A Neuro-Symbolic Benchmark Suite for Concept Quality and Reasoning Shortcuts

Samuele Bortolotti *samuele.bortolotti@unitn.it*

Emanuele Marconato *emanuele.marconato@unitn.it*

Tommaso Carraro *tcarraro@fbk.eu*

Paolo Morettin *paolo.morettin@unitn.it*

Emile van Krieken *Emile.van.Krieken@ed.ac.uk*

Antonio Vergari *avergari@ed.ac.uk*

Stefano Teso *stefano.teso@unitn.it*

Andrea Passerini *andrea.passerini@unitn.it*

12th November 2024 - NeurIPS 2024 Vancouver

Setting

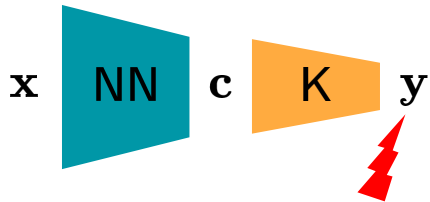


Figure: Neuro-Symbolic model: DeepProbLog (`DPL`) [1] & Logic Tensor Networks (`LTN`) [2]

[1] Manhaeve et al., DeepProbLog: Neural Probabilistic Logic Programming, NeurIPS (2018)
[2] Donadello *et al.*, Logic Tensor Networks, IEEE (2018)

Setting

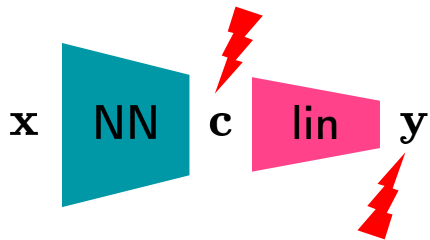


Figure: Concept bottleneck models (CBM) [3]

[3] Pang Wei Koh *et al.*, Concept bottleneck models, ICML (2020)

Setting

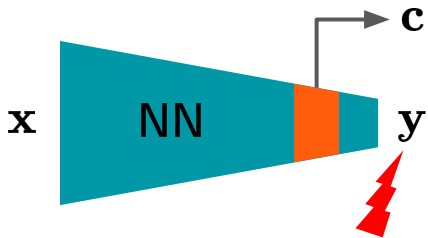
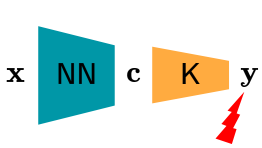


Figure: Neural Network (`NN`) & `CLIP` [4]

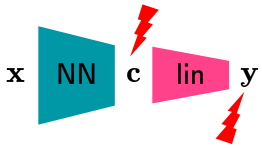
[4] Alec Radford *et al.*, Learning Transferable Visual Models From Natural Language Supervision, ICML (2021)

Setting

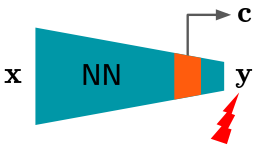
Goal: Study supervised models that classify samples *correctly* but for the *wrong concepts*.



(a) NeSy (DPL [1] & LTN [2])



(b) CBM [3]

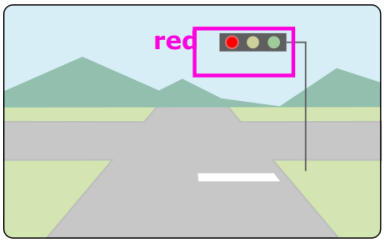


(c) NN & CLIP [4]

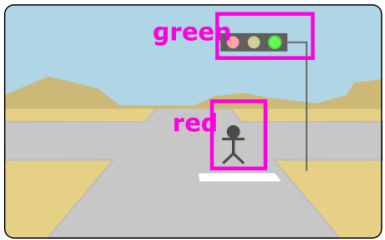
[1] Manhaeve et al., DeepProbLog: Neural Probabilistic Logic Programming, NeurIPS (2018)
[2] Donadello *et al.*, Logic Tensor Networks, IEEE (2018)
[3] Pang Wei Koh *et al.*, Concept bottleneck models, ICML (2020)
[4] Alec Radford *et al.*, Learning Transferable Visual Models From Natural Language Supervision, ICML (2021)

Reasoning Shortcuts

$$K_1 = (\text{pedestrian} \vee \text{red} \Rightarrow \text{stop})$$



$y = \text{stop} \quad \hat{y} = \text{stop}$ ✓

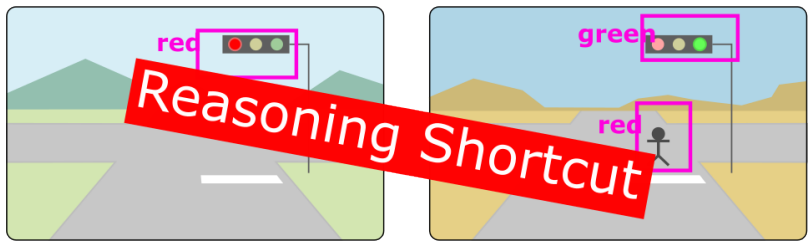


$y = \text{stop} \quad \hat{y} = \text{stop}$ ✓

- Task: predict `stop` vs.go using concepts "`pedestrian`", "`red`", and "`green`".

Reasoning Shortcuts

$$K_1 = (\text{pedestrian} \vee \text{red} \Rightarrow \text{stop})$$



$y = \text{stop}$ $\hat{y} = \text{stop}$ ✓ $y = \text{stop}$ $\hat{y} = \text{stop}$ ✓

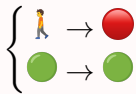
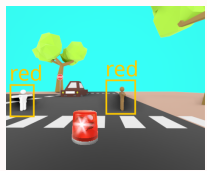
- Task: predict `stop` vs.go using concepts "`pedestrian`", "`red`", and "`green`".

Perfect accuracy by predicting pedestrians as red lights!

rsbench: *L&R tasks*

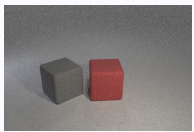
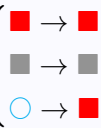
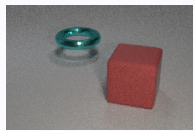
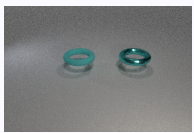
	Task	Data			Properties		
		Gen	OOD	ConL	Cplx **x**	Cplx K	Amb K
Arithmetic	MNMath (new)	✓	✓	✓	✗	✓	✗
	MNAdd-Half	✗	✓✓	✗	✗	✗	-
	MNAdd-EvenOdd	✗	✓✓	✓✓	✗	✗	-
Logic	MNLogic (new)	✓	✓	✓	✗	✓	✗
	Kand-Logic	✓	✓	✓	✗	✓	✓
	CLE4EVR	✓	✓	✓✓	✓	✗	✓
High Stakes	BDD-OIA	✗	✗	✗	✓	✓	✓
	SDD-OIA (new)	✓	✓✓	✓	✓	✓	✓

rsbench: *examples*

Task	Example	Shortcut	OOD Pred.
SDD-OIA	 = STOP		 = GO

Knowledge K *= the traffic laws.*

rsbench: *examples*

Task	Example		Shortcut	OOD Pred.
CLE4EVR		$= 0$		 $= 1$
		$= 1$		

Knowledge K *= same color and shape?*

rsbench: *features*

- **Challenging:** require complex perception and/or reasoning.
- **Versatile:** supports NeSy models, CBMs, post-hoc explainers.
- **Configurable:** can be easily configured with YAML/JSON files.
- **Intuitive:** straightforward to use:

```
from rsbench import MNLOGIC

dataset = MNLOGIC(args)
train(model, dataset)
test(model, dataset)
```

- **Model-level metrics**:
  - F1 and Accuracy
  - Concept level confusion matrix
  - Concept Collapse
- **Task-level metrics**:
  - Formally counts the # of potential RSs in any L&R task!

If you care about your concepts, come to our poster!

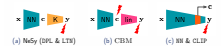
A Neuro-Symbolic Benchmark Suite for Concept Quality and Reasoning Shortcuts

S. Bortolotti[1] **E. Marconato**[1,2] T. Carraro[4,5] P. Morettin[1] E. v. Krieken[3] A. Vergari[3] S. Teso[1] A. Passerini[1]

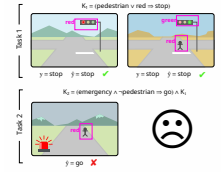
[1]University of Trento [2]University of Pisa [3]University of Edinburgh [4]Fondazione Bruno Kessler [5]University of Padova

REASONING SHORTCUTS

Goal: Study supervised models that classify samples *correctly* but for the *wrong concepts*.



Reasoning Shorcuts [1]: **NeSy predictors** [2], **Concept-based Models** [3] and **VLMs** like CLIP [4] solve Learning & Reasoning tasks by exploiting semantically misleading concepts.



L&R TASKS

Task	Data			Properties		
	Gen	OOD	Conf	Cpts x Cpts	K	Assn K
MNMath (easy)	✓	✓	✓	✗	✓	✗
+/× MNAdd-Half	✗	✓	✗	✗	✗	–
MNAdd-EvenOdd	✗	✓	✓	✗	✗	–
MNLogic (easy)	✓	✓	✓	✗	✓	✗
∧/∨ Kand-Logic	✓	✓	✓	✗	✓	✓
CLEAEVR	✓	✓	✓	✓	✓	✓
SDD-OIA (easy)	✗	✗	✗	✓	✓	✓
SDD-OIA (hard)	✓	✓	✓	✓	✓	✓

FEATURES

① **Challenging**: the # of RSs can be chosen **a priori** and counted using `countrss`, allows to control task difficulty.

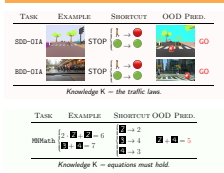
② **Configurable**: data sets & generators can be easily configured with `YAML`/`JSON` files.

③ **Intuitive**: straightforward to use:

```
from rsbench import MNLOGIC

dataset = MNLOGIC(args)
train(model, dataset)
test(model, dataset)
```

EXAMPLES



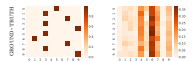
ASSESSING RS

① **Task-level**: `countrss` counts the # of potential RSs in any L&R task!

Example: *with 3 concepts and an exhaustive training set, MNLogic has 6 RSs if K is a conjunction and 24 if K is a XOR. This grows **exponentially** with the # of concepts!*

② **Model-level**: `rsbench` tasks induce RSs in all models!

Table 1. (L) DPL and (R) NN concept confusion matrix on MNAdd-EvenOdd



Quantitatively: Concept F1, accuracy and collapse

REFERENCES

[1] Marconato *et al.*, Analysis and Mitigation of RSs, NeurIPS (2023)
[2] Manhaeve *et al.*, DeepProbLog, NeurIPS (2018)
[3] Pang Wei Koh *et al.*, Concept bottleneck models, ICML (2020)
[4] Alec Radford *et al.*, CLIP, ICML (2021)

PAPER

CODE

WEBSITE

https://unitn-sml.github.io/rsbench/

TANGO

samuelebortolotti

samuele.bortolotti@unitn.it

POSTER SESSION #6! samuelebortolotti samuele.bortolotti@unitn.it