10.21 GRA Meeting

By: Evan Stosic

Overview

- Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance
- Lessons Learned and Future Directions for Security, Resilience and Artificial Intelligence in Cyber Physical Systems
- Questions/Comments

Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance

- 3 vulnerability points for data in AI systems:
  - Data-at-rest
  - Data-in-transit
  - Data-in-use
- This paper only covers cyber systems (not cyber-physical systems)
- Data-at-rest / data-in-transit: data assurance methods do not block or interfere a system's functions
  - Edge computing infrastructure
  - Encryption

Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance

- Data-in-use: resides in volatile memory (RAM), unencrypted and available to compromised applications, firmware, operating systems, and hypervisors
- Three classes of countermeasures:
  - Auto-repair without explicit detection
  - Auto-detection without auto-repair
  - Auto-detect followed by auto-repair
- Question: Does detection / repair significantly impact resilience?
  - Disruption will occur: how costly?

Trustworthy Distributed AI Systems: Robustness, Privacy, and Governance

- Question: When can we can only detect, but cannot auto-repair?
  - The problems can benefit significantly by detection but are hard to repair even by human experts, such as spam or out-of-distribution data
  - The problems can be auto-repaired, but the detection methods lack any built-in capability for auto-repair of the errors incurred due to disruption
  - The problems are attempted for auto-repairing, but the auto-repairing decision depends heavily on the detection methods
- On the modeling side, need to classify what types of attacks/error/problems we can only detect but cannot auto-repair
  - These are more likely to block/interfere with the pipeline

Types of Problems

Table 1. Robustness, privacy, and fairness threats covered in this paper.

		attack target		attack location		attack timing		attack effect
		data	model	client	server	training	inference	
irregular data	OOD	yes	no	no	yes	no	yes	misclassification
	imbalanced	yes	no	yes	no	yes	no	bias
contamination	evasion	yes	no	no	yes	no	yes	misclassification
	poisoning	yes	yes	yes	yes	yes	no	misclassification
	byzantine	no	yes	yes	yes	yes	no	misclassification
	adv. bandit	yes	no	yes	no	yes	no	non-optimal regret
privacy leakage	gradient leakage	yes	no	yes	yes	yes	no	data disclosure
	membership	yes	no	yes	yes	yes	yes	membership disclosure
	attributed	yes	no	yes	yes	no	yes	data disclosure
	extraction	no	yes	yes	yes	no	yes	model disclosure
bias	data collection	yes	no	yes	no	no	no	biased outcome
	data preprocessing	yes	no	yes	no	no	no	biased outcome
	data-driven learning	no	yes	no	yes	yes	no	biased outcome

Cyber Physical Systems

- Encryption insufficient for low-level components
  - Overhead is too high
  - Doesn’t help in terms of compromises at controller level
- Fault Tolerance
  - Byzantine Fault Tolerance (BFT)
- Event-Based Cryptography

Questions/Comments

- From what I was able to find, there was no unified standard about data assurance in AI pipelines, just various methods that one could apply
- To model the specific pipeline we've discussed, I would have to get my hands on this pipeline and see what methods make sense
- In general, is there some specific area that I'm missing currently when doing research? It seems to me at least that finding papers that address our specific concerns has been a challenge