Towards Correct-by-Construction Machine Learning via Differentiable Logics

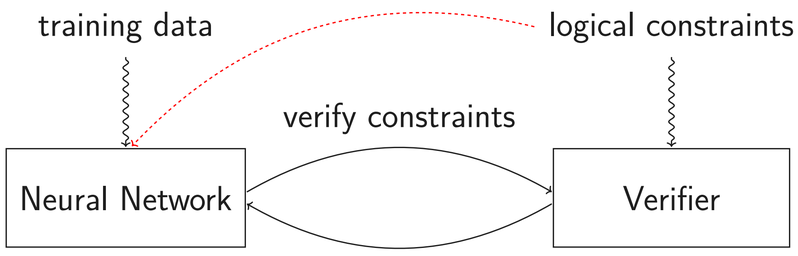
Thomas Flinkow, Barak A. Pearlmutter, Rosemary Monahan
Department of Computer Science, Maynooth University
thomas.flinkow@mu.ie

Engaging Content
Engaging People

Taighde Éireann
Research Ireland

1. Motivation: Correct-by-construction ML models

Task: train a neural network $\mathcal{N}$ to satisfy constraint ϕ.
Train: iteratively update weights using gradient descent and loss.
Verify: with α, β-CROWN, Marabou, NNV, ERAN, ...

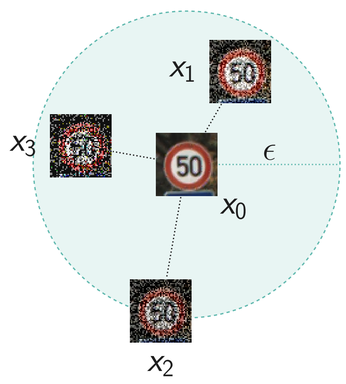


2. Background: Property-driven ML

Standard ML: Given data $\boldsymbol{x}$, target $\boldsymbol{y}$, and prediction loss $\mathcal{L}_{\text{P}}$, obtain optimal network weights $\boldsymbol{\theta}^+$ by

$$\boldsymbol{\theta}^+ = \arg\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}} \mathcal{L}_{\text{P}}(\boldsymbol{x}, \boldsymbol{y}).$$

Insight from DL2 (Deep Learning with Differentiable Logic): Approximate constraints of the form ϕ: $\forall \boldsymbol{x}.\, P(\boldsymbol{x}) \rightarrow Q(\boldsymbol{x})$ by finding a counterexample $\boldsymbol{x}^*$ that does *not satisfy* $Q(\boldsymbol{x})$ in the input space induced by $\mathcal{P}(\boldsymbol{x})$.



- Approximate counterexample *outside* of training set using PGD:

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}'\in\mathcal{S}} \mathcal{L}_{\phi}(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y})$$

- Use this counterexample in training:

$$\boldsymbol{\theta}^+ = \arg\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}} \left[\lambda \underbrace{\mathcal{L}_{\text{P}}(\boldsymbol{x}, \boldsymbol{y})}_{\text{prediction loss}} + (1-\lambda) \underbrace{\mathcal{L}_{\phi}(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{y})}_{\text{logical constraint loss}}\right].$$

3. Background: Differentiable Logics

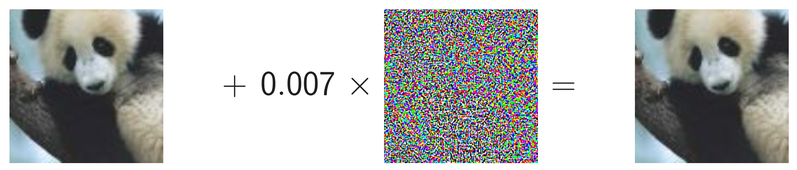
- **DL2:** $[\![\cdot]\!]_{\text{DL2}} : \Phi \rightarrow [0, \infty)$, where $[\![\top]\!]_{\text{DL2}} = 0$, and

$$[\![x \leq y]\!]_{\text{DL2}} := \max\{x - y, 0\}$$
$$[\![\phi \wedge \psi]\!]_{\text{DL2}} := [\![\phi]\!]_{\text{DL2}} + [\![\psi]\!]_{\text{DL2}}$$
$$[\![\phi \vee \psi]\!]_{\text{DL2}} := [\![\phi]\!]_{\text{DL2}} \cdot [\![\psi]\!]_{\text{DL2}}.$$

- **Fuzzy Logics:** $[\![\cdot]\!]_{L} : \Phi \rightarrow [0, 1]$, where $[\![\top]\!]_{L} = 1$ and $[\![\bot]\!]_{L} = 0$

Logic	Conjunction	Disjunction	Implication
Gödel	$\min\{x, y\}$	$\max\{x, y\}$	$\begin{cases} 1, & \text{if } x < y, \\ y, & \text{else.} \end{cases}$
Kleene-Dienes			$S(N(x), y)$
Łukasiewicz	$\max\{0, x + y - 1\}$	$\min\{1, x + y\}$	
Reichenbach			
Goguen	xy	$x + y - xy$	$\begin{cases} 1, & \text{if } x < y, \\ y^x, & \text{else.} \end{cases}$

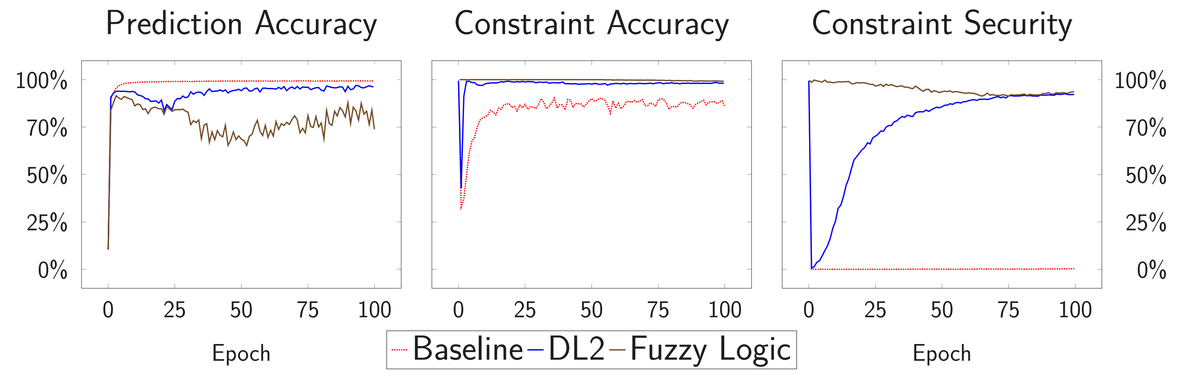
4. Training Experiment



'panda' 57.7 % confidence | adversarial noise | 'gibbon' 99.3 % confidence

Constraint: A neural network is *locally robust* in input $\boldsymbol{x}$, if

$$\underbrace{\forall \boldsymbol{x}'.\, \|\boldsymbol{x}' - \boldsymbol{x}\|_\infty \leq \varepsilon}_{\text{all elements in the input space close to } \boldsymbol{x}} \quad \text{implies} \quad \underbrace{\|\mathcal{N}(\boldsymbol{x}') - \mathcal{N}(\boldsymbol{x})\|_\infty \leq \delta}_{\text{the classification is roughly the same}}$$



Main finding: Property-driven training with *any* differentiable logic leads to significantly improved constraint satisfaction.

5. Verification Experiment

Verified constraint satisfaction on 500 randomly chosen images of the MNIST test set on networks trained with $\epsilon = 0.4$.

Logic	Prediction Accuracy	Constraint Security	Verified Satisfaction $\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
Baseline	96.50 %	79.68 %	0.68 % ($^3/_{444}$)	0 % ($^0/_{500}$)	0 % ($^0/_{500}$)
DL2	93.07 %	100 %	92.98 % ($^{384}/_{413}$)	55.29 % ($^{183}/_{331}$)	20.51 % ($^{73}/_{356}$)
Fuzzy Logic	94.87 %	100 %	92.70 % ($^{368}/_{397}$)	52.16 % ($^{157}/_{301}$)	9.22 % ($^{27}/_{293}$)

Main finding: Property-driven training leads to some formal guarantees, but significantly lower than what is reported in training.

6. Future Work: Formal Guarantees

Instead of *minimising* an *upper bound* on the loss, ...

(a) Adversarial Training (b) Certified Training

... *maximise* a *lower bound* on robustness!

Thomas Flinkow, Marco Casadio, Colin Kessler, Rosemary Monahan, Ekaterina Komendantskaya: **'A Generalised Framework for Property-Driven Machine Learning'**, Submitted to *8th International Symposium on AI Verification (SAIV)* (May 2025). preprint: arXiv:2505.00466

Thomas Flinkow, Barak A. Pearlmutter, Rosemary Monahan: **'Comparing differentiable logics for learning with logical constraints'**, In *Science of Computer Programming*, **244**, 103280 (Sep 2025). DOI: 10.1016/j.scico.2025.103280

This research was conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 20/FFP-P/8853 and Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Maynooth University [13/RC/2106_P2].

Maynooth University
National University of Ireland Maynooth