Towards Explainability in Knowledge Enhanced Neural Networks

Data Science Master Thesis

Riccardo Mazzieri

Supervisor: Luciano Serafini
Co-Supervisor: Alessandro Daniele

September 21, 2021

FONDAZIONE BRUNO KESSLER

UNIVERSITÀ DEGLI STUDI DI PADOVA

Outline

1. Introduction and motivations
2. Knowledge Enhanced Neural Networks (KENN)
3. **Contributions:** Experiments on collective classification
4. **Contributions:** Extracting explanations from KENN
5. Conclusions

Deep NNs have several flaws. For example:

- They are **data hungry**:
  - With few data, learning is not possible, even for simple logical reasoning tasks;
  - This motivates **Neural Symbolic Integration (NeSy)**.

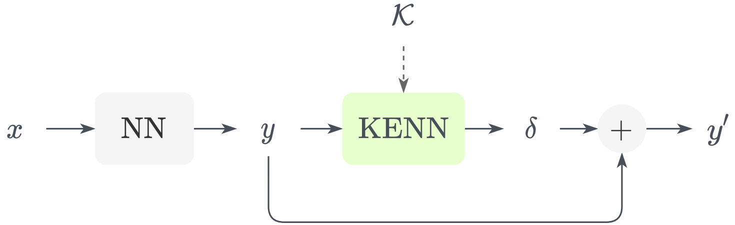
Introduction

Deep NNs have several flaws. For example:

- They are **data hungry**:
  - With few data, learning is not possible, even for simple logical reasoning tasks;
  - This motivates **Neural Symbolic Integration (NeSy)**.
- They are **black boxes**:
  - Predictions are not explainable, might lead to lack of trust in AI applications;
  - This motivates the research field of **Explainable AI (XAI)**.

Knowledge Enhanced Neural Networks

KENN[1] consists in a residual layer designed to improve the predictions of a base NN, by using logical prior knowledge, consisting in a set of FOL formulas $\mathcal{K}$.



[1] Daniele, Alessandro, and Luciano Serafini. "Knowledge enhanced neural networks." Pacific Rim International Conference on Artificial Intelligence. Springer, Cham, 2019.

Basic Terminology

Definition (The Language)

Our language will be a function-free first order language $\mathcal{L}$, defined by:

- A set of **constants**: $\mathcal{C} = \{a_1, \ldots, a_{|\mathcal{C}|}\}$;
- A set of **predicates**: $\mathcal{P} = \{P_1, \ldots, P_{|\mathcal{P}|}\}$;

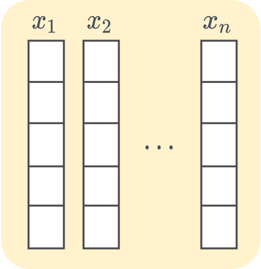
Definition (Clause)

A clause c is a formula expressed a disjunction of literals:

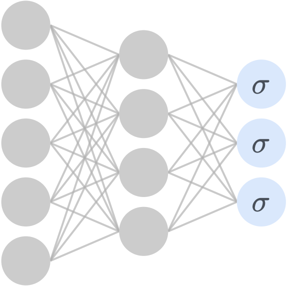
$$c := \bigvee_{i=1}^{k} l_i, \quad l_i \neq l_j \quad \forall i \neq j$$

Language Semantic

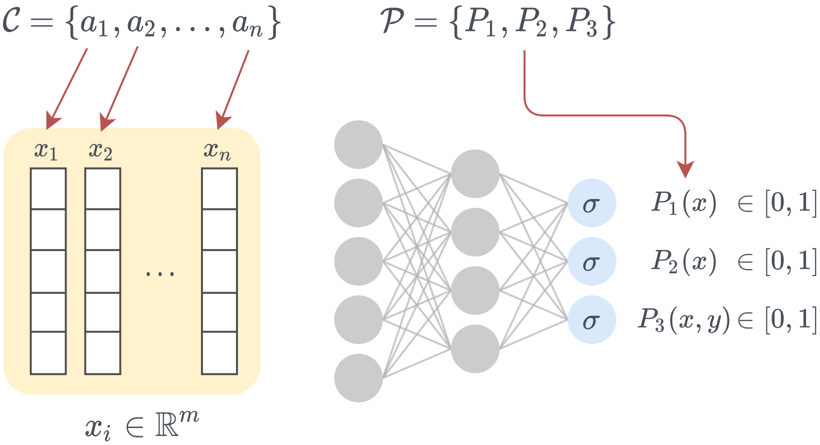
$$\mathcal{C} = \{a_1, a_2, \ldots, a_n\} \qquad \mathcal{P} = \{P_1, P_2, P_3\}$$



$x_i \in \mathbb{R}^m$



Language Semantic



KENN: Intuition

Given the vector of predictions of the NN y, KENN computes the final vector of predictions as follows:

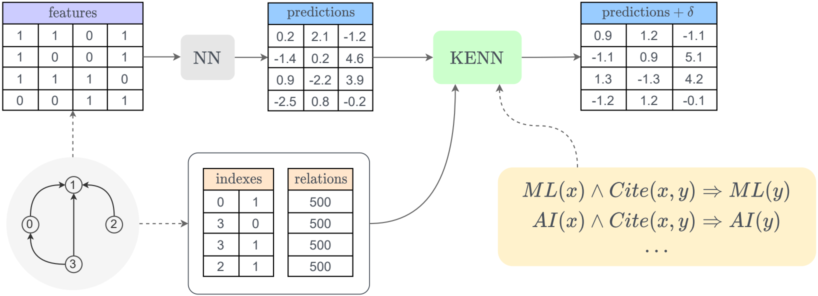
$$y' = y + \sum_{c \in \mathcal{K}} w_c \cdot \delta^c$$

where, for each $c \in \mathcal{C}$:

- δ^c improves the truth value of c, keeping $\|\delta^c\|_2$ minimal;
- $w_c \in \mathbb{R}$ is the **clause weight**, a learnable parameter that quantifies the importance of clause c.

Citeseer Experiments

- We tested KENN on a Collective Classification task;
- The **Citeseer Dataset** was used: citation network with 4732 citations (edges) between 3312 papers (nodes);
- The task is to predict the topic of each paper (6 possible topics).



Comparison with previous literature

- We also provide a comparison with two other NeSy models:
  - **Semantic Based Regularization**[2];
  - **Relational Neural Machines**[3];

[2] Diligenti, Michelangelo, Marco Gori, and Claudio Sacca. "Semantic-based regularization for learning and inference." Artificial Intelligence 244 (2017): 143-165.

[3] Marra, Giuseppe, et al. "Relational neural machines." arXiv preprint arXiv:2002.02193 (2020).

Comparison with previous literature

- We also provide a comparison with two other NeSy models:
  - **Semantic Based Regularization**[2];
  - **Relational Neural Machines**[3];
- The same base NN and the same base knowledge are used.

[2] Diligenti, Michelangelo, Marco Gori, and Claudio Sacca. "Semantic-based regularization for learning and inference." Artificial Intelligence 244 (2017): 143-165.

[3] Marra, Giuseppe, et al. "Relational neural machines." arXiv preprint arXiv:2002.02193 (2020).

Comparison with previous literature

- We also provide a comparison with two other NeSy models:
  - **Semantic Based Regularization**[2];
  - **Relational Neural Machines**[3];
- The same base NN and the same base knowledge are used.
- The main evaluation metric is the **relative improvement** over the base NN accuracy;

---

[2] Diligenti, Michelangelo, Marco Gori, and Claudio Sacca. "Semantic-based regularization for learning and inference." Artificial Intelligence 244 (2017): 143-165.

[3] Marra, Giuseppe, et al. "Relational neural machines." arXiv preprint arXiv:2002.02193 (2020).

Comparison with previous literature

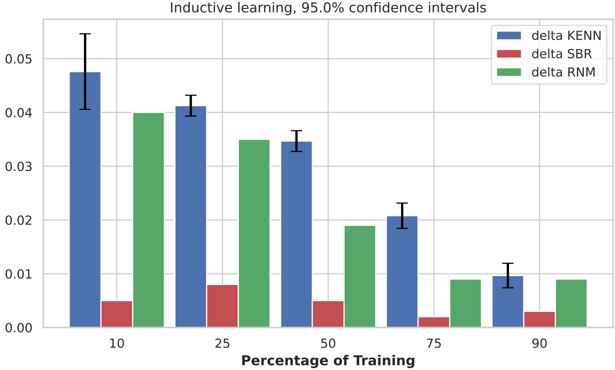
- We also provide a comparison with two other NeSy models:
  - **Semantic Based Regularization**[2];
  - **Relational Neural Machines**[3];
- The same base NN and the same base knowledge are used.
- The main evaluation metric is the **relative improvement** over the base NN accuracy;
- Same experiments are performed over different sizes of the training set.

---

[2] Diligenti, Michelangelo, Marco Gori, and Claudio Sacca. "Semantic-based regularization for learning and inference." Artificial Intelligence 244 (2017): 143-165.

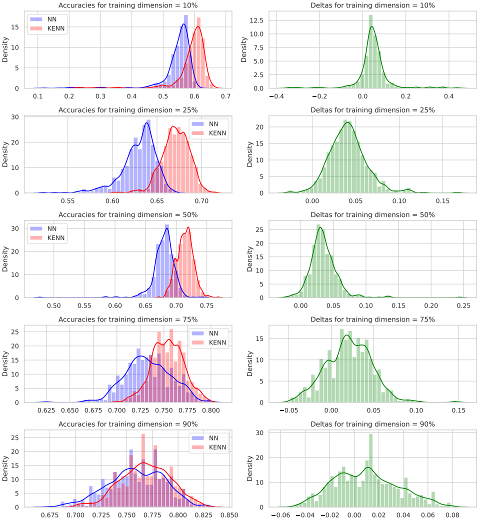
[3] Marra, Giuseppe, et al. "Relational neural machines." arXiv preprint arXiv:2002.02193 (2020).

Results

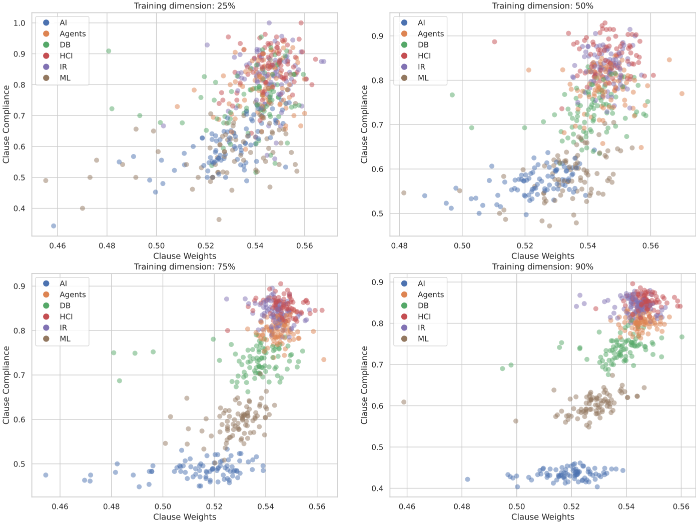

Inductive learning, 95.0% confidence intervals
delta KENN
delta SBR
delta RNM
0.00
0.01
0.02
0.03
0.04
0.05
10
25
50
75
90
Percentage of Training

Results

Università
degli Studi
di Padova



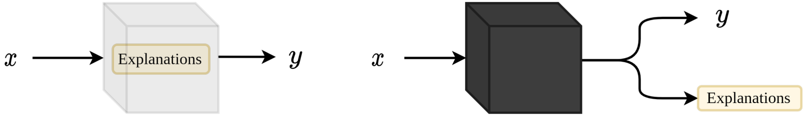
Clause Weights Learning



Explainability

In XAI, two main paradigms for explainability are distinguished:

- **Transparency**
- **Post-hoc explainability**



Explainability in KENN

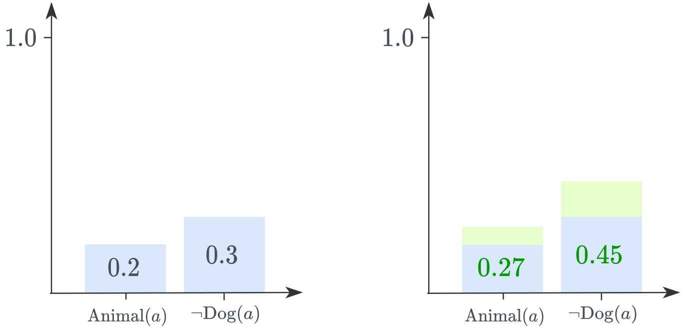
KENN can be considered a **partially transparent** model:

- A KENN layer will always be based on the prediction of a base NN, which will always be an inherently opaque model;
- On the contrary, everything happening inside the KENN layer is transparent;
- The explanations will only regard the knowledge enforcement stage.

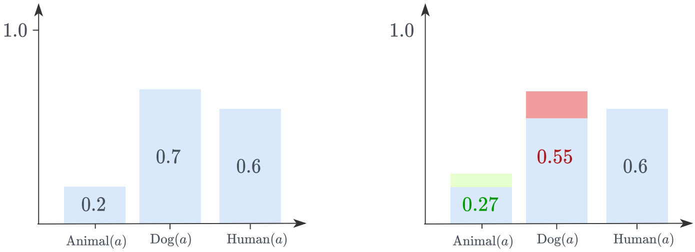
Local explanations from a single clause

$$\neg \, \mathrm{Dog}(a) \vee \mathrm{Animal}(a)$$

$$\neg\,\mathsf{Dog}(a) \lor \mathsf{Animal}(a)$$

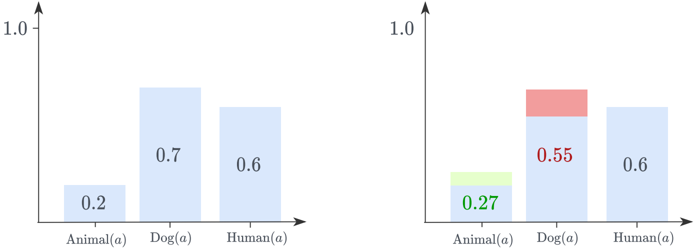


$\neg \, \mathsf{Dog}(a) \lor \mathsf{Animal}(a)$



Local explanations from a single clause

$$\neg \text{Dog}(a) \vee \text{Animal}(a)$$



Since the NN was confident that a is not an Animal, the truth value for a being a dog should decrease.

Assessing impact of more clauses

- In real use cases, we might have hundreds or thousands of clauses or samples $\Rightarrow$ one by one examination of each sample is not feasible;

Assessing impact of more clauses

- In real use cases, we might have hundreds or thousands of clauses or samples $\Rightarrow$ one by one examination of each sample is not feasible;
- We need ways to assess how the knowledge is modifying the base NN predictions, from a macroscopic point of view. Given any $\mathcal{C} \subseteq \mathcal{K}$ we might want to know:

Assessing impact of more clauses

- In real use cases, we might have hundreds or thousands of clauses or samples $\Rightarrow$ one by one examination of each sample is not feasible;
- We need ways to assess how the knowledge is modifying the base NN predictions, from a macroscopic point of view. Given any $\mathcal{C} \subseteq \mathcal{K}$ we might want to know:
  - if, and where those clauses provided a positive or negative contribution;

Assessing impact of more clauses

- In real use cases, we might have hundreds or thousands of clauses or samples $\Rightarrow$ one by one examination of each sample is not feasible;
- We need ways to assess how the knowledge is modifying the base NN predictions, from a macroscopic point of view. Given any $\mathcal{C} \subseteq \mathcal{K}$ we might want to know:
  - if, and where those clauses provided a positive or negative contribution;
  - if and where there is any conflict between the formulas inside $\mathcal{C}$.

Improvement Score

Given $\mathcal{C} \subseteq \mathcal{K}$, the improvement score quantifies the positive (or negative) contribution of $\mathcal{C}$ for sample x and is defined as follows:

$$IS(x, \mathcal{C}) = \sum_{i=1}^{m} \delta_i \cdot I_i.$$

Improvement Score

Given $\mathcal{C} \subseteq \mathcal{K}$, the improvement score quantifies the positive (or negative) contribution of $\mathcal{C}$ for sample x and is defined as follows:

$$IS(x, \mathcal{C}) = \sum_{i=1}^{m} \delta_i \cdot I_i.$$

$IS(x_1, \mathcal{C}) = -1.2$	$IS(x_2, \mathcal{C}) = 5.4$	$IS(x_3, \mathcal{C}) = 1.4$	$IS(x_4, \mathcal{C}) = -3.3$	$IS(x_5, \mathcal{C}) = 0.1$

Evaluation metrics

UNIVERSITÀ DEGLI STUDI DI PADOVA

Improvement Score

Given $\mathcal{C} \subseteq \mathcal{K}$, the improvement score quantifies the positive (or negative) contribution of $\mathcal{C}$ for sample x and is defined as follows:

$$IS(x, \mathcal{C}) = \sum_{i=1}^{m} \delta_i \cdot I_i.$$

$IS(x_1, \mathcal{C}) = -1.2$	$IS(x_2, \mathcal{C}) = 5.4$	$IS(x_3, \mathcal{C}) = 1.4$	$IS(x_4, \mathcal{C}) = -3.3$	$IS(x_5, \mathcal{C}) = 0.1$

$\big\downarrow$ sort

$IS(x_2, \mathcal{C}) = 5.4$	$IS(x_3, \mathcal{C}) = 1.4$	$IS(x_5, \mathcal{C}) = 0.1$	$IS(x_1, \mathcal{C}) = -1.2$	$IS(x_4, \mathcal{C}) = -3.3$

Evaluation metrics

Disagreement Score

We first define the disagreement vector:

$$DV(x, \mathcal{C}) = \sum_{c \in \mathcal{C}} |\delta_c| - \left| \sum_{c \in \mathcal{C}} \delta_c \right| .$$

Starting from $DV(x, C)$ we can finally define the disagreement score for a specific subset of predicates $\hat{\mathcal{P}} \subseteq \mathcal{P}$:

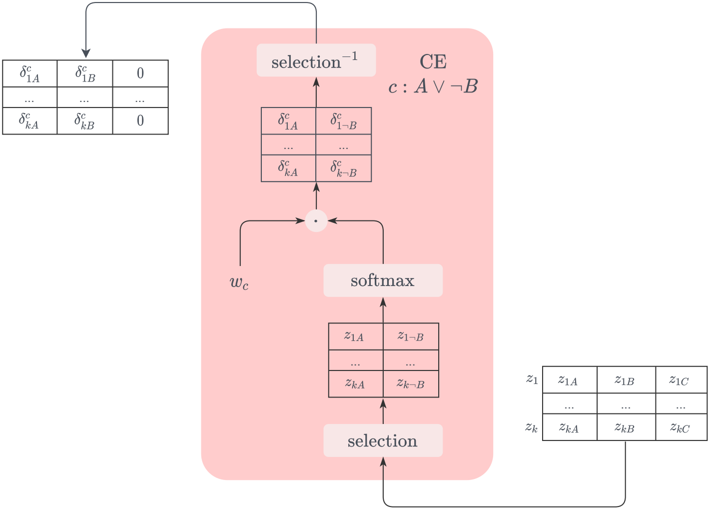
$$DS(x, \mathcal{C}, \hat{\mathcal{P}}) = \sum_{i \in \hat{\mathcal{P}}} DV(x, \mathcal{C})_i .$$

Conclusions

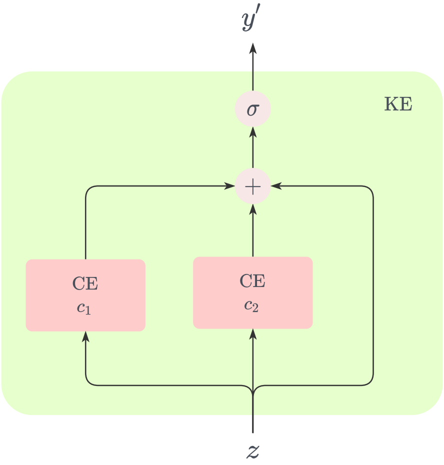
Università degli Studi di Padova

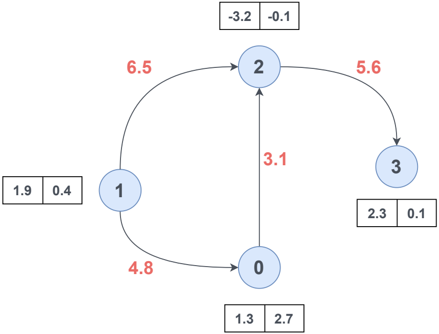
1. Experimental results show that KENN outperforms other NeSy methods for the collective classification task;
2. Further experiments show a correlation between the clause weights and the satisfaction of the clause in the training data;
3. KENN is inherently a transparent NN layer: explanations can be easily extracted in a understandable and human readable form;
4. We proposed two evaluation metrics which can be used for debugging purposes.

Thank you for your attention


selection^{-1}
CE
$c : A \vee \neg B$
δ^c_{1A} δ^c_{1B} 0
... ...
δ^c_{kA} δ^c_{kB} 0
δ^c_{1A} $\delta^c_{1\neg B}$
... ...
δ^c_{kA} $\delta^c_{k\neg B}$
$\cdot$
w_c
softmax
z_{1A} $z_{1\neg B}$
... ...
z_{kA} $z_{k\neg B}$
selection
z_1 z_{1A} z_{1B} z_{1C}
... ...
z_k z_{kA} z_{kB} z_{kC}

Appendix: Knowledge Enhancer





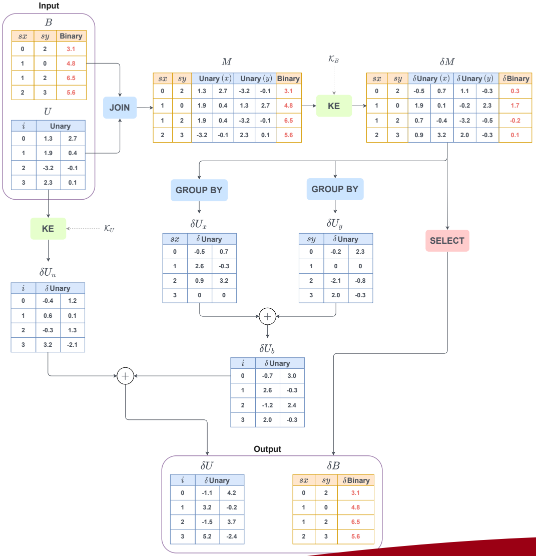
U

i	Unary	
0	1.3	2.7
1	1.9	0.4
2	-3.2	-0.1
3	2.3	0.1

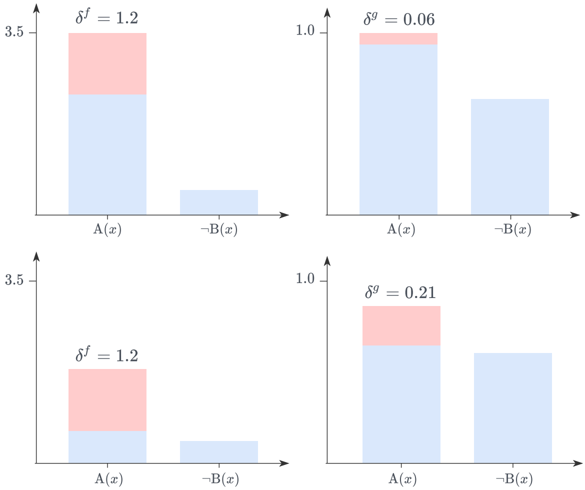
B

sx	sy	Binary
0	2	3.1
1	0	4.8
1	2	6.5
2	3	5.6

Appendix: KENN for relational data



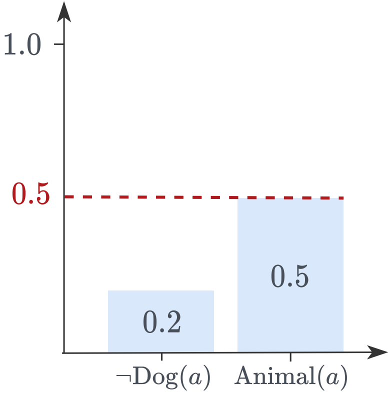
Appendix: preactivations vs activations



Example: truth value of a clause

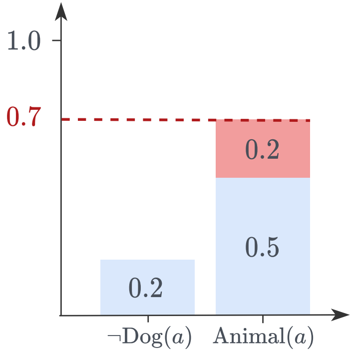
$$z = (0.5, 0.8, 0.3) \xrightarrow[c\,:\,\neg\,\mathrm{Dog}(a)\,\vee\,\mathrm{Animal}(a)]{\text{selection}} z_c = (0.2, 0.5)$$

Example: truth value of a clause

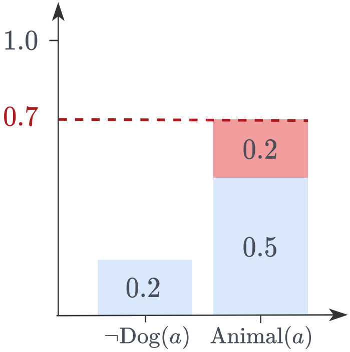


$\perp_{\max}(0.2, 0.5) = \max(0.2, 0.5) = 0.5$

Increasing satisfaction of a single clause



Increasing satisfaction of a single clause



$$\delta_s^{w_c}(z_c) = w_c \cdot \mathrm{softmax}(z_c)$$