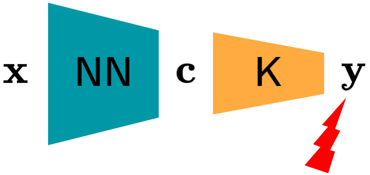
BEARS Make Neuro-Symbolic Models Aware of their Reasoning Shortcuts

Emanuele Marconato[1,2,*], Samuele Bortolotti[1,*], Emile van Krieken[3,*], Antonio Vergari[3], Andrea Passerini[1], Stefano Teso[1]

[1] University of Trento, [2] University of Pisa, [3] University of Edinburgh, [*] Equal contribution

NeSy predictors like **DeepProbLog** [1] and **LTN** [2] combine perception and reasoning.



$\mathbf{x}$: input, $\mathbf{y}$: labels
$\mathbf{c}$: concepts (discrete)
NN: encoder w/ params θ
K: prior knowledge

Trained to achieve maximum log-likelihood (MLL):

$$\underset{\theta}{\text{argmax}} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \log p_\theta(\mathbf{y} \mid \mathbf{x}; \mathsf{K})$$

[1] Manhaeve et al., DeepProbLog: Neural Probabilistic Logic Programming, NeurIPS (2018)
[2] Donadello et al., Logic Tensor Networks for Semantic Image Interpretation, IJCAI (2017)

REASONING SHORTCUTS (RSs)

NeSy models can achieve high accuracy by learning unintended concepts! [3]

$\equiv$ Non-identifiability of latent concepts

OPEN PROBLEMS

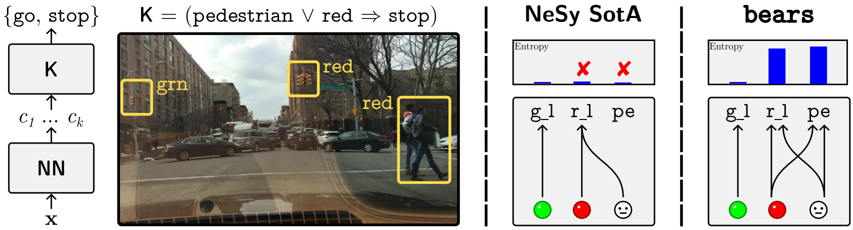
- RSs compromise **OOD generalization**
- **Effective mitigation**, like concept supervision [3], is often **impractical**.
- **Over-confidence** in predicted concepts: impossible to spot the wrong ones!

[3] Marconato et al., Not All Neuro-Symbolic Concepts are created Equal: Analysis and Mitigation of Reasoning Shortcuts, NeurIPS (2023)

🐻 BEARS: BE AWARE OF REASONING SHORTCUTS! 🐻

- combines **Deep Ensembles** [4] **+ diversification** and optimizes for all desiderata:

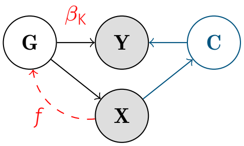
$$\mathcal{L} = \mathcal{L}(\mathbf{x}, \mathbf{y}; \mathsf{K}, \theta_{t+1}) \; + \gamma_1 \cdot \mathrm{KL}\Big(p_{\theta_{t+1}}(\mathbf{C} \mid \mathbf{x}) \; || \; \frac{1}{t} \sum_{j=1}^{t} p_{\theta_j}(\mathbf{C} \mid \mathbf{x})\Big) + \gamma_2 \cdot H(p_{\theta_{t+1}}(\mathbf{C} \mid \mathbf{x}))$$



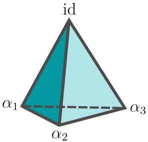
THEORY

Under two assumptions (in **red** below) we show:

1. All (stochastic) RSs live in a simplex
2. Average different RSs = entropy maximization
3. Ensembles + KL = average different RSs



Data Generation Process



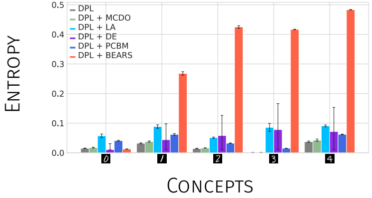
Simplex of Optima

POSTER # 691

EXPERIMENTS

① **An example from `MnAddHalf`** Solve the sum between two digits, *e.g.*,

 +  = 5.



② **Active learning with `bears`**