NIST's Adversarial Machine Learning

Module 4: Discussion and Remaining Challenges

4. Discussion and Remaining Challenges

4.1 The Scale Challenge

- With growing models, the amount of **training data** also increases proportionally, posing a huge challenge.
- No single organization or even a nation possesses the **full data** used for training an LLM.
- **Scale-related issues** also include the ability to generate synthetic content at scale and its possible negative impact on LLMs.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.1 The Scale Challenge

- Open-source **data poisoning tools** increase the risk of large-scale attacks on image training data. While meant to protect copyright, these tools can be harmful if misused.
- The existence of powerful models allows for the generation of massive amounts of unmarked **synthetic content**. Watermarking may alleviate this issue, but unmarked synthetic content can hamper subsequently trained LLMs, possibly causing model collapse.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.2 Theoretical Limitations on Adversarial Robustness

- **Designing mitigations** for AI system attacks is a challenge due to the lack of information-theoretically secure machine learning algorithms.
- A key challenge in the AML field is the ability to **detect when the model is under attack**. Techniques to detect adversarial examples is equivalent to robust classification, which is hard to construct.
- Detection of **out-of-distribution (OOD) inputs** is a crucial challenge in AML too, given adversarial examples may be either from the expected data distribution or OOD ones.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.2 Theoretical Limitations on Adversarial Robustness

- Data and model sanitization techniques can reduce the impact of poisoning attacks. They should be combined with **cryptographic techniques** for origin and integrity attestation, as recommended by the National Security Commission on AI.
- Prompt injections are specific attacks targeted at **chatbots**, imposing rigor to prevent adverse behavior. But limitations require deploying other cybersecurity mechanisms.
- As development of AI-enabled chatbots grow, **risk management** throughout technology life cycle and pre-deployment testing is essential.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.2 Theoretical Limitations on Adversarial Robustness

- Emerging technology like chatbots should be deployed only in apps that have **high degree of trust** with consistent monitoring.
- With increasing deployment of chatbots online, adversaries seek to discover and exploit **vulnerabilities**, while tech companies aim to improve designs against such attacks.
- Identification and mitigation of risks such as bias, discrimination, harmful content generation, privacy violations is crucial.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.2 Theoretical Limitations on Adversarial Robustness

- Robust training techniques offer different approaches to **theoretically certified defenses** against data poisoning attacks, but more research is needed to make them handle OOD inputs and large-scale models.
- There is a lack of **reliable benchmarks** for AML mitigation testing, making proposed mitigations incomparable. Development of standardized benchmarks is essential for reliable insights.
- Formal methods verification, while expensive, can provide **security** and safety assurances, especially for high-risk applications.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.2 Theoretical Limitations on Adversarial Robustness

- AI technology outpaces the development of **mitigation techniques**, leading to privacy attacks and attracting adversaries to expose weaknesses.
- Challenges include: finding ways to **mitigate potential exploits** of memorized data, prevent inference of training data membership or other properties, and protect ML models from intellectual property theft.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.3 The Open vs. Closed Model Dilemma

- Open source has established itself as an **indispensable methodology** for developing software today.
  - With benefits such as democratizing access, leveling the playing field, and enhancing **scientific reproducibility**, it is a powerful tool that bridges performance gaps with closed models.
- On the contrary, there are concerns over the misuse of **open AI technology** by those with malicious intent.
  - This brings to light the question: Should the unrestricted use of open models be allowed?

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.3 The Open vs. Closed Model Dilemma

- Similar question has been proposed in other fields like **cryptography** and bioengineering, each with distinct outcomes.
    - Cryptography risks have been accepted by society leading to strong, publicly available cryptographic algorithms.
    - Conversely, **bioengineering risks** are considered too severe, disallowing open access to the technology.
- The open vs. closed model dilemma in **AI** is being actively debated in the community of stakeholders and should be resolved before models become more powerful and managing becomes uncontrollable.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.4 Supply Chain Challenges

- There is an observed trend in **AML literature** of designing new attacks with higher power and stealthier behavior, bringing challenges to applications using open models downstream the supply chain.
- **DARPA** in collaboration with **NIST** started a program, TrojAI, focusing on defense of AI systems from intentional, malicious Trojan attacks.
- A new class of attacks: **information-theoretically undetectable Trojans**.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.5 Tradeoffs Between the Attributes of Trustworthy AI

- The **trustworthiness** of an AI system is dependent on all its characteristics; any AI system with a vulnerability or bias, despite being accurate, is less likely to be trusted.
- Trade-offs exist between various AI attributes such as **explainability and adversarial robustness**, **privacy and fairness**; optimizing AI for one attribute can lead to underperformance in others
- The exact portrayal of trade-offs between different attributes of trustworthy AI is an open research problem growing in importance

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.5 Tradeoffs Between the Attributes of Trustworthy AI

- **Organizations** need to navigate these trade-offs, deciding what to prioritize based on the AI system, use case, and numerous considerations about the
  - Economic
  - Environmental
  - Social
  - Cultural
  - Political
  - Global implications

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.6 Multimodal Models: Are They More Robust?

- **Multimodal Models:** While they have great potential for achieving high performance, they are not necessarily robust against adversarial perturbations of a single modality even with the redundancy of information across different modalities.
- Researchers have devised efficient mechanisms for constructing simultaneous attacks on multiple modalities.
- Mitigation techniques that only rely on **single modality** perturbations are not likely to be robust.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.6 Multimodal Models: Are They More Robust?

- The existence of simultaneous attacks on multimodal models suggests that mitigation techniques only focusing on single modality perturbations are not likely robust.
- In real life, attackers don't limit themselves to attacks within a given security model but employ any attack available to them indicating that **multimodal models** might not offer improved performance against adversarial attacks.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

4. Discussion and Remaining Challenges

4.7 Quantized Models

- Quantization is a **technique** for efficiently deploying models to edge platforms, reducing computational and memory costs by using low-precision data types.
- Quantized models are susceptible to **adversarial attacks** due to error amplification from reduced computational precision.
- Mitigation techniques for PredAI models exist, but the effects of quantization on GenAI models have been less explored.

Source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.ipd.pdf

QuantUniversity, LLC

www.quantuniversity.com

Thank you!

Contact

Email: info@qusandbox.com

www.QuantUniversity.com