Tutorial on Explainable AI – Part I

Pascal Hitzler
Data Semantics Laboratory (DaSe Lab)
Data Science and Security Cluster (DSSC)
Wright State University
http://www.pascal-hitzler.de

This tutorial

- Derek Doran – got a project meeting scheduled which he couldn't miss.
- Ning Xie – got an industry internship
- Freddy Lecue – got a new job
- Md Kamruzzaman Sarker – got an industry internship (but managed to arrange to come here anyway).

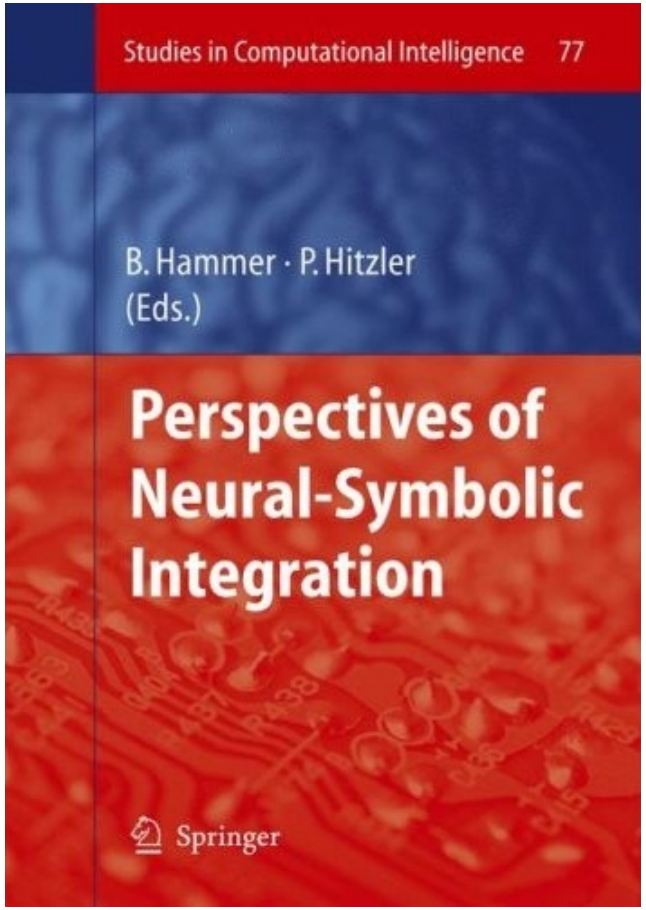
- I'm the substitute, I'll do my very best.

  - Part I (Pascal Hitzler): On Neural-symbolic Integration and some of our own ongoing work on Explainable AI and Semantic Web.
  - Part II (Md Kamruzzaman Sarker): Other work on explainable AI and how Semantic Web could improve the state of the art.

Some Background

Workshop Series on Neural-Symbolic Learning and Reasoning Since 2005.
http://neural-symbolic.org/

Perspectives on Neural-Symbolic Integration
Barbara Hammer and Pascal Hitzler (eds)
Springer, 2007



Neural-Symbolic Learning and Reasoning: A Survey and Interpretation

Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kuehnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, Gerson Zaverucha

https://arxiv.org/abs/1711.03902 (2017)

Data Semantics Laboratory

DaSe Lab

Monireh Ebrahimi **Md Kamruzzaman Sarker** **Federico Bianchi (not shown)**

WRIGHT STATE UNIVERSITY

Neural-Symbolic? Symbolic-Subsymbolic?

Neural

- Refers to computational abstractions of (natural) neural network systems.
- Prominently includes Artificial Neural Networks and Deep Learning as machine learning paradigms.
- More generally sometimes referred to as *connectionist systems*.

- Prominent applications come from the machine learning world.

- And of course, there is the current deep learning hype.

Symbolic

- Refers to (computational) symbol manipulations of all kind.

- Graphs and trees, traversal, data structure operations.
- Knowledge representation in explicit symbolic form (data base, ontology, knowledge graph)
- Inductive and statistical inference.
- Formal logical (deductive or abductive) reasoning.

- Prominent applications all over computer science, including expert systems (and their modern versions), information systems, data management, added value of data annotation, etc.

- Semantic Web data is inherently symbolic.

Neural-Symbolic

Computer Science perspective:

- **Connectionist machine learning systems are**
  - **very powerful for some machine learning problems**
  - **robust to data noise**
  - **very hard to understand or explain**
  - **really poor at symbol manipulation**
  - **unclear how to effectively use background (domain) knowledge**
- **Symbolic systems are**
  - **Usually rather poor regarding machine learning problems**
  - **Intolerant to data noise**
  - **Relatively easy to analyse and understand**
  - **Really good at symbol manipulation**
  - **Designed to work with other (background) knowledge**

Neural-Symbolic

Computer Science perspective:

- **Let's try to get the best of both worlds:**
  - **very powerful machine learning paradigm**
  - **robust to data noise**
  - **easy to understand and assess by humans**
  - **good at symbol manipulation**
  - **work seamlessly with background (domain) knowledge**

- **How to do that?**
  - **Endow connectionist systems with symbolic components?**
  - **Add connectionist learning to symbolic reasoners?**

Neural-Symbolic

DaSe Lab

SSC

Note:

- **Deep Learning systems are a far cry from how natural neural networks work.**
- **There are things that our brain can do, and things that symbolic approaches can do, where we do not have the faintest idea how to solve them through deep learning (or any other connectionist learning approach).**

- **The argument that we "just don't have enough training data" speaks (understandably) to the current hype, but is a hope that is unfounded: While this may be the case in some cases, there is no rationale to overgeneralize.**
  [Note: if we had "enough computational power," we could also solve all reasonable-size NP-complete problems in an instant.]

The Interface Issue

- Symbolic knowledge comes as logical theories (sets of formulas over a logic)
- Subsymbolic systems process tuples of real/float numbers (vectors, matrices, tensors)

- How do you interface?
- How do you map between the symbolic world and the subsymbolic world?

Some key problems that need to be overcome:

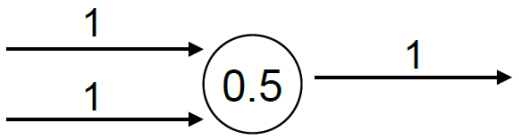
- Logic is full of highly structured objects, how to represent them in Real Space?
- How to represent variable bindings in a distributed setting?
- The required length of logical deduction chain is not known up front.

Representation Issues

McCulloch & Pitts, 1943

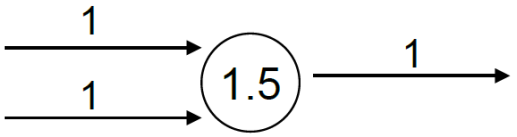
- McCulloch & Pitts 1943
  - Neurons with binary activation functions.
  - Modelling of propositional connectives.
  - Networks equivalent to finite automata.

Values 0 („false“) and 1 („true“) being propagated.



disjunction

Simultaneous update of all nodes in network.



conjunction

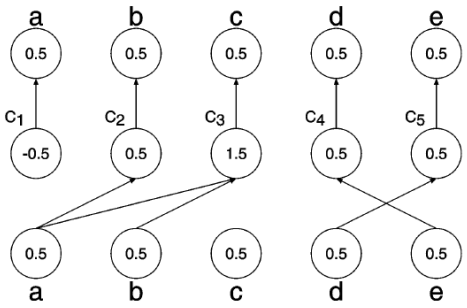
negation

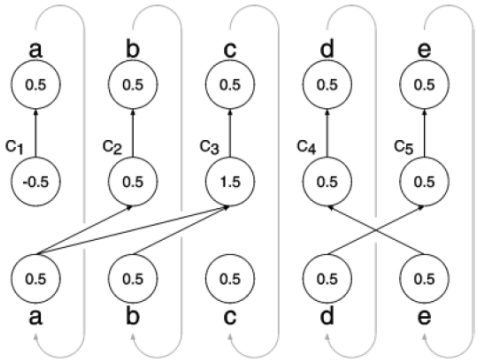
McCulloch & Pitts follow-on

- Hölldobler & Kalinke 1994
  - Extends the approach by McCulloch & Pitts.
  - Representation of propositional logic programs and their semantics.
  - „Massively parallel reasoning.“

logic program → core net → recurrent net

$a \leftarrow$
$b \leftarrow a$
$c \leftarrow a \wedge b$
$d \leftarrow e$
$e \leftarrow d$

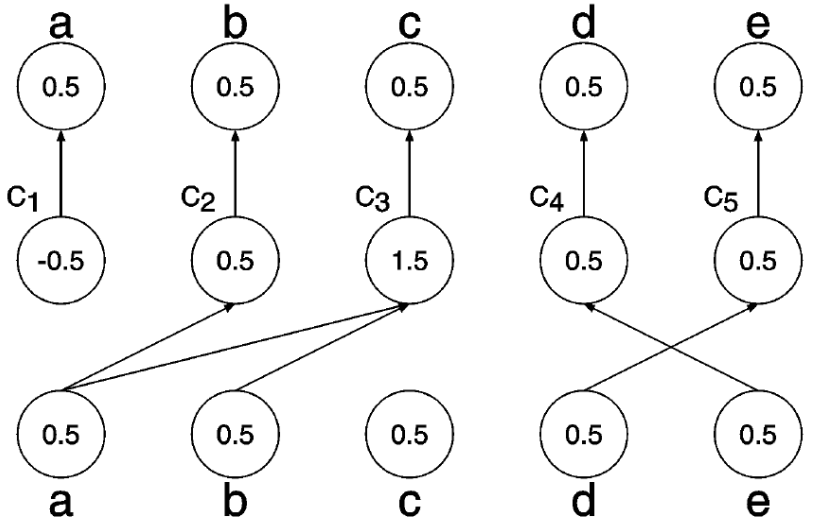




McCulloch & Pitts follow-on

Logic program P → core net

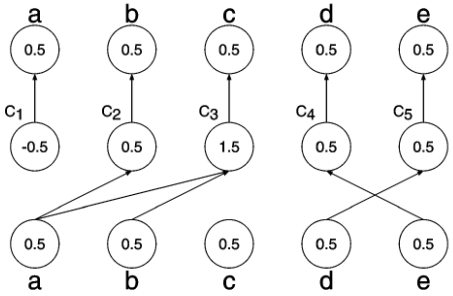
$a \leftarrow$
$b \leftarrow a$
$c \leftarrow a \wedge b$
$d \leftarrow e$
$e \leftarrow d$



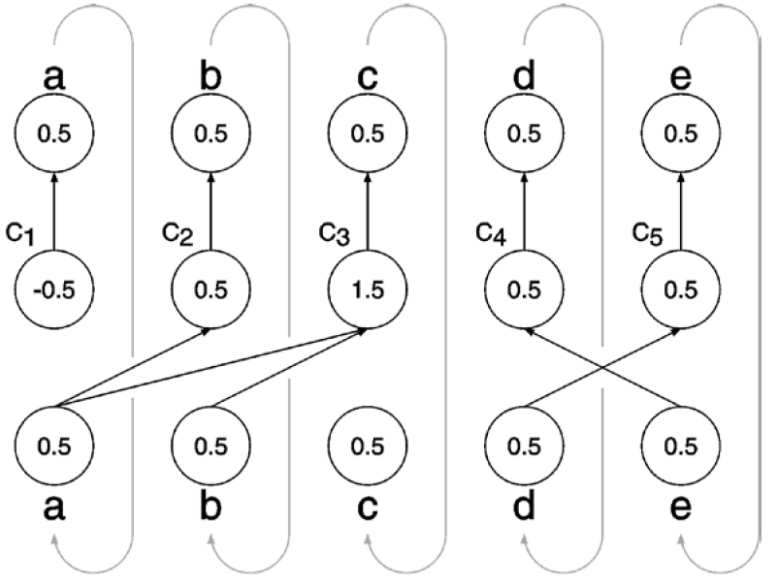
- Update „along implication“.
- Corresponds to computing the semantic operator T_P.
- T_P represents meaning (semantics) of P through its fixed points.

McCulloch & Pitts follow-on

core net



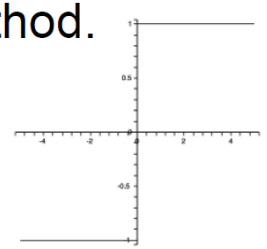
recurrent net

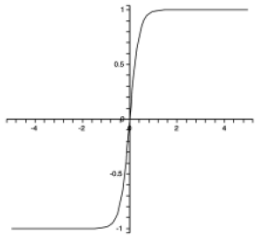


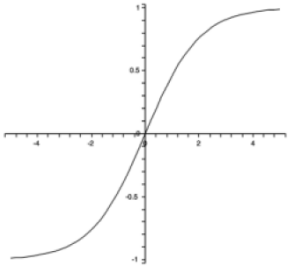
- Repeated updates along layers corresponds to iterations of the semantic operator.
- Semantics of the program (= fixed point of the operator) can be computed in a parallel manner.

McCulloch & Pitts follow-on

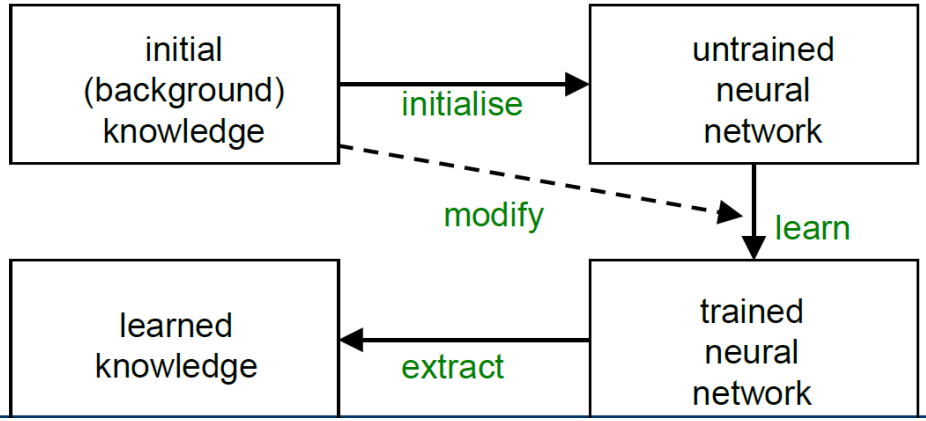
- Garcez & Zaverucha 1999
  Garcez, Broda & Gabbay 2001
- Development of a learning paradigm from the Core Method.
- Required: differentiable activation function.
  - Allows learning with standard methods.
  - Backpropagation algorithm.







- Establishing the *neural-symbolic learning cycle*.



The catch

- This is all propositional.
- There's only that much you can do with propositional logic.
- In particular, in terms of knowledge representation and reasoning, propositional logic doesn't really get you anything useful.

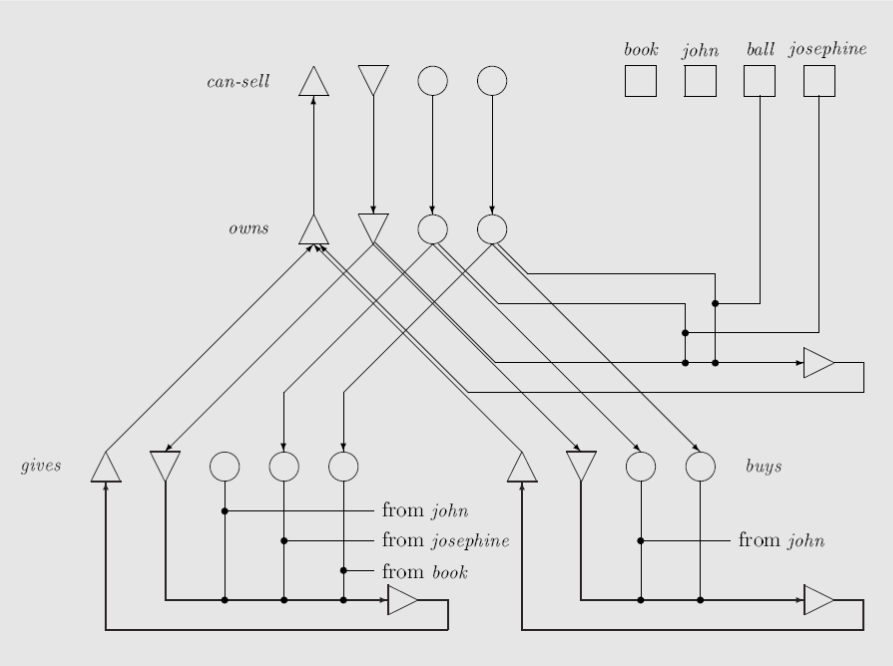
E.g.

- RDF (knowledge graphs) is already much closer to datalog than to propositional logic.
- OWL (knowledge graph schemas) is a fragment of first-order predicate logic.

Variable Binding

DaSe Lab

SHRUTI



Shastri & Ajjanagadde 1993

Variable binding
via time synchronization.

Reflexive (i.e. fast)
reasoning possible.

Picture: Hölldobler,
Introduction to Computational Logic, 2001

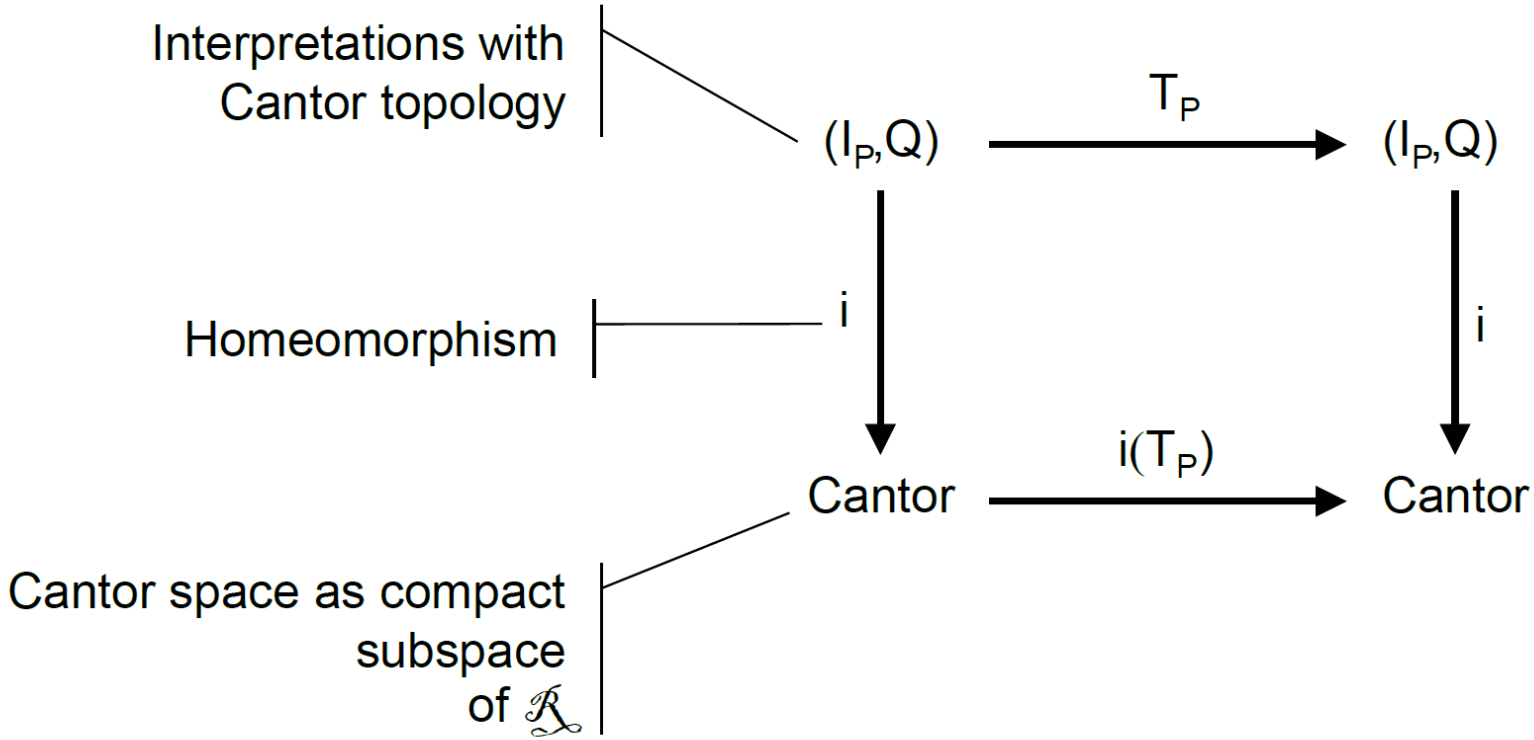
gives(X,Y,Z) → owns(Y,Z)
buys(X,Y) → owns(X,Y)
owns(X,Y) → can-sell(X,Y)

gives(john,josephine,book)
(∃X) buys(john,X)
owns(josephine,ball)

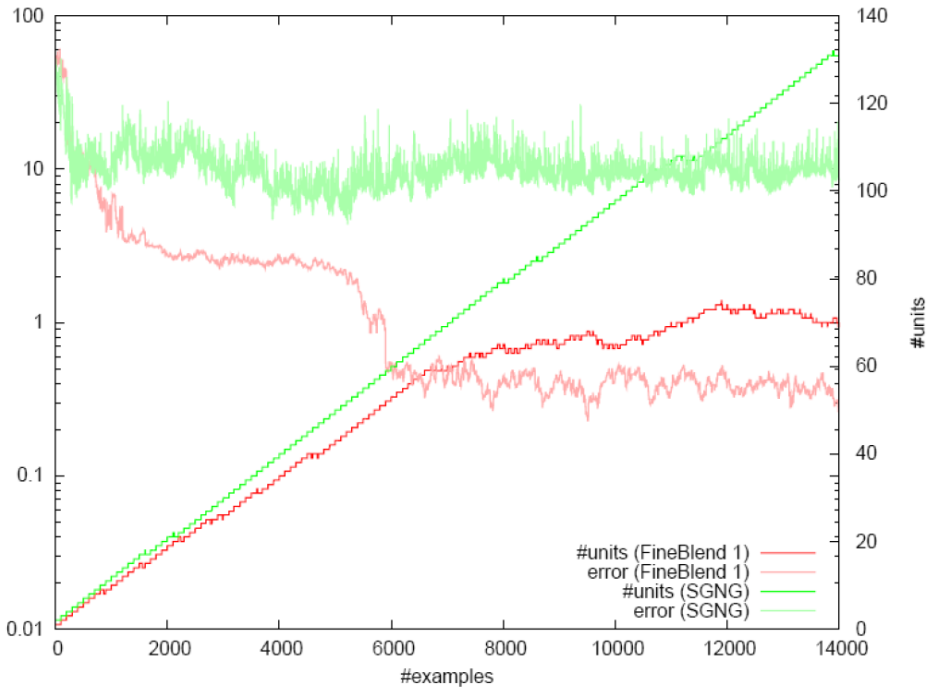
Problems:

- **It's still essentially datalog.** *** It doesn't really learn.**
- **It has a globally bounded reasoning depth.**

Logic in Real Space



Logic in Real Space

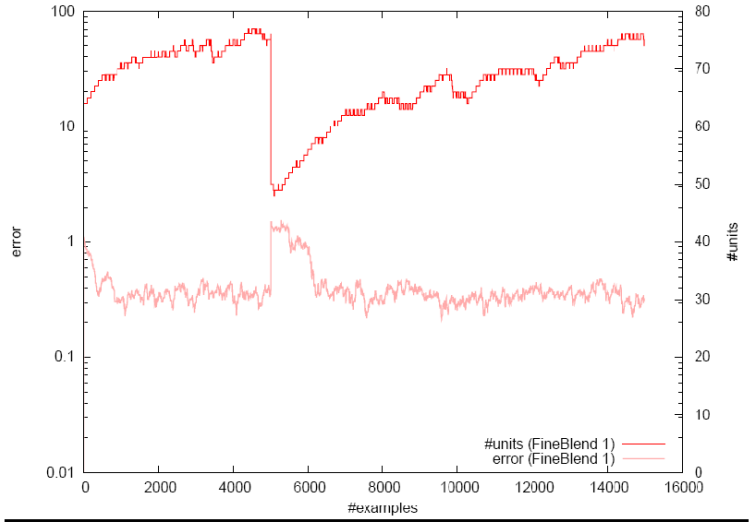


Architecture is mix of radial basis function network and neural gas approach.

target:
$e(0).$
$e(s(X)) \leftarrow o(X).$
$o(X) \leftarrow \neg e(X)$

initial:
$e(s(X)) \leftarrow \neg o(X)$
$e(X) \leftarrow e(X)$

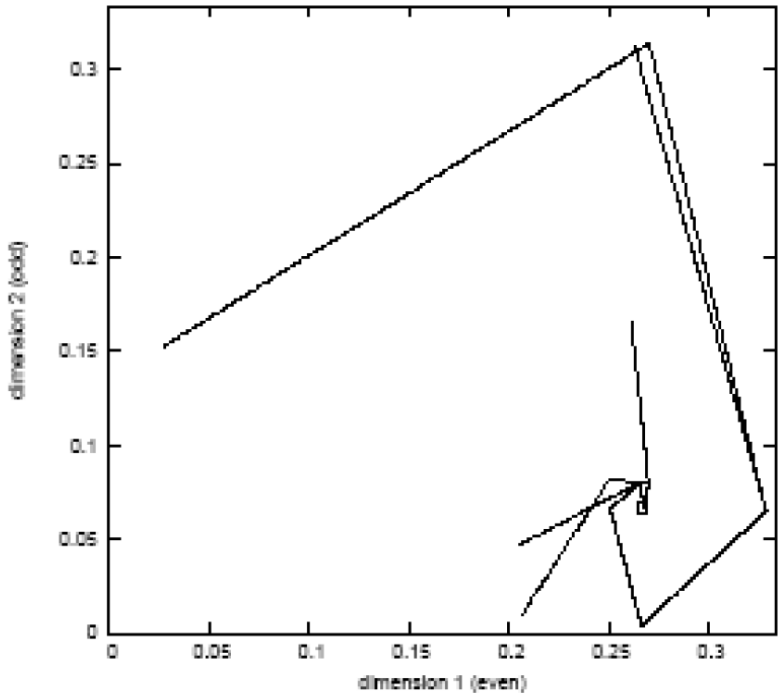
unit failure

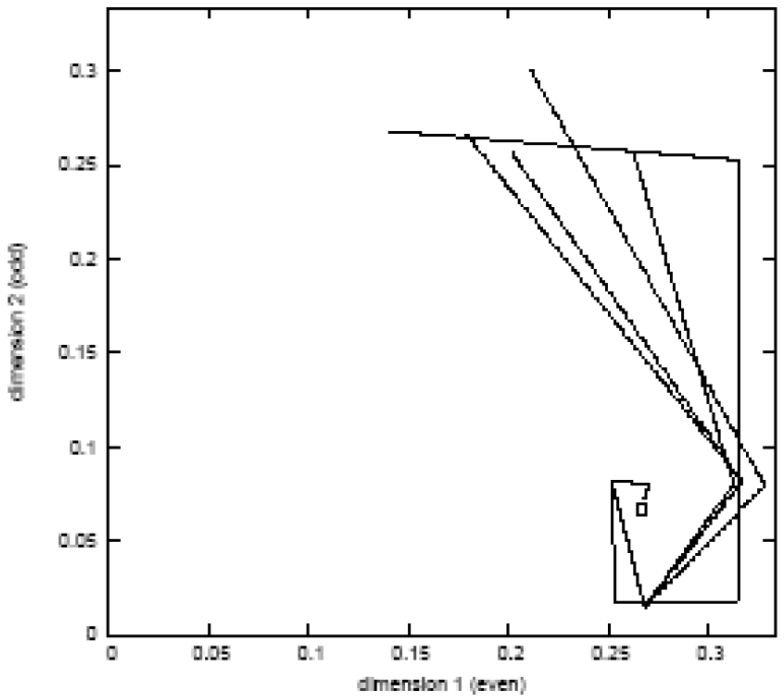


Logic in Real Space

We observe convergence to unique supported model of the program.

Bader, Hitzler, Hölldobler, Witzel, IJCAI-07





But it works only for toy size problems.
The theoretically required embedding into real numbers cannot scale.

RDFS Deductive Reasoning via Deep Memory Networks

Monireh Ebrahimi, Md Kamruzzaman Sarker, Aaron Eberhart, Federico Bianchi, Ning Xie, Derek Doran, Pascal Hitzler, under review

This would fit here, but Monireh is presenting about it this afternoon in the Deep Learning session, and I don't want to replicate it ☺

Fuzzy Deductive Reasoning via Logic Tensor Networks

Federico Bianchi, Pascal Hitzler

Logic Tensor Networks

Based on Neural Tensor Networks.

Logic Tensor Networks are due to Serafini and Garcez (2016). They have been used for image analysis under background knowledge.

Their capabilities for deductive reasoning have not been sufficiently explored.

Underlying logic: First-order predicate, fuzzyfied.

Every language primitive becomes a vector/matrix/tensor.

Terms/Atoms/Formulas are embedded as corresponding tensor/matrix/vector multiplications over the primitives.

Embeddings of primitives are learned s.t. the truth values of all formulas in the given theory are maximized.

A-priori Limitations

- **Not clear how to adapt this such that you can transfer to unseen input theories.**
- **Scalability is an issue.**
- **While apparently designed for deductive reasoning, the inventors hardly report on this issue.**

Transitive closure

- $\forall a, b, c \in A : (sub(a, b) \wedge sub(b, c)) \rightarrow sub(a, c)$
- $\forall a \in A : \neg sub(a, a)$
- $\forall a, b : sub(a, b) \rightarrow \neg sub(b, a)$

Satisfiability	MAE	Matthews	F1	Precision	Recall
0.99	**0.12** (0.12)	**0.58** (0.45)	**0.64** (0.51)	**0.60** (0.47)	0.68 (0.55)
0.56	0.51 (0.52)	0.09 (0.06)	0.27 (0.20)	0.20 (0.11)	**0.95** (0.93)
Random	0.50 (0.50)	0.00 (0.00)	0.22 (0.17)	0.14 (0.10)	0.50 (0.50)

parentheses: only newly entailed part of KB
MAE: mean absolute error;
Matthews: Matthews coefficient (for unbalanced classes)
top: top performing model, layer size and embeddings: 20, mean aggregator

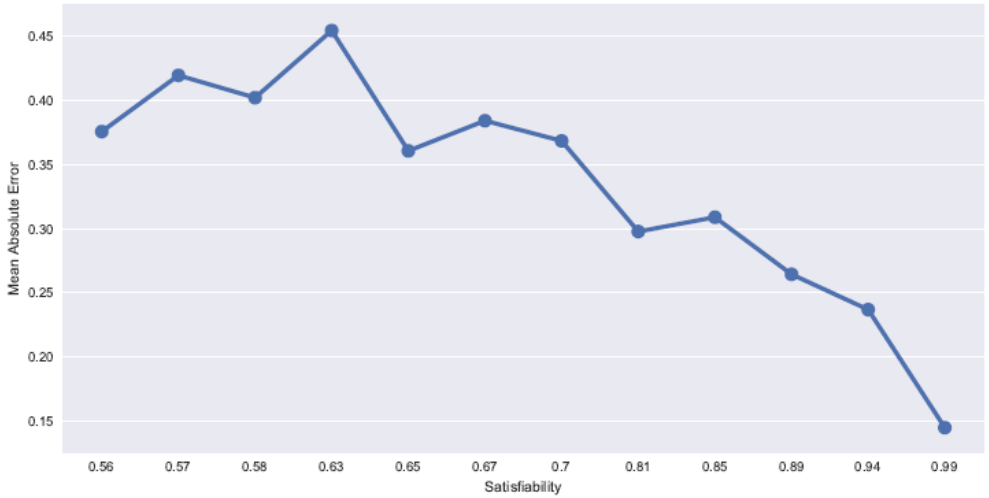
Bottom: one of the worst performing models.

Multi-hop inferences difficult.

More take-aways from experiments

DaSe Lab

- **Error decreases with increasing satisfiability.**



- **Adding redundant formulas to the input KB decreases error.**

Figure 3: Average MAE for the ancestors tasks on rounded level of satisfiability. MAE decreases with the increase of satisfiability.

Type	MAE	Matthews	F1	Precision	Recall
Six Axioms	0.16 (0.17)	0.73 (0.61)	0.77 (0.62)	0.64 (0.47)	**0.96** (0.92)
Eight Axioms	**0.14** (0.14)	**0.83** (0.69)	**0.85** (0.72)	**0.80** (0.66)	0.89 (0.79)

More take-aways from experiments

- **Higher arity of predicates significantly increases learning time.**

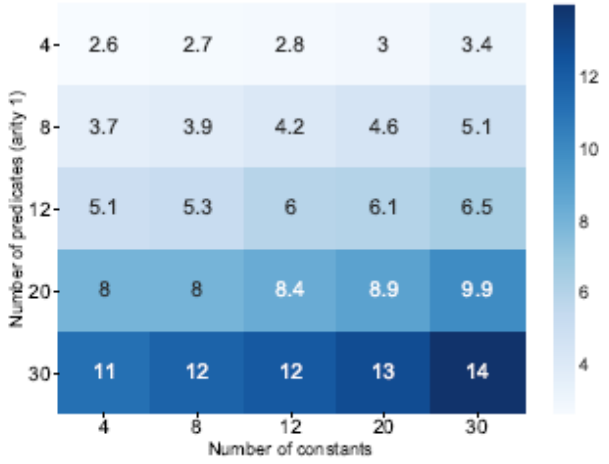


Figure 5: Computational times in seconds for predicates of arity one and constants

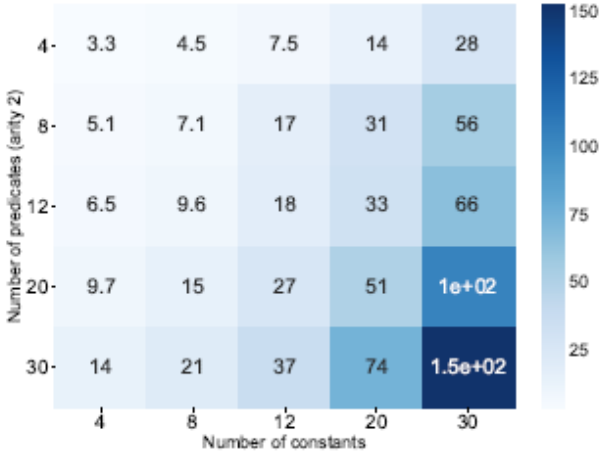


Figure 6: Computational times in seconds for predicates of arity two and constants

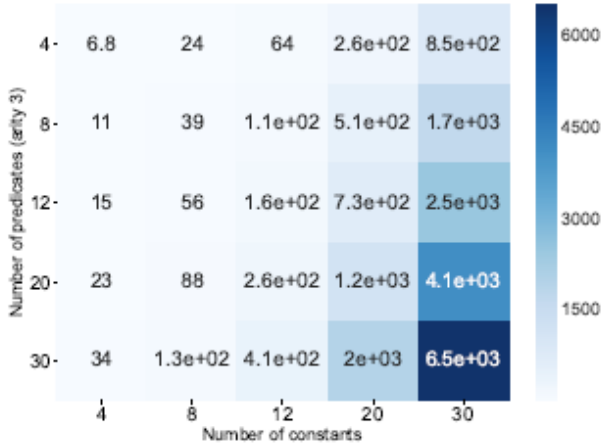


Figure 7: Computational times in seconds for predicates of arity three and constants

More take-aways from experiments

- Model seems to often end up in local minima. This may be addressable using known approaches.
- LTNs seem to predict many false positives, while they are better regarding true negatives. This may be just because of the test knowledge bases we used, but needs to be looked at.
- Overfitting is a problem, but it doesn't seem straightforward to address this for LTNs. [e.g. cross-validation may need completeness information, which may bias the network]
- Increasing layers and embedding size makes optimizing parameters much more difficult.

- Hence, there's a path for more investigations, we're only starting to understand this.

Explaining Deep Learning via Symbolic Background Knowledge

Md Kamruzzaman Sarker, Ning Xie, Derek Doran, Mike Raymer, Pascal Hitzler

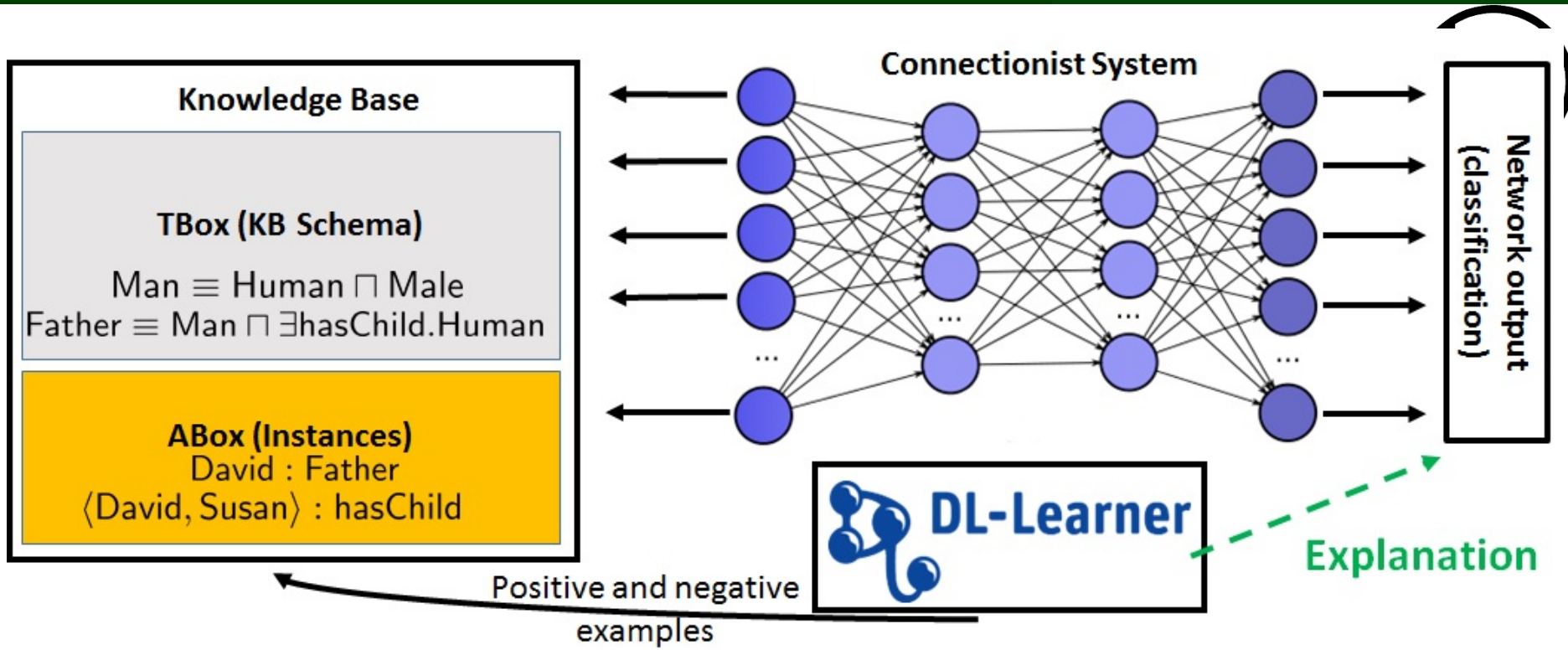
Explainable AI

- **Explain behavior of trained (deep) NNs.**

- **Idea:**
  - **Use background knowledge in the form of linked data and ontologies to help explain.**
  - **Link inputs and outputs to background knowledge.**
  - **Use a symbolic learning system (e.g., DL-Learner) to generate an explanatory theory.**

- **We're just starting on this, I report on very first experiments.**

Explainable AI



Using SUMO

Testing on ADE20k image dataset / scene recognition.

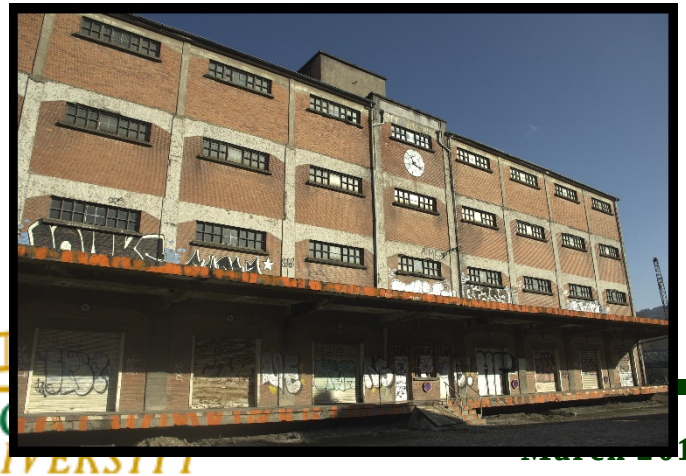
Workshop paper at NeSy'2017 with preliminary results.

Proof of Concept Experiment

DaSe Lab

SSC

Positive:



Negative:

Images

Come from the MIT ADE20k dataset
http://groups.csail.mit.edu/vision/datasets/ADE20K/
They come with annotations of objects in the picture:

```
001 # 0 # 0 # sky # sky # ""
002 # 0 # 0 # road, route # road # ""
005 # 0 # 0 # sidewalk, pavement # sidewalk # ""
006 # 0 # 0 # building, edifice # building # ""
007 # 0 # 0 # truck, motortruck # truck # ""
008 # 0 # 0 # hovel, hut, hutch, shack, shanty # hut # ""
009 # 0 # 0 # pallet # pallet # ""
011 # 0 # 0 # box # boxes # ""
001 # 1 # 0 # door # door # ""
002 # 1 # 0 # window # window # ""
009 # 1 # 0 # wheel # wheel # ""
```

Mapping to SUMO

Simple approach: for each known object in image, create an individual for the ontology which is in the appropriate SUMO class:

contains road1
contains window1
contains door1
contains wheel1
contains sidewalk1
contains truck1
contains box1
contains building1

SUMO

- **Suggested Merged Upper Ontology**
  http://www.adampease.org/OP/

- **Approx. 25,000 common terms covering a wide range of domains**

- **Centrally, a relatively naïve class hierarchy.**

- **Objects in image annotations became individuals (constants), which were then typed using SUMO classes.**

DL-Learner input

Positive:

- **img1:** road, window, door, wheel, sidewalk, truck, box, building
- **img2:** tree, road, window, timber, building, lumber
- **img3:** hand, sidewalk, clock, steps, door, face, building, window, road

Negative:

- **img4:** shelf, ceiling, floor
- **img5:** box, floor, wall, ceiling, product
- **img6:** ceiling, wall, shelf, floor, product

DL-Learner results include:

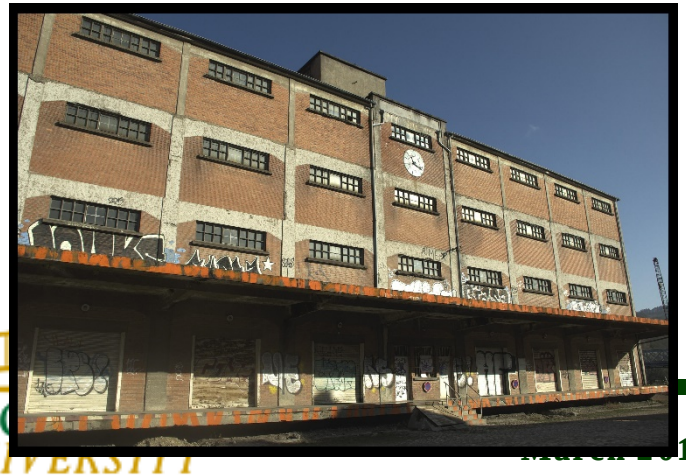
$$\exists \text{contains.Transitway}$$

$$\exists \text{contains.LandArea}$$

Proof of Concept Experiment

DaSe Lab

Positive:



Negative:

$\exists contains.Transitway$

$\exists contains.LandArea$

First 10 DL-Learner responses

$\exists$contains.Window	(1)	$\exists$contains.LandTransitway	(6)
$\exists$contains.Transitway	(2)	$\exists$contains.LandArea	(7)
$\exists$contains.SelfConnectedObject	(3)	$\exists$contains.Building	(8)
$\exists$contains.Roadway	(4)	$\forall$contains.$\neg$Floor	(9)
$\exists$contains.Road	(5)	$\forall$contains.$\neg$Ceiling	(10)

Experiment 2

DaSe Lab

SSC

Positive (selection):

Negative (selection):

$$\exists \text{contains}.(\text{DurableGood} \sqcap \neg \text{ForestProduct})$$

Experiment 3

DaSe Lab

Positive:

Negative:

$$\forall contains.(\neg Furniture \sqcap \neg IndustrialSupply)$$

Experiment 4

DaSe Lab

SSC

Positive (selection):

Negative (selection):

$\exists$contains.SentientAgent

Experiment 5

Positive:




$\exists$contains.BodyOfWater

Negative (selection):



DL-Learner efficiency problem

- **DL-Learner was too slow – we needed several hours for each computation, and couldn't explore and/or scale up.**
- **We thus implemented our own system, ECII (Efficient Concept Induction from Instances) which trades some correctness for speed. [Sarker, Hitzler, AAAI-19, to appear]**

Experiment Name	Number of Logical Axioms	Runtime (sec)					Accuracy (α_3)		Accuracy α_2			
		DL[a]	DL FIC(1)[b]	DL FIC(2)[c]	ECII DF[d]	ECII KCT[e]	DL[a]	ECII DF[d]	DL FIC(1)[b]	DL FIC(2)[c]	ECII DF[d]	ECII KCT[e]
Yinyang_examples	157	0.065	0.0131	0.019	0.089	0.143	1.000	0.610	1.000	1.000	0.799	1.000
Trains	273	0.01	0.020	0.047	0.05	0.095	1.000	1.000	1.000	1.000	1.000	1.000
Forte	341	2.5	1.169	6.145	0.95	0.331	0.965	0.642	0.875	0.875	0.733	1.000
Poker	1,368	0.066	0.714	0.817	1	0.281	1.000	1.000	0.981	0.984	1.000	1.000
Moral Reasoner	4,666	0.1	3.106	4.154	5.47	6.873	1.000	0.785	1.000	1.000	1.000	1.000
ADE20k I	4,714	577.3[f]	4.268	31.887	1.966	23.775	0.926	0.416	0.263	0.814	0.744	1.000
ADE20k II	7,300	983.4[f]	16.187	307.65	20.8	293.44	1.000	0.673	0.413	0.413	0.846	0.900
ADE20k III	12,193	4,500[g]	13.202	263.217	51	238.8	0.375	0.937	0.375	0.375	0.930	0.937
ADE20k IV	47,468	4,500[g]	93.658	523.673	116	423.349	0.375	NA	0.608	0.608	0.660	0.608

[a] DL : DL-Learner
[b] DL FIC (1) : DL-Learner fast instance check with runtime capped at execution time of ECII DF
[c] DL FIC (2) : DL-Learner fast instance check with runtime capped at execution time of ECII KCT
[d] ECII DF : ECII default parameters
[e] ECII KCT : ECII keep common types and other default parameters
[f] Runtimes for DL-Learner were capped at 600 seconds.
[g] Runtimes for DL-Learner were capped at 4,500 seconds.

ECII vs. DL-Learner

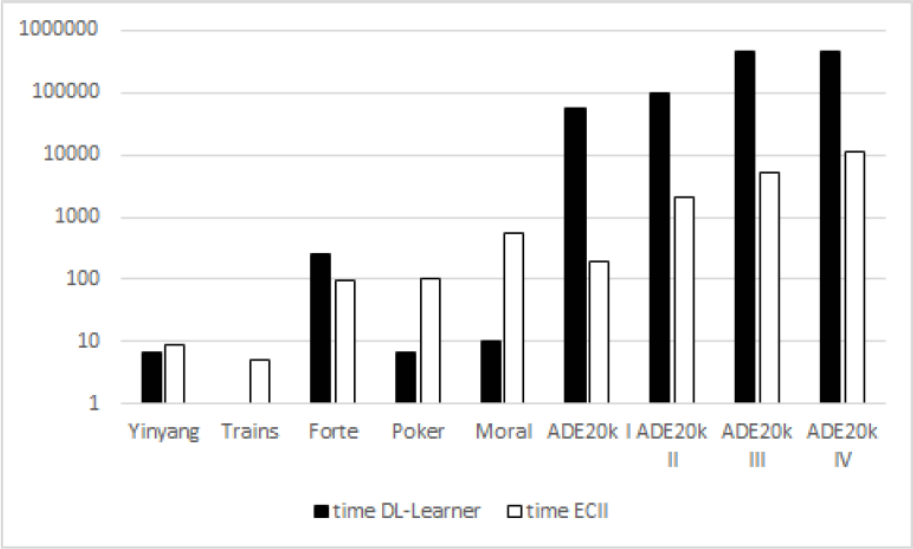


Figure 1: Runtime comparison between DL-Learner and ECII. The vertical scale is logarithmic in hundredths of seconds, and note that DL-Learner runtime has been capped at 4,500 seconds for ADE20k III and IV. For ADE20k I it was capped at each run at 600 seconds.

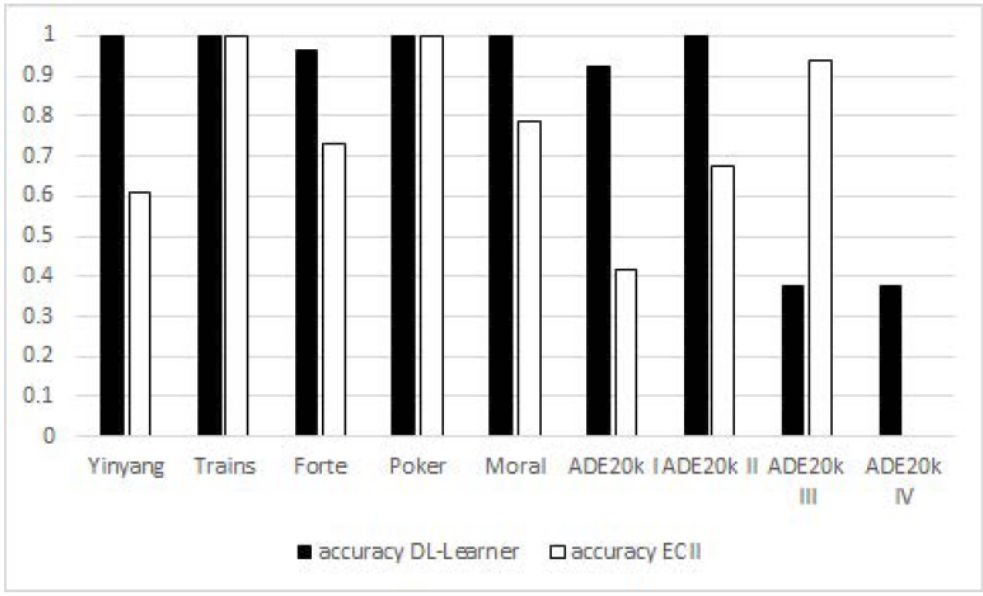


Figure 2: Accuracy (α_3) comparison between DL-Learner and ECII. For ADE20k IV it was not possible to compute an accuracy score within 3 hours for ECII as the input ontology was too large.

Reasons for Improvement

- DL-Learner loops the following steps:
  1. Generate several (refined) candidate solutions.
  2. Test candidate solutions by calling a reasoner.
  3. Keep only the best solution(s).
- This results in many reasoner calls, which are expensive.

- ECII optimizes by introducing several (approximate) simplifications:
  - Partially materialize reasoning up-front: only one reasoner call required.
  - Allow only solutions of a restricted form/syntax.
  - Compose solution from pieces which are independently verified against the materialized data.

Next:

- **We're just now starting to run full-scale experiments with ECII in the described setting.**

Conclusions

Conclusions

- Briding the symbolic-subsymbolic gap is still a major quest.

- We looked at
  - RDFS reasoning using memory networks (very good)
  - Logic Tensor Networks for first-order predicate logic (unclear)
  - Background knowledge for explainable AI (first steps suggest optimism)

- Possible direct transfers:
  - To other types of inference (e.g., common-sense reasoning, natural language reasoning)
  - Explaining other machine learning paradigms.

Thanks!

References

Barbara Hammer and Pascal Hitzler (eds), Perspectives on Neural-Symbolic Integration. Springer, 2007

Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kuehnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, Gerson Zaverucha, Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. https://arxiv.org/abs/1711.03902 (2017)

McCulloch, W.S. & Pitts, W. Bulletin of Mathematical Biophysics (1943) 5: 115.

P. Hitzler, S. Hölldobler and A. K. Seda. Logic Programs and Connectionist Networks. Journal of Applied Logic, 2(3), 2004, 245-272.

References

Artur S. d'Avila Garcez, Gerson Zaverucha, The Connectionist Inductive Learning and Logic Programming System. Appl. Intell. 11(1): 59-77 (1999)

Artur S. d'Avila Garcez, Krysia Broda, Dov M. Gabbay, Symbolic knowledge extraction from trained neural networks: A sound approach. Artificial Intelligence 125(1-2): 155-207 (2001)

J. McCarthy. Epistemological challenges for connectionism. Behavioral and Brain Sciences, 11 (1): 44, 1988

Lokendra Shastri, SHRUTI: A Neurally Motivated Architecture for Rapid, Scalable Inference. Perspectives of Neural-Symbolic Integration 2007: 183-203

References

Sebastian Bader, Pascal Hitzler, Steffen Hölldobler, Connectionist model generation: A first-order approach. Neurocomputing 71(13-15): 2420-2432 (2008)

Bassem Makni, James Hendler, Deep learning for noise-tolerant RDFS reasoning. Under review at Semantic Web journal.

Md. Kamruzzaman Sarker, Ning Xie, Derek Doran, Michael Raymer, Pascal Hitzler, Explaining Trained Neural Networks with Semantic Web Technologies: First Steps. In: Tarek R. Besold, Artur S. d'Avila Garcez, Isaac Noble (eds.), Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning, NeSy 2017, London, UK, July 17-18, 2017. CEUR Workshop Proceedings 2003, CEUR-WS.org 2017

References

Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph, Foundations of Semantic Web Technologies. Textbooks in Computing, Chapman and Hall/CRC Press, 2010.

Sebastian Bader, Pascal Hitzler, Dimensions of neural-symbolic integration – a structured survey. In: S. Artemov, H. Barringer, A. S. d'Avila Garcez, L. C. Lamb and J. Woods (eds). We Will Show Them: Essays in Honour of Dov Gabbay, Volume 1. International Federation for Computational Logic, College Publications, 2005, pp. 167-194.

Monireh Ebrahimi, Md Kamruzzaman Sarker, Federico Bianchi, Ning Xie, Derek Doran, Pascal Hitzler, Reasoning over RDF Knowledge Bases using Deep Learning. arXiv:1811.04132, November 2018.

References

Md Kamruzzaman Sarker, Pascal Hitzler, Efficient Concept Induction for Description Logics. AAAI-19, to appear.

Federico Bianchi, Pascal Hitzler, On the Capabilities of Logic Tensor Networks for Deductive Reasoning. Unpublished Manuscript, November 2018.