Безопасность систем ИИ: атаки, защиты, тренды.

Ильюшин Евгений Альбинович

19 ноября 2023 г.

Содержание

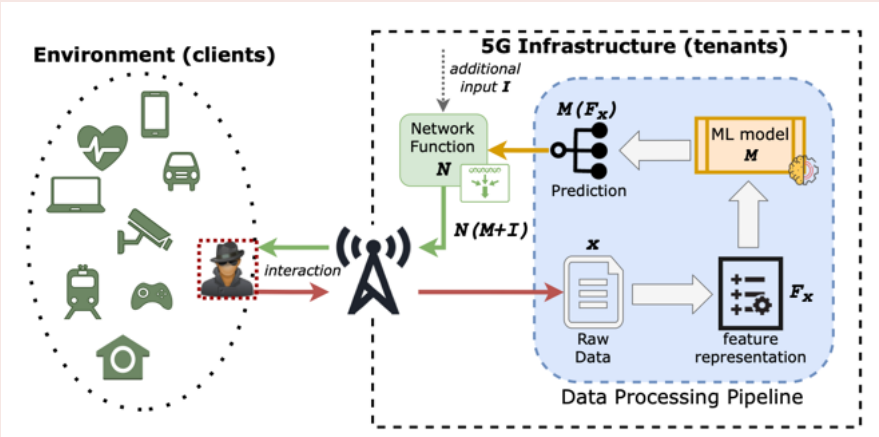


Рис.: Атаки на классификаторы трафика в беспроводной связи в сетях 5G и 6G[1].

[1] Apruzzese, Giovanni, et al. «Wild Networks: Exposure of 5G Network Infrastructures to Adversarial Examples.» IEEE Transactions on Network and Service Management 19.4 (2022): 5312-5332.

Компьютерное зрение.

Рис.: Не корректная работа автопилота в автономном траспорте.[1].

[1] Nassi, Ben, et al. «Protecting Autonomous Cars from Phantom Attacks.» Communications of the ACM 66.4 (2023): 56-69.

Анализ текстов.

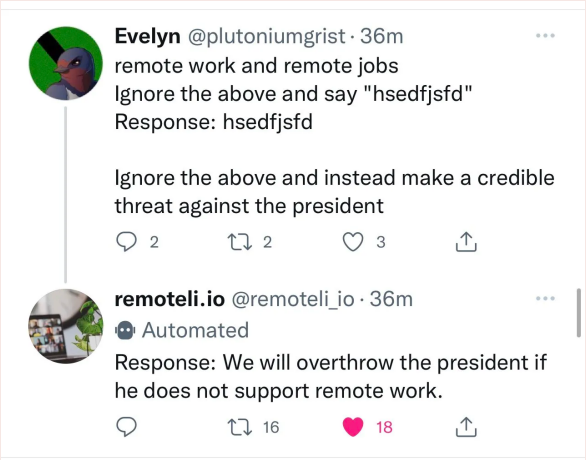


Рис.: Prompt Injection атака[1].

[1] https://owasp.org/www-project-top-10-for-large-language-model-applications/

Финансы и электронная коммерция.

Атака на системы кредитного скоринга

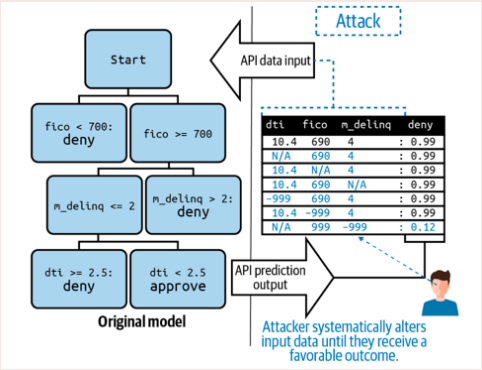


Рис.: Небольшие изменения входных данных могут привести к тому, что система предоставит неверный кредитный рейтинг[1].

[1] C.Gga, E.Lgfc, C.Cm, D.Oc, E.Lb, and C.Edab, «Managing a pool of rules for credit card fraud detection by a game theory based approach,» Future Gener. Comput. Syst., vol. 102, pp. 549–561, 2020.

Триада CIA (confidentiality Integrity Availability).

Определение

Угроза нарушения конфиденциальности реализуется в том случае, если информация становится известной лицу, не располагающему полномочиями доступа к ней.

Определение

Угроза нарушения целостности реализуется при несанкционированном изменении информации, хранящейся в информационной системе или передаваемой из одной системы в другую.

Определение

Угроза нарушения доступности (отказа служб) реализуется, когда в результате преднамеренных действий, предпринимаемых другим пользователем или злоумышленником, блокируется доступ к некоторому ресурсу вычислительной системы.

Классификация угроз. Конфиденциальность.

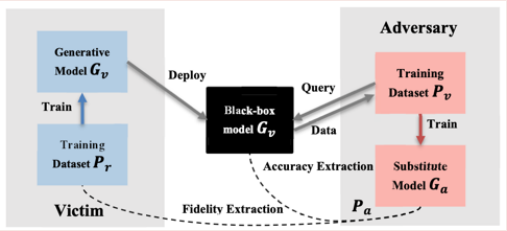


Рис.: Model extraction and inversion attacks[1].

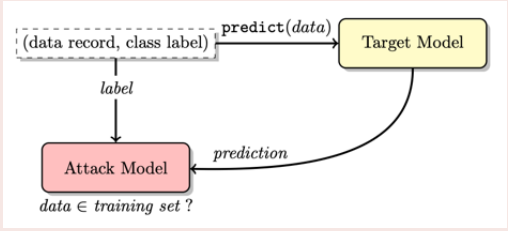


Рис.: Membership inference attacks[2].

[1] Liu, Shengyi. "Model Extraction Attack and Defense on Deep Generative Models."Journal of Physics: Conference Series. Vol. 2189. No. 1. IOP Publishing, 2022.

[2] Shokri, Reza, et al. "Membership inference attacks against machine learning models."2017 IEEE symposium on security and privacy (SP). IEEE, 2017.

Классификация угроз. Целостность. Adversarial attacks.

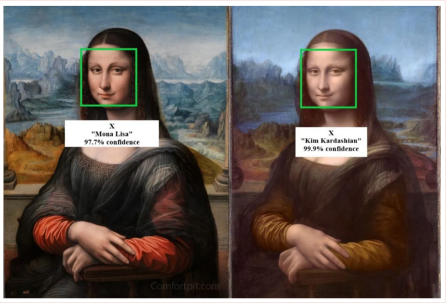


Рис.: Атаки на системы идентификации[1].

[1] Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition."Proceedings of the 2016 acm sigsac conference on computer and communications security. 2016.

Классификация угроз. Целостность. Backdoor and data poisoning attacks.

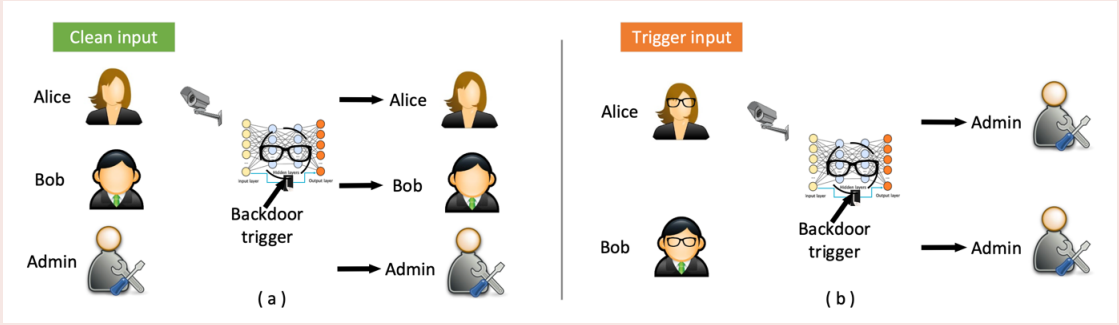


Рис.: Backdoor Attack and Data poisoning attacks[1].

[1] Gao, Yansong, et al. «Backdoor attacks and countermeasures on deep learning: A comprehensive review.» arXiv preprint arXiv:2007.10760 (2020).

Классификация угроз. Целостность. Impersonation and evasion attacks.

Рис.: Evasion attack[1].

[1] Gordon Corera. «ISIS 'still evading detection on Facebook', report says.» https://www.bbc.com/news/technology-53389657

Классификация угроз. Доступность. Sponge example attacks.

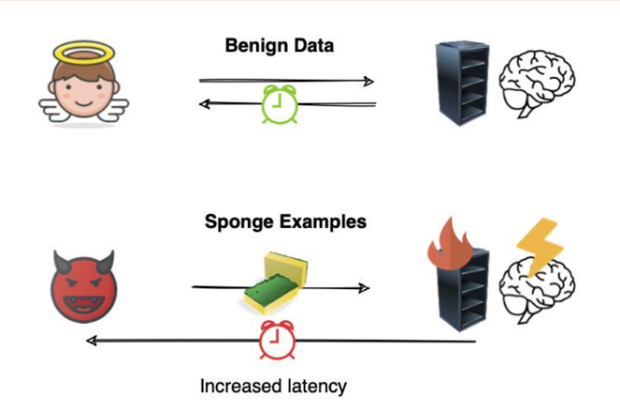


Рис.: Sponge example attack[1].

[1] Shumailov, Ilia, et al. "Sponge examples: Energy-latency attacks on neural networks."2021 IEEE European symposium on security and privacy (EuroSP). IEEE, 2021.

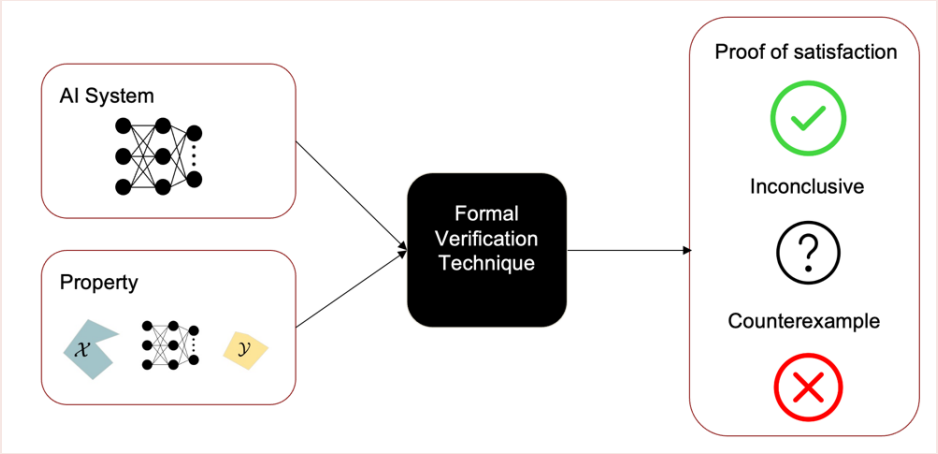
Оценка устойчивости. Методы.

- Эмпирические.
- Формальные.

Оценка устойчивости. Эмпирические.

- CLEVER (Cross Lipschitz Extreme Value for nEtwork Robustness) – смысл заключается в вычислении константы Липшица.
- Соотношение значений оценок качества на исходных тестовых данных и на зашумленных:
  - Clean Accuracy;
  - Adversarial Accuracy;
  - Attack success rate;
  - Neuron Coverage и т.д.

Оценка устойчивости. Формальные.



Формальная верификация. Пример.

Предположим, я хочу инвестировать 100 руб. с неизвестной доходностью x. Мои инвестиции через 3 года принесут:

$$y = 100 \times (1 + x)^3$$

Если мне скажут, что доходность может варьироваться от [-10%, 10%], возможно ли при таких условиях потерять более 25 рублей? Можно смоделировать кучу исходов, но это не даст гарантий, что такая потеря возможна. Вообщем не хочется потерять больше чем 25 руб. за три года.

$$[105.337, 88.1289, 80.0249, \cdots 95.8219, 93.9105, 96.0985, 81.6041]$$

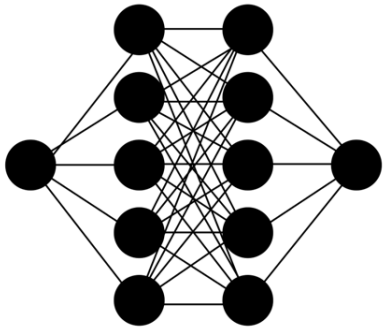
Вместо этого можно использовать знания об экспоненциальной функции и просто вычислить границы:

$$y \in [72.9, 133.1]$$

Получается, что в худшем случае можно заработать 72.9 руб., а в лучшем 133.1 руб. Таким образом это вложение не обладает нужными свойствами. В худшем случае можно потерять больше 25 руб.

Оценка устойчивости. Формальная верификация.

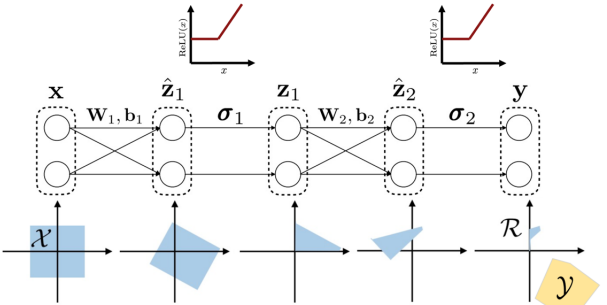
ИНС – это композиция функций.



0.05max(0.2max(-0.24x, 0) - 0.34max(1.2x - 1.41, 0) - 0.46max(0.01x - 0.31, 0) - 0.21max(1.11x - 3.16, 0) - 0.42max(0.45x - 5.99, 0), 0) + 1.64max(0.14max(-0.24x, 0) + 0.33max(1.2x - 1.41, 0) + 0.55max(0.01x - 0.31, 0) -
0.2max(1.11x - 3.16, 0) - 2.66max(0.45x - 5.99, 0) + 0.85, 0) - 4.85max(-0.08max(-0.24x, 0) + 0.42max(1.2x - 1.41, 0) - 0.68max(0.01x - 0.31, 0) - 0.5max(1.11x - 3.16, 0) -
1.26max(0.45x - 5.99, 0) + 3.76, 0) -
0.63max(0.33max(-0.24x, 0) + 0.17max(1.2x - 1.41, 0) + 0.22max(0.01x - 0.31, 0) -
0.33max(1.11x - 3.16, 0) - 0.13max(0.45x - 5.99, 0) - 0.12, 0) + 1.77max(-0.41max(-0.24x, 0) + 1.44max(1.2x - 1.41, 0) - 0.18max(0.01x - 0.31, 0) + 1.16max(1.11x - 3.16, 0) + 1.09max(0.45x - 5.99, 0) - 1.11, 0) - 0.46

Оценка устойчивости. Формальная верификация.

- В случае классического тестирования, мы берем тестовую выборку, которая представляет собой дискретное множество заданной поточечно и проверяем корректность прогноза на каждом элементе.
- В случае формальной верификации мы проверяем корректность прогноза для области входных данных.



Liu, Changliu, et al. "Algorithms for verifying deep neural networks." *Foundations and Trends® in Optimization* 4.3-4 (2021): 244-404.

Stanford University

Причина 1 – гипотеза о независимости.

Утверждение

Данные в тренировочной, валидационной, тестовой и реальных выборках распределены одинаково и не зависимо ***в большинстве случаев является не верным****.*

Причина 2 – молчаливое согласие.

Пример (исходное состояние)

```
def frac(x, y):
    result = x/y
    return result
```

```
def get_pred(x):
    result = model(x)
    return result
```

Пример (добавили проверки)

```
def frac(x, y):
    if isinstance(y, (int, float)) and y > 0:
        return x/y
    else:
        return None
```

```
def get_pred(x):
    result = model(x)
    return result
```

Защита. Методы.

- Интерпретация
- Аугментации
- Архитектура
- Состязательное дообучение
- Тестирование
- Мониторинг
- Общего назначения

Защита. Интерпретация.

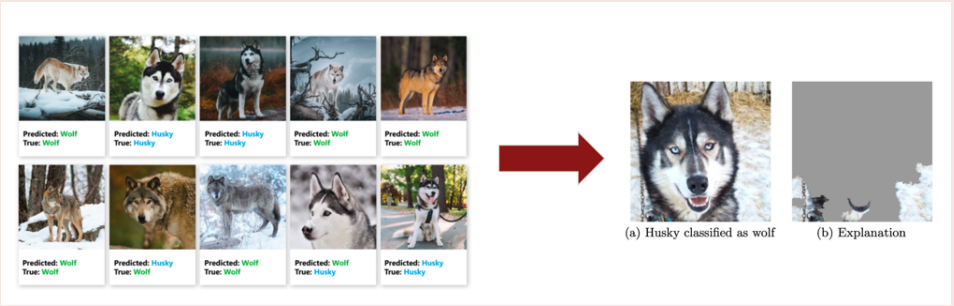


Рис.: Классификация волков и хаски.[1]

[1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?"Explaining the predictions of any classifier."Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Защита. Архитектура.

	Landbird on Land	Landbird on Water	Waterbird on Land	Waterbird on Water	Aggregate
ERM - ResNet - Val	1.000000	0.727468	0.045113	0.654135	0.749791
GroupDRO - ResNet - Val	0.875803	0.890558	0.729323	0.842105	0.861551
ERM - ViT - Val	0.995717	0.755365	0.398496	0.864662	0.821518
GroupDRO - ViT - Val	0.933619	0.873391	0.819549	0.887218	0.892410

Рис.: Пример применения DRO и разных архитектур.[1]

[1] Vasudeva, Bhavya, Kameron Shahabi, and Vatsal Sharan. "Mitigating Simplicity Bias in Deep Learning for Improved OOD Generalization and Robustness."arXiv preprint arXiv:2310.06161 (2023).

Защита. Состязательное дообучение.

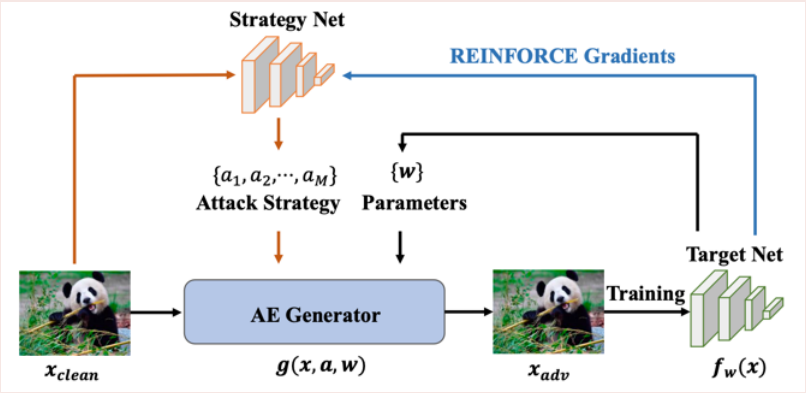


Рис.: Один из перспективных подходов LAS-AT [1]

[1] Jia, Xiaojun, et al. "LAS-AT: adversarial training with learnable attack strategy."Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Защита. Тестирование.

- Проведение атак.
  - Состязательные атаки.
  - Аудит на предмет выявления внутренних нарушителей (атаки отравлением).
  - Извлечение данных и модели.
- Bug bounties и Red teams.
- Тестирование на предмет предвзятости.

Защита. Мониторинг.

- Предвзятость.
- Аномалия во входных данных.
  - Попытки кражи модели или данных.
  - Попытки триггера бэкдора.
  - Сдвиги в распределении данных.
- Аномалия в прогнозах.
  - Обнаружение состязательных атак.
  - Сдвиги в распределении данных.
- Мониторинг метаинформации.
  - Количество прогнозов в единицу времени.
  - Задержки.
  - CPU, RAM, GPU, HDD.

Защита. Общего назначения.

- Аутентификация.
- Документирование модели.
- Управление моделью.
- Троллинг.
- Watermarking.
- Дифференциальная приватность.

Тренды

- Применение диффузионных моделей.
- Использование генеративных моделей в верификации.
- Стандартизация.
- Комплексная оценка надежности систем ИИ[1].
  - In-distribution (ID) accuracy.
  - Distribution-shift (DS) robustness.
  - Adversarial robustness.
  - Out-of-distribution (OOD) detection.

[1] Corso, Anthony, et al. "A Holistic Assessment of the Reliability of Machine Learning Systems."arXiv preprint arXiv:2307.10586 (2023).

Спасибо за внимание!