Neuro-Symbolic Reasoning Shortcuts: Mitigation Strategies and their Limitations

Emanuele Marconato[1,2], Stefano Teso[1,3] and Andrea Passerini[1]

[1]*Department of Engineering and Information Science (DISI), University of Trento, Italy*

[2]*Department of Computer Science (DI), University of Pisa, Italy*

[3]*Centre for Mind/Brain Sciences (CIMeC), University of Trento, Italy.*

Abstract

Neuro-symbolic predictors learn a mapping from sub-symbolic inputs to higher-level concepts and then carry out (probabilistic) logical inference on this intermediate representation. This setup offers clear advantages in terms of consistency to symbolic prior knowledge, and is often believed to provide interpretability benefits in that – by virtue of complying with the knowledge – the learned concepts can be better understood by human stakeholders. However, it was recently shown that this setup is affected by *reasoning shortcuts* whereby predictions attain high accuracy by leveraging concepts with unintended semantics [1, 2], yielding poor out-of-distribution performance and compromising interpretability. In this short paper, we establish a formal link between reasoning shortcuts and the optima of the loss function, and identify situations in which reasoning shortcuts can arise. Based on this, we discuss limitations of natural mitigation strategies such as reconstruction and concept supervision.

1. Introduction

Neuro-symbolic (NeSy) integration of learning and reasoning is a key challenge in AI. NeSy *predictors* achieve integration by learning a *neural network* mapping low-level representations (e.g., MNIST images) to high-level symbolic concepts (e.g., digits), and then predicting a label (e.g., the sum) by *reasoning* over concepts and prior knowledge [3]. Most works on the topic focus on how to best integrate knowledge into the loop, cf. [4]. The issue of *concept quality* is, however, generally neglected. Loosely speaking, the consensus is that knowledge ensures learning high quality concepts and that issues with these should be viewed as "learning artifacts".

This is not the case. Recently, Li et al. [2] and Marconato et al. [1] have shown that NeSy predictors can learn *reasoning shortcuts* (RSs), that is, mappings from inputs to concepts that yield high accuracy on the training set by predicting the *wrong* concepts. While RSs – by definition – do not hinder the model's accuracy on the training task, they prevent identification of concepts with the "right" semantics, and as such compromise *generalization* beyond the training distribution and *interpretability* [1]. As an example, consider MNIST Addition [3]. Here, the model has to determine the sum of two MNIST digits, under the constraint that the sum is correct. Given the examples "$0 + 1 = 1$" and "$0 + 2 = 2$", there exist two alternative
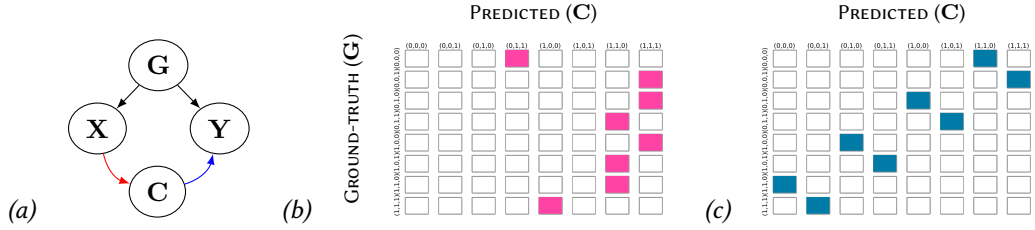
Figure 1: *(a)* Graphical model of our setup: the **black** arrows encode the data generation process, the **red** arrow indicate the learned concept distribution, and the **blue** arrow the reasoning module. *(b)* Confusion matrices of (Boolean) concepts learned by DeepProblog for our three bits XOR task without any mitigation strategy, and *(c)* with a reconstruction term, cf. Eq. (5). In both cases the learned concepts are consistent with the knowledge and in the second one they also manage to reconstruct the input. The confusion matrices immediately show that, despite this, the learned concepts are reasoning shortcuts.

solutions: the intended one (⓿ → 0, ❶ → 1, ❷ → 2) and a RS (⓿ → 1, ❶ → 0, ❷ → 1). Both of them ensure the sum is correct, but only one of them captures the correct semantics.

This begs the question: *under what conditions do reasoning shortcuts appear, and what strategies can be used to mitigate them?* In this short paper, we outline answers to these questions. First, we go beyond existing works and show how to *count* the number of RSs affecting a NeSy prediction task. Based on this result, we show that, in the general case, it is *impossible* to identify the correct concepts from label supervision only. We also consider two mitigation strategies, namely reconstruction and concept supervision, and study their effects and limitations.

2. Neuro-symbolic task construction

We consider a NeSy prediction task where, given sub-symbolic inputs $\mathbf{X}$, the goal is to infer one or more labels $\mathbf{Y} \in \{0,1\}^{\ell}$ consistent with a given propositional formula K encoding prior knowledge. We focus on DeepProbLog [3], a representative and sound framework for such tasks. From a probabilistic perspective, DeepProbLog: (i) Extracts k concepts $\mathbf{C} \in \{0,1\}^{k}$ from a $\mathbf{X}$ via a neural network $p_{\theta}(\mathbf{C} \mid \mathbf{X})$, and (ii) Models the distribution over the labels $\mathbf{Y}$ as a uniform $u_{\mathsf{K}}(\mathbf{y} \mid \mathbf{c}) = \mathbb{1}\{(\mathbf{y}, \mathbf{c}) \models \mathsf{K}\}$. The label distribution is obtained by marginalizing $\mathbf{C}$:

$$p_{\theta}(\mathbf{y} \mid \mathbf{x}; \mathsf{K}) = \sum_{\mathbf{c}} u_{\mathsf{K}}(\mathbf{y} \mid \mathbf{c}) p_{\theta}(\mathbf{c} \mid \mathbf{x}) \tag{1}$$

DeepProbLog is then trained via *maximum likelihood*.

In order to understand when doing so recovers concepts $\mathbf{C}$ with the "correct semantics", we have to first define the unobserved generative mechanism underlying the training data whose concepts we wish to identify. Motivated by work on identifiability in (causal) representation learning [5, 6, 7, 8], we assume there exist k ground-truth concepts $\mathbf{G} \in \{0,1\}^{k}$ spanning a space $\mathcal{G}$, and that the examples $(\mathbf{X}, \mathbf{Y}) = (f(\mathbf{G}), h(\mathbf{G}))$ are generated by an invertible function $f : \mathcal{G} \to \mathcal{X} \subset \mathbb{R}^{d}$ and a surjective function $h : \mathcal{G} \to \mathcal{Y}$, with $|\mathcal{Y}| \leq |\mathcal{G}|$. Here, h plays the role of the ground-truth reasoning module that infers the label $\mathbf{Y}$ from the ground-truth concepts $\mathbf{G}$ according to K, while f generates the observations themselves.[1] Cf. Fig. 1 for an

---

[1]Due to space constraints, we assume $\mathbf{X}$ depends on $\mathbf{G}$ only. In practice, it might also depend on additional "stylistic"

illustration. In the next sections, we will show how maximum likelihood training can recover the mechanism $g \circ f^{-1}$, but not the ground-truth mapping from inputs to concepts f^{-1}, *i.e.*, the "correct semantics".

3. Reasoning shortcuts and mitigation strategies

We consider training points $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\mathbf{X}, \mathbf{Y}}$, each originated by corresponding ground-truth concepts $\mathbf{g} \in \mathcal{D}_{\mathbf{G}}$.[2] Our starting point is the log-likelihood, which constitutes the objective of training:

$$\mathcal{L}(\theta) := \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\mathbf{X}, \mathbf{Y}}} \log p_\theta(\mathbf{y} \mid \mathbf{x}; \mathsf{K}) \equiv \sum_{\mathbf{g} \in \mathcal{D}_{\mathbf{G}}} \log p_\theta \big(h(\mathbf{g}) \mid f(\mathbf{g}); \mathsf{K} \big) \qquad (2)$$

Notice that all optima of Eq. (2) satisfy $p_\theta(\mathbf{y} \mid \mathbf{x}; \mathsf{K}) = 1$ for all examples. By Eq. (1), this entails that any $\mathbf{c} \sim p_\theta(\mathbf{c} \mid \mathbf{x})$ must satisfy the knowledge K, that is, $(\mathbf{c}, \mathbf{y}) \models \mathsf{K}$ (see [1, Theorem 3.2]). How many alternative distributions $p_\theta(\mathbf{c} \mid \mathbf{g}) := p_\theta(\mathbf{c} \mid f(\mathbf{g}))$ do attain maximum likelihood? Since p_θ is a neural network, there may be infinitely many, yet all of them except one are RSs. This is sufficient to show that *RSs cannot be discriminated from the ground-truth concept distribution based on likelihood alone* [1].

Importantly, it turns out all optimal distributions $p_\theta(\mathbf{c} \mid \mathbf{x})$ are convex combinations of the *deterministic* optima (*det-opts*), that is, those distributions $p_\theta(\mathbf{c} \mid \mathbf{g})$ mapping each $\mathbf{g}$ to a unique $\mathbf{c}$ with probability one. If the likelihood admits a single *det-opt*, this is also the *only* solution and – by construction – it recovers the ground-truth concepts. *RSs arise when there are two or more det-opts.* How many det-ops are there? Let $S_{\mathbf{y}} = \big\{ \mathbf{c} : (\mathbf{c}, \mathbf{y}) \models \mathsf{K} \big\}$ be the set of $\mathbf{c}$'s that K assigns to label $\mathbf{y}$. Notice that if $p_\theta(\mathbf{c} \mid \mathbf{g})$ attains maximum likelihood, then any $\mathbf{c} \sim p_\theta(\mathbf{c} \mid \mathbf{g})$ falls within $S_{h(\mathbf{g})}$. In this sense, a *det-opt* implicitly maps each vector $\mathbf{g} \in \mathcal{G}$ to a vector $\mathbf{c} \in S_{h(\mathbf{g})}$. This gives us a mechanism to count *det-opts*: for each $\mathbf{g}$ there are exactly $|S_{h(\mathbf{g})}|$ vectors $\mathbf{c}$ that it can be mapped to, meaning that number of *det-opts* for Eq. (2) is:

$$\#\text{det-opts}(\mathcal{L}) = \prod_{\mathbf{y} \in \mathcal{Y}} |S_{\mathbf{y}}|^{|S_{\mathbf{y}}|} \qquad (3)$$

As a consequence, the ground-truth concepts can only be retrieved if $|S_{\mathbf{y}}| = 1$, *i.e.*, each label $\mathbf{y}$ can be deduced from a unique $\mathbf{c}$. This is seldom the case in NeSy tasks, meaning that maximizing the likelihood of the labels $\mathbf{Y}$ cannot rule out RSs in general.

In the following, we discuss two natural mitigation strategies and their impact in reducing the total number of *det-ops*.

Reconstruction is insufficient. Given the likelihood is incapable of discriminating intended and RS solutions, one option is to augment it with a term encouraging learned concepts $\mathbf{C}$ to capture information necessary to reconstruct the input $\mathbf{X}$, for instance:

$$\mathcal{R}(\theta) = \sum_{\mathbf{x} \in \mathcal{D}_{\mathbf{X}}} \Big[\sum_{\mathbf{c}} p_\theta(\mathbf{c} \mid \mathbf{x}) \log p_\psi(\mathbf{x} \mid \mathbf{c}) \Big] \equiv \sum_{\mathbf{g} \in \mathcal{D}_{\mathbf{G}}} \Big[\sum_{\mathbf{c}} p_\theta(\mathbf{c} \mid \mathbf{g}) \log p_\psi(\mathbf{g} \mid \mathbf{c}) \Big] \quad (4)$$

---

factors of variation (e.g., font) [9]. Our results apply to this more complex case with minimal modifications.
[2]We assume the training examples are *noiseless* and cover all possible combinations of ground-truth factors $\mathbf{G}$, as even this "ideal" setting admits RSs.

Here, $p_\psi(\mathbf{x} \mid \mathbf{c})$ is the distribution output by a neural decoder with parameters ψ, and we introduced $p_\psi(\mathbf{g} \mid \mathbf{c}) := p_\psi(f(\mathbf{g}) \mid \mathbf{c})$. The optima of Eq. (4) must satisfy $p_\psi(\mathbf{g} \mid \mathbf{c}) = 1$ for all $\mathbf{c} \sim p_\theta(\mathbf{c} \mid \mathbf{g})$. In other words, restricting again to *det-ops* for the encoder, the only *det-ops* that ensure perfect reconstruction are those mapping distinct $\mathbf{g}$'s to distinct $\mathbf{c}$'s, *i.e.*, that ensure the encoder is injective. How many such *det-opts* are there? Notice that these *det-opts* can be enumerated by taking each $\mathbf{g} \in \mathcal{G}$ in turn and mapping it to an arbitrary $\mathbf{c}$ in $S_{h(\mathbf{g})}$ *without replacement* (to ensure injectivity), until all $\mathbf{g}$'s have been mapped. This entails that the number of *det-opts* – under perfect reconstruction – becomes:

$$\text{\#det-opts}(\mathcal{L} + \mathcal{R}) = \prod_{\mathbf{y} \in \mathcal{Y}} |S_\mathbf{y}|! \tag{5}$$

Once again, unless $|S_\mathbf{y}| = 1$ for all $\mathbf{y}$'s, there are multiple possible solutions, most of which are RSs. In other words, *adding a reconstruction term can be insufficient to completely rule out learning reasoning shortcuts.*

The effect of concept supervision.   Next, we consider a scenario where *concept supervision* is provided (for all concepts) for at least some examples $(\mathbf{x}, \mathbf{g}) \in \overline{\mathcal{D}}_{\mathbf{X},\mathbf{G}}$. We consider the L_2 loss for fitting the supervision, for simplicity:

$$\mathcal{C}(\theta) \propto \sum_{(\mathbf{x},\mathbf{g}) \in \overline{\mathcal{D}}_{\mathbf{X},\mathbf{G}}} (\mathbf{c} - \mathbf{g})^2 p_\theta(\mathbf{c} \mid \mathbf{x}) \equiv \sum_{\mathbf{g} \in \overline{\mathcal{D}}_\mathbf{G}} \sum_\mathbf{c} (\mathbf{c} - \mathbf{g})^2 p_\theta(\mathbf{c} \mid \mathbf{g}) \tag{6}$$

The only concept distributions $p_\theta(\mathbf{c} \mid \mathbf{g})$ minimizing Eq. (6) are those that allocate all probability mass to the annotated concepts. Now, let $\nu_\mathbf{y}$ be the number of vectors $\mathbf{c} \in S_\mathbf{y}$ for which we have supervision $\mathbf{g}$, for a total of $|\overline{\mathcal{D}}_\mathbf{G}| = \sum_\mathbf{y} \nu_\mathbf{y}$. The situation is analogous to Eq. (3) and Eq. (5), except that now for exactly $\nu_\mathbf{y}$ vectors $\mathbf{c}$ we know exactly what $\mathbf{g}$ they should be mapped to, leaving the remaining $|S_\mathbf{y}| - \nu_\mathbf{y}$ vectors dangling. This gives:

$$\text{\#det-opts}(\mathcal{L} + \mathcal{C}) = \prod_{\mathbf{y} \in \mathcal{Y}} |S_\mathbf{y}|^{|S_\mathbf{y}| - \nu_\mathbf{y}}, \qquad \text{\#det-opts}(\mathcal{L} + \mathcal{R} + \mathcal{C}) = \prod_{\mathbf{y} \in \mathcal{Y}} (|S_\mathbf{y}| - \nu_\mathbf{y})! \tag{7}$$

Here, the first term counts how many *det-opts* optimize both the label likelihood and the concept supervision, and the second one those optimizing the likelihood, reconstruction and concept supervision. This shows providing concept supervision can dramatically reduce the number of *det-opts* but also that *a substantial amount is necessary to rule out all RSs.*

4. Empirical Verification

We outline a toy experiment showing how reasoning shortcuts affect even a simple NeSy task. Let $\mathbf{g} = (g_1, g_2, g_3)$ be three bits and consider the task of predicting their parity, that is, $y = g_1 \oplus g_2 \oplus g_3$. Each label $y \in \{0, 1\}$ can be deduced from 4 possible concept vectors $\mathbf{g}$. We train two MLPs, one encoding directly $\mathbf{g}$ into $p_\theta(\mathbf{c} \mid \mathbf{g})$, and another decoding $\mathbf{c}$ into $p_\psi(\mathbf{g} \mid \mathbf{c})$. Labels are predicted as per Eq. (1). Given the problem at hand, the total number of *det-opts* given by Eq. (3) is $\text{\#det-opts}(\mathcal{L}) = (4^4 \cdot 4^4)$, and that given by Eq. (5) is $\text{\#det-opts}(\mathcal{L} + \mathcal{R}) = (4! \cdot 4!)$. Empirically, what happens is that without concept supervision, the model picks up reasoning shortcuts to solve the task. Fig. 1 shows two such RSs, both optimal, obtained by our model when optimizing *(b)* only the likelihood, and *(c)* both the likelihood and the reconstruction term. In both cases, the solutions fail to recover the ground-truth concepts.

Conclusion. Our results altogether show that the ground-truth concepts are hard, if not impossible, to recover empirically, and that two natural mitigation strategies do not completely address the problem. In particular, the amount of concept supervision required grows linearly with the number of possible concept combinations. We envisage well-tuned strategies based on targeted concept-supervision, combined with additional restrictions on the model itself (and specifically *disentanglement* between concepts [10]), will likely facilitate (provable) identification of the ground-truth concepts. This is left to future work.

References

[1] E. Marconato, G. Bontempo, E. Ficarra, S. Calderara, A. Passerini, S. Teso, Neuro symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal, arXiv preprint arXiv:2302.01242 (2023).

[2] Z. Li, Z. Liu, Y. Yao, J. Xu, T. Chen, X. Ma, L. Jian, et al., Learning with logical constraints but without shortcut satisfaction, in: The Eleventh International Conference on Learning Representations, 2023.

[3] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, L. De Raedt, DeepProbLog: Neural Probabilistic Logic Programming, NeurIPS (2018).

[4] L. De Raedt, S. Dumančić, R. Manhaeve, G. Marra, From statistical relational to neural-symbolic artificial intelligence, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 4943–4950.

[5] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in: ICML, 2019.

[6] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Toward causal representation learning, Proceedings of the IEEE (2021).

[7] I. Khemakhem, D. Kingma, R. Monti, A. Hyvarinen, Variational autoencoders and nonlinear ICA: A unifying framework, in: AISTATS, 2020.

[8] K. Ahuja, D. Mahajan, V. Syrgkanis, I. Mitliagkas, Towards efficient representation identification in supervised learning, in: Conference on Causal Learning and Reasoning, PMLR, 2022, pp. 19–43.

[9] J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, F. Locatello, Self-supervised learning with data augmentations provably isolates content from style, in: NeurIPS, 2021.

[10] R. Suter, D. Miladinovic, B. Schölkopf, S. Bauer, Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness, in: International Conference on Machine Learning, PMLR, 2019, pp. 6056–6065.