерь, чем самим частям реальных обрабатываемых данных из генератора.

Метод оборонительной дистилляции имеет особенности при защите моделей ансамблей. Так, если в решающей модели будет использоваться алгоритм с привилегированной информацией, то раздельное функционирование модели-ученика и модели-учителя может привести к коллизиям в процессе нормальной работы базовой модели

Дискуссия (2)

Метод JSMA также обладает следующей важной особенностью. Он не может функционировать одно временно с моделью МО. Поэтому правильная организация потоков данных в пайплайне построения модели МО поможет полностью исключить негативный эффект от внедрения JSMA в качестве вредоносного компонента.

Метод One Pixel Attack весьма неэффективен, если в базовой конструкции модели машинного обучения используется два и более слоя пулинга.

Заключение

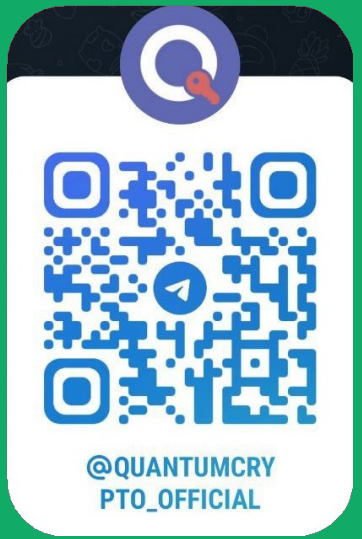
Следует отметить, что развитие технологий искусственного интеллекта идет непрерывно и быстро. При этом технологии машинного обучения с применением нейронных сетей становятся все более популярными. Однако при использовании систем ИИ возникают угрозы для информационной безопасности, которые до сих пор остаются малоизученными. Уязвимости систем ИИ могут быть использованы для достижения злонамеренных целей, поэтому необходимо обеспечить дополнительную специальную защиту, чтобы гарантировать безопасность систем ИИ. Развитие моделей угроз для систем искусственного интеллекта будет способствовать улучшению безопасности при интеграции ИИ в критически важные сферы деятельности человека.

Контактная информация

- **Электронная почта:**

  vasn2020@mail.ru
- **Телефон:**

  +7 911 793 35 14



Литература

1. Huang, X.; Kroening, D.; Ruan, W.; Sharp, J.; Sun, Y.; Thamo, E.; Wu, M.; Yi, X. A survey of safety and trustwor-thiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Computer Science Review 2020, 37, 100270. https://doi.org/10.1016/j.cosrev.2020.100270.
2. Martins, N.; Cruz, J.M.; Cruz, T.; Henriques Abreu, P. Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review. IEEE Access 2020, 8, 35403-35419. https://doi.org/10.1109/ACCESS.2020.2974752.
3. Oseni, A.; Moustafa, N.; Janicke, H.; Liu, P.; Tari, Z.; Vasilakos, A. Security and Privacy for Artificial Intelligence: Opportunities and Challenges. ArXiv abs/2102.04661, 2020. https://doi.org/10.48550/arXiv.2102.04661.
4. Xu, H.; Ma, Y.; Liu, H.C.; Deb, D.; Liu, H.; Tang, J.-L.; Jain, A.K. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. Int. J. Autom. Comput. 2020, 17, 151–178. https://doi.org/10.1007/s11633-019-1211-x.
5. Ren, K.; Zheng, T.; Qin, Zh.; Liu, X. Adversarial Attacks and Defenses in Deep Learning. Engineering 2020, 6, 346–360. https://doi.org/10.1016/j.eng.2019.12.012.
6. Zhou, X.; Canady, R.; Li, Y.; Koutsoukos, X.; Gokhale, A. Overcoming Stealthy Adversarial Attacks on Power Grid Load Predictions Through Dynamic Data Repair. In Dynamic Data Driven Applications Systems. DDDAS 2020. Lecture Notes in Computer Science, 2020, vol. 12312, pp. 102–109. https://doi.org/10.1007/978-3-030-61725-7_14.
7. Akhtar, N.; Mian, A.; Kardan, N.; Shah, M. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. IEEE Access 2021, 9, 155161-155196. https://doi.org/10.1109/ACCESS.2021.3127960.
8. Zhang, H.; Liu, B.; Wu, H. Smart Grid Cyber-Physical Attack and Defense: A Review. IEEE Access 2021, 9, 29641-29659. https://doi.org/10.1109/ACCESS.2021.3058628.
9. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. A survey on adversarial attacks and defences. CAAI Trans. Intell. Technol 2021, 6, 25-45. https://doi.org/10.1049/cit2.12028.
10. Rosenberg, I.; Shabtai, A.; Elovici, Y.; Rokach, L. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. ACM Comput. Surv. 2021, 54, 108. https://doi.org/10.1145/3453158.
11. Tian, J.; Wang, B.; Li, J.; Konstantinou, C. Adversarial attack and defense methods for neural network based state estimation in smart grid. IET Renewable Power Generation 2021, 16, 3507-3518. https://doi.org/10.1049/rpg2.12334.
12. Kong, Z.; Xue, J.; Wang, Y.; Huang, L.; Niu, Z.; Li, F.; Meng, W. A Survey on Adversarial Attack in the Age of Ar-tificial Intelligence. Wireless Communications and Mobile Computing 2021, 2021, 4907754. https://doi.org/10.1155/2021/4907754.
13. Zhou, Sh.; Liu, Ch.; Ye, D.; Zhu, T.; Zhou, W.; Yu, Ph.S. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity. ACM Comput. Surv. 2022, 55, 163. https://doi.org/10.1145/3547330.

Литература

14. Khamaiseh, S.Y.; Bagagem, D.; Al-Alaj, A.; Mancino, M.; Alomari, H.W. Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification. IEEE Access 2022, 10, 102266-102291. https://doi.org/10.1109/ACCESS.2022.3208131.
15. Liang, H.; He, E.; Zhao, Y.; Jia, Z.; Li, H. Adversarial Attack and Defense: A Survey. Electronics 2022, 11, 1283. https://doi.org/10.3390/electronics11081283.
16. Tian, Q.; Zhang, S.; Mao, Sh.; Lin, Y. Adversarial attacks and defenses for digital communication signals identifi-cation. Digital Communications and Networks 2022, in press. https://doi.org/10.1016/j.dcan.2022.10.010.
17. Anastasiou, Th.; Karagiorgou, S.; Petrou, P.; Papamartzivanos, D.; Giannetsos, Th.; Tsirigotaki, G.; Keizer, J. To-wards Robustifying Image Classifiers against the Perils of Adversarial Attacks on Artificial Intelligence Systems. Sensors 2022, 22, 6905. https://doi.org/10.3390/s22186905.
18. Li, Y.; Cheng, M.; Hsieh, Ch.-J.; Lee, Th.C.M. A Review of Adversarial Attack and Defense for Classification Methods. The American Statistician 2022, 76, 329-345. https://doi.org/10.1080/00031305.2021.2006781.
19. Tian, J.; Wang, B.; Li, J.; Wang, Z. Adversarial Attacks and Defense for CNN Based Power Quality Recognition in Smart Grid. IEEE Transactions on Network Science and Engineering 2022, 9, 807-819. https://doi.org/10.1109/TNSE.2021.3135565.
20. Li, H.; Namiot, D. A Survey of Adversarial Attacks and Defenses for Image Data on Deep Learning. International Journal of Open Information Technologies 2022, 10, 9-16. http://injoit.org/index.php/j1/article/view/1301/1220.
21. Girdhar, M.; Hong, J.; Moore, J. Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Ad-versarial Attacks and Defense Models. IEEE Open Journal of Vehicular Technology 2023, 4, 417-437. https://doi.org/10.1109/OJVT.2023.3265363.
22. Goyal, Sh.; Doddapaneni, S.; Khapra, M.M.; Ravindran, B. A Survey of Adversarial Defenses and Robustness in NLP. ACM Comput. Surv. 2023, 55, 332. https://doi.org/10.1145/3593042.
23. Al-Khassawneh, Y.A. A Review of Artificial Intelligence in Security and Privacy: Research Advances, Applications, Opportunities, and Challenges. Indonesian Journal of Science and Technology 2023, 8, 79–96. https://doi.org/10.17509/IJOST.V8I1.52709.
24. He, K.; Kim, D.D.; Asghar, M.R. Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey. IEEE Communications Surveys & Tutorials 2023, 25, 538-566. https://doi.org/10.1109/COMST.2022.3233793.
25. Sun, L.; Dou, Y.; Yang, C.; Zhang, K.; Wang, J.; Yu, Ph.S.; He, L.; Li, B. Adversarial Attack and Defense on Graph Data: A Survey. IEEE Transactions on Knowledge and Data Engineering 2023, 35, 7693-7711. https://doi.org/10.1109/TKDE.2022.3201243.
26. Qureshi, A.U.H.; Larijani, H.; Yousefi, M.; Adeel, A.; Mtetwa, N. An Adversarial Approach for Intrusion Detection Systems Using Jacobian Saliency Map Attacks (JSMA) Algorithm. Comput. 2020, 9, 58. https://doi.org/10.3390/COMPUTERS9030058