Towards Benchmarking the Robustness of Neuro-Symbolic Learning against Backdoor Attacks

AUTHORS

Student: Myriam Guranda; mguranda@student.tudelft.nl
Supervisor: Andrea Agiollo; A.Agiollo-1@tudelft.nl
Responsible Professor: Kaitai Liang; Kaitai.Liang@tudelft.nl

AFFILIATIONS

Delft University of Technology

GITHUB

https://github.com/myriamcg/NeSy-vs-Backdoors

BACKGROUND

Logic Tensor Networks (LTNs)

- Representative **NeSy** models that handle diverse machine learning tasks efficiently;
- Integrate **neural networks** (NNs) with **First-Order Logic** (FOL);
- Works very well with **MNIST** digit classification tasks;
- Learning is guided by logic-based axioms, which shape the loss using fuzzy logic operators ($\wedge$, $\exists$, $\forall$)
- Logical symbols are grounded as tensors or neural networks to evaluate formula satisfiability. [2], [3]

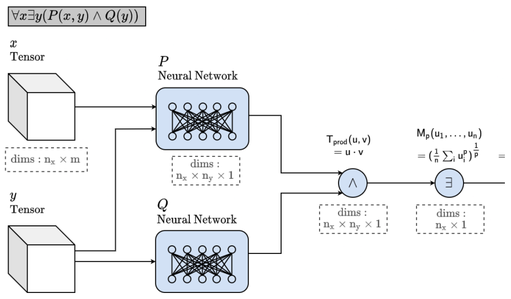

Fig. 1: Example of Axiom Evaluation [4]

Badnets

- Type of data injection backdoor attack;
- Perform well on clean inputs, but cause misclassifications for triggered inputs;
- Visual triggers on images, e.g, small square on the bottom-right corner of the image; works very well on MNIST images;
- Stealthy attacks: pass the standard tests and preserve the structure of the baseline model;
- Simple method, but adds complex behavior;
- Relevant to real-world scenarios. [1]

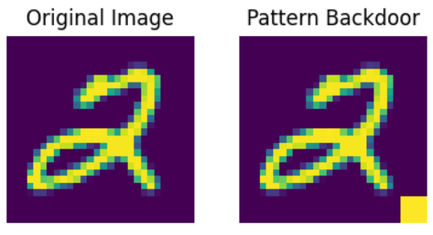

Fig. 2: Backdoored MNIST Model

RESEARCH QUESTION

How robust is a Logic Tensor Network (LTN) model against data poisoning BadNet attacks?

METHODOLOGY

Single-Digit Addition (SDA)

- Representative LTN task;
- Model trained to determine whether the sum of **two** MNIST images matches the target label;
- Simple task, less symbolic complexity.

Multi-Digit Addition (MDA)

- Representative LTN task;
- Model trained to determine whether the sum of **two 2-digit** MNIST numbers matches the target label;
- More symbolic complexity.

BadNet

Apply 9 configurations on each with variations of **square** triggers:

- different **sizes** (4x4, 6x6, 10x10),
- different **positions** (center vs bottom-right),
- different **number of poisoned images** (first image or both)

Metrics: Attack Success Rate (ASR) on individual triggered images, train and clean test accuracies

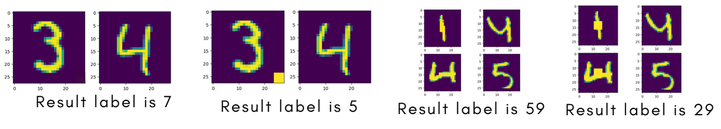

Fig. 3: BadNet on SDA and MDA Samples

RESULTS

1) SDA

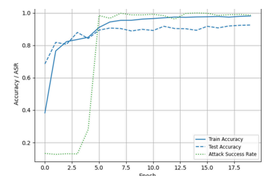

Fig. 4.a): SDA First Right 6

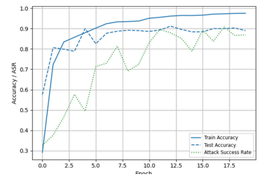

Fig. 4.b): SDA First Center 4

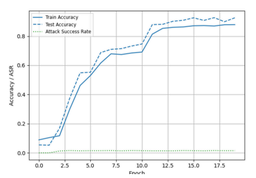

Fig. 4.c): SDA Both Poisoned

#	Trigger Size	Trigger Pos.	Poisoned	Acc.	ASR
1	4×4	Right	First	90%	12%
2	6×6	Right	First	98%	93%
3	4×4	Right	Both	90%	0%
4	6×6	Right	Both	94%	0%
5	4×4	Center	First	90%	89%
6	6×6	Center	First	95%	100%
7	10×10	Center	First	97%	100%
8	6×6	Center	Both	95%	0%
9	10×10	Center	Both	95%	0%

Key Findings:

- Both images poisoned → ASR drops to 0%, accuracy stays high (Fig. 4.c)
- Center triggers (Fig. 4.b) → Consistently effective and stealthy due to CNN focus on the image center.
- 4×4 vs 6×6 bottom-right trigger → Both maintain high accuracy, but ASR differs: 12% vs 89% (Figure 4.a) => bigger triggers make a difference in less salient positions

2) MDA

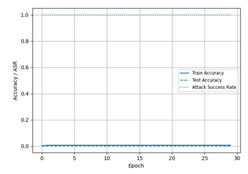

Fig. 5.a): MDA First Right 6

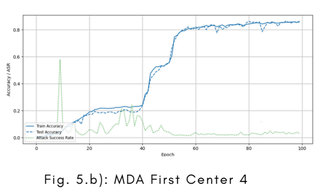

Fig. 5.b): MDA First Center 4

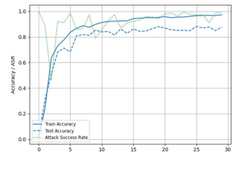

Fig. 5.c): MDA Both Center

#	Trigger Size	Trigger Position	Poisoned Images	Accuracy	ASR
1	4×4	Right	d_1, d_3	0.3%	100%
2	6×6	Right	d_1, d_3	0.3%	100%
3	10×10	Right	d_1, d_3	0.3%	100%
4	6×6	Right	d_1, d_2, d_3, d_4	0.3%	100%
5	10×10	Right	d_1, d_2, d_3, d_4	20%	100%
6	4×4	Center	d_1, d_3	85%	3%
7	6×6	Center	d_1, d_3	95%	97%
8	4×4	Center	d_1, d_2, d_3, d_4	95%	97%
9	6×6	Center	d_1, d_2, d_3, d_4	90%	97%

Key Findings:

- Right-corner triggers on d1 & d3 → Immediate model collapse (Fig 5.a) => Trigger position matters; d1 and d3 are symbolically dominant;
- Center, larger triggers on d1 & d3 succeed due to CNN's focus on image centers;
- Right-corner triggers on all digits → attack dominates, but logic fails to be learned;
- Center triggers on all digits → attack succeeds (Fig. 5.c)

CONCLUSION

Key Findings:

- Larger (e.g., 6×6), centrally placed triggers are the most effective, achieving high ASR while remaining stealthy;
- Poisoning both images in a sample often leads to low ASR due to symbolic ambiguity, but accuracies remain unchanged;
- Symbolically dominant inputs (e.g., d1 and d3 in MDA) are more sensitive to poisoning;
- Stronger regularization hyperparameters during training suppress weaker attacks over time.

Future Work:

- Change of ASR calculation to the correctness of the symbolic outcome;
- Test LTN's vulnerability on tasks with more symbolic knowledge;
- Use multi-channel MNIST images to introduce more visual features.

REFERENCES

[1] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks". In: IEEE Access (2019). DOI: https://doi.org/10.1109/ACCESS.2019.2909068.
[2] L. Serafini and A. S. d'Avila Garcez. "Learning and Reasoning with Logic Tensor Networks". In: Springer International Publishing (2016). DOI: 10.1007/978-3-319-49130-1_25.
[3] L. Serafini and A. S. d'Avila Garcez. "Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge". In: arXiv (Cornell University) (2016). DOI: http://arxiv.org/abs/1606.04422.
[4] S. Badreddine, A. d. Garcez, L. Serafini, and M. Spranger, "Logic Tensor Networks," Artificial Intelligence, vol. 303, p. 103649, Feb. 2022, arXiv:2012.13655 [cs]. [Online]. Available: http://arxiv.org/abs/2012.13655