Benchmarking the Robustness of Neuro-Symbolic Learning Against Backdoor Attacks

Author: Andrei Chiru - A.Chiru@student.tudelft.nl; Supervisor: Andrea Agiollo – a.agiollo-1@tudelft.nl; Responsible professor: Kaitai Liang – kaitai.liang@tudelft.nl

1. Background

Neuro-Symbolic Model

- Learns data patterns using neural networks.
- Follows logical relationships from symbolic reasoning engines.
- Considered more resilient than NN [1].

Logic Tensor Networks

- Uses first-order logic.
- Writes formulas as differentiable functions on tensors.

Backdoor Attacks

- Adversary injects hidden trigger in dataset.
- Model outputs attacker-chosen label when the trigger is present.

Clean-Label Data Poisoning

- Only modifies data without labels
- Invisible to human inspection
- Considered PGD [2] based (targeted) implementation and blending based (naïve) implementation

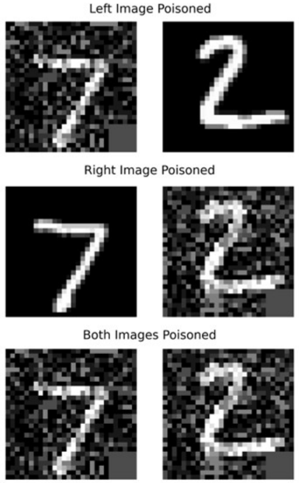
2. Research Question

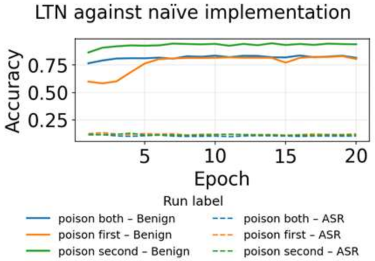
How Robust is Neural-Symbolic Model Logic Tensor Networks Against Label-Consistent Data Poisoning Backdoor Attacks?

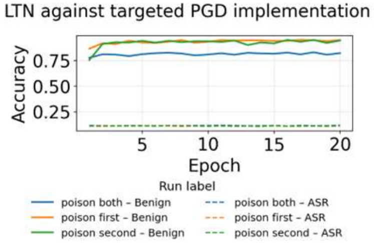
Metrics

- Benign accuracy – how the model performs on normal data
- Attach success rate (ASR) – how well the model predicts the trigger when poisoned

3. Attack Type



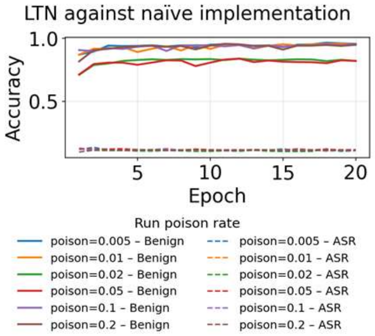


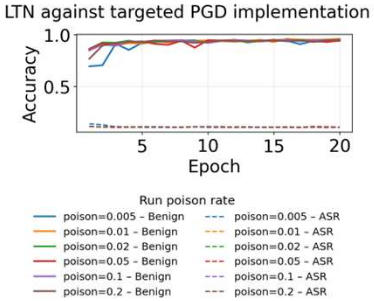


Key Findings

- Poisoning the first image in Naïve harms LTN more, since it's more critical to modulo.
- Poisoning both images in PGD disrupts modulo more - both inputs must be correct.

4. Poison Rate Implications

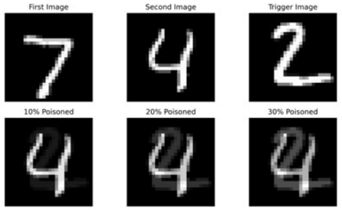


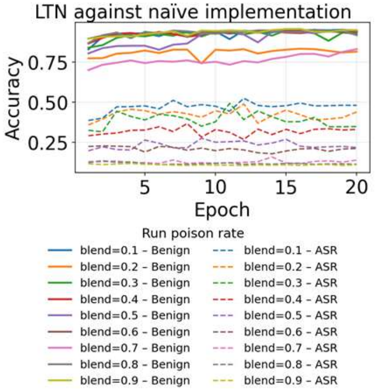


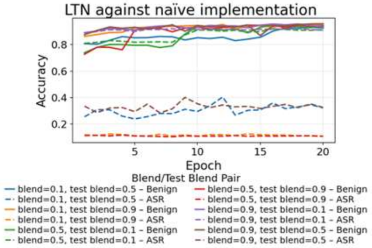
Key Findings

- Naïve attack performance decreases at 2 – 5 % poison rates, while PGD is indifferent to rate changes.
- Pushing poison beyond 10 % gives no extra gain.

5. Trigger Blend Impact



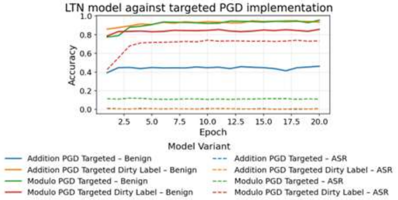




Key Findings

- Small blends (0.1–0.3) expose the trigger, lifting ASR but hurting accuracy.
- Near-invisible 0.9 blends restore accuracy yet kill the attack.

6. Poisoning Labels Effect



Key Findings

- Dirty-label poison increases modulo ASR from 10 to 75%, but leaves addition unchanged.
- Benign accuracy drops to 50% without re-labelling but stays at 95% with it in the addition task

7. Conclusion

- Task-dependent backdoor attacks matter.
- Raising poisoning rate did not change metrics.
- Naïve implementation is not effective on those two tasks.
- Targeted PGD impacted the modulo task slightly, but was not effective against addition.
- Dirty-label backdoor attack has higher attack success rate than clean-label, but it is less stealthy.

8. References

[1] R. Kumar and R. Singh, "A hybrid neuro-symbolic framework for real-time detection of adversarial attacks in cybersecurity", International Research Journal of Engineering and Technology (IRJET), vol. 10, no. 4, 2023. [Online]. Available:https://www.irejournals.com/formatedpaper/1706618.pdf

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2018.