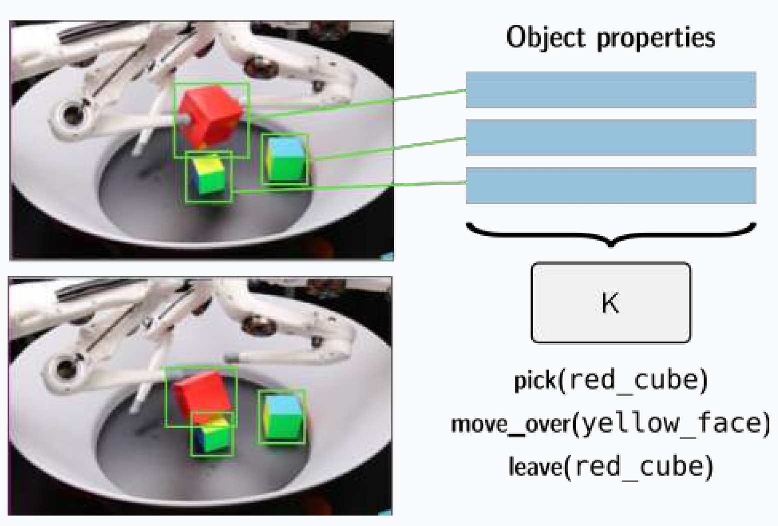
Neuro-Symbolic Continual Learning: Knowledge, Reasoning Shortcuts, and Concept Rehearsal

E. Marconato [1,2] G. Bontempo [2,3] E. Ficarra [3] S. Calderara [3] A. Passerini [1] S. Teso [1]

[1]University of Trento [2]University of Pisa [3]University of Modena-Reggio Emilia

1. Neuro-Symbolic Reasoning



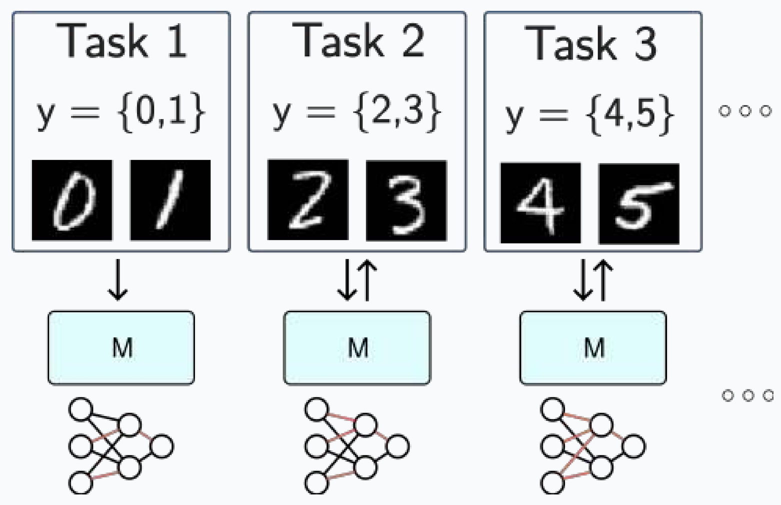
1. The labels $\mathbf{Y}$ depend on *concepts* $\mathbf{C}$, in accordance to knowledge K
2. Concepts $\mathbf{C}$ are learned *input* $\mathbf{X}$, via neural networks.

NeSy predictors are static.

2. Continual Learning

1. Learning is divided in **tasks**
2. *Fine-tuning* the models leads to **Catastrophic Forgetting**;
3. **CL Strategies** make use of *memory and/or regularization* to prevent it.

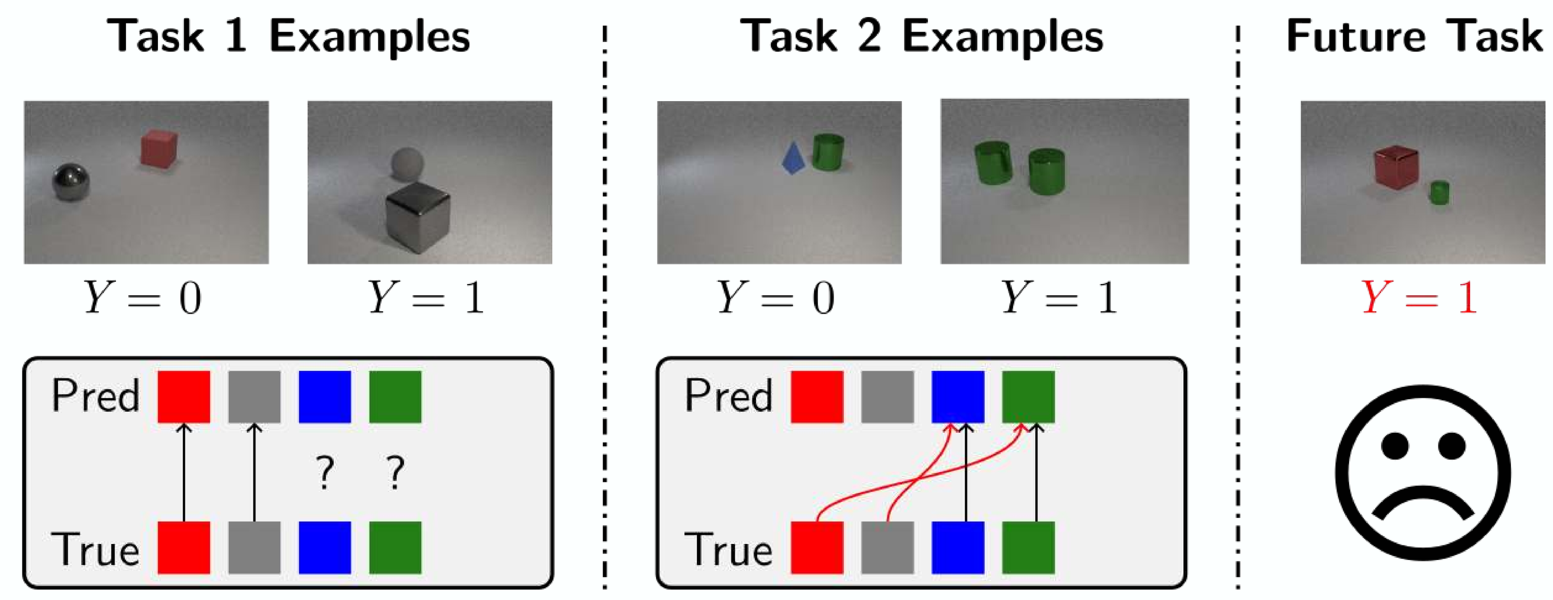
No method makes use of *prior knowledge*.



3. Neuro-Symbolic Continual Learning

We introduce **Neuro-Symbolic Continual Learning** (NeSy-CL), a new machine learning problem, where the machine has to:

1. **Learn** over a sequence of NeSy tasks;
2. **Acquire** high-quality concepts, avoiding **reasoning shortcuts**;
3. **Preserve** the knowledge on the concepts and labels.

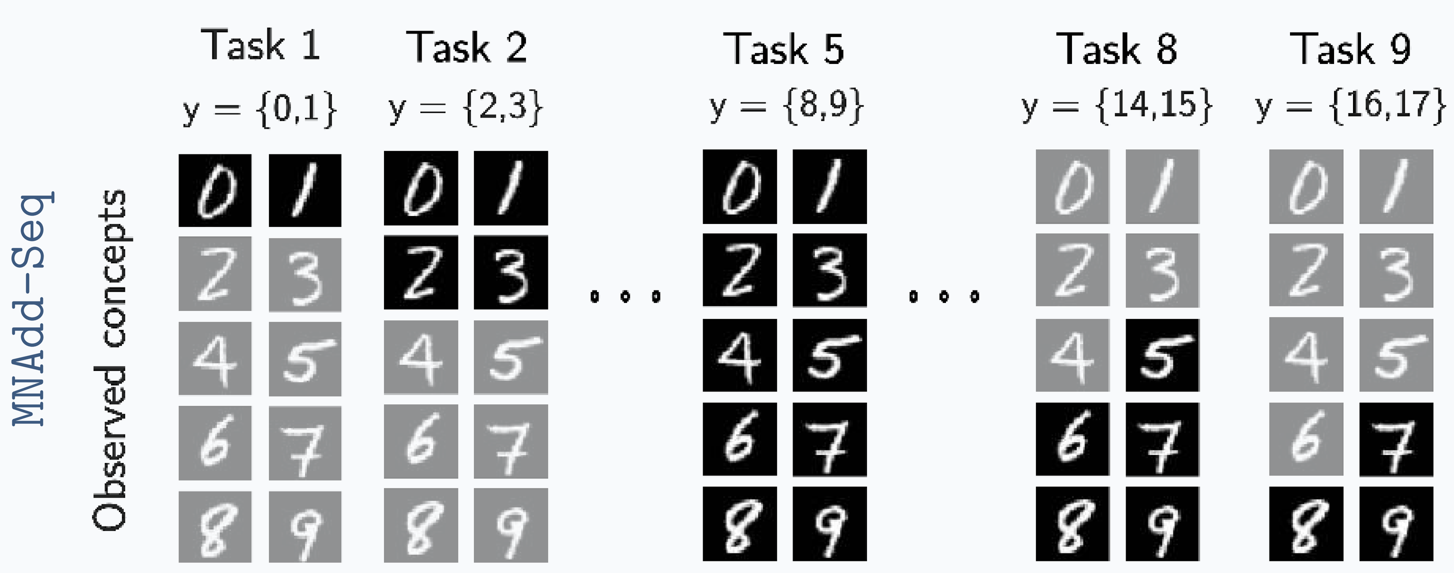


4. Problem Statement

We denote input $\mathbf{x} \in \mathbb{R}^d$, concepts $\mathbf{c} \in \mathbb{N}^k$, labels $\mathbf{y} \in \mathbb{N}^n$, and prior knowledge K. Data are distributed according to:

$$p^{(t)}(\mathbf{X}, \mathbf{C}, \mathbf{Y}; \mathsf{K}^{(t)}) := p^{(t)}(\mathbf{Y} \mid \mathbf{C}; \mathsf{K}^{(t)}) \cdot p(\mathbf{C} \mid \mathbf{X}) \cdot p^{(t)}(\mathbf{X})$$

Example: `MNIST-Addition` consists of sums between digits, i.e. 0 + 1 = 1.
We extend it to a **class/concept-incremental** NeSy-CL **benchmark**.



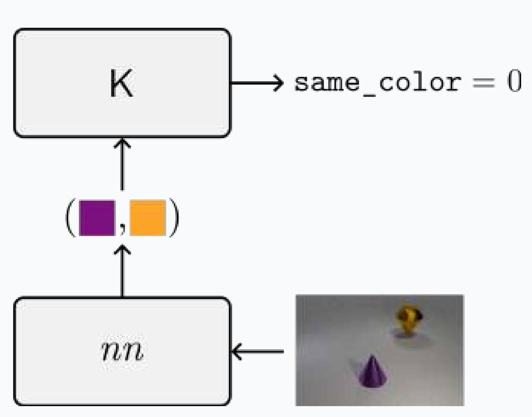
5. Reasoning with DeepProbLog

DeepProbLog = **probabilistic logic programming** + **neural predicates**:

$$p_\theta(y|\mathbf{x}; \mathsf{K}) = \sum_{\mathbf{c}} u_{\mathsf{K}}(y|\mathbf{c}) \cdot p_\theta(\mathbf{c}|\mathbf{x})$$

$$u_{\mathsf{K}}(y|\mathbf{c}) = \frac{1}{Z(\mathbf{c}; \mathsf{K})} \cdot \mathbb{1}\{(\mathbf{c}, \mathbf{y}) \models \mathsf{K}\}$$

For each task, we learn the parameters θ via **maximum likelihood**.



6. NeSy predictors fall prey of Reasoning Shortcuts

A **Reasoning Shortcut** is an optimal solution with **incorrect concepts**.

Theorem (3.2): A model with parameters θ attains maximal likelihood, i.e. $\theta \in \Theta^*(\mathsf{K}, \mathcal{D})$, *if and only if*, for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$, it holds that $p_\theta(\mathbf{C} \mid \mathbf{x}) \models \mathsf{K}[\mathbf{Y}/\mathbf{y}]$.

Example: Consider `MNIST-Addition` with only 0 + 1 = 1 and 0 + 2 = 2. Then:

$$0 \to 0, \quad 1 \to 1, \quad 2 \to 2$$

$$0 \to 1, \quad 1 \to 0, \quad 2 \to 1$$

are two **optimal solutions**.
Mitigation with **concept supervision**. **Q3:** We show that only few is sufficient!

7. COOL: a new strategy for Continual Learning

We prove (**Theorem 4.1**) that in NeSy-CL we must **remember the learnt concept distribution**. COOL optimizes for this:

$$\mathcal{L}_{\text{COOL}} := \frac{1}{N_{mb}} \sum_{(\mathbf{x}, \tilde{\mathbf{q}}_c, \mathbf{y}) \in \mathcal{M}} \Big[\alpha \cdot \mathsf{KL}\big(p_\theta(\mathbf{C} \mid \mathbf{x}) \,\|\, \tilde{\mathbf{q}}_c\big) - \beta \cdot \log p_\theta(\mathbf{Y} = \mathbf{y} \mid \mathbf{x}; \mathsf{K}^{(t)})\Big]$$

8. Experimental Verification

New NeSy-CL benchmarks:

1) MNAdd-Seq
2) MNAdd-Shortcut*
3) CLE4EVR* from CLEVR

CLE4EVR benchmark

Task	Colors	Shapes
1	red, gray	sphere, cube
2	green, blue	cylinder, tetrahedron
3	brown, purple	cone, triangular prism
4	yellow, cyan	pyramid, toroid
5	orange, pink	diamond, star prism

$\mathbf{Y} \in \{$**same color**, **same shape**, **both**, **neither**$\}$

Q1: Knowledge helps

We compare different strategies paired to **Concept Bottleneck Models** and **DeepProbLog**

	Strategy	Class-IL $\mathbf{Y}$ (↑)	Class-IL $\mathbf{C}$ (↑)	FWT (↑)
CBM @ 10%	Naïve	11.71 ± 0.8	36.2 ± 2.6	7.5 ± 0.3
	Restart	10.78 ± 0.1	29.7 ± 0.1	7.3 ± 0.2
	LwF	18.08 ± 1.8	63.2 ± 4.4	-4.7 ± 1.1
	EWC	11.57 ± 0.6	37.4 ± 0.6	7.6 ± 0.4
	ER	13.29 ± 0.4	43.5 ± 2.0	13.4 ± 1.6
	DER	18.63 ± 2.5	53.1 ± 1.7	15.7 ± 0.9
	DER++	18.17 ± 1.6	54.1 ± 3.0	16.6 ± 1.8
	COOL	$\mathbf{38.0 \pm 1.9}$	$\mathbf{78.1 \pm 2.5}$	$\mathbf{29.0 \pm 4.8}$
DeepProbLog	Naïve	6.9 ± 0.2	6.7 ± 0.4	6.2 ± 0.2
	Restart	9.6 ± 0.3	0.2 ± 0.1	6.9 ± 0.8
	LwF	6.8 ± 0.5	10.8 ± 4.6	18.3 ± 0.2
	EWC	6.8 ± 0.4	7.8 ± 0.6	6.1 ± 0.3
	ER	44.3 ± 9.7	62.0 ± 8.6	8.2 ± 4.1
	DER	68.3 ± 9.4	81.3 ± 6.9	44.5 ± 23.7
	DER++	62.2 ± 5.4	77.1 ± 4.2	27.1 ± 5.2
	COOL	$\mathbf{71.9 \pm 2.9}$	$\mathbf{84.5 \pm 1.9}$	$\mathbf{83.2 \pm 0.9}$

Q2: Addressing Reasoning Shortcuts

- We show that **supervising concepts** + **replay strategies** is not enough.
- Only COOL **avoids** reasoning shortcuts

CLE4EVR