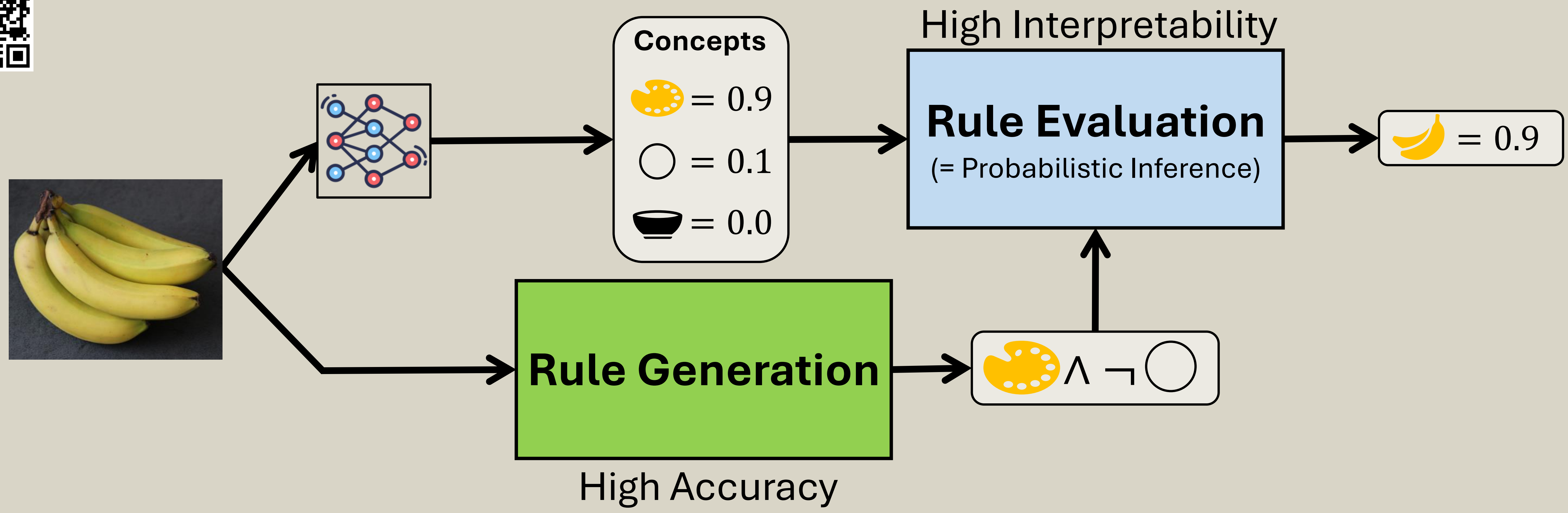
Neurosymbolic Concept-Based Reasoners Go Beyond the **Accuracy-Interpretability** Trade-Off of Concept Bottleneck Models

David Debot and Giuseppe Marra

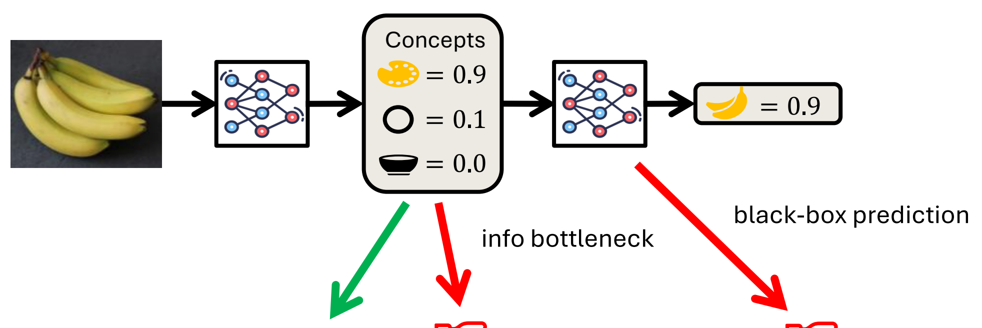
Integrating Local and Global Interpretability for Deep Concept-Based Reasoning Models



Concept Bottleneck Models [1]

CBNMs = intrinsically explainable models that first predict **concepts** and then predict a downstream task with them

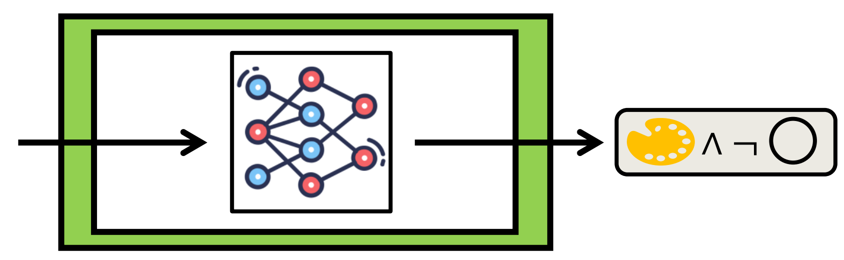
= high-level, human-understandable features related to the task



High-level, human-understandable explanations | **Low accuracy** | **Low interpretability**

Deep Concept Reasoner [2]

Rule generation = neural prediction of a rule



Highly accurate no matter the concepts

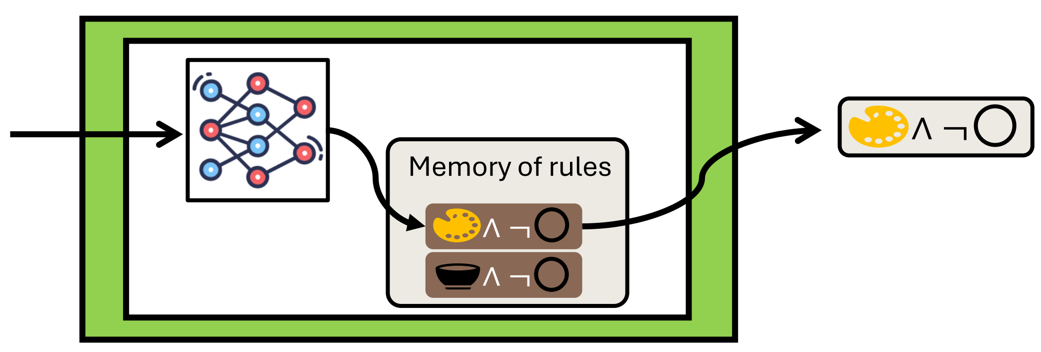
Requires extensive tuning to obtain meaningful rules

Locally interpretable: interpret *how* concepts are used *for a specific input*

but not globally

Concept-based Memory Reasoner [3]

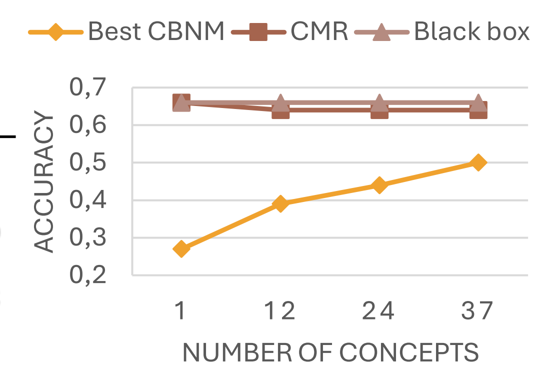
Rule generation = neural selection in a learned memory of rules



CMR is highly accurate no matter the employed concepts

Theorem: CMR is a universal binary approximator if $n_R \geq 3$

	MNIST+	MNIST+*	CELEBA	CEBAB
Best CBNM	97.41 ± 0.55	77.63 ± 0.44	50.24 ± 0.34	83.80 ± 0.01
Black box	83.26 ± 8.71	83.26 ± 8.71	65.72 ± 0.70	88.67 ± 0.19
CMR	97.52 ± 0.30	95.47 ± 0.47	63.56 ± 0.48	85.14 ± 0.43



CMR's global interpretability allows verification of properties

All decision rules in memory are transparent ⇒ model properties can be verified **before deployment**

"Does the property hold no matter which rule is selected?"

memory + selection   property to check

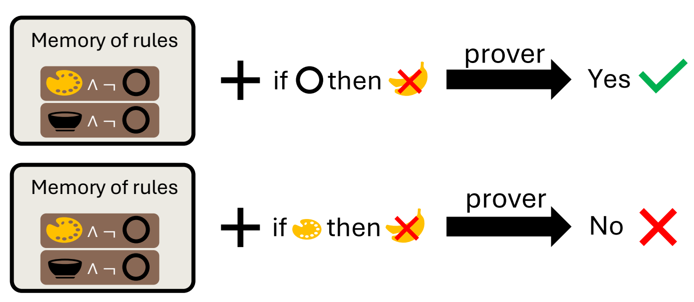
Given theory T and formula φ, compute $T \vDash \varphi$

or   or

Model checking: $(\forall M, M \vDash T \Rightarrow M \vDash \varphi) \Leftrightarrow T \vDash \varphi$

Theorem proving: $T \vdash \varphi \Rightarrow T \vDash \varphi$

SAT solving: $SAT(F_T \wedge \neg\varphi) \Leftrightarrow T \nvDash \varphi$

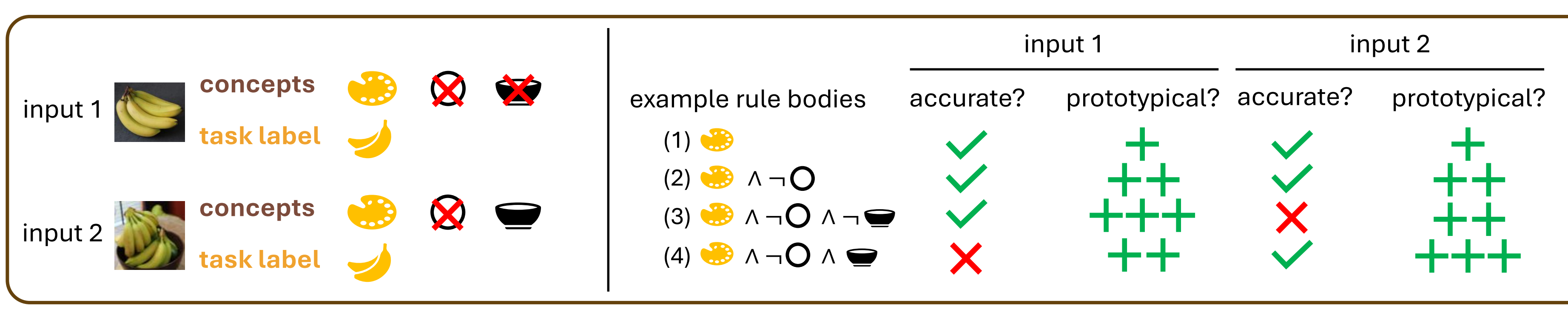


CMR's rule learning allows human interaction during training

The way CMR learns rules allows for human interaction in multiple ways = **"rule interventions"**

Rule intervention	Can be used for	Example
Add rules manually to the memory	Incorporating expert knowledge	Add ← ∧
Forbid a concept from being in a rule	Debiasing	Avoid using in rule 1
Force a concept to be in a rule	Enforcing safety	Use in all rules

CMR learns meaningful rules that are both accurate and prototypical of concept activations



Possibility 1: space for 1 rule in memory

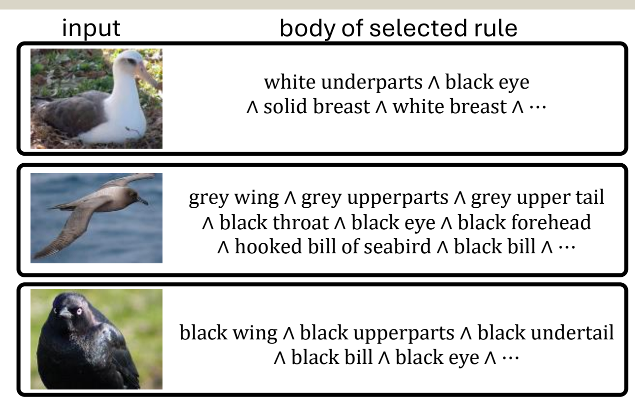
⇒ learn rule (2)

Possibility 2: space for 2+ rules in memory

⇒ learn rules (3) and (4)

⇒ select (3) for input 1

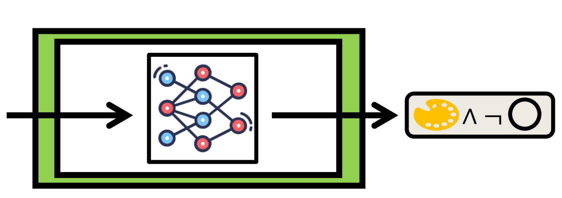
⇒ select (4) for input 2



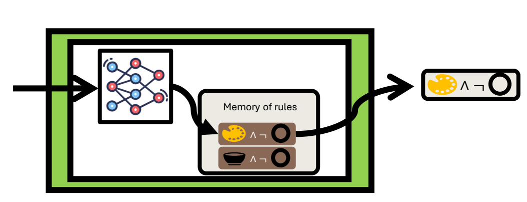
Highly accurate no matter the concepts | **Globally interpretable**: all decision rules transparent | **Prototypes** as meaningful rules | **Difficult optimization problem**

Unified Concept Reasoner [4]

Use



as a **variational approximation** of



	Rule correspondence
Positive class	98.9 ± 0.3
Negative class	74.2 ± 2.0

	MNIST+ accuracy
UCR (DCR head)	97.8 ± 0.2
UCR (CMR head)	95.7 ± 0.8

Training objective: 2 components

1. Apply CMR's objective on DCR's rule generation (i.e. **accurate + prototypical**)
2. **Maximize rule correspondence** between DCR's and CMR's rule generation (KL divergence)

Highly accurate no matter the concepts | **Globally interpretable** | **Prototypes** as meaningful rules | **Easier optimization**

References

[1] Concept Bottleneck Models — Koh et al.

[2] Interpretable Neural-Symbolic Concept Reasoning — Barbiero et al.

[3] Interpretable Concept-Based Memory Reasoning — Debot et al.

[4] Integrating Local and Global Interpretability for Deep Concept-Based Reasoning Models — Debot et al.

DTAI Declarative Languages & Artificial Intelligence | KU LEUVEN | LEUVEN.AI