Logic Based Explanations for Neural Networks

João Leite

(with M. de Sousa Ribeiro, J. Ferreira, R. Gonçalves)

NOVALINCS
LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

NOVA SCHOOL OF SCIENCE & TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE

NOVA UNIVERSITY LISBON

Logic Based Explanations for Neural Networks

- Existing neural network-based systems proved to be:
  - Capable of analyzing and classifying text, image, video and speech;
  - Able to act autonomously, making decisions previously made by humans.



The need for humans to understand their reasoning becomes evident

Logic Based Explanations for Neural Networks

Justifications or Explanations

- Allow users to build trust in a model and its results;
- Increase the chances of users acting based on models' outputs;
- Lead to better assessment of when a system is right or wrong.

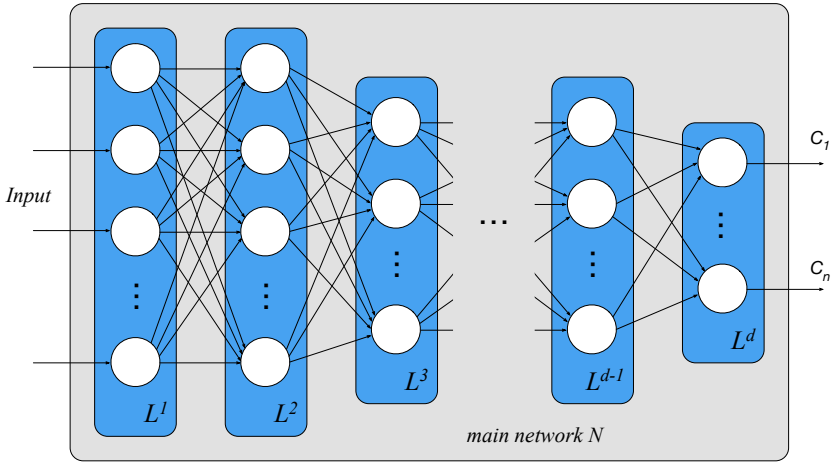
Ideally, neural networks would have the ability to explain or justify their results

Logic Based Explanations for Neural Networks

However…

- Neural networks use representations based on high-dimensional data.
- Do not provide human interpretable indications of why a specific output was produced

We need a human-understandable language



Logic Based Explanations for Neural Networks

Ontology / Background Theory

- Provides a conceptualization of a domain, describing how concepts are related with each other

- Usually specified using a logic-based language with a precise semantics
  - E.g. ASP or DL

$$
\begin{aligned}
\text{Train} &\equiv \exists \text{has}.(\text{Wagon} \sqcup \text{Locomotive}) \\
\text{WarTrain} &\sqsupseteq \exists \text{has}.\text{ReinforcedCar} \sqcap \exists \text{has}.\text{PassengerCar} \\
\text{EmptyTrain} &\equiv \forall \text{has}.(\text{EmptyWagon} \sqcup \text{Locomotive}) \sqcap \exists \text{has}.\text{EmptyWagon} \\
\text{PassengerTrain} &\sqsupseteq \exists \text{has}.(\text{PassengerCar} \sqcap \text{LongWagon}) \sqcup (\geq 2\ \text{has}.\text{PassengerCar}) \\
\text{LongFreightTrain} &\equiv \text{LongTrain} \sqcap \text{FreightTrain} \\
\text{LongTrain} &\sqsupseteq (\geq 2\ \text{has}.\text{LongWagon}) \sqcup (\geq 3\ \text{has}.\text{Wagon}) \\
\text{FreightTrain} &\sqsupseteq (\geq 2\ \text{has}.\text{FreightWagon}) \\
\text{RuralTrain} &\sqsupseteq \exists \text{has}.\text{EmptyWagon} \sqcap \exists \text{has}.(\text{PassengerCar} \sqcup \text{FreightWagon}) \\
&\quad \sqcap \neg \exists \text{has}.\text{LongWagon} \\
\text{MixedTrain} &\sqsupseteq \exists \text{has}.\text{FreightWagon} \sqcap \exists \text{has}.\text{PassengerCar} \sqcap \exists \text{has}.\text{EmptyWagon} \\
\text{TypeA} &\equiv \text{WarTrain} \sqcup \text{EmptyTrain} \\
\text{TypeB} &\equiv \text{PassengerTrain} \sqcup \text{LongFreightTrain} \\
\text{TypeC} &\equiv \text{RuralTrain} \sqcup \text{MixedTrain}
\end{aligned}
$$

Logic Based Explanations for Neural Networks

de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

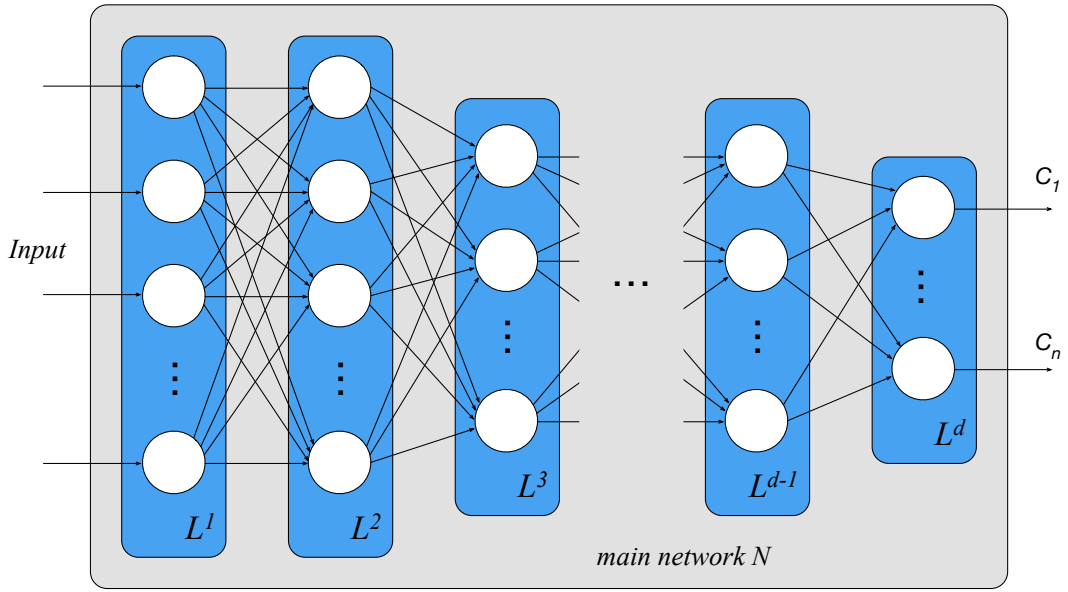
Background Knowledge

$C_1 \equiv C_{M1} \sqcup (C_{M2} \sqcap \neg C_{Mm})$

...

$C_n \equiv \neg C_{M2} \sqcup C_{Mm}$

How to relate both Systems



Logic Based Explanations for Neural Networks

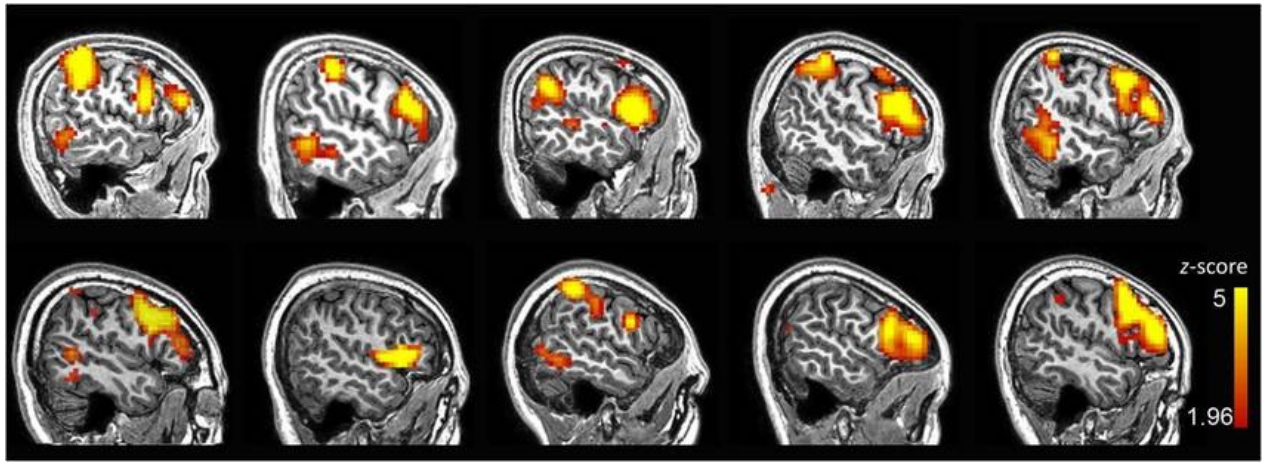
de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

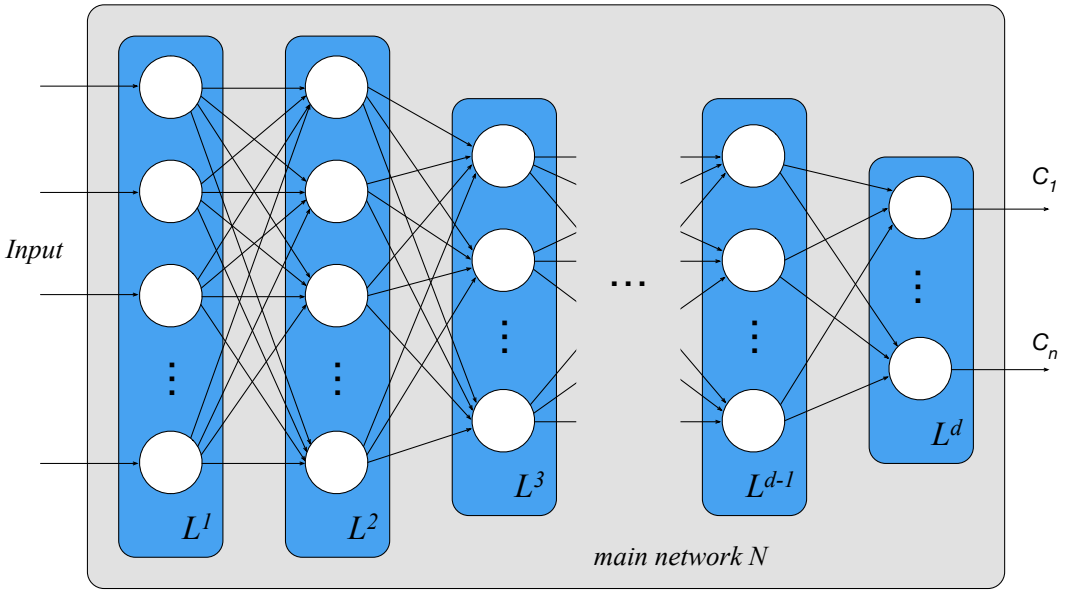
Background Knowledge

$C_1 \equiv C_{M1} \sqcup (C_{M2} \sqcap \neg C_{Mm})$

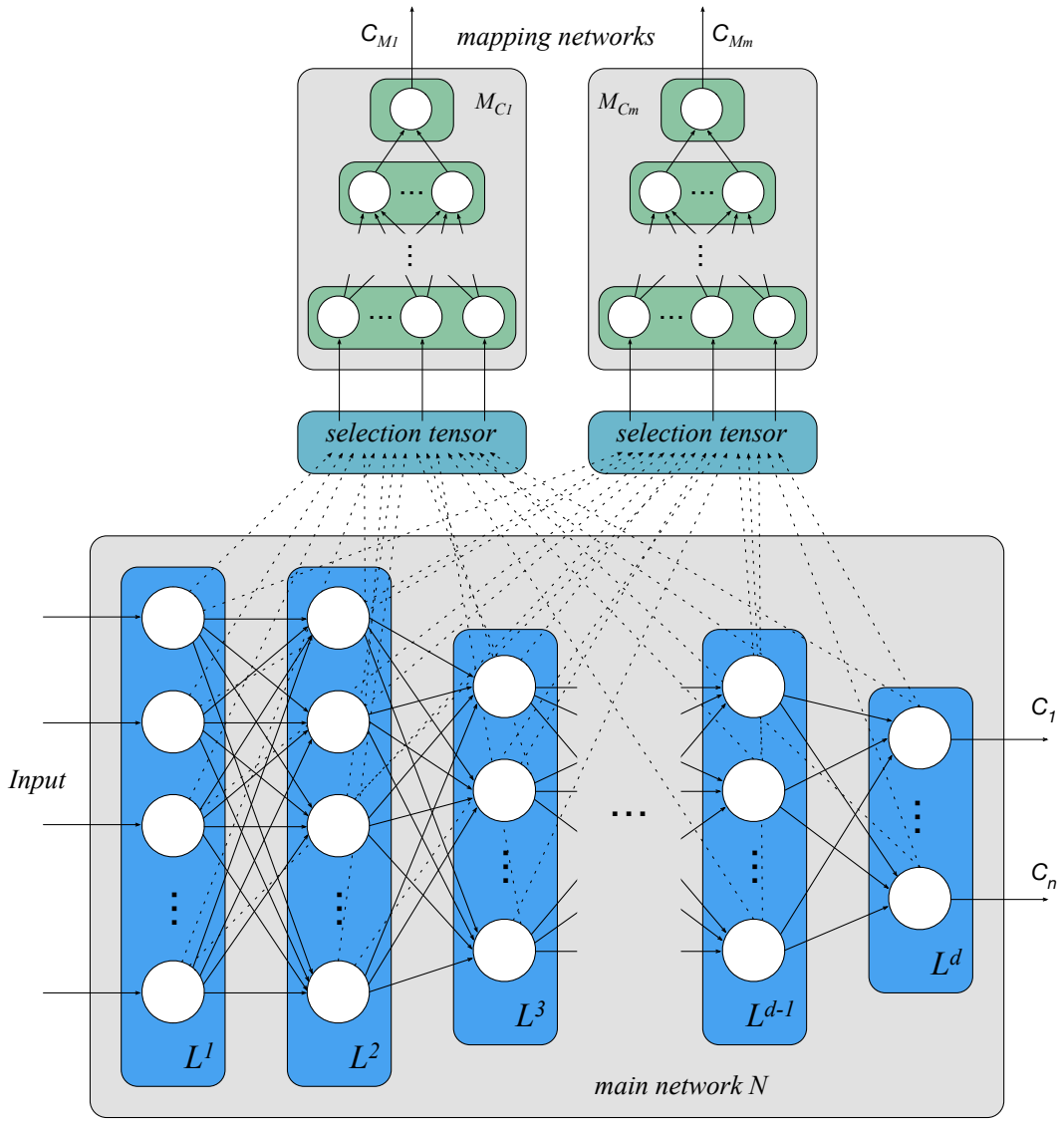
...

$C_n \equiv \neg C_{M2} \sqcup C_{Mm}$



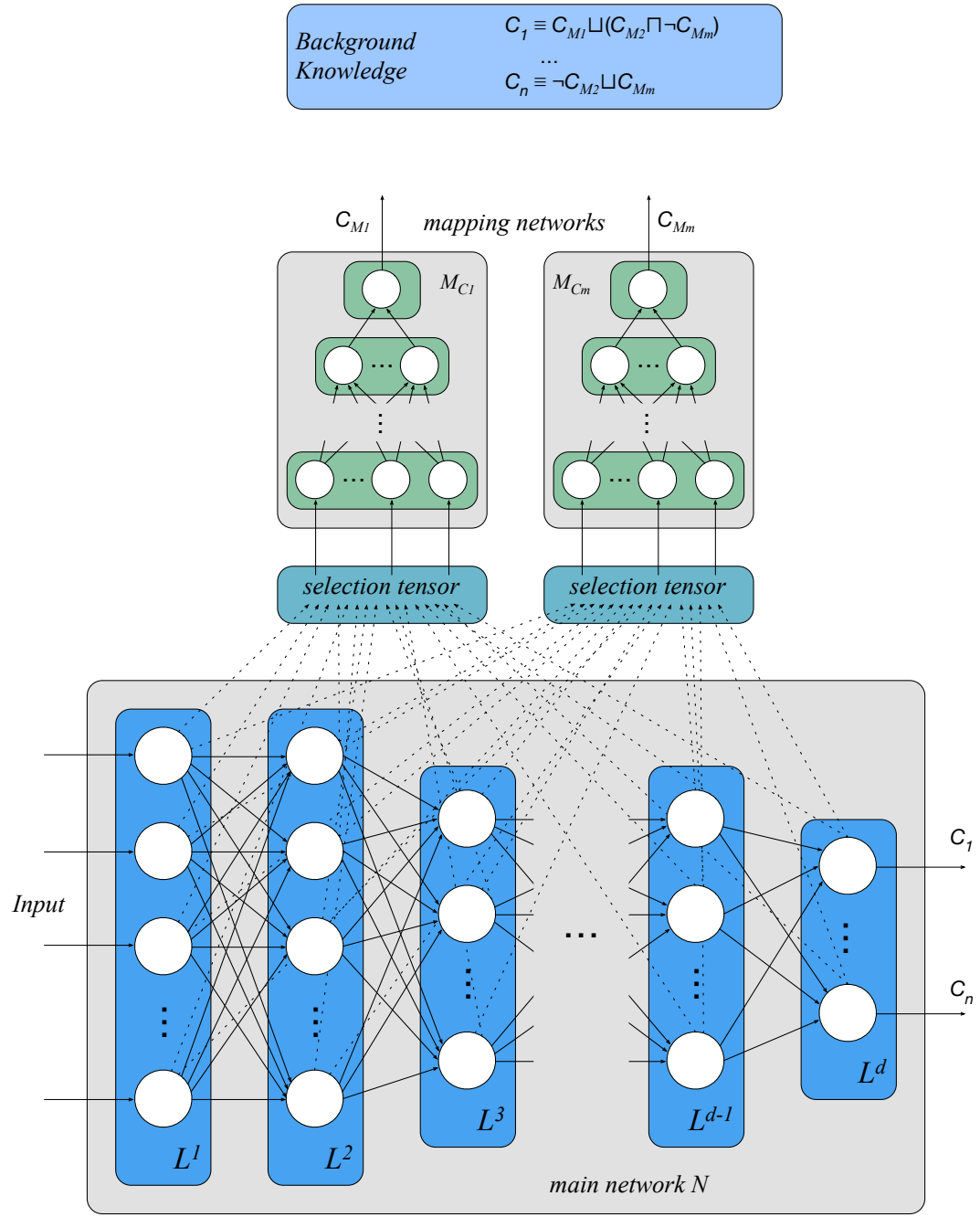


Logic Based Explanations for Neural Networks



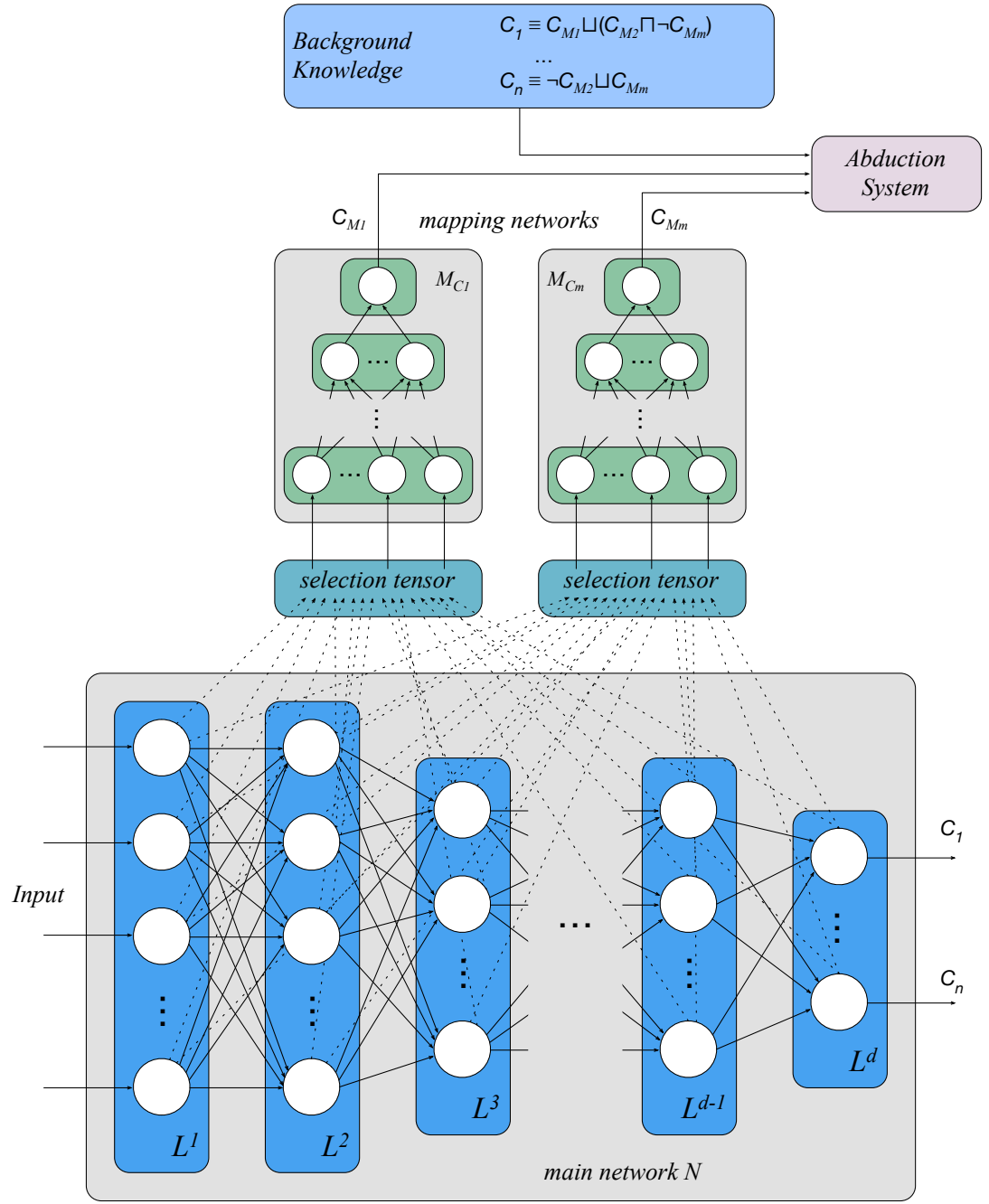
de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

Logic Based Explanations for Neural Networks



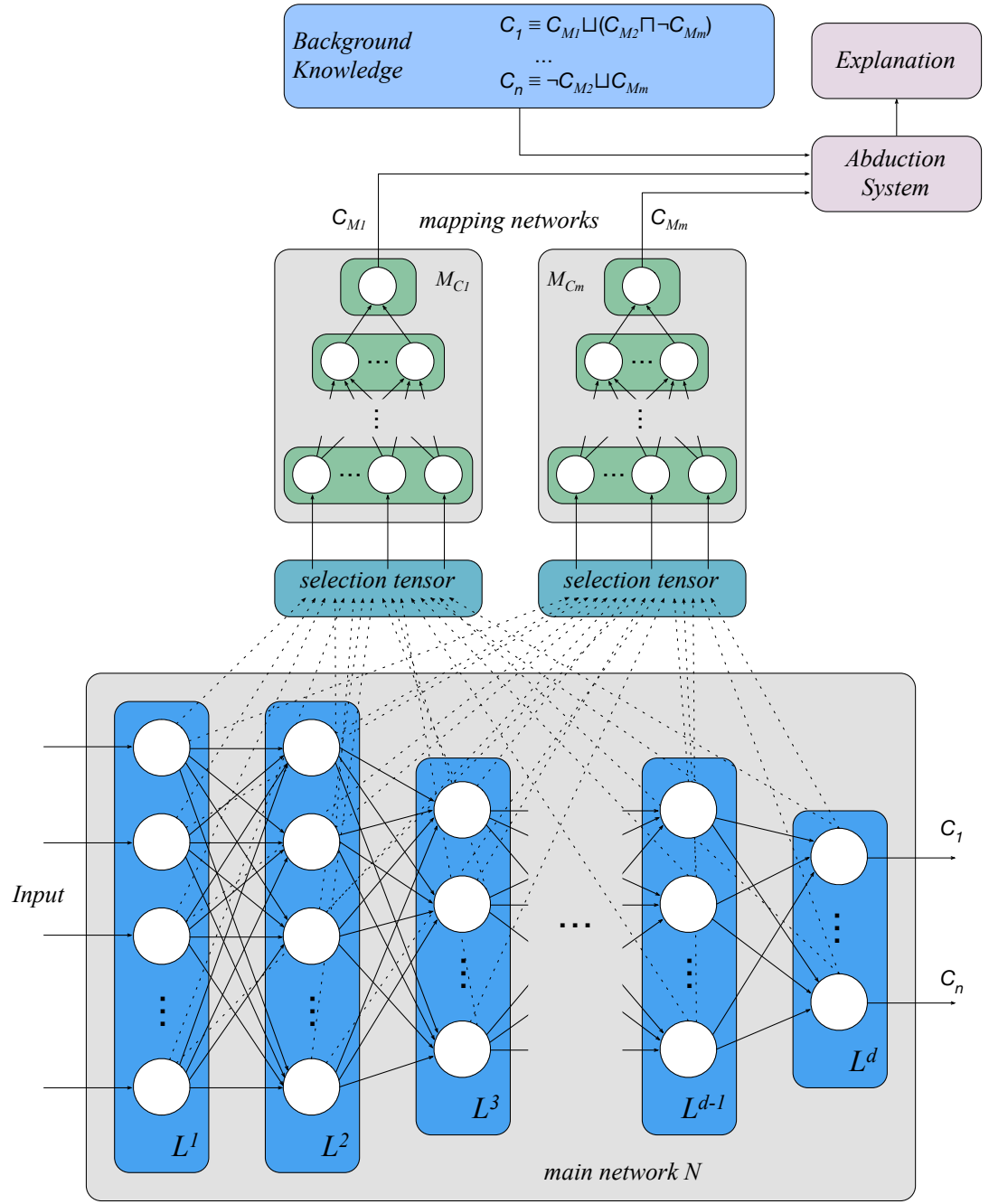
de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

Logic Based Explanations for Neural Networks



de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

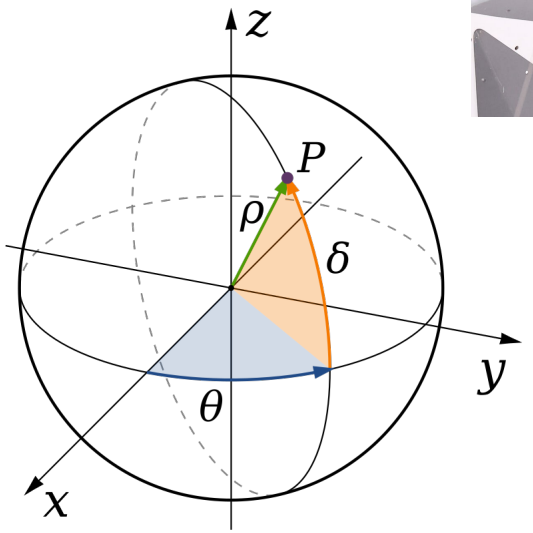
Logic Based Explanations for Neural Networks



de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

Logic Based Explanations for Neural Networks

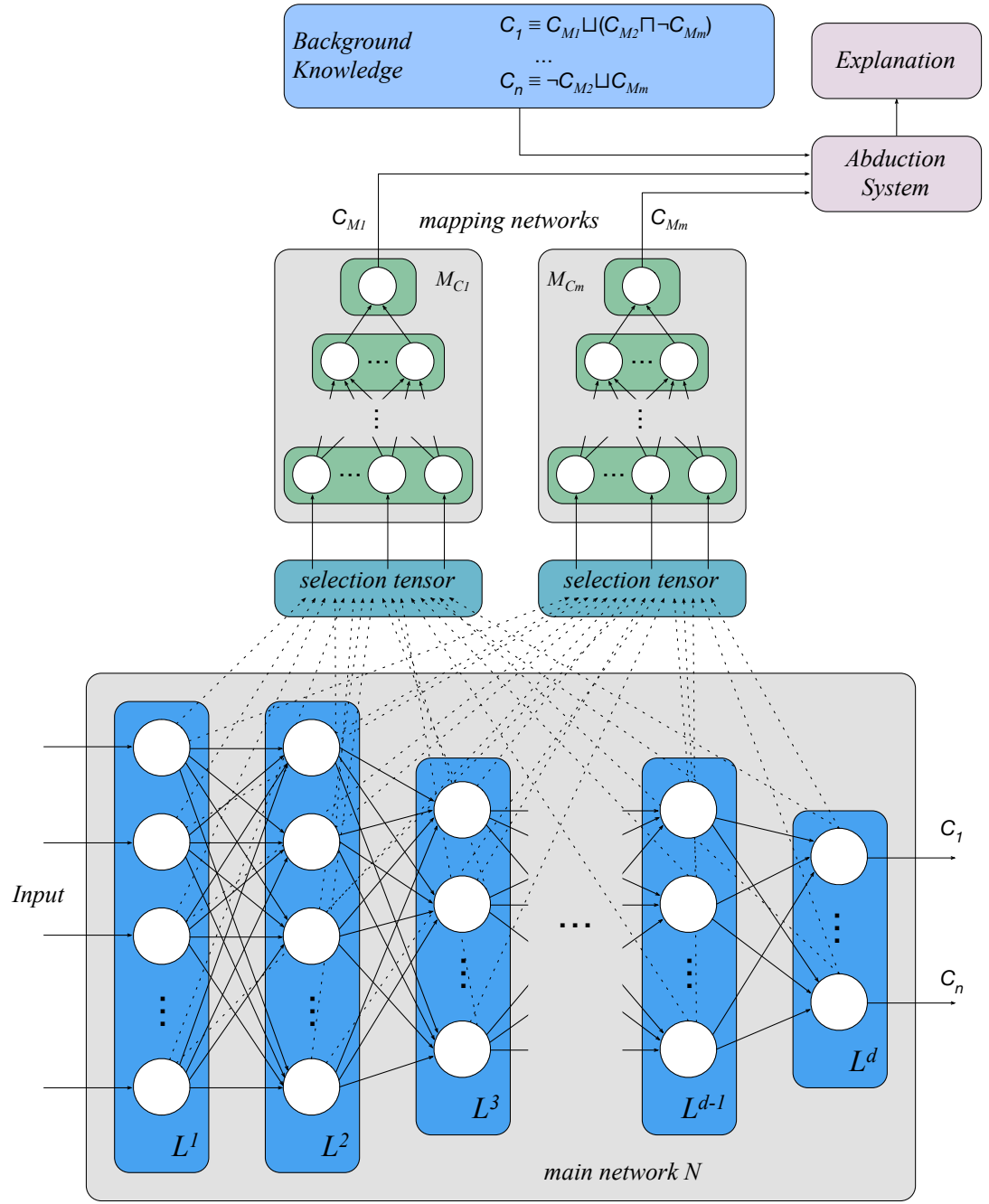
- Do different languages lead to different justifications?



- Based on the cartesian coordinate system.
- Based on the spherical coordinate system.

de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

Logic Based Explanations for Neural Networks



de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

Logic Based Explanations for Neural Networks

de Sousa Ribeiro & L, Aligning Artificial Neural Networks and Ontologies towards Explainable AI, In AAAI'21

Relevant and Non-Relevant Concepts

- If a concept is relevant, will we be able to extract it?

Cost of the Mappings

- Is there a benefit to using the activations of the main network?

Meaning of the Extracted Concepts

- Do the extracted concepts correspond to our understanding?

Origin of the Extracted Concepts

- Is it possible to pinpoint the neurons necessary to extract a given concept?

Justifications

- How good are the justifications obtained?

Logic Based Explanations for Neural Networks

XTRAINS Dataset
bitbucket.org/xtrains/dataset

Logic Based Explanations for Neural Networks

XTRAINS Dataset
bitbucket.org/xtrains/dataset

Type A

- trains having either a wagon with at least a circle inside and **a wagon with two walls in each side**, or no wagons with geometric figures inside them.

Type B

- trains having a long wagon or two wagons with at least a circle inside, or trains having at least two long wagons, or three wagons, with at least two of which with a geometric figure inside.

Type C

- trains having a wagon with no geometric figure inside, and either a wagon with a circle inside and a wagon with a geometric figure inside that is not a circle, or no long wagons and a wagon with a figure inside.

Logic Based Explanations for Neural Networks

XTRAINS Dataset
bitbucket.org/xtrains/dataset

XTRAINS Ontology (partial)

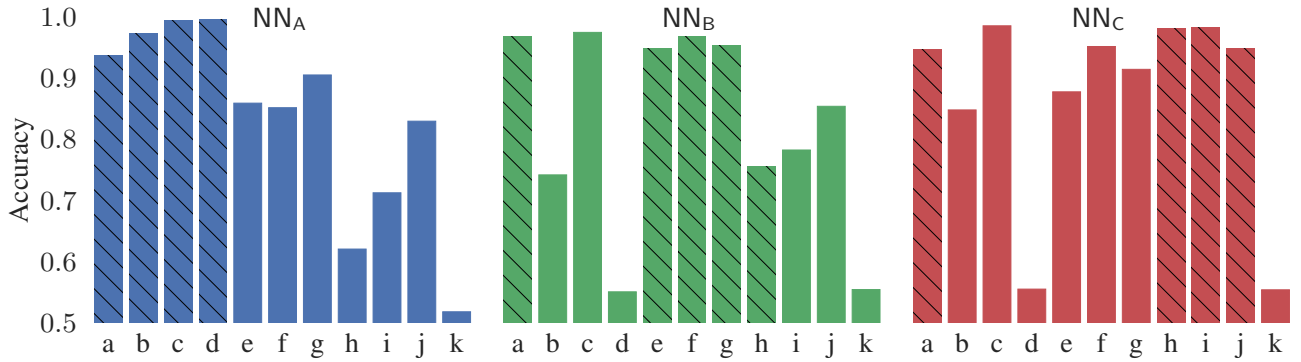
$$
\begin{aligned}
\text{Train} &\equiv \exists \text{has}.(\text{Wagon} \sqcup \text{Locomotive}) \\
\text{WarTrain} &\sqsupseteq \exists \text{has}.\text{ReinforcedCar} \sqcap \exists \text{has}.\text{PassengerCar} \\
\text{EmptyTrain} &\equiv \forall \text{has}.(\text{EmptyWagon} \sqcup \text{Locomotive}) \sqcap \exists \text{has}.\text{EmptyWagon} \\
\text{PassengerTrain} &\sqsupseteq \exists \text{has}.(\text{PassengerCar} \sqcap \text{LongWagon}) \sqcup (\geq 2\ \text{has}.\text{PassengerCar}) \\
\text{LongFreightTrain} &\equiv \text{LongTrain} \sqcap \text{FreightTrain} \\
\text{LongTrain} &\sqsupseteq (\geq 2\ \text{has}.\text{LongWagon}) \sqcup (\geq 3\ \text{has}.\text{Wagon}) \\
\text{FreightTrain} &\sqsupseteq (\geq 2\ \text{has}.\text{FreightWagon}) \\
\text{RuralTrain} &\sqsupseteq \exists \text{has}.\text{EmptyWagon} \sqcap \exists \text{has}.(\text{PassengerCar} \sqcup \text{FreightWagon}) \\
&\quad \sqcap \neg \exists \text{has}.\text{LongWagon} \\
\text{MixedTrain} &\sqsupseteq \exists \text{has}.\text{FreightWagon} \sqcap \exists \text{has}.\text{PassengerCar} \sqcap \exists \text{has}.\text{EmptyWagon} \\
\text{TypeA} &\equiv \text{WarTrain} \sqcup \text{EmptyTrain} \\
\text{TypeB} &\equiv \text{PassengerTrain} \sqcup \text{LongFreightTrain} \\
\text{TypeC} &\equiv \text{RuralTrain} \sqcup \text{MixedTrain}
\end{aligned}
$$

Logic Based Explanations for Neural Networks

Relevant and Non-Relevant Concepts

Relevant and Non-Relevant Concepts

- If a concept is relevant, will we be able to extract it?



a - ∃has.FreightWagon
b - WarTrain
c - EmptyTrain
d - ∃has.ReinforcedCar
e - PassengerTrain
f - LongTrain
g - FreightTrain
h - ∃has.LongWagon
i - RuralTrain
j - MixedTrain
k - ∃has.OpenRoofCar

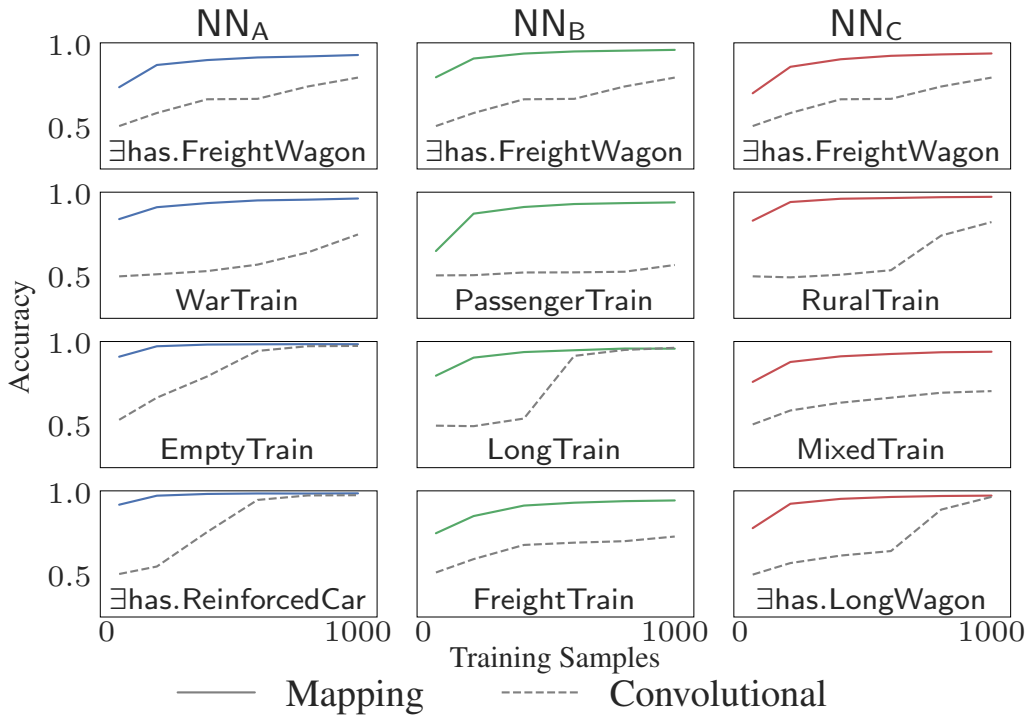
Relevant concepts are extracted with the highest accuracy values

Logic Based Explanations for Neural Networks

Cost of the Mappings

Cost of the Mappings

- Is there a benefit to using the activations of the main network?



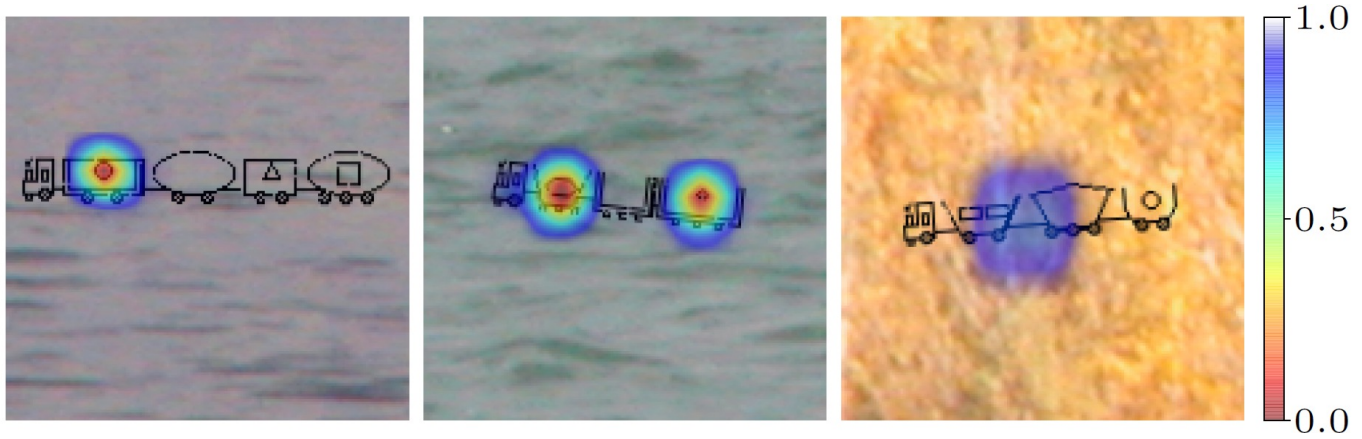
Mapping networks require few labeled training data

Logic Based Explanations for Neural Networks

Meaning of the Extracted Concepts

Meaning of the Extracted Concepts

- Do the extracted concepts correspond to our understanding?



Mapping networks properly identify the visual features embodying each concept

Logic Based Explanations for Neural Networks

Origin of the Extracted Concepts

Origin of the Extracted Concepts

- Is it possible to pinpoint the neurons necessary to extract a given concept?

Input Reduce procedure

- Searches for the smallest set of inputs achieving the highest accuracy to extract a given concept;
- On average, decreased the mapping networks' input by 95%;
- Resulting mapping networks achieved similar accuracies to those trained using all neurons from the dense layers of the main network.

Logic Based Explanations for Neural Networks

Origin of the Extracted Concepts

Origin of the Extracted Concepts

- Is it possible to pinpoint the neurons necessary to extract a given concept?

	Output Concept	Dense Layers		Input Reduce	
		Accuracy	Features	Accuracy	Features
NN_A	∃has.FreightWagon	0.9367	10480	0.9263	453
	WarTrain	0.9719	10480	0.9930	4
	EmptyTrain	0.9937	10480	0.9942	2
	∃has.ReinforcedCar	0.9950	10480	0.9928	4
NN_B	∃has.FreightWagon	0.9676	10464	0.9629	2374
	PassengerTrain	0.9485	10464	0.9433	1107
	LongTrain	0.9670	10464	0.9701	534
	FreightTrain	0.9523	10464	0.9493	1247
NN_C	∃has.FreightWagon	0.9459	10608	0.9500	519
	RuralTrain	0.9820	10608	0.9916	7
	MixedTrain	0.9484	10608	0.9750	14
	∃has.LongWagon	0.9813	10608	0.9814	12

It is possible to pinpoint the neurons necessary to extract a given concept

Logic Based Explanations for Neural Networks

Justifications

Justifications

- How good are the justifications obtained?

$\mathsf{LongTrain}(\mathsf{i}_1)$

$\mathsf{FreightTrain}(\mathsf{i}_1)$

$\mathsf{LongFreightTrain} \equiv \mathsf{LongTrain} \sqcap \mathsf{FreightTrain}$

$\mathsf{TypeB} \equiv \mathsf{PassengerTrain} \sqcup \mathsf{LongFreightTrain}$

	All Correct	Some Correct	None Correct	No Justifications
$\mathsf{NN_A}$	85.5%	14.3%	0.2%	0.0%
$\mathsf{NN_B}$	94.2%	2.1%	0.7%	3.0%
$\mathsf{NN_C}$	90.6%	8.9%	0.1%	0.4%

The resulting justifications were correct in most cases

Logic Based Explanations for Neural Networks

Relevant and Non-Relevant Concepts

- Relevant concepts are extracted with the highest accuracy values

Cost of the Mappings

- Mapping networks require few labeled training data

Meaning of the Extracted Concepts

- Mapping networks properly identify the visual features embodying each concept

Origin of the Extracted Concepts

- It is possible to pinpoint the neurons necessary to extract a given concept

Justifications

- The resulting justifications were correct in 90% of the cases

Logic Based Explanations for Neural Networks

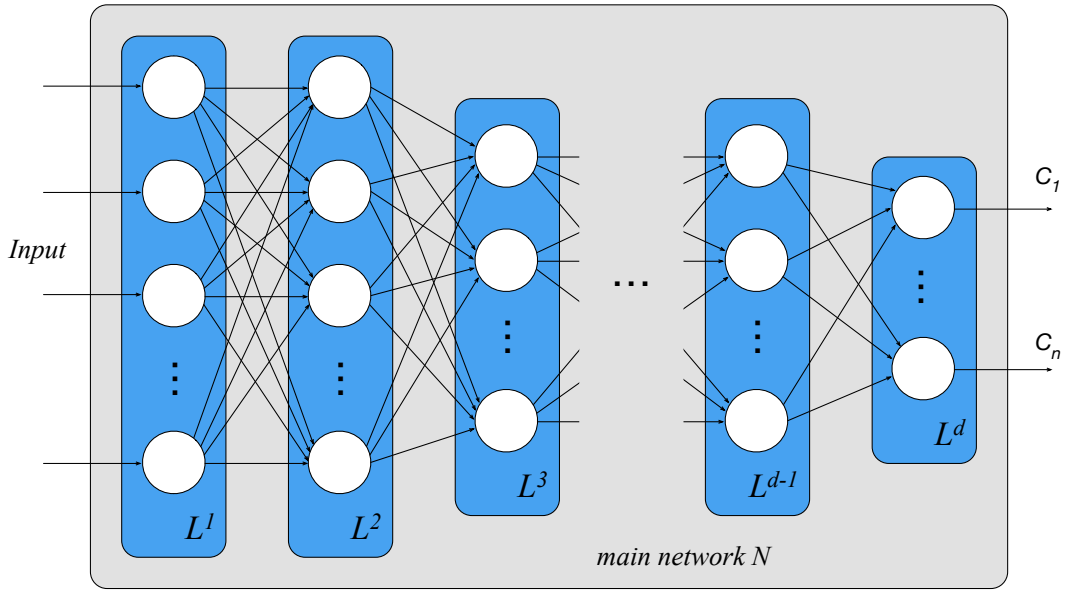
Background Knowledge

$C_1 \equiv C_{M1} \sqcup (C_{M2} \sqcap \neg C_{Mm})$

...

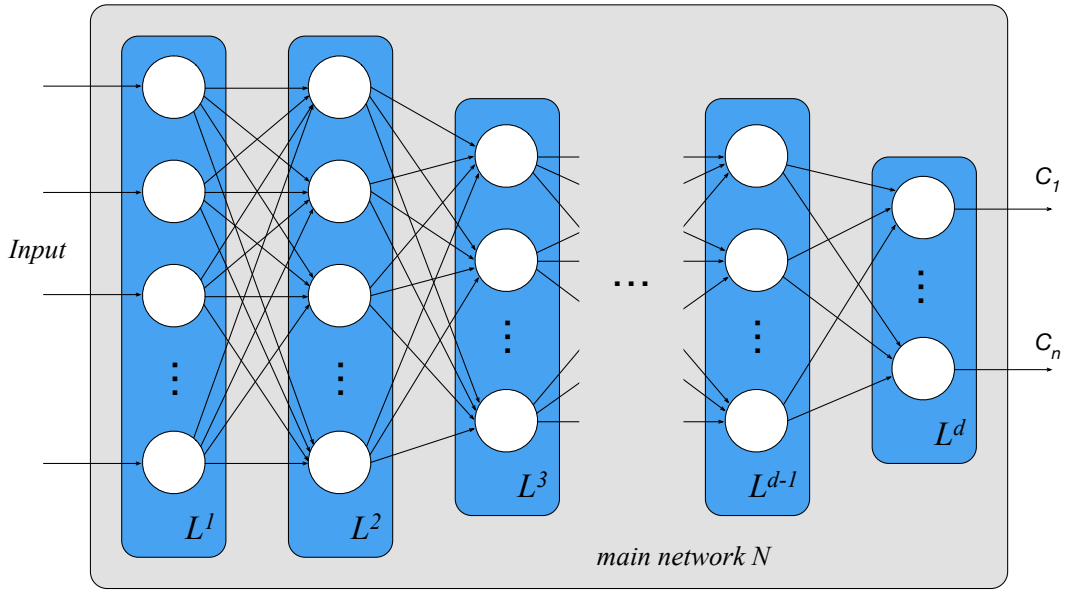
$C_n \equiv \neg C_{M2} \sqcup C_{Mm}$

What if we do not have the ontology?

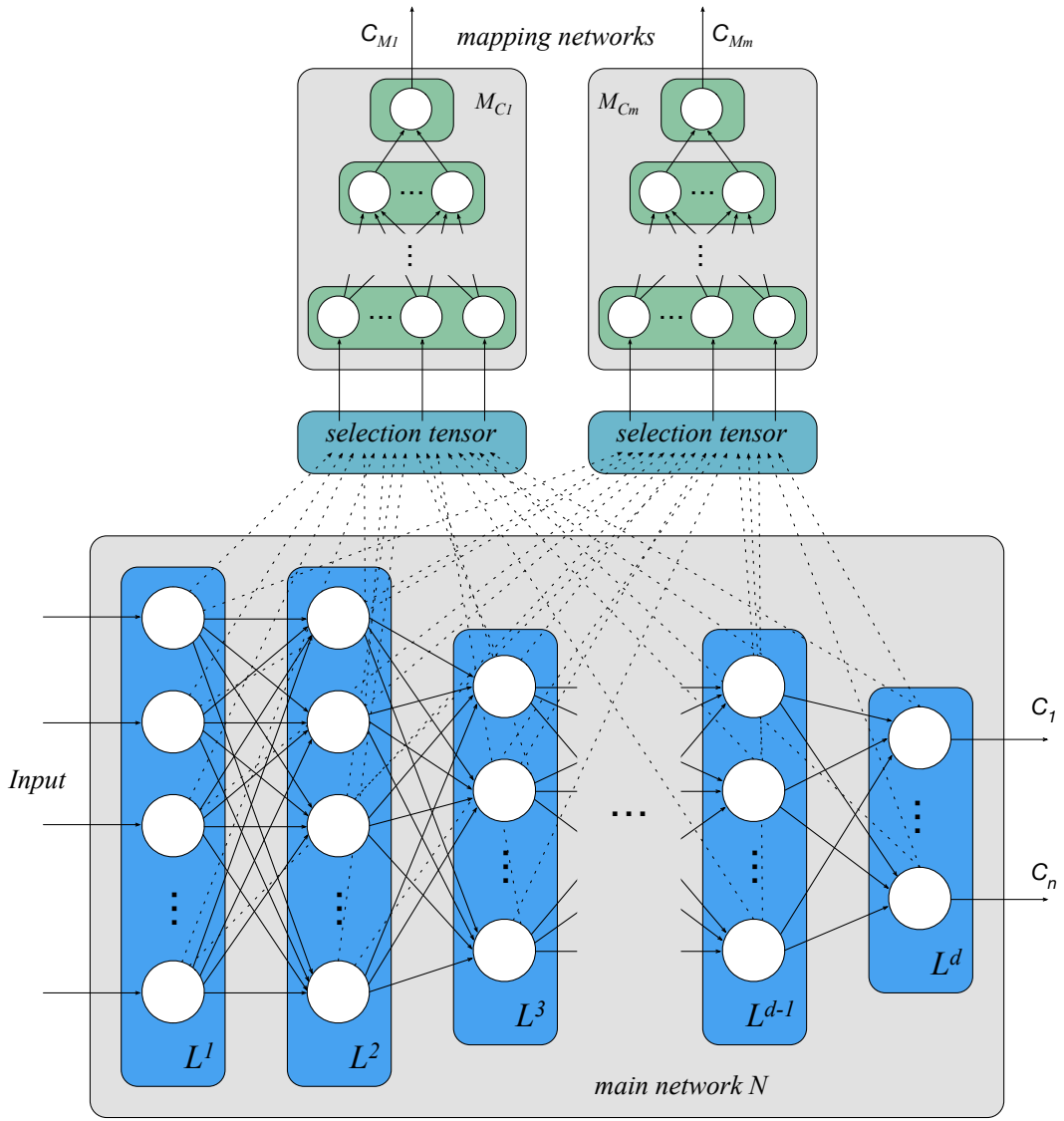


Logic Based Explanations for Neural Networks

Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

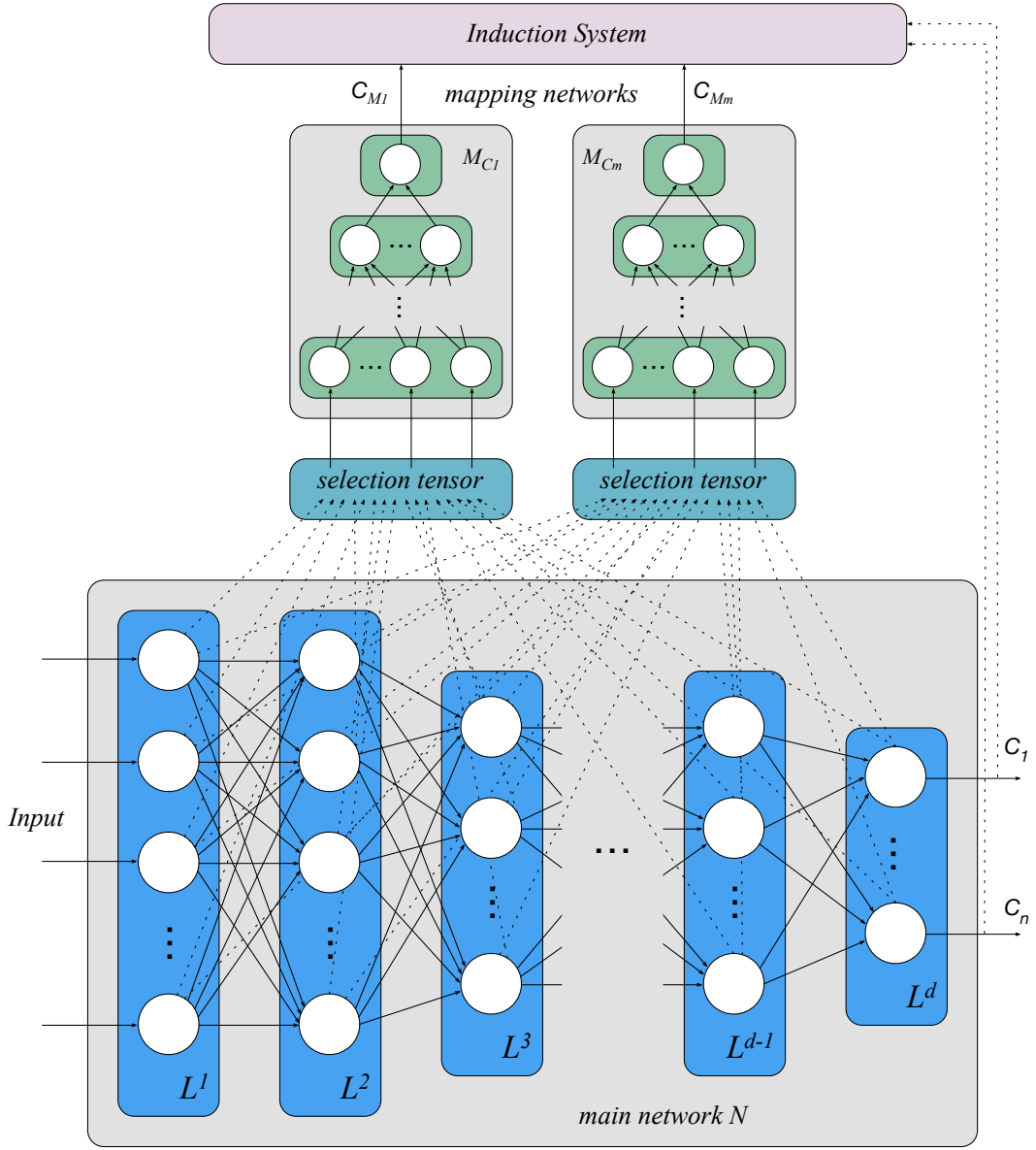


Logic Based Explanations for Neural Networks



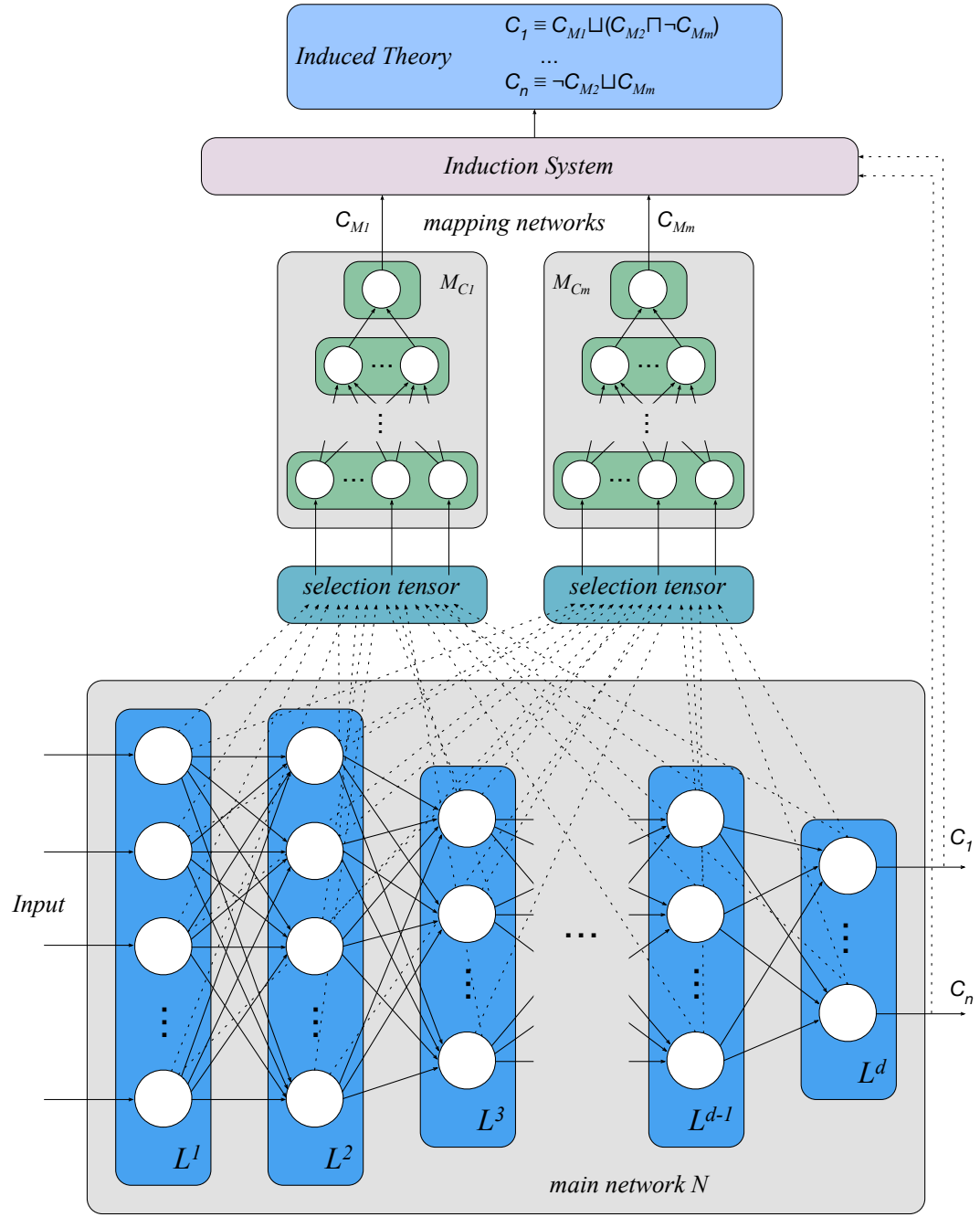
Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Logic Based Explanations for Neural Networks



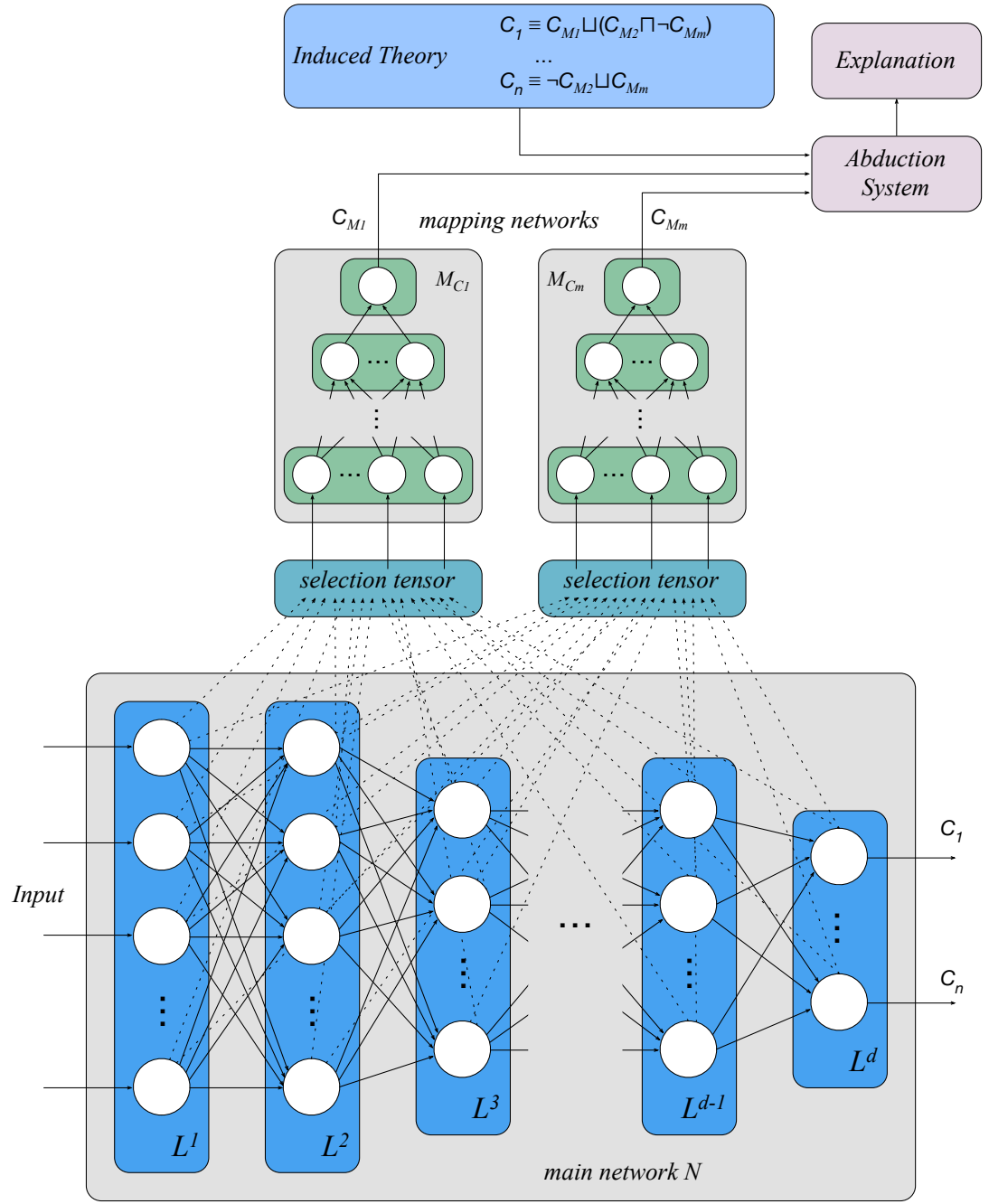
Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Logic Based Explanations for Neural Networks



Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Logic Based Explanations for Neural Networks



Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Logic Based Explanations for Neural Networks

Learning the Ontology

Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Dataset's Ontology

$\mathsf{TypeA} \equiv \mathsf{WarTrain} \sqcup \mathsf{EmptyTrain}$

$\mathsf{TypeB} \equiv \mathsf{PassengerTrain} \sqcup \mathsf{LongFreightTrain}$

$\mathsf{TypeC} \equiv \mathsf{MixedTrain} \sqcup \mathsf{RuralTrain}$

Learned Ontology

$\mathsf{TypeA} \equiv \mathsf{WarTrain} \sqcup \mathsf{EmptyTrain}$

$\mathsf{TypeB} \equiv (\mathsf{FreightTrain} \sqcap \mathsf{LongTrain}) \sqcup (\mathsf{PassengerTrain} \sqcap \neg\mathsf{EmptyTrain})$

$\mathsf{TypeC} \equiv \mathsf{MixedTrain} \sqcup \mathsf{RuralTrain}$

Mapped Concepts

$\mathsf{EmptyTrain}, \mathsf{FreightTrain}, \mathsf{LongTrain}$

$\mathsf{MixedTrain}, \mathsf{PassengerTrain}, \mathsf{RuralTrain}$

$\mathsf{WarTrain}, \exists\mathsf{has.FreightWagon}, \exists\mathsf{has.LongWagon}$

$\exists\mathsf{has.OpenRoof}, \exists\mathsf{has.ReinforcedCar}$

Fidelity Scores

	F_{Main}	$F_{XTrains}$
M_A	$99.72 \pm 0.18\%$	$99.92 \pm 0.35\%$
M_B	$98.71 \pm 0.31\%$	$99.83 \pm 0.76\%$
M_C	$99.33 \pm 0.32\%$	$99.52 \pm 1.52\%$

Logic Based Explanations for Neural Networks

Levels of Abstraction

Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Train-Level Concepts

Mapped Concepts

EmptyTrain, LongFreightTrain, MixedTrain, PassengerTrain, RuralTrain, WarTrain

Learned ontologies

$\text{TypeA} \equiv \text{EmptyTrain} \sqcup \text{WarTrain}$

$\text{TypeB} \equiv \text{LongFreightTrain} \sqcup \text{PassengerTrain}$

$\text{TypeC} \equiv \text{MixedTrain} \sqcup \text{RuralTrain}$

Fidelity Scores

	F_{Main}	$F_{XTrains}$
M_A	$99.76 \pm 0.19\%$	$100.00 \pm 0.0\%$
M_A	$98.82 \pm 0.39\%$	$100.00 \pm 0.0\%$
M_A	$99.46 \pm 0.20\%$	$100.00 \pm 0.0\%$

Wagon-Level Concepts

Mapped Concepts

$\exists \text{has.EmptyWagon}$, $\exists \text{has.FreightWagon}$, $\exists \text{has.LongWagon}$, $\exists \text{has.}(\text{LongWagon} \sqcap \text{PassengerCar})$, $\exists \text{has.PassengerCar}$, $\exists \text{has.ReinforcedCar}$, $\geq 2\text{has.FreightWagon}$, $\geq 2\text{has.LongWagon}$, $\geq 2\text{has.PassengerCar}$, $\geq 3\text{has.Wagon}$

Learned ontologies

$\text{TypeA} \equiv (\exists \text{has.PassengerCar} \sqcup \neg\exists \text{has.FreightWagon}) \sqcap (\exists \text{has.ReinforcedCar} \sqcup \neg\exists \text{has.PassengerCar})$

$\text{TypeB} \equiv \exists \text{has.}(\text{LongWagon} \sqcap \text{PassengerCar}) \sqcup (\geq 3\text{has.Wagon} \sqcap \geq 2\text{has.FreightWagon})$

$\text{TypeC} \equiv \exists \text{has.EmptyWagon} \sqcap (\neg\exists \text{has.LongWagon} \sqcup (\geq 3\text{has.Wagon} \sqcap \exists \text{has.PassengerCar}))$

Fidelity Scores

	F_{Main}	$F_{XTrains}$
M_A	$94.55 \pm 10.14\%$	$94.44 \pm 11.12\%$
M_A	$97.50 \pm 0.52\%$	$96.73 \pm 1.18\%$
M_A	$98.16 \pm 0.36\%$	$99.02 \pm 0.56\%$

Logic Based Explanations for Neural Networks

Insufficient Concepts

Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Dataset's Ontology

$\text{TypeA} \equiv \text{WarTrain} \sqcup \text{EmptyTrain}$

$\text{TypeB} \equiv \text{PassengerTrain} \sqcup \text{LongFreightTrain}$

$\text{TypeC} \equiv \text{MixedTrain} \sqcup \text{RuralTrain}$

Learned ontology

$\text{TypeA} \equiv \top$

$\text{TypeB} \equiv \geq 3\text{has.Wagon} \sqcup \text{LongFreightTrain}$

$\text{TypeC} \equiv (\geq 3\text{has.Wagon} \sqcap \exists\text{has.PassengerCar}) \sqcup \sqcup (\neg(\geq 2\text{has.PassengerCar}) \sqcap \sqcap \neg\exists\text{has.LongWagon})$

- Sampled 20 random sets of 5 concepts
- Trained mapping networks for each main network
- Learned Theories
- Average Fidelity scores
  - 72.6% (F_{Main})
  - 71.9% ($F_{XTrains}$)

Fidelity Scores

	F_{Main}	$F_{XTrains}$
M_A	$50.16 \pm 0.26\%$	$50.00 \pm 0.00\%$
M_B	$92.53 \pm 2.75\%$	$92.70 \pm 1.43\%$
M_C	$76.78 \pm 1.99\%$	$76.99 \pm 0.94\%$

Mapped Concepts

$\text{LongFreightTrain}, \exists\text{has.LongWagon}, \exists\text{has.PassengerCar}, \geq 3\text{has.Wagon}, \geq 2\text{has.PassengerCar}$

Logic Based Explanations for Neural Networks

Importance of the Mappings

Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Modified Dataset

Modified Dataset's Ontology

$$\text{On} \equiv \text{TopLeftOn} \sqcup \text{TopRightOn}$$
$$\text{On} \equiv \text{BottomLeftOn} \sqcup \text{BottomRightOn}$$

Baseline

- Trained 50 main networks
- Learned Theories using dataset labels
- Always obtained*

$$\text{On} \equiv \text{BottomLeftOn} \sqcup \text{BottomRightOn}$$

Using Mapping Networks

- 22%

  $\text{On} \equiv \text{TopLeftOn} \sqcup \text{TopRightOn}$
- 42%

  $\text{On} \equiv \text{BottomLeftOn} \sqcup \text{BottomRightOn}$
- 36%

  Other Combinations

Logic Based Explanations for Neural Networks

Conclusions

- Our method
  - Induces theories that are faithful to a neural network's classifications
  - Deals with unnecessary concepts
  - Induces theories at different levels of abstraction
  - Is applicable even when few labeled data is available

Future Work

- More challenging datasets
- Recurrent Networks
- Neuron Selection
- Concept finding
- Concept whitening

Ferreira, de Sousa Ribeiro, Gonçalves and L, Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, In KR'22

Logic Based Explanations for Neural Networks

João Leite

(with M. de Sousa Ribeiro, J. Ferreira, R. Gonçalves)

NOVALINCS
LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

NOVA SCHOOL OF SCIENCE & TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE

NOVA UNIVERSITY LISBON