Некрасов Е.А., Петренко С.А. Угрозы безопасности и атаки в ИИ // Международный журнал информационных технологий и энергоэффективности. – 2024. – Т. 9 № 8(46) с. 125–130

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.056

УГРОЗЫ БЕЗОПАСНОСТИ И АТАКИ В ИИ

[1]Некрасов Е.А., Петренко С.А.
ФГБОУ ВО «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ», г. Москва, Россия (119454, г. Москва, Пр-т Вернадского, д. 78, стр.4), e-mail: [1]evgeniinekr@yandex.ru

Целью данной статьи является понимание нынешнего сценария атак ИИ и угроз безопасности и конфиденциальности ИИ. Риски, связанные с атаками на ИИ, медленно но верно становятся все более очевидными, что приводит к множеству финансовых и социальных потерь. Состязательные атаки, атаки на инверсию модели, атаки с отравлением, атаки с извлечением данных и атаки на вывод членства — вот некоторые категории, под которые в этой статье будут подпадать различные типы атак на модели искусственного интеллекта. Таким образом, эта статья послужит классификацией различных подобных атак на модели искусственного интеллекта.

Ключевые слова: Машинное обучение, безопасность, конфиденциальность, враждебные атаки, искусственный интеллект.

SECURITY THREATS AND ATTACKS IN AI

[1]Nekrasov E.A., **Petrenko S.A.**
MIREA - RUSSIAN TECHNOLOGICAL UNIVERSITY, Moscow, Russia (119454, Moscow, avenue. Vernadsky, 78, b. 4), e-mail: [1]evgeniinekr@yandex.ru

The purpose of this article is to understand the current scenario of AI attacks and threats to AI security and privacy. The risks associated with attacks on AI are slowly but surely becoming more obvious, which leads to a lot of financial and social losses. Adversarial attacks, model inversion attacks, poisoning attacks, data extraction attacks, and membership withdrawal attacks are some of the categories that various types of attacks on artificial intelligence models will fall under in this article. Thus, this article will serve as a classification of various similar attacks on artificial intelligence models.

Keywords: Machine learning, security, privacy, hostile attacks, artificial intelligence.

Введение

Искусственный интеллект (ИИ) стремительно развивается и интегрируется в различные приложения и сервисы, что приводит к увеличению числа атак ИИ и рисков для безопасности и конфиденциальности ИИ. В результате крайне важно исследовать текущее состояние атак ИИ и риски, которые они представляют для безопасности и конфиденциальности ИИ. Основываясь на последних интернет-данных и литературе, можно выделить несколько преобладающих моделей и опасностей, связанных с безопасностью и конфиденциальностью ИИ [3,4,5]. Чтобы снизить эти риски, предприятия должны внедрять надежные протоколы безопасности, такие как строгий контроль доступа, усовершенствованные механизмы шифрования и частые оценки безопасности. Кроме того, они должны гарантировать, что их системы ИИ прозрачны и подлежат надзору, соответствуют законам и нормативным актам в области конфиденциальности. В этой представлен подробный обзор атак ИИ и рисков для безопасности и конфиденциальности ИИ. В документе подчеркивается важность разработки

безопасных и надежных моделей ИИ для обеспечения конфиденциальности и сохранности конфиденциальных данных. Цель этого документа - информировать о потенциальных рисках и проблемах, связанных с защитой систем ИИ, путем изучения различных типов атак ИИ [4].

1. Атаки

Основными видами атак на модели искусственного интеллекта и машинного обучения являются атаки с состязательными воздействиями, атаки с инверсией модели, атаки с внедрением искаженных данных, атаки с извлечением данных и атаки с определением принадлежности данных. В течении последнего времени эти виды атак подверглись детальному изучению.

Состязательные атаки являются одними из наиболее распространенных атак на модели искусственного интеллекта и машинного обучения. Состязательные атаки предназначены для внесения небольших изменений во входные данные, что приводит к неправильной классификации входных данных моделью. Состязательные атаки могут быть целенаправленными или нецелевыми. Целенаправленные атаки заставляют модель выдавать конкретный неверный результат, в то время как нецелевые атаки приводят к тому, что модель выдает неверный результат. Состязательные атаки могут привести к значительным потерям во многих приложениях, таких как самоуправляемые автомобили и медицинские диагнозы . Атаки с использованием инверсии модели - это еще одна атака на модели искусственного интеллекта и машинного обучения. Целью этих атак является извлечение информации о данных обучения из модели. При атаках с использованием инверсии модели злоумышленник может использовать выходные данные модели для восстановления входных данных, чтобы ввести в заблуждение исходную модель машинного обучения. Эти атаки могут привести к утечке конфиденциальной информации, которая может быть использована для различных неэтичных действий. Атаки с отравлением - это еще один распространенный тип атак на модели искусственного интеллекта и машинного обучения. При атаках с отравлением злоумышленник вводит вредоносные данные в набор обучающих данных, чтобы манипулировать поведением модели во время тестирования. Эти атаки может быть сложно обнаружить и смягчить, поскольку на этапе обучения зараженные данные могут быть незаметны. При атаке с извлечением данных злоумышленник, не имеющий предварительных знаний о модели, пытается извлечь конфиденциальные данные, используемые для обучения модели. Атаки с извлечением данных предназначены для извлечения информации об обучающих данных из модели. Злоумышленник может использовать выходные данные модели для получения информации об обучающих данных. Эти атаки могут привести к утечке конфиденциальной информации. Атаки с выводом членства предназначены для определения того, использовалась ли конкретная точка данных в обучающем наборе данных. При атаках с выводом членства злоумышленник может использовать выходные данные модели, чтобы определить, использовалась ли конкретная точка данных в наборе обучающих данных. В этой атаке злоумышленник пытается получить обучающие данные из прогноза, полученного в результате ответа модели. Эти атаки могут привести к утечке конфиденциальной информации.

Исследователи предложили различные методы предотвращения таких атак, такие как состязательное обучение, очистка данных и сокращение модели. Состязательное обучение включает в себя обучение модели на состязательных примерах для повышения ее устойчивости к противоборствующим атакам. Очистка данных включает в себя фильтрацию

обучающего набора данных для удаления вредоносных или нерелевантных данных. Очистка модели включает в себя удаление ненужных функций или соединений из модели для снижения сложности и повышения надежности.

2. Защита моделей искусственного интеллекта

Защита моделей искусственного интеллекта (ИИ) стала ключевым аспектом разработки, учитывая возможность их использования злоумышленниками. Обеспечение безопасности моделей ИИ включает защиту от различных атак, таких как состязательные атаки, атаки на изменение модели, отравляющие атаки, атаки на извлечение данных и атаки на вывод членства. Для защиты моделей ИИ используются различные подходы, включая состязательное обучение, которое делает модель более устойчивой к атакам, обучая ее на чистых и состязательных примерах. Также применяются методы обнаружения аномалий для определения атак, путем мониторинга входных и выходных данных модели и выявления неожиданного поведения или закономерностей, которые могут указывать на атаку.

2.1. Подходы

Защита моделей искусственного интеллекта стала ключевой задачей из-за расширения использования ИИ в разных отраслях. Для обеспечения безопасности моделей ИИ существует много способов. Вот некоторые из этих подходов:

Состязательное обучение – такой подход добавляет состязательные примеры в обучающие данные для повышения устойчивости модели к атакам. Он прост и эффективен, но может быть затратным в вычислительном плане и не гарантирует полной защиты от всех атак. Очистка входных данных – этот подход предполагает предварительную обработку данных для удаления вредоносных элементов. Он эффективен и требует низких вычислительных затрат, но может быть неэффективен против сложных атак, способных обойти очистку. Такой подход, как объяснимость модели, предполагает повышение прозрачности и интерпретируемости модели для выявления и предотвращения атак. Он может выявить уязвимости и повысить доверие к модели, но реализация может быть сложной, и он может быть неэффективен против атак, использующих слабые места в архитектуре модели. Диверсификация моделей предполагает обучение нескольких моделей с разными архитектурами или параметрами для повышения надежности системы. Он эффективен против множества атак и может улучшить точность, но требует значительных вычислительных ресурсов и может быть неприменим для всех приложений. Интегрированное обучение, предполагает обучение модели с использованием данных из разных источников без обмена данными, что повышает конфиденциальность и снижает риск атак.

Защита моделей ИИ включает обеспечение конфиденциальности данных, используя методы, такие как дифференциальная конфиденциальность, которая добавляет случайный шум к данным. Проблемы защиты включают трудность интерпретации моделей глубокого обучения и сложность разработки эффективных средств защиты. Для решения этих проблем предлагаются лучшие практики, такие как регулярное обновление и тестирование защиты, использование объяснимых методов ИИ и интеграция безопасности в весь процесс разработки ИИ.

Защита моделей ИИ – это сложная, но важная область исследований для обеспечения надежности систем ИИ в разных областях.

3. Методы обнаружения атаки AI или ML

Методы обнаружения атак на модели AI или ML – это инструменты, которые определяют, была ли модель атакована. Они важны для повышения безопасности моделей AI и ML, выявляя потенциальные атаки и уменьшая их последствия.

Использование этих методов позволяет обнаруживать атаки на ранней стадии, снижать ущерб и повышать безопасность и надежность моделей. Организации могут активно отслеживать свои модели и выявлять подозрительные действия, что улучшает их способность быстро реагировать на инциденты безопасности. Эти методы также укрепляют доверие между пользователями и заинтересованными сторонами, показывая приверженность защите конфиденциальных данных и обеспечению точности и надежности моделей AI и ML. Существуют различные типы методов обнаружения.

Защитная дистилляция – это метод обнаружения атак на глубокие нейронные сети, предложенный для защиты от противоречивых примеров, которые заставляют модель машинного обучения делать неверные прогнозы. Метод заключается в обучении второй нейронной сети (дистиллированной модели) для аппроксимации выходных данных исходной модели (модели учителя). Регуляризация – это метод обнаружения для защиты моделей машинного обучения от атак, цель которого – ограничить сложность модели и предотвратить переобучение. Существуют разные типы регуляризации, такие как L1, L2 и отсев, каждый из которых применяет различные штрафы к параметрам модели во время обучения для создания более простых моделей. Регуляризация эффективна для повышения устойчивости моделей к различным атакам, но может быть недостаточной и должна использоваться в сочетании с другими методами обнаружения и мерами безопасности. Обнаружение отклонений в данных – это метод выявления атак на модели AI и ML, который фокусируется на наблюдениях, значительно отличающихся от других в наборе данных, что может указывать на аномалии или потенциальные атаки. Выбросы могут быть обнаружены с помощью статистических методов и методов машинного обучения, таких как кластеризация, классификация и регрессионный анализ. Надежные статистические методы – это важный подход к обнаружению аномалий или выбросов в данных, которые могут вызвать атаки на модели AI и ML. Эти методы используют статистические модели, устойчивые к выбросам, и способны точно выявлять отклонения от ожидаемых закономерностей в данных. Дифференциальная конфиденциальность – это метод защиты конфиденциальной информации при обработке данных, который добавляет случайный шум к данным для скрытия личной информации, сохраняя при этом полезную информацию. В контексте безопасности AI и ML дифференциальная конфиденциальность может предотвратить получение злоумышленниками конфиденциальной информации из обучающих данных или выходных данных модели. Механизмы рандомизированного реагирования – это метод обнаружения атак на AI или ML, который сохраняет конфиденциальность данных при предоставлении статистической информации. Он вносит случайность в данные, скрывая их истинное значение, но сохраняя распределение, похожее на исходные данные.

Атаки ИИ и риски для безопасности и конфиденциальности ИИ вызывают все большую обеспокоенность по мере того, как модели ИИ становятся все более распространенными. В этом обзорном документе рассматриваются различные типы атак ИИ и риски для безопасности и конфиденциальности ИИ. Важно разработать надежные и защищенные модели

искусственного интеллекта, устойчивые к атакам, чтобы обеспечить конфиденциальность и сохранность конфиденциальных данных.

Список литературы

1. C. Campbell, K. Plangger, S. Sands, and J. Kietzmann, “Preparing for an era of deepfakes and ai-generated ads: A framework for understanding responses to manipulated advertising,” Journal of Advertising, vol. 51, no. 1, pp. 22–38, 2022.
2. B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, “The emerging threat of ai-driven cyber attacks: A review,” Applied Artificial Intelligence, vol. 36, no. 1, p. 2037254, 2022.
3. B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, “Trustworthy ai: From principles to practices,” ACM Computing Surveys, vol. 55, no. 9, pp. 1–46, 2023.
4. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2015.
5. N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, “Sok: Security and privacy in machine learning,” 04 2018, pp. 399–414.
6. N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57.
7. F. Tramer, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing machine learning models via prediction apis,” 2016.
8. B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndi c, P. Laskov, G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in Machine Learning and Knowledge Discovery in Databases, H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezn ˇ y, Eds. Berlin, ´ Heidelberg: Springer Berlin Heidelberg, 2013, pp. 387–402.
9. K. Shaukat Dar, S. Luo, V. Varadharajan, I. Hameed, and M. Xu, “A survey on machine learning techniques for cyber security in the last decade,” IEEE Access, 11 2020.
10. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” 2017.
11. M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?”: Explaining the predictions of any classifier,” 2016.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?”: Explaining the predictions of any classifier,” 2016.
13. Z. C. Lipton, “The mythos of model interpretability,” 2017.
14. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, pp. 1929–1958, 06 2014.
15. C. Dwork, “Differential privacy: A survey of results,” in Theory and Applications of Models of Computation, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19.

References

1. C. Campbell, K. Plangger, S. Sands, and J. Kietzmann, “Preparing for an era of deepfakes and ai-generated ads: A framework for understanding responses to manipulated advertising,” Journal of Advertising, vol. 51, No. 1, pp. 22-38, 2022.

2. B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, “The emerging threat of ai-driven cyber attacks: A review,” Applied Artificial Intelligence, vol. 36, No. 1, p. 2037254, 2022.
3. B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, “Trustworthy ai: From principles to practices,” ACM Computing Surveys, vol. 55, no. 9, pp. 1–46, 2023.
4. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2015.
5. N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, “Sok: Security and privacy in machine learning,” 04 2018, pp. 399–414.
6. N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57.
7. F. Tramer, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing machine learning models via prediction apis,” 2016.
8. B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndi c, P. Laskov, G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in Machine Learning and Knowledge Discovery in Databases, H. Blockeel, K. Kersting, S. Nijssen, and F. Zelezn ˇ y, Eds. Berlin, ´ Heidelberg: Springer Berlin Heidelberg, 2013, pp. 387–402.
9. K. Shaukat Dar, S. Luo, V. Varadharajan, I. Hameed, and M. Xu, “A survey on machine learning techniques for cyber security in the last decade,” IEEE Access, 11 2020.
10. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” 2017.
11. M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?”: Explaining the predictions of any classifier,” 2016.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?”: Explaining the predictions of any classifier,” 2016.
13. Z. C. Lipton, “The mythos of model interpretability,” 2017.
14. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, pp. 1929–1958, 06 2014.
15. C. Dwork, “Differential privacy: A survey of results,” in Theory and Applications of Models of Computation, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19.