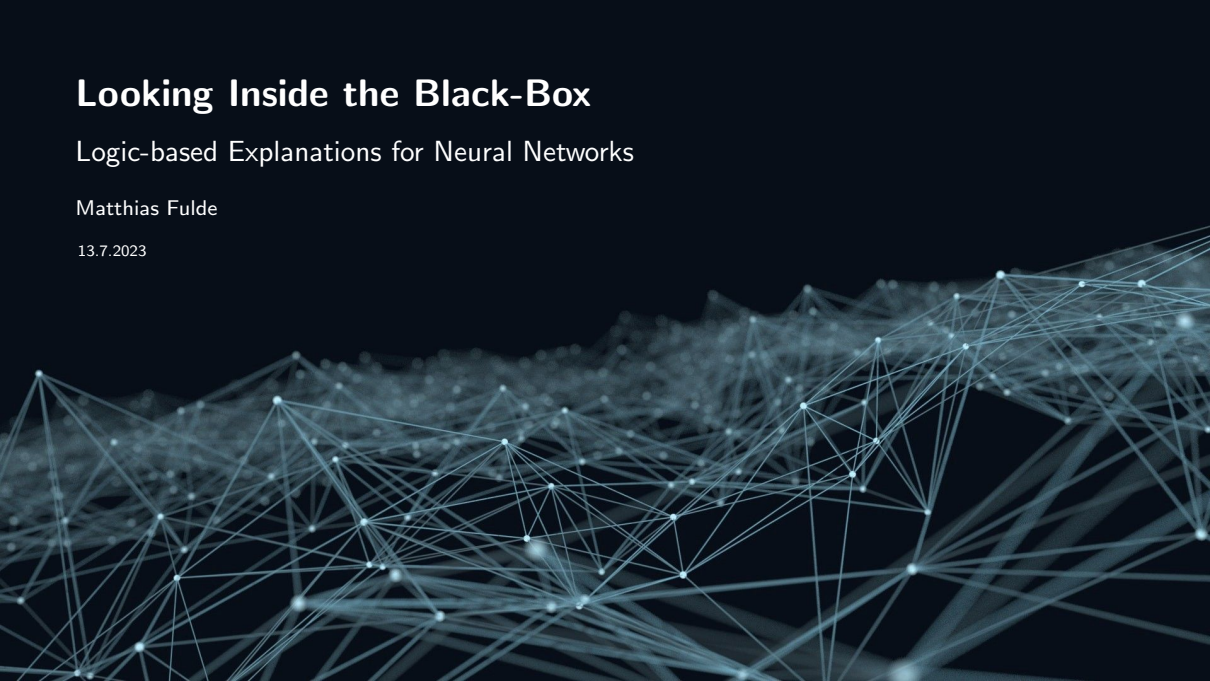
Looking Inside the Black-Box

Logic-based Explanations for Neural Networks

Matthias Fulde

13.7.2023



Original Paper

Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning
Special Session on KR and Machine Learning

Looking Inside the Black-Box: Logic-based Explanations for Neural Networks

João Ferreira , **Manuel de Sousa Ribeiro** , **Ricardo Gonçalves** , **João Leite**
NOVA LINCS, NOVA University Lisbon, Portugal
{jmdi.ferreira, mad.ribeiro}@campus.fct.unl.pt, {rjrg, jleite}@fct.unl.pt

Roadmap

- Introduction
- Methods
- Results
- Discussion
- Further Results

Introduction

Explainable Artificial Intelligence

- For many real world applications of artificial intelligence it is important that humans can understand the reasoning behind the decisions or predictions made by an algorithm

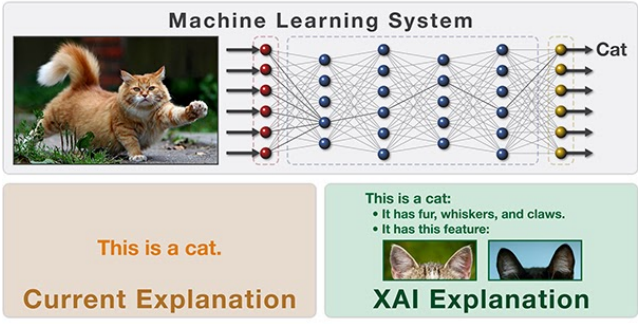


Figure from Explainable Artificial Intelligence (XAI)

Example: Medical Diagnosis

- Machine learning algorithms used for medical diagnoses must be explainable so that doctors and patients trust the predictions

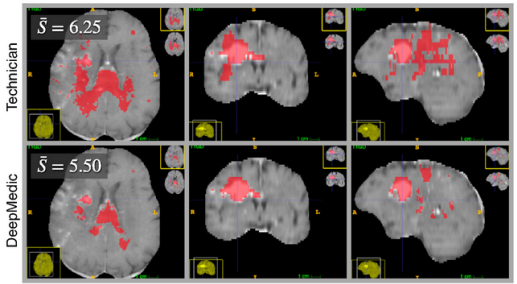


Figure from Deep neural network to locate and segment brain tumors outperformed the expert technicians who created the training data, Mitchell et al., 2020

Example: Autonomous Driving

- Machine learning systems for autonomous driving eventually have to make life and death decisions that should be explainable

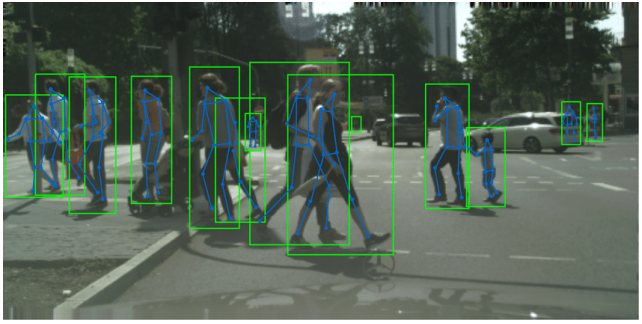


Figure from ClueNet : A Deep Framework for Occluded Pedestrian Pose Estimation, Perla et al., 2019

Problem: Neural Networks are Black-Box Models

- The inner decision processes of artificial neural networks are generally opaque
- Computations in neural networks are entirely numerical, lacking any human-understandable symbolic representations of concepts

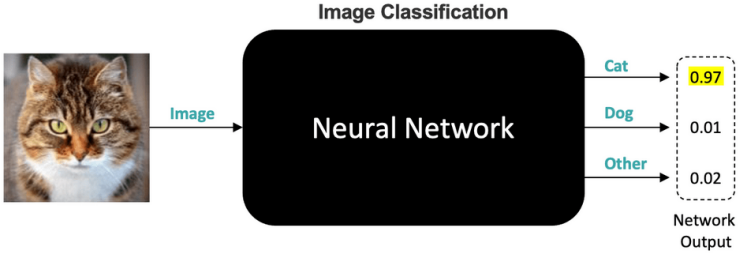


Figure from Getting Startet with Neural Networks

Explaining Neural Networks

- The authors of this paper present a method to obtain human-understandable explanations for the internal decision processes of neural networks
- Given a neural network whose predictions should be explained, the idea is to
  - Identify a set of concepts from the respective domain that are constructive to or related to the predicted concepts of the network, so that they can be used for explanation
  - Label part of the input data for the network with the identified concepts and train a set of small mapping networks to predict the concepts from the neural activations of the main network
  - Use the predicted concepts of the main network and the predicted concepts from the mapping networks to induce a logical theory that explains the network output in terms of the mapped concepts

Methods

Overview

- The method consists of the following parts:
  - A specification of the class of neural networks taken into consideration
  - A description of how to extract neural activations and map them to predefined concepts of interest
  - A description of the logical language which is used to represent the explanations of the decision processes of the network
  - A description of the induction framework that is used to induce the logical theories serving as explanations

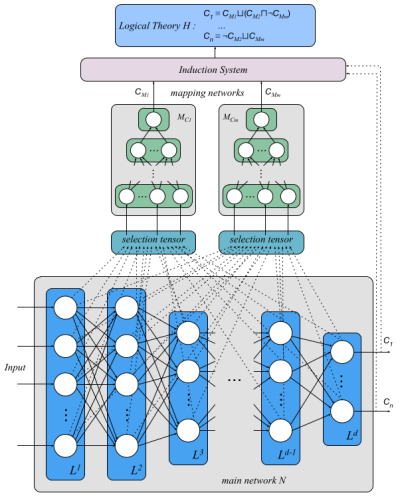


Figure from Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, Ferreira et al., 2022

Feed-forward Neural Networks

- General class of neural networks taken into consideration by the authors are feed-forward networks
- These networks form directed acyclic graphs where information flows only in one direction, from input to output

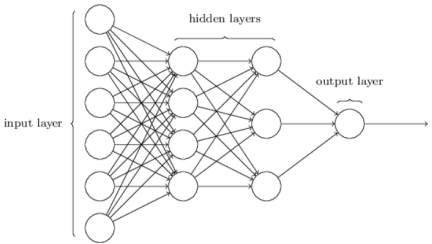


Figure from Neural Networks and Deep Learning, Nielsen, 2015

Layers and Composition

- Artificial neurons in a network N are grouped into layers where neurons within a layer L are not connected to each other

$$N = (L^1, \ldots, L^d)$$

- Each layer L decribes a non-linear mapping between tensor spaces

$$L \colon \mathbb{I}_L \to \mathbb{O}_L$$

- The computation of the network N is the function composition of its layers

$$F_N(\mathbf{X}) = L^d(L^{d-1}(\ldots L^2(L^1(\mathbf{X}))\ldots))$$

Neural Activations

- Elements of a layer's output tensor represent the neurons and the values they take on for some particular input are called the neuron's activations
- Neural activations are computed as sums of input values weighted by learnable parameters, followed by a non-linear function

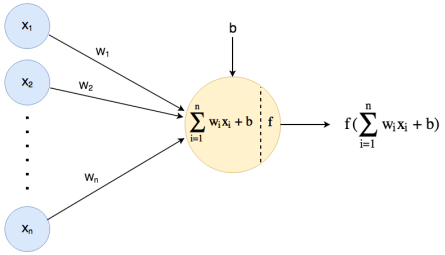


Figure from Activation functions in artificial neural networks

Classification Networks

- Network N whose output can be used to predict whether a particular concept C from a fixed set of predefined concepts $\mathcal{C}_N$ is present in the input
- Allows to define prediction function

$$\text{concept}_N : \mathbb{O}_N \times \mathcal{C}_N \to \{0, 1\}$$

such that

$$\text{concept}_N(F_N(\mathbf{X}), C) = 1$$

indicates that concept C is predicted by network N to be present in the corresponding input

Concepts and Labelling

- Each concept $C \in \mathcal{C}_N$ for which the network N was trained is assumed to be composed of or related to other concepts of interest
- The proposed method requires to identify a set of concepts $\mathcal{C}$ that can be used to explain the concepts $\mathcal{C}_N$
- Domain experts have to annotate part of the training data for network N with concepts $\mathcal{C}$

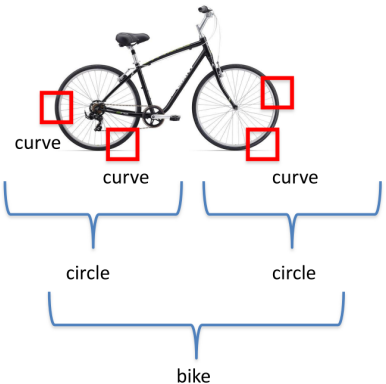


Figure from Deep Learning and Convolutional Neural Networks lecture, Gemma Roig, 2020

Mapping Networks

- For each concept $C \in \mathcal{C}$, a small classification network M is trained
- It predicts from the neural activations of the main network N, whether the concept C was present in the input given to N
- Assuming an activation vector $\mathbf{a} \in \mathbb{R}^k$ as input, we can define again a prediction function

$$\text{concept}_M \colon \mathbb{O}_M \to \{0, 1\}$$

such that

$$\text{concept}_M(F_M(\mathbf{a})) = 1$$

indicates that concept C is predicted by the mapping network M to be present in the input that generated the activations

Activation Selection

- The activations of the network N are given by the output values of its layers

$$(\mathbf{O}^1, \ldots, \mathbf{O}^d) \in \mathbb{O}_{L^1} \times \cdots \times \mathbb{O}_{L^d}$$

- Using all activations as input to mapping networks M is computationally intractable, due to large number of neurons
- Instead define binary selection masks $\mathbf{S}^1, \ldots, \mathbf{S}^d$ with elements in $\{0, 1\}$ and same shape as the output tensors and compute input to M as

$$\mathbf{a} = \mathrm{vec}(\mathbf{S}^1 \odot \mathbf{O}^1, \ldots, \mathbf{S}^d \odot \mathbf{O}^d)$$

where $\odot$ denotes the pointwise product

In practice, only the selected activations are kept and all others are discarded

Logical Language

- A logical language $\mathcal{L}$ is required that allows for the representation of concepts, relations between concepts, and individuals
- The language $\mathcal{L}$ should be defined over concepts $\mathcal{C} \cup \mathcal{C}_N$, and a set of constants

$$c_N = \{c_{\mathbf{X}} | \mathbf{X} \in \mathbb{I}_N\}$$

  representing individual samples from input domain of the network N
- Basic formulas are atoms $C(c)$ with concepts C and individual c, asserting that concept C is present in individual c
- A semantic consequence relation $\models$ over the language $\mathcal{L}$ is assumed

Induction Framework

- An induction framework over $(\mathcal{L}, \models)$ is required for generating the theories explaining the network concepts $\mathcal{C}_N$ in terms of the concepts $\mathcal{C}$
- The framework takes as background knowledge BK a set of atoms defined over the concepts of interest $\mathcal{C}$
- In addition it takes sets of positive and negative examples from the input domain of the network N, defined as

$$\text{Pos} = \bigcup_{C \in \mathcal{C}_N} \{C(c_{\mathbf{X}}) | \mathbf{X} \in \mathbb{I}_N, \text{concept}_N(F_N(\mathbf{X}), C) = 1\}$$

$$\text{Neg} = \bigcup_{C \in \mathcal{C}_N} \{C(c_{\mathbf{X}}) | \mathbf{X} \in \mathbb{I}_N, \text{concept}_N(F_N(\mathbf{X}), C) = 0\}$$

Only those concepts in $\mathcal{C}$ are considered for which the accuracy of the mapping network M is sufficiently high

Inducing Hypotheses

- Task of the induction framework is to induce a hypothesis $H \subset \mathcal{L}$
- A hypothesis is a set of formulas that satisfies

$$BK \cup H \models C(c), \quad \forall C(c) \in \mathrm{Pos}$$

$$BK \cup H \not\models C(c), \quad \forall C(c) \in \mathrm{Neg}$$

- Thus we can express the concepts $C \in \mathcal{C}_N$ in terms of the concepts $C \in \mathcal{C}$ in a way that is consitent with the predictions of the main network N

Results

Synthetic Classification Tasks

- All experiments conducted are classification tasks over a synthetic image dataset
- Overall, the following experimental settings were explored:
  - A proof of concept using a carefully selected set of predefined concepts $\mathcal{C}$ known to be constructive of the classification concepts $\mathcal{C}_N$
  - A test how well the proposed method performs when faced with concepts $\mathcal{C}$ at different levels of abstraction
  - A test how well the proposed method performs when faced with concepts $\mathcal{C}$ that are insufficient to describe the decision processes of the network N
  - A test how the accuracy of the mapping networks M affects the results
  - An ablation study replacing the output of the mapping networks with concept labels from the dataset

Dataset

- For all experiments, the synthetic XTRAINS images dataset was used
- Each sample is a $152 \times 152 \times 3$ color image showing different sketched trains in different positions in front of a colored background

- The dataset is accompanied by an onthology represented using description logics providing definitions such as

$$\text{Train} \equiv \exists \text{has}.(\text{Wagon} \sqcup \text{Locomotive})$$

Figure from Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, Ferreira et al., 2022

Main Networks

- For all but the last experiment, three main networks $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$ were trained for classification
- Networks have different architectures but are all convolutional neural networks with batch normalization, dropout, and pooling layers, where the final layers form a fully-connected network

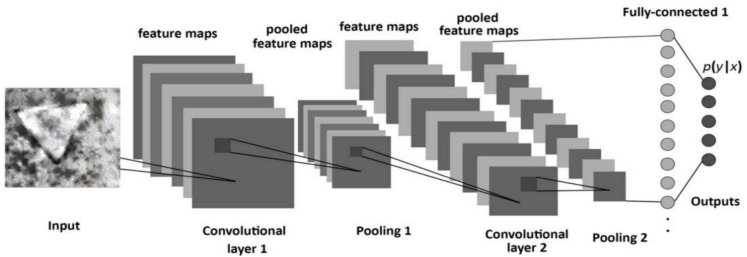


Figure from Does Artificial Intelligence dream of non-terrestrial techno-signatures?, De la Torre, 2019

Tasks

- The networks $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$ were trained to identify a single concept each, describing a particular type of trains
- The respective concepts TypeA, TypeB, and TypeC, are defined in terms of the presence or absence of particular geometric features
- For example TypeA includes 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
- Top level definition in terms of the language describing the dataset's onthology

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

In this case $\mathrm{WarTrain}$ and $\mathrm{EmptyTrain}$ are themselves defined in terms of simpler concepts

Training

- The main networks were trained on a subset of $25\,000$ images and evaluated on a subset of $10\,000$ images from the XTRAINS dataset
- The mapping networks, which consist only of an input and output layer with sigmoid activation, were trained on a balanced set of 800 images and evaluated on a set of 1000 images
- All networks were trained using the Adam optimization algorithm with a learning rate of 0.001 and the binary cross-entropy loss function

$$L = -\frac{1}{N}\sum_{n=1}^{N}\left[t_n \ln(p_n) + (1 - t_n)\ln(1 - p_n)\right]$$

Here, N is the minibatch size, $t_n \in \{0, 1\}$ is the label, and $p_n \in (0, 1)$ is the normalized network prediction

Induction of Theories

- For the background knowledge BK of the induction framework, only concepts were cosidered where the corresponding mapping network achieved an accuracy of at least $\alpha = 90\%$
- As logical language, onthologies over description logics were used to allow for comparsion with the dataset's associated onthology
- As induction system, the DL-Learner framework was used, choosing an algorithm that is biased to minimize the induced theory

Reference DL-Learner: Learning concepts in description logics, Lehmann, 2009

Evaluation Metrics

- For quantitative evaluation of the quality of the induced theories, two fidelity measures F_{Main} and $F_{XTrains}$ have been used
- F_{Main}
  Ratio of samples where the classifications of the main network coincide with those obtained from the induced theory together with the knowledge obtained from the outputs of the mapping networks
- $F_{XTrains}$
  Ratio of samples where the classifications of the XTRAINS labels (for the main network's output concepts $\mathcal{C}_N$) coincide with those obtained from the induced theory together with the knowledge obtained from the labels of the mapped concepts ($\mathcal{C}$)

Quotes from Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, Ferreira et al., 2022

Proof of Concept

- For each main network $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$ mapping networks for the same set of 11 concepts $\mathcal{C}$ were trained
- As qualitative results, the following theories were induced

$$\begin{aligned}
\mathrm{TypeA} &\equiv \mathrm{WarTrain} \sqcup \mathrm{EmptyTrain} \\
\mathrm{TypeB} &\equiv (\mathrm{FrightTrain} \sqcap \mathrm{LongTrain}) \sqcup (\mathrm{PassengerTrain} \sqcap \neg \mathrm{EmptyTrain}) \\
\mathrm{TypeC} &\equiv \mathrm{MixedTrain} \sqcup \mathrm{RuralTrain}
\end{aligned}$$

- The definitions of concepts TypeA and TypeC are equivalent to to those in the dataset's onthology, while the definition for TypeB is a subclass

Proof of Concept

- The experiment was conducted 20 times in total
- As quantitative results, the following averaged fidelity scores for the three main networks $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$ were obtained

	F_{Main}	$F_{XTrains}$
$\mathcal{M}_\mathsf{A}$	$99.72 \pm 0.18\%$	$99.92 \pm 0.35\%$
$\mathcal{M}_\mathsf{B}$	$98.71 \pm 0.31\%$	$99.83 \pm 0.76\%$
$\mathcal{M}_\mathsf{C}$	$99.33 \pm 0.32\%$	$99.52 \pm 1.52\%$

Table from Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, Ferreira et al., 2022

Complete Results

- For a synthetic classification task, the authors showed empirically
  - That human-understandable explanations for the decision processes of neural networks can be obtained by inducing logical theories based on concepts predicted from the activations of a network
  - That the method works for different levels of abstraction in the concepts that the explanations are based on
  - That the quality of the induced theories depends on the degree to which the selected concepts are adequate for explaining the predictions of the network
  - That the accuracy of the mapping networks is positively correlated with the quality of the induced theories
  - That mapping networks processing the activations of the main network are necessary in order to properly reflect the inner decisions processes of the main network

Discussion

Limitations

- While the authors show that the proposed method can work, the results are limited to a synthetic problem
- Since no experiments under realistic assumptions have been conducted, it is not clear how generalizable the method is
- In order to be useful in practice, it has to work with far larger and more diverse datasets and much more complex networks

Accuracy

- While the used dataset has the appearence of naturalistic data, the features relevant for the performed classification task lie on a very low-dimensional latent manifold and are almost discrete
- This makes mapping into a separable space by the network easy and allows for unrealistic high accuracies
- Since accuracy and quality of the results have been shown to be positively correlated, it seems questionable whether satisfying explanations can be obtained for many real-world problems

Figure from Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, Ferreira et al., 2022

Labelling

- The authors claim that only few data is necessary that has to be labelled by domain experts for training the mapping networks with the concepts of interest
- The most powerful models for which interpretability would be desired the most, are trained on such vast amounts of data, though, that attempting to define and annotate with a sufficient number of concepts seems hopeless

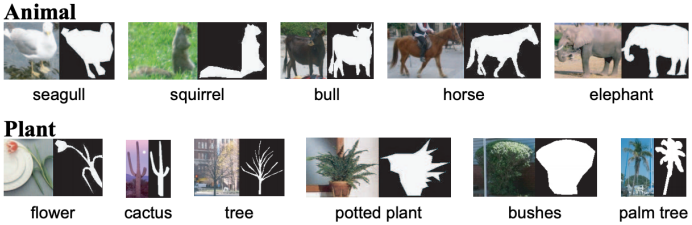


Figure from LabelMe: A database and web-based tool for image annotation, Russel et al., 2008

Activation Selection

- The authors acknowledge that only a subset of neural activations can be used as input to the mapping networks, for computational reasons
- But they don't provide a strategy for selecting the activations for the input to the mapping networks
- However, a poor choice of activations, missing the areas in the network where a concept is encoded, will likely decrease the accuracy of the corresponding mapping network significantly
- This in turn will have the concept removed from the background knowledge and decrease the quality of the induced theories

Future Work

- In order for the proposed method to gain any traction, it is absolutely mandatory that experiments under realistic assumptions are added
- Further analysis of the computational requirements of the method has to be conducted and the aforementioned issues have to be addressed
- As indicated by the authors, exploring the possibility to induce probabilistic theories based on the accuracy of the mapping networks could yield better explanations of the inner workings of neural networks

Thank you for listening!

Further Results

Levels of Abstraction

- Further experiments were performed to asses the quality of the induced theories when high-level or low-level concepts $\mathcal{C}$ were used for explanation
- Train-level concepts:

$$\{\text{EmptyTrain}, \text{LongFreightTrain}, \\ \text{MixedTrain}, \text{PassengerTrain}, \text{RuralTrain}, \text{WarTrain}\}$$

- Wagon-level concepts:

$$\{\exists \text{has}.\text{EmptyWagon}, \exists \text{has}.\text{FreightWagon}, \\ \exists \text{has}.\text{LongWagon}, \exists \text{has}.(\text{LongWagon} \sqcap \text{PassengerCar}), \ldots\}$$

Levels of Abstraction

- Qualitatively, the induced theories from the train-level concepts were logically equivalent to the definitions in the dataset's onthology
- Regarding the wagon-level concepts, the induced definition for TypeB was a subclass of the onthology's definition, while the other definitions were neither subclasses nor superclasses to the corresponding definitions in the dataset's onthology

Levels of Abstraction

- The experiment was again conducted 20 times in total
- As quantitative results, the following averaged fidelity scores for the three main networks $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$ were obtained

		F_{Main}	$F_{XTrains}$
Train-level	$\mathcal{M}_\mathsf{A}$	$99.76 \pm 0.19\%$	$100.00 \pm 0.0\%$
	$\mathcal{M}_\mathsf{B}$	$98.82 \pm 0.39\%$	$100.00 \pm 0.0\%$
	$\mathcal{M}_\mathsf{C}$	$99.46 \pm 0.20\%$	$100.00 \pm 0.0\%$
Wagon-level	$\mathcal{M}_\mathsf{A}$	$94.55 \pm 10.14\%$	$94.44 \pm 11.12\%$
	$\mathcal{M}_\mathsf{B}$	$97.50 \pm 0.52\%$	$96.73 \pm 1.18\%$
	$\mathcal{M}_\mathsf{C}$	$98.16 \pm 0.36\%$	$99.02 \pm 0.56\%$

Table from Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, Ferreira et al., 2022

Insufficient Concepts

- Choosing concepts $\mathcal{C}$ that are insufficient to describe the concepts $\mathcal{C}_N$ of the main networks should result in lower quality theories
- Otherwise the previously obtained results could be attributed to spurious correlations in the data rather than accurate descriptions of the main network's decision process in terms of the concepts $\mathcal{C}$
- For this test, 20 random sets of 5 concepts among all concepts defined in the XTRAINS ontology were used to define $\mathcal{C}$
- As expected by the authors, the average fidelity scores dropped significantly to

$$F_{Main} = 72.6\% \qquad F_{XTrains} = 71.9\%$$

Insufficient Concepts

- The experiment was again conducted 20 times in total
- As quantitative results, the following averaged fidelity scores for the three main networks $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$ were obtained

	F_{Main}	$F_{XTrains}$
$\mathcal{M}_\mathsf{A}$	$50.16 \pm 0.26\%$	$50.00 \pm 0.00\%$
$\mathcal{M}_\mathsf{B}$	$92.53 \pm 2.75\%$	$92.70 \pm 1.43\%$
$\mathcal{M}_\mathsf{C}$	$76.78 \pm 1.99\%$	$76.99 \pm 0.94\%$

Table from Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, Ferreira et al., 2022

Accuracy of the Mapping Networks

- Another experiment was performed to test whether the quality of the resulting theories mostly depends on the accuracy of the mapping networks
- The same set of 11 concepts as in the initial experiment was used to induce theories while varying the amount of data used to train the mapping networks between 50 and 1200 samples
- The amount of data used to induce the theories remained constant at 3000 samples

Accuracy of the Mapping Networks

- Again a subset $\mathcal{C}_\alpha \subset \mathcal{C}$ of concepts was used for which the respective mapping networks achieved an accuracy of at least $\alpha = 90\%$
- A Pearson's correlation test on fidelity F_{Main} and the average accuracy of the mapping networks yielded a strong positive correlation of $r = 0.8161$ with a p-value of $p < 0.0001$
- This indicates that when the mapping networks' accuracy increases, the quality of the induced theories increases as well

Ablation Study

- In order to assess whether the mapping networks are necessary to induce theories explaining the actual decision processes of a main network, an ablation study was conducted
- The mapping networks were removed from the procedure and instead of their predicted labels, fixed labels from the dataset were used for inducing logical theories
- For this experiment, the XTRAINS dataset was augmented and a new task was constructed

Dataset Augmentation

- The XTRAINS images were augmented to include four traffic signals in varying top left, top right, bottom left, and bottom right positions in the image

- A traffic signal is said to be *on* if it contains any symbol in it
- The images in the dataset are labelled with concepts

$$\{\text{On, TopLeftOn, TopRightOn, BottomLeftOn, BottomRightOn}\}$$

Figure from Looking Inside the Black-Box: Logic-based Explanations for Neural Networks, Ferreira et al., 2022

Classification Task

- The dataset is defined such that if one of the top signals is *on*, then also one of the bottom signals, and vice versa
- A main network is tasked to predict the concept On, which is present in an image, if any of the signals is on
- By construction of the dataset, for predicting the concept, it is sufficient for a network to either look at the top or bottom signals
- For this experiment, 50 main networks with an accuracy of at least 90% were trained

Comparsion

- When using the dataset labels for inducing the theories over the main networks predictions, always the same theory was obtained

$$\mathrm{On} = \mathrm{BottomLeftOn} \sqcup \mathrm{BottomRightOn}$$

- Conducting the same experiment with labels from trained mapping networks, the obtained theories were much more diverse
- 22% of the main networks learned to classify their outputs just by considering whether the two top signals were on while 42% looked at the two bottom signals
- This suggests that the mapping networks are necessary in order to reflect the actual decision processes of the main network

For label based theory, the choice is due to the determinsim of the DL-Learner framework

Thank you for listening!