TECHNISCHE UNIVERSITÄT DARMSTADT

DeepRED – Rule Extraction from Deep Neural Networks

Jan Ruben Zilke, *Eneldo Loza Mencía*, Frederik Janssen
Knowledge Engineering Group, TU Darmstadt
j.zilke@mail.de
{eneldo,janssen}@ke.tu-darmstadt.de

Outline

Comprehension and Extraction from Neural Networks

DeepRED: Rule extraction from Deep Neural Networks

Experimental Results

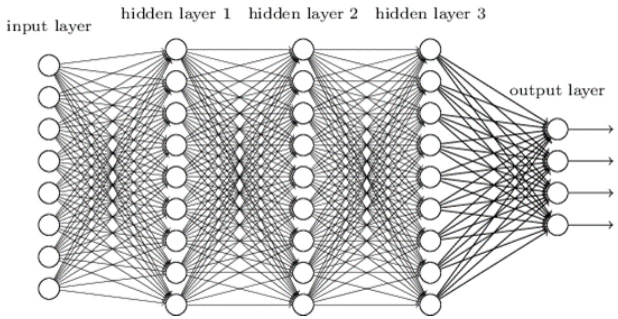
Conclusions

Comprehending Neural Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

NNs are widely used for classification

- current hype about Deep Neural Networks (DNN)
- outperform previous state-of-the-art approaches in many domains
- DNNs might represent complex, abstract concepts in hidden nodes



Understanding how a NN comes to its decision is not trivial

- we only know the network’s structure and its weights
- predictive model: usually NNs seen and used as a black box
- learned higher level concepts remain hidden
  - exception: visual domain

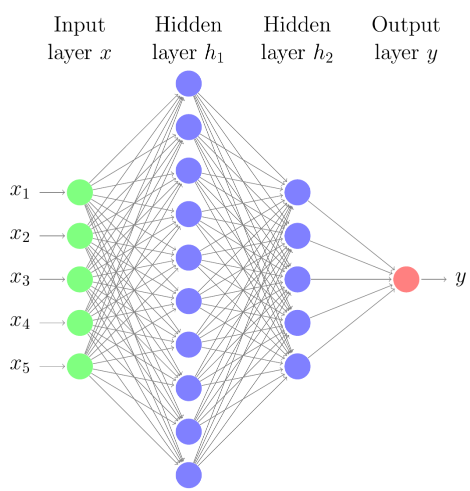
Comprehensible Decision Systems

TECHNISCHE UNIVERSITÄT DARMSTADT

Comprehensible description of a NN's behaviour sometimes essential

- safety critial domains, e.g. medicine, power stations, autonomous driving, financial markets

Solution: → **represent NN's behaviour as decision rules**

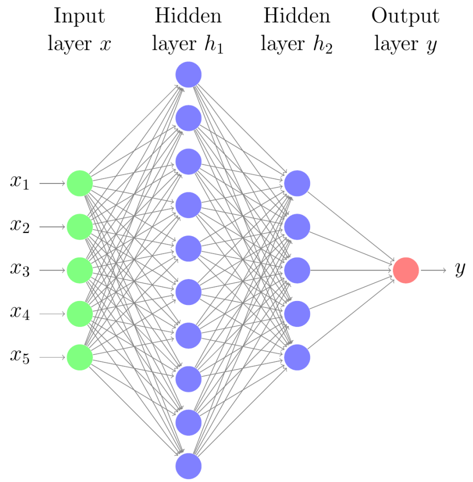



```
IF X1<0.5 AND X2>0.75 THEN OUT=1
IF X1>0.9 THEN OUT=1
IF X1>0.5 AND X1<0.9 AND X3>0.2 THEN OUT=1
IF X2>0.2 AND X3<0.5 AND X5<0.5 THEN OUT=1
IF X2>0.4 AND X3<0.7 THEN OUT=1
IF X2<0.2 THEN OUT=1
IF X4>0.8 THEN OUT=1
IF X3<0.7 AND X3>0.2 AND X4<0.3 THEN OUT=1
```

Comprehensible Decision Systems

Rules are considered to be *comprehensible* and *interpretable*

- symbolic rule model can be inspected
  - discover relations between inputs and target concept
  - experts can check critical rules, e.g.: IF … THEN *emergency braking*
- taken decisions can be explained by firing rules
  - firing rule reveals decisive attributes and the training examples from which the rule was learned

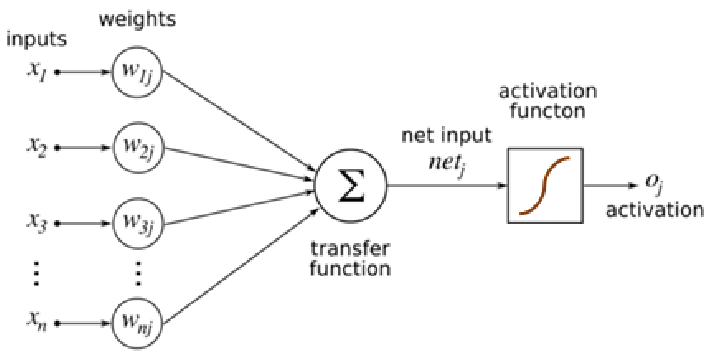



```
IF X1<0.5 AND X2>0.75 THEN OUT=1
IF X1>0.9 THEN OUT=1
IF X1>0.5 AND X1<0.9 AND X3>0.2 THEN OUT=1
IF X2>0.2 AND X3<0.5 AND X5<0.5 THEN OUT=1
IF X2>0.4 AND X3<0.7 THEN OUT=1
IF X2<0.2 THEN OUT=1
IF X4>0.8 THEN OUT=1
IF X3<0.7 AND X3>0.2 AND X4<0.3 THEN OUT=1
```

Extracting Rules from Neural Networks

Rule extraction strategies

- Decompositional (considering NN's structure)



IF X1=hi OR X2=hi OR X3=hi THEN OUT=hi

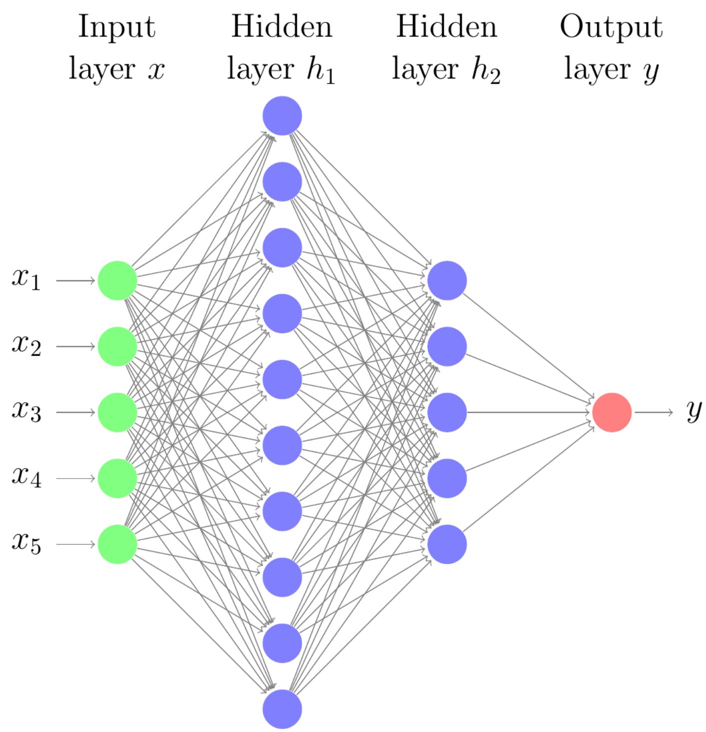
Extracting Rules from Neural Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

Rule extraction strategies

- Decompositional (considering NN's structure)
- Pedagogical (NN as black box)

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



o
1
0
1
0
1
1
1
…

Extracting Rules from Neural Networks

Rule extraction strategies

- Decompositional (considering NN's structure)
- Pedagogical (NN as black box)

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

```
IF X1<0.5 AND X2>0.75 THEN OUT=1
IF X1>0.9 THEN OUT=1
IF X1>0.5 AND X1<0.9 AND X3>0.2 THEN OUT=1
IF X2>0.2 AND X3<0.5 AND X5<0.5 THEN OUT=1
IF X2>0.4 AND X3<0.7 THEN OUT=1
IF X2<0.2 THEN OUT=1
IF X4>0.8 THEN OUT=1
IF X3<0.7 AND X3>0.2 AND X4<0.3 THEN OUT=1
```

o
1
0
1
0
1
1
1
…

Extracting Rules from Neural Networks

Rule extraction strategies

- Decompositional (considering NN's structure)
- Pedagogical (NN as black box)
- Eclectic (mixture of both)

Models

- previous research in the 90s focussed on extracting rules from flat NNs
- types of extracted rules (DNFs, decision tree, fuzzy rules, ...)

DeepRED: Extraction of Rules from Deep Neural Networks

Goals

- make hidden features accessible (in contrast to pedagogical)
- exploit deep structure to improve efficacy of rule extraction and induction process

Based on CRED

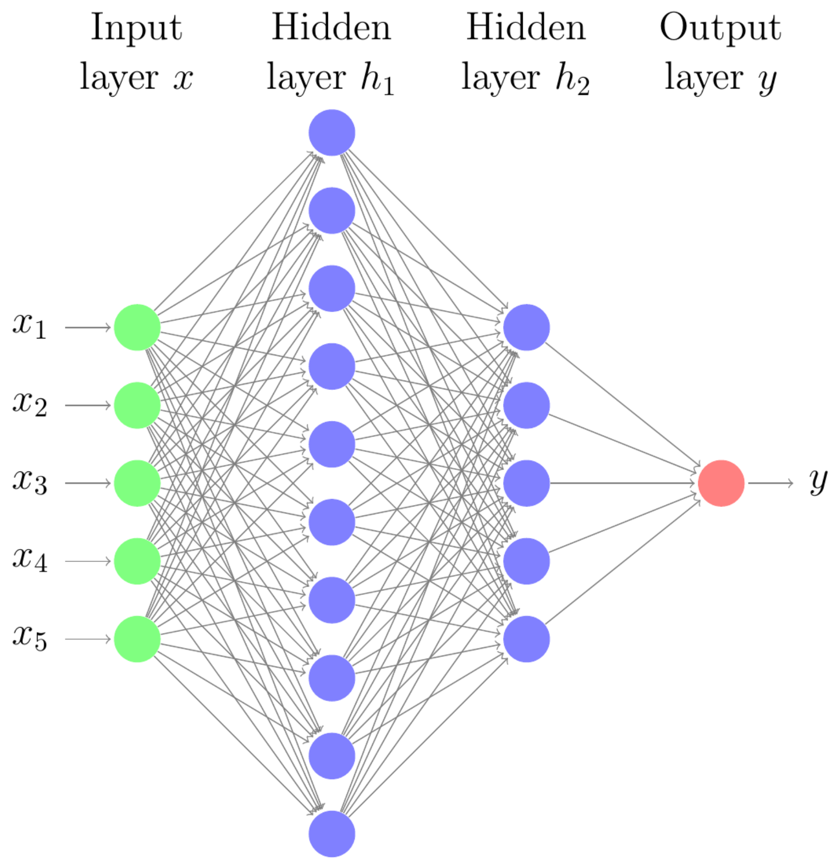
- *Continuous/discrete Rule Extractor via Decision tree induction* (CRED) [Sato and Tsukimoto, 2001]
- only supports NNs with one hidden layer
- uses C4.5 to induce rules

DeepRED extends CRED to arbitrary number of layers

- roughly speaking: apply CRED layer by layer
- decomposible w.r.t. neurons, pedagogical w.r.t. neurons' behaviour

Pedagogical Baseline

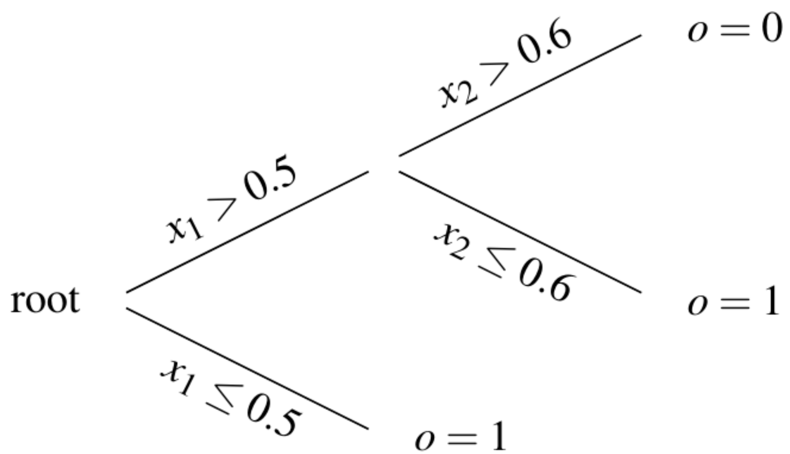
x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



o
1
0
1
0
1
1
1
…

Pedagogical Baseline

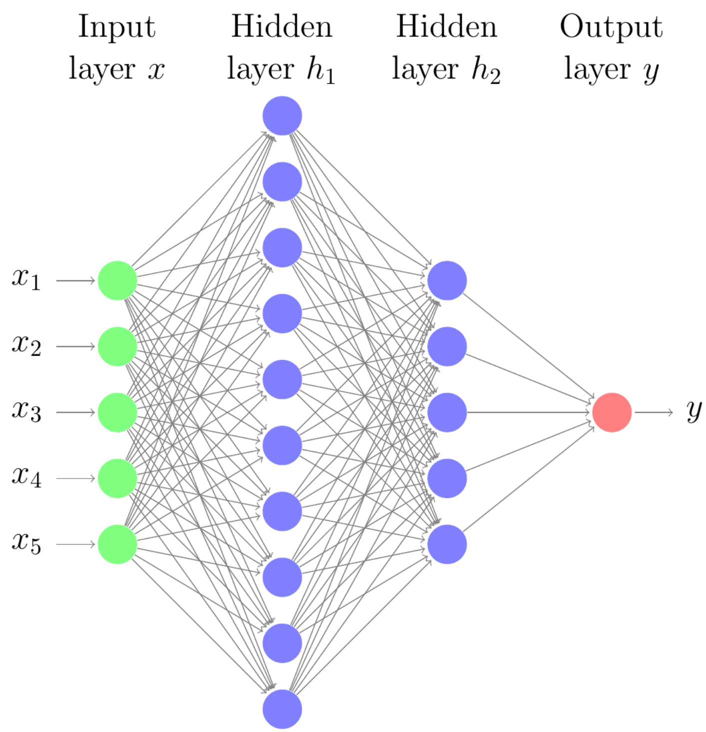
x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



o
1
0
1
0
1
1
1
…

Extracting Rules from Neural Networks

$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	$\mathbf{x_5}$
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



o
1
0
1
0
1
1
1
…

Extracting Rules from Neural Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

Goals of extracting rules from (deep) neural networks

- make hidden logic and features accessible
- exploit deep structure to improve efficacy of rule extraction and induction process

Solution by DeepRED: → **Mimic internal logic of NN at each layer and neuron**

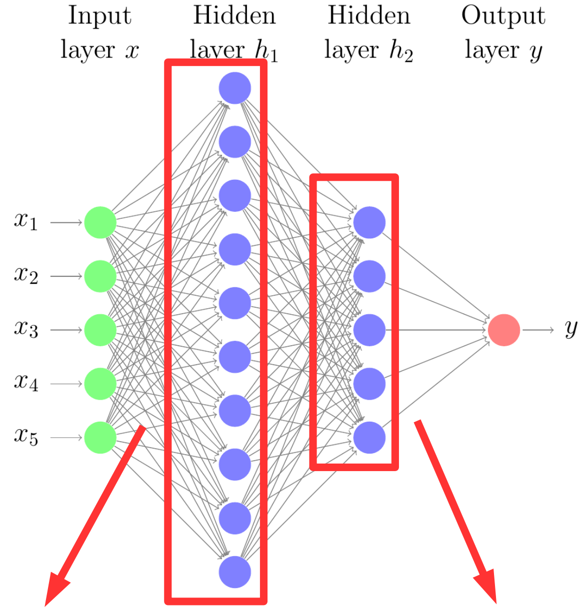
x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

```
IF X1<0.5 AND X2>0.75 THEN OUT=1
IF X1>0.9 THEN OUT=1
IF X1>0.5 AND X1<0.9 AND X3>0.2 THEN OUT=1
IF X2>0.2 AND X3<0.5 AND X5<0.5 THEN OUT=1
IF X2>0.4 AND X3<0.7 THEN OUT=1
IF X2<0.2 THEN OUT=1
IF X4>0.8 THEN OUT=1
IF X3<0.7 AND X3>0.2 AND X4<0.3 THEN OUT=1
```

o
1
0
1
0
1
1
1
…

DeepRED: Extraction of Rules from Deep Neural Networks

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



o
1
0
1
0
1
1
1
…

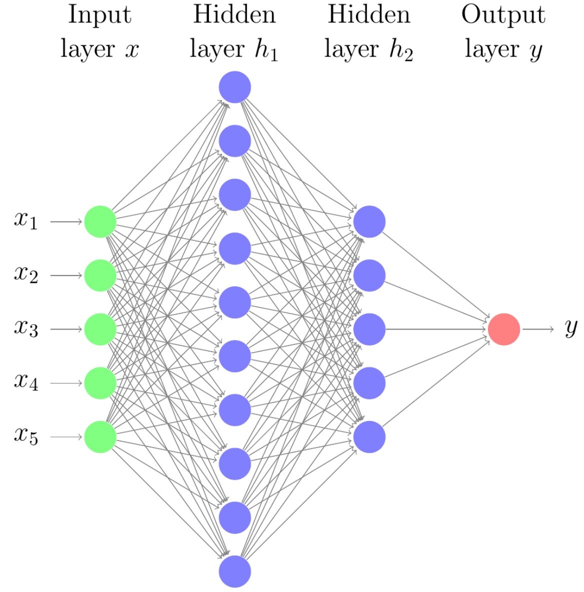
$h_{1,1}$	$h_{1,2}$	…	$h_{1,10}$
0.865	0.079	…	0.818
0.050	0.675	…	0.613
0.767	0.485	…	0.020
0.388	0.160	…	0.491
0.555	0.767	…	0.606
0.312	0.231	…	0.376
0.770	0.211	…	0.805
…	…	…	…

$h_{2,1}$	$h_{2,2}$	…	$h_{2,5}$
0.034	0.635	…	0.928
0.089	0.049	…	0.435
0.057	0.369	…	0.233
0.346	0.462	…	0.181
0.834	0.945	…	0.354
0.443	0.644	…	0.892
0.778	0.691	…	0.708
…	…	…	…

DeepRED: Extraction of Rules from Deep Neural Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



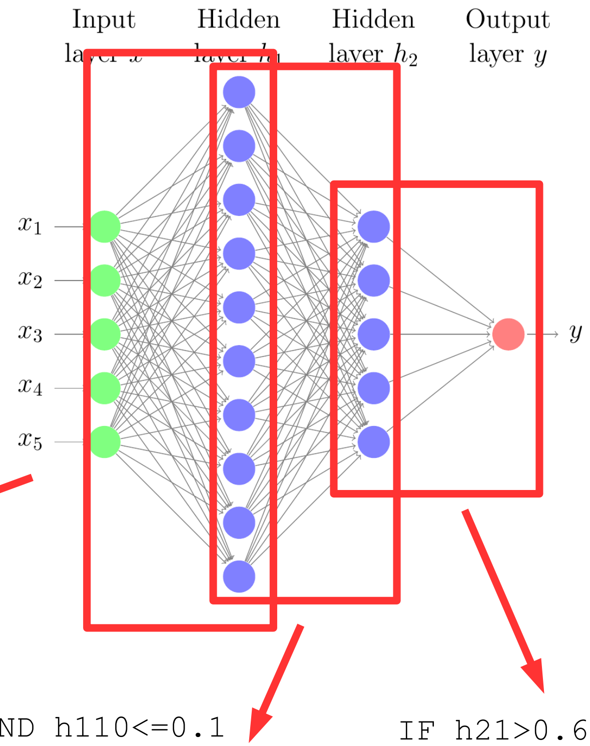
o
1
0
1
0
1
1
1
…

$h_{1,1}$	$h_{1,2}$	…	$h_{1,10}$
0.865	0.079	…	0.818
0.050	0.675	…	0.613
0.767	0.485	…	0.020
0.388	0.160	…	0.491
0.555	0.767	…	0.606
0.312	0.231	…	0.376
0.770	0.211	…	0.805
…	…	…	…

$h_{2,1}$	$h_{2,2}$	…	$h_{2,5}$
0.034	0.635	…	0.928
0.089	0.049	…	0.435
0.057	0.369	…	0.233
0.346	0.462	…	0.181
0.834	0.945	…	0.354
0.443	0.644	…	0.892
0.778	0.691	…	0.708
…	…	…	…

DeepRED: Extraction of Rules from Deep Neural Networks

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



o
1
0
1
0
1
1
1
…

```
IF x1>0.5  AND x2>0.6
    THEN h11<=0.4
IF x1>0.5  AND x2<=0.6
    THEN h11>0.4
IF x1<=0.5  …
...
```

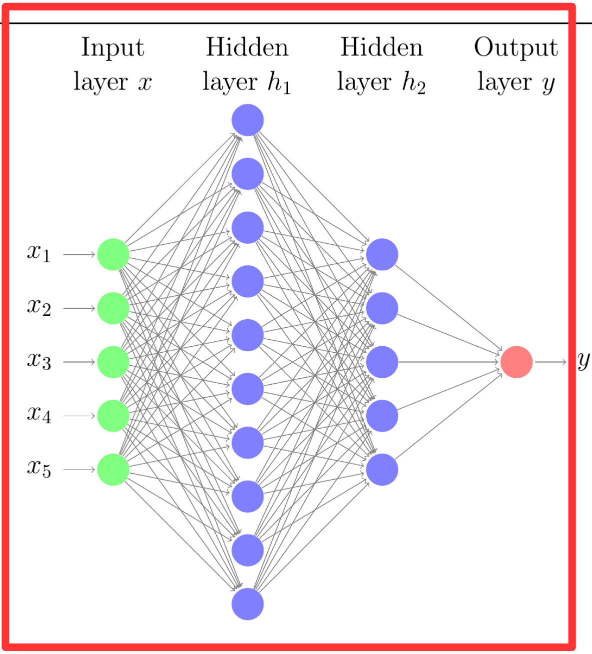
```
IF h12>0.4 AND h110<=0.1
   THEN h23<=0.5
IF h12>0.4 AND h110>0.1
   THEN h24>0.3
IF h12<=0.4 AND h11<=0.4
   THEN h21>0.6
IF h12<=0.4 AND h11 >0.1
   THEN h21<=0.6
```

```
IF h21>0.6 AND h24>0.3
    THEN o=0
IF h21>0.6 AND h24<=0.3
    THEN o=1
IF h21<=0.6
    THEN o=1
```

DeepRED: Extraction of Rules from Deep Neural Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

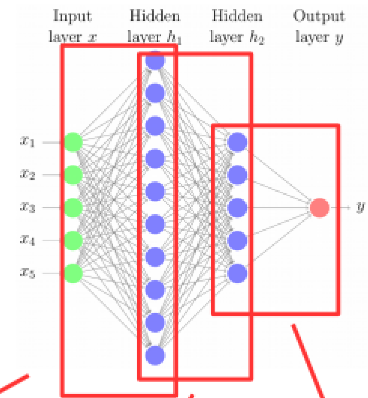


o
1
0
1
0
1
1
1
…

```
IF x1<0.5 AND x2>0.75 THEN o=1
IF x1>0.9 THEN o=1
IF x1>0.5 AND x1<0.9 AND x3>0.2 THEN o=1
IF x2>0.2 AND x3<0.5 AND x5<0.5 THEN o=1
IF x2>0.4 AND x3<0.7 THEN o=1
IF x2<0.2 THEN o=1
IF x4>0.8 THEN o=1
IF x3<0.7 AND x3>0.2 AND x4<0.3 THEN o=1
```

TECHNISCHE UNIVERSITÄT DARMSTADT

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



o
1
0
1
0
1
1
1
…

Mimicking
+ Sparseness Pruning
+ Activation Polarization

```
IF x1>0.5  AND x2>0.6
    THEN h11<=0.4
IF x1>0.5  AND x2<=0.6
    THEN h11>0.4
IF x1<=0.5  …
...
```

```
IF h12>0.4 AND h110<=0.1
   THEN h23<=0.5
IF h12>0.4 AND h110>0.1
   THEN h24>0.3
IF h12<=0.4 AND h11<=0.4
   THEN h21>0.6
IF h12<=0.4 AND h11 >0.1
   THEN h21<=0.6
```

```
IF h21>0.6 AND h24>0.3
    THEN o=0
IF h21>0.6 AND h24<=0.3
    THEN o=1
IF h21<=0.6
    THEN o=1
```

Substitution
+ Simplification

```
IF x1<0.5 AND x2>0.75 THEN o=1
IF x1>0.9 THEN o=1
IF x1>0.5 AND x1<0.9 AND x3>0.2 THEN o=1
IF x2>0.2 AND x3<0.5 AND x5<0.5 THEN o=1
IF x2>0.4 AND x3<0.7 THEN o=1
IF x2<0.2 THEN o=1
IF x4>0.8 THEN o=1
IF x3<0.7 AND x3>0.2 AND x4<0.3 THEN o=1
```

Experimental setup

TECHNISCHE UNIVERSITÄT DARMSTADT

Datasets and DNNs used

	#attributes	#training ex.	#test ex.	NN structure	acc(training)	acc(test)
MNIST	784	12056	2195	784-10-5-2	99.6%	98.8%
letter	16	1239	438	16-40-30-26	96.9%	97.3%
artif-I	5	20000	10000	5-10-5-2	99.5%	99.4%
artif-II	5	3348	1652	5-10-5-2	99.4%	99.0%
XOR	8	150	106	8-8-4-4-2-2-2	100%	100%

Evaluation measures

- fidelity on test set: accuracy on mimicking NN's behaviour
- number of terms: tries to assess comprehensibility of found rule set

Algorithm setup

- 36 combinations of varying C4.5 parameters, pruning parameters and train set sizes

Can DeepRED make use of complex concepts hidden in NNs?

TECHNISCHE UNIVERSITÄT DARMSTADT

artif-I

- artificial dataset randomly drawn
- output defined by rule set which cannot easily be realized by decision trees
  - contains pairwise comparisons between inputs

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

```
IF x1 = x2 THEN out=1
IF x1 > x2 AND x3 > 0.4 THEN out=1
IF x3 > x4 AND x4 > x5 AND x2 > 0 THEN out=1
ELSE out=0
```

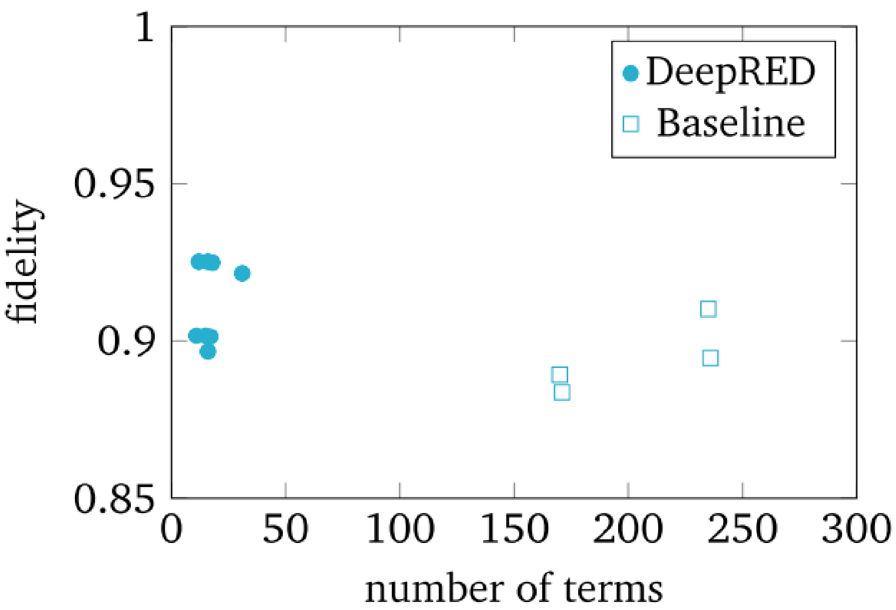
o
1
0
1
0
1
1
1
…

Can DeepRED make use of complex concepts hidden in NNs?

TECHNISCHE UNIVERSITÄT DARMSTADT

artif-I

- artificial dataset randomly drawn
- output defined by rule set which cannot easily be realized by decision trees
  - contains pairwise comparisons between inputs



Results

- DeepRED outperforms pedagogical baseline
  - especially in comprehensibility dimension
- hidden concepts lead to compactne

Can DeepRED make use of complex concepts hidden in NNs?

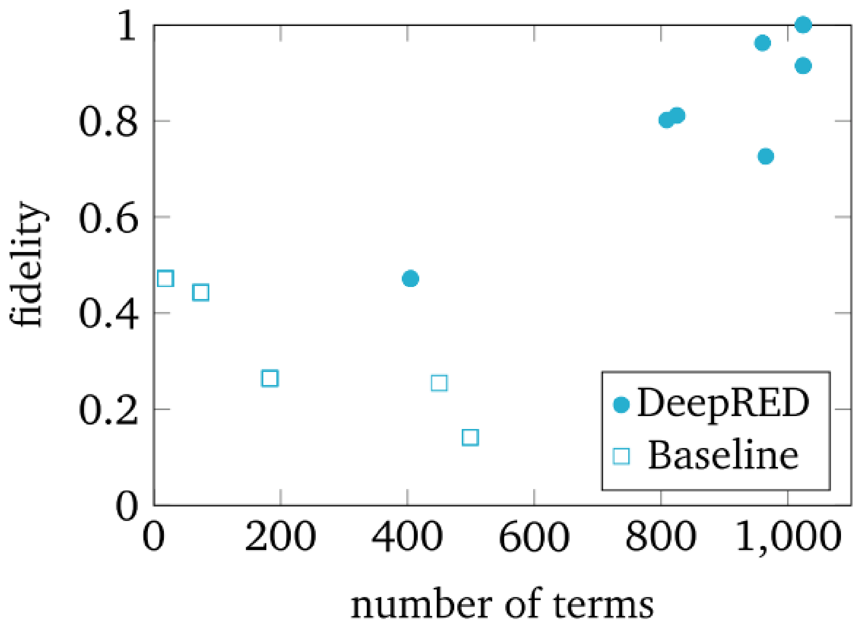
TECHNISCHE UNIVERSITÄT DARMSTADT

XOR

- parity function: $x \in \{0,1\}^8 \rightarrow \mathrm{XOR}(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$
- 2^8 examples split into 150 training and 106 test examples
- top-down approaches (e.g. C4.5) usually need all examples to learn consistent model

Results

- as expected, baseline fails
- DeepRED is able to extract rules that classify all or almost all test examples correctly



Can DeepRED make use of complex concepts hidden in NNs?

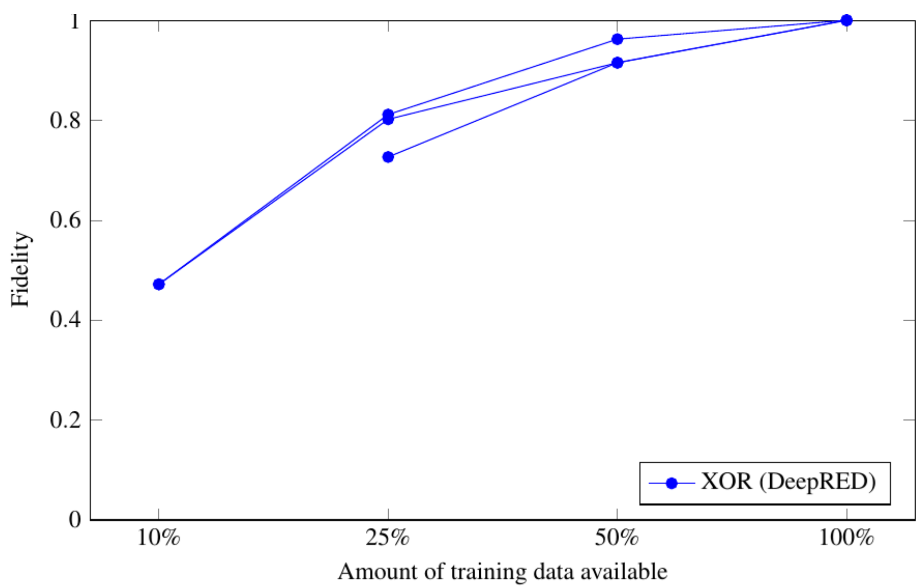
TECHNISCHE UNIVERSITÄT DARMSTADT

XOR

- parity function: $x \in \{0,1\}^8 \rightarrow XOR(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$
- 2^8 examples split into 150 training and 106 test examples
- top-down approaches (e.g. C4.5) usually need all examples to learn consistent model

Results

- even with only 75 training examples DeepRED extracts meaningful rules (>90% fidelity)
- DeepRED effectively captures inherent concepts otherwise non accessible



More insights

TECHNISCHE UNIVERSITÄT DARMSTADT

Limitations

- artif-II
  - *can* easily be realized by decision tree
  - baseline finds more comprehensible rules with very good fidelity

Pruning

- removal of up to 10% inputs possible without substantial decrease in fidelity
- but reduction in number of conditions of several magnitudes

Training set size

- DeepRED quite stable w.r.t. reduction of training set

Conclusions

DeepRED

- to our knowledge, first attempt on extracting rules form deep neural networks
  - important step towards making NN's decisions transparent
- outperforms pedagogical baselines for most of the analyzed cases
- DeepRED benefits from deep architecture of NNs when addressing data with complex concepts

Questions?

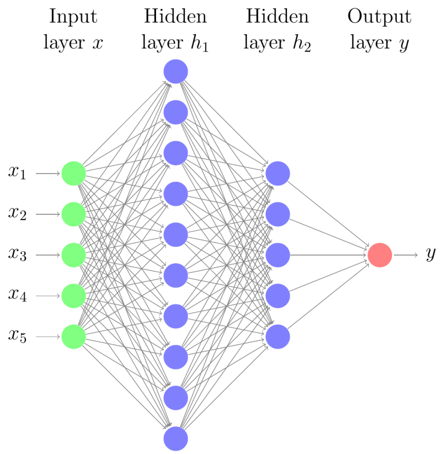
TECHNISCHE UNIVERSITÄT DARMSTADT

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

$h_{1,1}$	$h_{1,2}$	…	$h_{1,10}$
0.865	0.079	…	0.818
0.050	0.675	…	0.613
0.767	0.485	…	0.020
0.388	0.160	…	0.491
0.555	0.767	…	0.606
0.312	0.231	…	0.376
0.770	0.211	…	0.805
…	…	…	…

$h_{2,1}$	$h_{2,2}$	…	$h_{2,5}$
0.034	0.635	…	0.928
0.089	0.049	…	0.435
0.057	0.369	…	0.233
0.346	0.462	…	0.181
0.834	0.945	…	0.354
0.443	0.644	…	0.892
0.778	0.691	…	0.708
…	…	…	…

o
1
0
1
0
1
1
1
…




```
IF x1 = x2 THEN out=1
IF x1 > x2 AND x3 > 0.4 THEN out=1
IF x3 > x4 AND x4 > x5 AND x2 > 0 THEN out=1
IF x4=look OR x4=see THEN out=1
ELSE out=0
```