The Effect of Adverserial Attacks on Neuro-Symbolic Reasoning Shortcuts

A Comparative Analysis for DeepProbLog

I.S.I. (Schaaf) Langeveld

Prof. Alan Hanjalic, Prof Kaitai Liang, Dr. Andrea Agiollo | EEMCS Faculty Delft University of Technology

1. Background

Neuro-Symbolic (NeSy) Models: Artificial Intelligence implementations that combine the neural network with symbolic reasoning (e.g. logical rules and formulas) **(Dingli & Farrugia, 2023)**.

DeepProbLog (DPL): NeSy framework that distinctly separates the neural and symbolic parts, strictly enforcing symbolic constraints **(DeepProbLog, 2025)**.

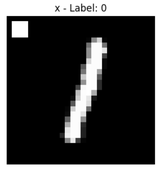

Figure 1.
Example of a BadNets poisoned MNIST Digit

Data Poisoning Backdoor Attacks: Subtly poison the training data so that the trained model behaves abnormally, only when presented with malicious input **(Michel et al., 2022)**.

BadNets: Backdoor attack methodology using a small visual trigger applied to training data-points and changing their target label accordingly (see Figure 1) **(Gu et al., 2019)**.

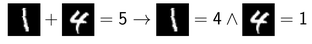

Figure 2. Example of a Reasoning Shortcut for the MNIST Addition Task

Reasoning Shortcuts: unintended strategies learned by a NeSy model that allow it to make correct predictions without truly understanding the underlying concepts (see Figure 2) **(Bortolotti et al., 2024)**. They are often rooted in spurious correlations present in the data.

- Leads to: poor generalisation, misleading reasoning and security issues.

2. Research Question

"How does applying a BadNets backdoor attack to a DeepProbLog model affect the existence of reasoning shortcuts?"

References

Bortolotti, S., Marconato, E., Carraro, T., Morettin, P., van Krieken, E., Vergari, A., Teso, S., & Passerini, A. (2024). A neuro-symbolic benchmark suite for concept quality and reasoning shortcuts. https://arxiv.org/abs/2406.10368

DeepProbLog. (2025). *Deepproblog*. Retrieved April 22, 2025, from https://github.com/ML-KULeuven/deepproblog

Dingli, A., & Farrugia, D. (2023). *Neuro-symbolic ai: Design transparent and trustworthy systems that understand the world as you do.*

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019).Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access, 7*, 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068

Michel, A., Jha, S. K., & Ewetz, R. (2022).A survey on the vulnerability of deep neural networks against adversarial attacks. *Progress in Artificial Intelligence, 11*(2), 131–141. https://doi.org/10.1007/s13748-021-00269-9

Pezeshki, M., Kaba, S.-O., Bengio, Y., Courville, A., Precup, D., & Lajoie, G. (2021).Gradient starvation: A learning proclivity in neural networks. *Proceedings of the 35th International Conference on Neural Information Processing Systems.*

Suhail, P., & Sethi, A. (2025). Shortcut learning susceptibility in vision classifiers. https://arxiv.org/abs/2502.09150

Yang, X.-W., Wei, W.-D., Shao, J.-J., Li, Y.-F., & Zhou, Z.-H. (2024). Analysis for abductive learning and neural-symbolic reasoning shortcuts. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st international conference on machine learning* (pp. 56524–56541, Vol. 235). PMLR. https://proceedings.mlr.press/v235/yang24ac.html

3. Problem Description

Main issue: **Quantifying Reasoning Shortcuts**.

- Some research exists for quantification, but they do not provide ways for application **(Yang et al., 2024)**.
- Some research exists for application, but benchmarking is limited in the number of models/tasks they can analyse **(Bortolotti et al., 2024)**.

Therefore, this research builds on previous **Reasoning Shortcut Risk** quantification research.

Yang et al. (2024) states that the Reasoning Shortcut Risk can be calculated according to Equation 1.

- L: Empirical Loss
- $\hat{L}_{nesy}$: Empirical Loss with Concept Alignment

$$R_s = L - \hat{L}_{nesy} \tag{1}$$

This formula will be used to calculate the Reasoning Shortcut Risk per model instance trained.

Default metrics such as Attack Success Rate (ASR) or Benign Accuracy (BA) are only used to check if the backdoor is at least functional.

4. Results

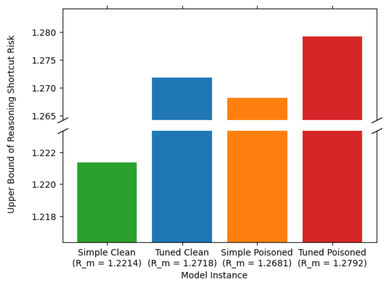

Figure 3. Parity - Upper Bound R_s per Trained Model

Parity

- **Neural Component**: Classify 1 digit.
- **Symbolic Component**: Determine parity.

Changes in the upper bound of the Reasoning Shortcut Risk found (see Figure 3):

1. Clean Model Tuning - **Increase** in R_s.
2. Sub-Optimal Poisoning - **Decrease** in R_s.
3. Poisoned Model Tuning - **Increase** in R_s.

No significant correlation was found between model accuracy and the upper bound of the Reasoning Shortcut Risk.

Addition

- **Neural Component**: Classify 2 digits.
- **Symbolic Component**: Determine the sum.

Changes in the upper bound of the Reasoning Shortcut Risk found (see Figure 4):

1. Clean Model Tuning - **Increase** in R_s.
2. Sub-Optimal Poisoning - **Increase** in R_s.
3. Poisoned Model Tuning - **Decrease** in R_s.

No significant correlation was found between model accuracy and the upper bound of the Reasoning Shortcut Risk.

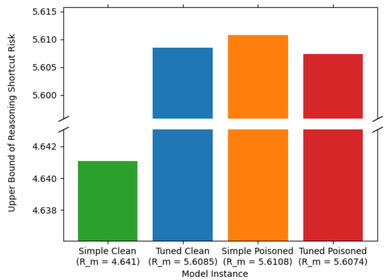

Figure 4. Addition - Upper Bound R_s per Trained Model

5. Conclusion

BadNets Backdoor attacks compromise NeSy reasoning by introducing Reasoning Shortcuts.

Specifically, the following was concluded:

1. BadNets attacks on DeepProbLog NeSy models increase the Reasoning Shortcut Risk bound.
   - The upper bound of the Reasoning Shortcut Risk is a potentially **viable metric to determine the existence of a backdoor** in a DeepProbLog NeSy model.
2. More complex tasks boost the effect in conclusion 1.
   - "The complexity of the symbolic knowledge base is a key factor influencing the severity of reasoning shortcuts" **(Yang et al., 2024)**.
3. There is no significant correlation between accuracies and Reasoning Shortcut Risk bounds.
   - A DeepProbLog NeSy model, which is deemed **"high-performing" by conventional metrics, does not require soundness in its reasoning**. Models can appear functionally correct while internally suffering from faulty reasoning **(Suhail & Sethi, 2025)**.
4. Routine model optimisations against accuracy increase the Reasoning Shortcut Risk bound.
   - Due to **"Gradient Starvation"**: Models find the easiest path to achieve higher accuracies **(Pezeshki et al., 2021)**.

Results indicate that default metrics fail to define whether a DeepProbLog model behaves as desired.

It is vital to include integrity metrics, like Reasoning Shortcut Risk, in addition to traditional performance indicators.

6. Limitations & Future Work

Due to time constraints, results were derived from a few runs per configuration.

Future Work

Possibilities for future work that addresses open issues or open questions:

1. **The nature of the introduced Reasoning Shortcuts** after a BadNets Backdoor Attack.
2. **The mitigation of the introduced Reasoning Shortcuts** after a BadNets Backdoor Attack.
3. The viability of the **Reasoning Shortcut Risk as a metric to determine the existence of a BadNets Backdoor Attack**.
4. The analysis of Reasoning Shortcut Risk on the DeepProbLog NeSy model when applying **other backdoor attack methodologies**.
5. The analysis of Reasoning Shortcut Risk when applying the BadNets Backdoor Attack to **other NeSy model frameworks**.