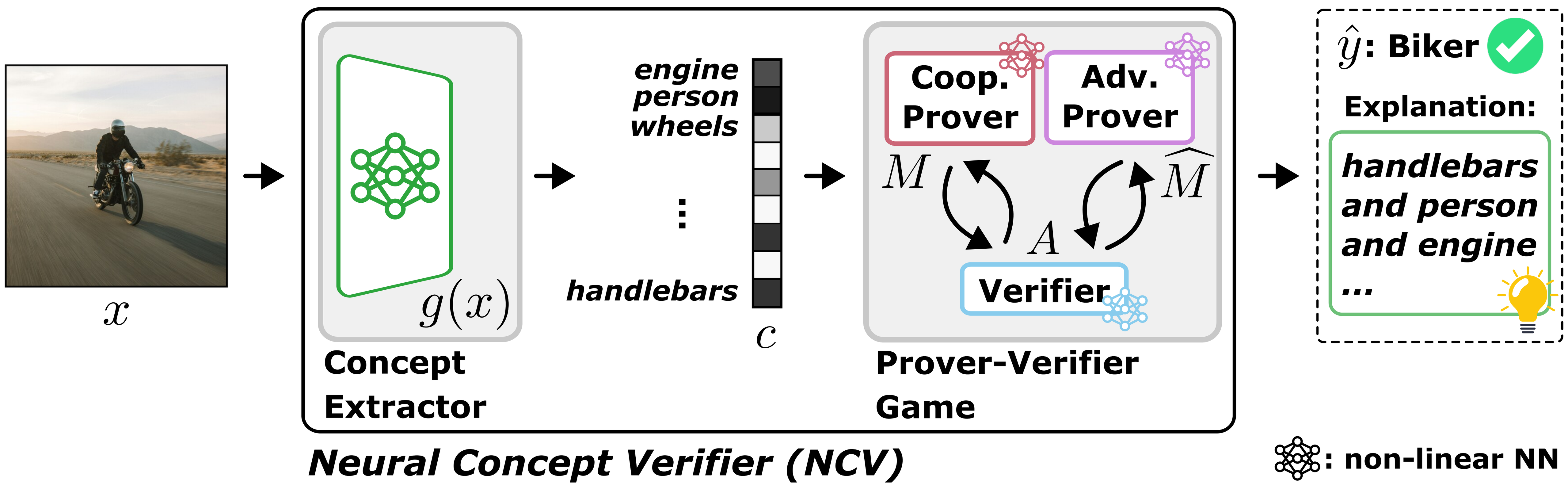
Neural Concept Verifier: Scaling Prover-Verifier Games via Concept Encodings

Berkant Turan[1,2], Suhrab Asadulla[1,2], David Steinmann[3,4], Wolfgang Stammer[3,4], Sebastian Pokutta[1,2]

[1] Zuse Institute Berlin [2] TU Berlin [3] TU Darmstadt [4] Hessian.AI

Neural Concept Verifier

A scalable Prover–Verifier framework using sparse concept encodings for expressive, verifiable, and robust image classification.



Overview

The **Neural Concept Verifier (NCV)**, a framework combining Prover-Verifier Games (PVGs) with concept bottlenecks for interpretable, non-linear classification

- Utilizes **minimally supervised concept extraction pipelines**
- Incorporates a **verifiable prediction protocol**
- Scaling PVGs to **high-dimensional** inputs while preserving **interpretability**
- Empirical results show that NCV improves **verifiability and performance** compared to pixel-level baselines

The NCV Framework

Decomposes an image classification task into several components:

1. A **concept extractor** $g : \mathscr{X} \to \mathscr{C}$, maps each input x to a high-level concept encoding $\mathbf{c} \in \mathbb{R}^C$, where $C \in \mathbb{N}$ is the number of discovered concepts.
2. A pair of **prover agents** $M, \widehat{M} : \mathscr{C} \to \{0,1\}^C$, generate sparse binary masks to select m concepts each.
   - M (Merlin, **cooperative** Prover) supports classification,
   - $\widehat{M}$ (Morgana, **adversarial** Prover) attempts to mislead it.
3. A **verifier** (Arthur) $A : \mathbb{R}^C \to \mathscr{Y}$, predicts the final label **only using the masked concept encodings**.

We instantiate the concept extractor with:

- **CLIP-Sim:** Utilizes CLIP, producing dense, semantically aligned concept encodings.
- **Neural Concept Binder (NCB):** Recent unsupervised object-centric concept encoder enabling symbolic concept encodings.

Outlook

- **Contribution:** NCV framework as a promising step toward performative, verifiable models.
- **Future work:** Extend NCV to alternative PVG architectures and new domains (e.g., NLP, structured data).

Experimental Evaluations

*NCV delivers **high accuracy** and **robustness** via **verifiable, concept-based interpretability**, outperforming baselines on both synthetic and real-world benchmarks.*

Model	Feature Space	Completeness (Accuracy)	Soundness (Robustness)
CLEVR-Hans3			
ResNet-18	pixel space	97.87 ± 0.24	n/a
Pixel-MAC	pixel space	96.59 ± 0.72	99.99 ± 0.01
CBM	NCB	95.44 ± 0.08	n/a
Ours	NCB	$\mathbf{98.92 \pm 0.32}$	$\mathbf{100.00 \pm 0.00}$
CLEVR-Hans7			
ResNet-18	pixel space	98.71 ± 0.24	n/a
Pixel-MAC	pixel space	97.61 ± 0.38	99.88 ± 0.28
CBM	NCB	89.12 ± 0.12	n/a
Ours	NCB	$\mathbf{97.89 \pm 0.31}$	$\mathbf{100.00 \pm 0.00}$
CIFAR-100			
ResNet-18	pixel space	79.73 ± 0.36	n/a
Pixel-MAC	pixel space	15.27 ± 4.78	96.31 ± 4.12
CBM	SpLiCE	75.42 ± 0.04	n/a
Ours	CLIP-Sim	$\mathbf{83.32 \pm 0.28}$	$\mathbf{99.99 \pm 0.01}$
ImageNet			
ResNet-18	pixel space	66.16 ± 0.41	n/a
Pixel-MAC	pixel space	35.06 ± 3.20	99.65 ± 0.26
CBM	SpLiCE	$\mathbf{68.59 \pm 0.01}$	n/a
Ours	CLIP-Sim	67.04 ± 0.16	$\mathbf{99.94 \pm 0.02}$

*NCV improves **generalization under distribution shift**, achieving the **smallest shortcut gap** and **highest test accuracy** on **CLEVR-Hans7**.*

Ratio Clean (Samples)	Model	Val Acc (w/ shortcut)	Test Acc (w/o shortcut)	Val–Test Gap (↓)
0% (0)	CBM (lin.)	90.37 ± 0.10	85.27 ± 0.15	**5.10**
	CBM (nonlin.)	98.09 ± 0.24	90.69 ± 1.17	7.40
	Ours	$\mathbf{98.38 \pm 0.18}$	$\mathbf{92.23 \pm 0.67}$	6.15
5% (525)	CBM (lin.)	90.37 ± 0.15	86.37 ± 0.18	4.00
	CBM (nonlin.)	98.32 ± 0.22	95.19 ± 0.80	3.13
	Ours	$\mathbf{98.47 \pm 0.24}$	$\mathbf{96.24 \pm 0.71}$	**2.23**
20% (2100)	CBM (lin.)	89.93 ± 0.29	87.21 ± 0.31	2.72
	CBM (nonlin.)	98.21 ± 0.29	97.00 ± 0.49	1.21
	Ours	$\mathbf{98.63 \pm 0.13}$	$\mathbf{97.74 \pm 0.28}$	**0.89**