Machine Learning with Differentiable Logics

Thomas Flinkow[1], Barak A. Pearlmutter[1,2], Rosemary Monahan[1,2]

[1]Department of Computer Science, Maynooth University
[2]Hamilton Institute, Maynooth University

1. Introduction

A variety of formal verifiers (such as Reluplex, NNV, α, β-CROWN) have been developed to guarantee correct behaviour and safety of neural networks. However, verifiers typically assume a trained network with fixed weights.

A promising approach to create correct-by-construction machine-learnt models is to encode background knowledge (safety and correctness properties) as logical constraints that must be satisfied during network training.

$$\mathcal{L} = \mathcal{L}_{\mathsf{CE}} + \lambda \mathcal{L}_{\mathsf{L}}(\phi)$$

Here, the logical constraint ϕ is used in combination with standard cross-entropy loss and can be minimised with standard optimisers.

2. Logical Constraints

An example constraint that encodes relations between classes "cars are more similar to a truck than to a cat" and "birds are more similar to a plane than to a dog":

$$\begin{aligned}\phi := \quad & (\mathcal{N}(x)_{\texttt{car}} \geq 0.1) \rightarrow (\mathcal{N}(x)_{\texttt{truck}} \geq \mathcal{N}(x)_{\texttt{cat}}) \\ \wedge\ & (\mathcal{N}(x)_{\texttt{bird}} \geq 0.1) \rightarrow (\mathcal{N}(x)_{\texttt{plane}} \geq \mathcal{N}(x)_{\texttt{dog}})\end{aligned}$$

This logical constraint would then need to be translated into real-valued loss for use in the loss function.

3. Experimental Setup

Our weakly-supervised comparison experiment (`https://github.com/tflinkow/dl-comparison`) is implemented in PyTorch and evaluates training on Fashion-MNIST, CIFAR-10, and GTSRB.

We used the class-similarity constraint to encode semantic relationships between classes and for GTSRB, we used a group constraint expressing that certain class groups either have very high or very low probability.

We investigate multiple possible translations based on DL2 and fuzzy logics.

4. DL2 [1]

"Deep Learning with Differentiable Logics" is a system for training and querying neural networks with logics.

- $\mathcal{L}$ maps logical constraints into $[0, \infty)$,
- $\mathcal{L}(\phi) = 0$ iff ϕ is satisfied,
- $\mathcal{L}(\phi)$ is differentiable almost everywhere.

Recursive definition of loss translation:

$$\begin{aligned}\mathcal{L}_{\mathsf{DL2}}(x \leq y) &:= \max(x - y, 0) \\ \mathcal{L}_{\mathsf{DL2}}(x \neq y) &:= \xi[x = y] \\ \mathcal{L}_{\mathsf{DL2}}(x \wedge y) &:= \mathcal{L}_{\mathsf{DL2}}(x) + \mathcal{L}_{\mathsf{DL2}}(y) \\ \mathcal{L}_{\mathsf{DL2}}(x \vee y) &:= \mathcal{L}_{\mathsf{DL2}}(x) \cdot \mathcal{L}_{\mathsf{DL2}}(y).\end{aligned}$$

5. Fuzzy Logics

Fuzzy logics map satisfaction of logical formulae into $[0, 1]$, where 0 denotes absolute falsehood and 1 absolute truth.

They use triangular norms for conjunction and their co-norms for disjunction, and are usually differentiable almost everywhere, making them a natural choice for use in loss functions.

Table: A few t-norms and t-conorms.

T-norm (Conjunction)	T-conorm (Disjunction)
$T_{\mathsf{G}}(x, y) = \min(x, y)$	$S_{\mathsf{G}}(x, y) = \max(x, y)$
$T_{\mathsf{\L K}}(x, y) = \max(0, x + y - 1)$	$S_{\mathsf{\L K}}(x, y) = \min(1, x + y)$
$T_{\mathsf{P}}(x, y) = xy$	$S_{\mathsf{PS}}(x, y) = x + y - xy$

Table: A few fuzzy implications.

Name	Implication
Gödel	$I_{\mathsf{G}}(x, y) = \begin{cases} 1, & \text{if } x < y \\ y \end{cases}$
Kleene-Dienes	$I_{\mathsf{KD}}(x, y) = \max(1 - x, y)$
Łukasiewicz	$I_{\mathsf{\L K}}(x, y) = \min(1 - x + y, 1)$
Goguen	$I_{\mathsf{GG}}(x, y) = \begin{cases} 1, & \text{if } x < y \\ y/x \end{cases}$
Reichenbach	$I_{\mathsf{RC}}(x, y) = 1 - x + xy$

Prior work by [2] suggests a *sigmoidal* implication that would perform well due to favorable derivatives in the corners:

$$(I(x, y))_s := \frac{(1 + \exp(s/2))\sigma(sI(x, y) - s/2) - 1}{\exp(s/2) - 1}$$

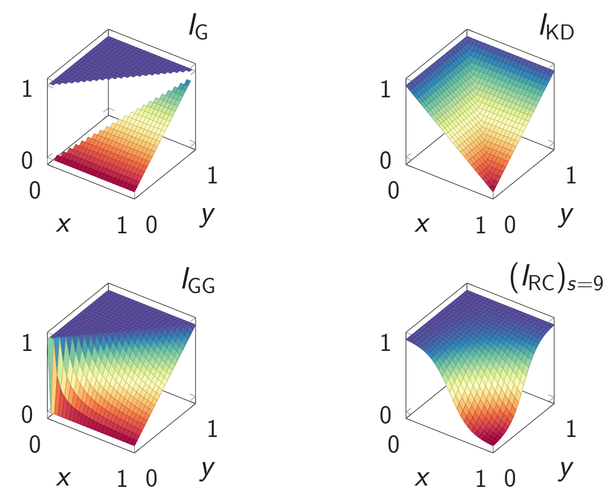

Figure: Different fuzzy logic implications $I(x, y)$ mapping the logical statement $x \rightarrow y$ into $[0, 1]$.

6. Mapping of Atomic Terms

In DL2, atomic terms are either comparisons or inequalities for which it provides loss translations. Fuzzy logics typically do not define comparison operators. In [3], the authors introduce $\mathcal{L}_{\mathsf{FL}}(x \leq y) := 1 - \max\left(\frac{x-y}{x+y}, 0\right)$, and we change this mapping from $\mathcal{L}_{\mathsf{FL}} : [0, 1]^2 \rightarrow [0, 1]$ to $\mathcal{L}_{\mathsf{FL}} : \mathbb{R}^2 \rightarrow [0, 1]$, allowing us to forgo the need for an external oracle.

$$\mathcal{L}_{\mathsf{FL}}(x \leq y) := 1 - \frac{\max(x - y, 0)}{|x| + |y| + \epsilon}$$

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 20/FFP-P/8853.

7. Results

Main finding: training with logical constraints can significantly improve constraint accuracy, at the expense of prediction accuracy.

Table: P/C is prediction / constraint accuracy in %.

	Fashion-MNIST			CIFAR-10		
	P	C	λ	P	C	λ
Baseline	77.55	84.31	–	79.06	48.65	–
DL2	77.88	89.15	0.6	78.55	52.21	0.4
I_{G}	63.46	91.30	3.0	77.65	81.56	1.2
I_{KD}	75.39	80.59	0.8	78.82	72.94	0.6
$I_{\mathsf{\L K}}$	64.64	97.28	4.0	76.06	87.75	6.0
I_{GG}	67.25	95.54	3.0	74.83	88.67	10.0
I_{RC}	76.79	92.56	0.8	79.14	80.51	0.8
$(I_{\mathsf{RC}})_{s=9}$	77.06	93.63	0.8	78.30	80.87	0.8
$(I_{\mathsf{RC}})_{\phi=x^2}$	**76.88**	**95.94**	1.0	**78.31**	**90.74**	1.6
I_{YG}	74.14	80.15	1.0	77.81	73.19	0.8

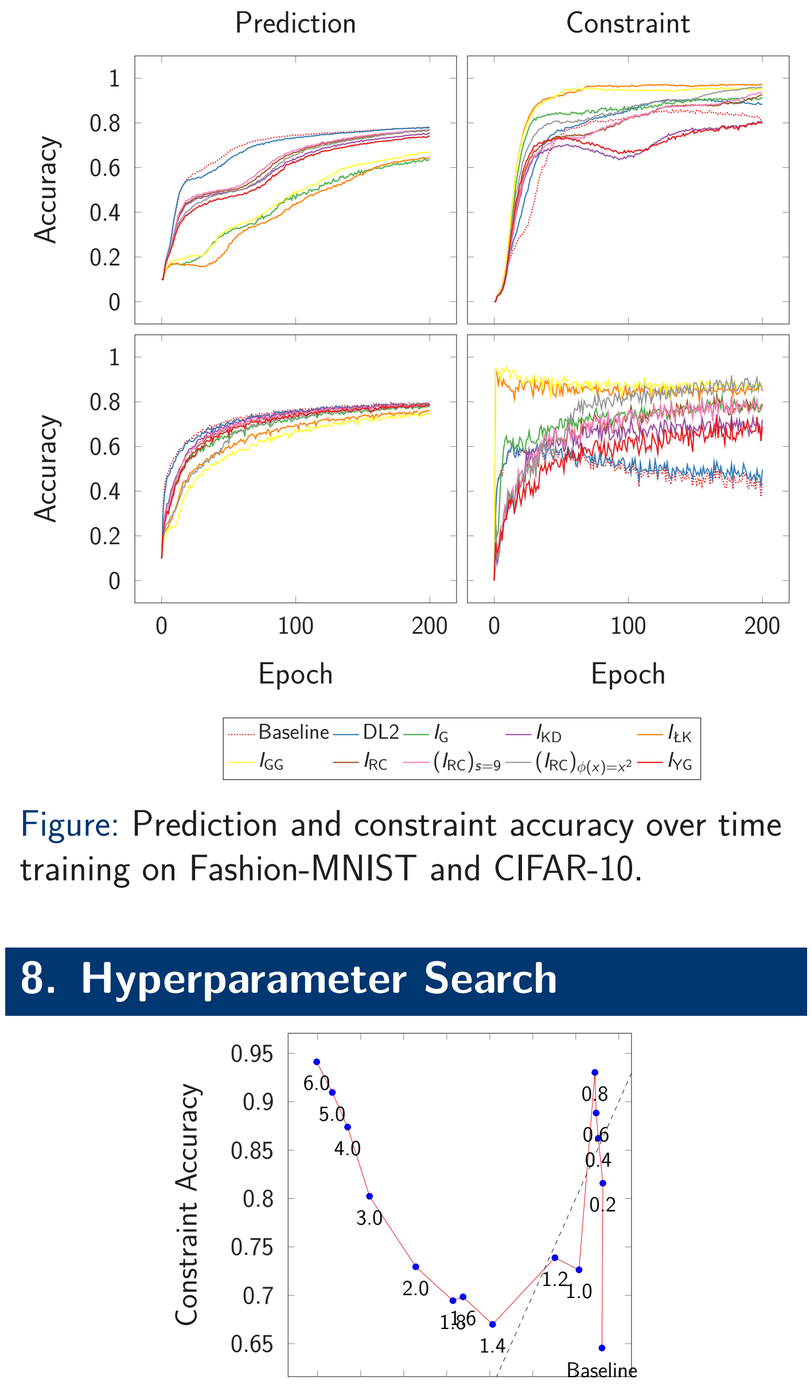

Figure: Prediction and constraint accuracy over time training on Fashion-MNIST and CIFAR-10.

8. Hyperparameter Search

Figure: Effectiveness depends heavily on λ in a non-obvious, non-monotonous way.

References

[1] Marc Fischer, Mislav Balunovic, Dana Drachsler-Cohen, Timon Gehr, Ce Zhang, and Martin Vechev.
DL2: Training and Querying Neural Networks with Logic.
In *Proceedings of the 36th International Conference on Machine Learning*, pages 1931–1941. PMLR, May 2019.

[2] Emile van Krieken, Erman Acar, and Frank van Harmelen.
Analyzing Differentiable Fuzzy Logic Operators.
Artificial Intelligence, 302:103602, January 2022.

[3] Natalia Ślusarz, Ekaterina Komendantskaya, Matthew Daggitt, Robert Stewart, and Kathrin Stark.
Logic of Differentiable Logics: Towards a Uniform Semantics of DL.
In *EPiC Series in Computing*, volume 94, pages 473–493. EasyChair, June 2023.