

Explainable Deep Learning using Concept Induction



Pascal Hitzler

Data Semantics Laboratory (DaSe Lab)
Kansas State University

<http://www.daselab.org>

- **Neurosymbolic Artificial Intelligence**
- **Concept Induction**
- **Explainability Framework**
- **Explaining Hidden Neuron Activations**
- **Are Concept Induction Explanations Meaningful To Humans?**
- **Improving Deep Learning Through Concept Induction**

Some Background

**Workshop Series on Neural-Symbolic Learning and Reasoning, since 2005.
Joint with Artur d'Avila Garcez.**

<http://neural-symbolic.org/>

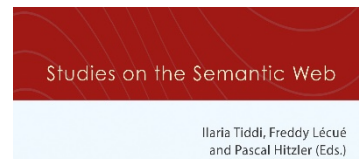
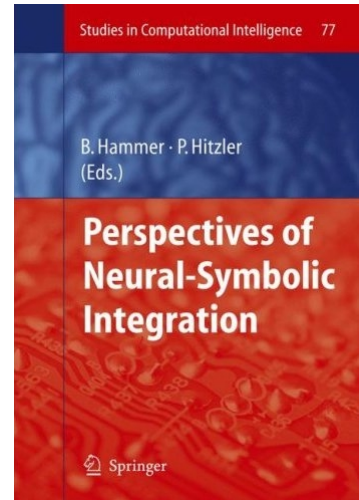
**Barbara Hammer and Pascal Hitzler (eds), Perspectives of
Neural-Symbolic Integration, Springer, 2007**

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)

**Ilaria Tiddi, Freddy Lecue, Pascal Hitzler (eds.), Knowledge Graphs
for eXplainable Artificial Intelligence: Foundations, Applications and
Challenges. Studies on the Semantic Web Vol. 47, IOS Press, 2020.**



Neuro-symbolic AI



**Publications on neuro-symbolic AI in major conferences
(research papers only):**

| conference | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | total |
|------------|------|------|------|------|------|------|------|------|------|------|-------|
| ICML | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 2 | 5 | 6 | 17 |
| NeurIPS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 2 | 4 | 10 |
| AAAI | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 4 |
| IJCAI | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 2 | 7 |
| ICLR | N/A | N/A | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 3 | 6 |
| total | 1 | 0 | 0 | 0 | 0 | 2 | 6 | 10 | 9 | 16 | 44 |

See

**Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, Pascal Hitzler
Neuro-Symbolic Artificial Integration: Current Trends
AI Communications 34 (3), 197-209, 2022.**

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)
- 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 / knowledge graphs are inherently 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
- **But how to do this best?**

2022 Book

Neuro-symbolic Artificial Intelligence: The State of the Art

Pascal Hitzler and Md Kamruzzaman Sarker, editors

Frontiers in AI and Applications Vol. 342, IOS Press, Amsterdam, 2022

<https://www.iospress.com/catalog/books/neuro-symbolic-artificial-intelligence-the-state-of-the-art>

Frontiers
in
Artificial
Intelligence
and
Applications

NEURO-SYMBOLIC
ARTIFICIAL
INTELLIGENCE:
THE STATE
OF THE ART

Edited by
Pascal Hitzler
Md Kamruzzaman Sarker

IOS Press

| | | |
|--|-----|--|
| Preface: The 3rd AI wave is coming, and it needs a theory <i>Frank van Harmelen</i> | v | |
| Introduction <i>Pascal Hitzler and Md Kamruzzaman Sarker</i> | ix | |
| Chapter 1. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation <i>Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C. Lamb, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon and Gerson Zaverucha</i> | 1 | Chapter 9. Spike-Based Symbolic Computations on Bit Strings and Numbers <i>Ceca Kraišniković, Wolfgang Maass and Robert Legenstein</i> 214 |
| Chapter 2. Symbolic Reasoning in Latent Space: Classical Planning as an Example <i>Masataro Asai, Hiroshi Kajino, Alex Fukunaga and Christian Muise</i> | 52 | Chapter 10. Explainable Neuro-Symbolic Hierarchical Reinforcement Learning <i>Daoming Lyu, Fangkai Yang, Hugh Kwon, Bo Liu, Wen Dong and Levent Yilmaz</i> 235 |
| Chapter 3. Logic Meets Learning: From Aristotle to Neural Networks <i>Vaishak Belle</i> | 78 | Chapter 11. Neuro-Symbolic Semantic Reasoning <i>Bassem Makni, Monireh Ebrahimi, Dagmar Gromann and Aaron Eberhart</i> 253 |
| Chapter 4. Graph Reasoning Networks and Applications <i>Qingxing Cao, Wentao Wan, Xiaodan Liang and Liang Lin</i> | 103 | Chapter 12. Learning Reasoning Strategies in End-to-End Differentiable Proving <i>Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette and Tim Rocktäschel</i> 280 |
| Chapter 5. Answering Natural-Language Questions with Neuro-Symbolic Knowledge Bases <i>Haitian Sun, Pat Verga and William W. Cohen</i> | 126 | Chapter 13. Generalizable Neuro-Symbolic Systems for Commonsense Question Answering <i>Alessandro Oltramari, Jonathan Francis, Filip Ilievski, Kaixin Ma and Roshanak Mirzaee</i> 294 |
| Chapter 6. Tractable Boolean and Arithmetic Circuits <i>Adnan Darwiche</i> | 146 | Chapter 14. Combining Probabilistic Logic and Deep Learning for Self-Supervised Learning <i>Hoifung Poon, Hai Wang and Hunter Lang</i> 311 |
| Chapter 7. Neuro-Symbolic AI = Neural + Logical + Probabilistic AI <i>Robin Manhaeve, Giuseppe Marra, Thomas Demeester, Sebastijan Dumančić, Angelika Kimmig and Luc De Raedt</i> | 173 | Chapter 15. Human-Centered Concept Explanations for Neural Networks <i>Chih-Kuan Yeh, Been Kim and Pradeep Ravikumar</i> 337 |
| Chapter 8. A Constraint-Based Approach to Learning and Reasoning <i>Michelangelo Diligenti, Francesco Giannini, Marco Gori, Marco Maggini and Giuseppe Marra</i> | 192 | Chapter 16. Abductive Learning <i>Zhi-Hua Zhou and Yu-Xuan Huang</i> 353 |
| | | Chapter 17. Logic Tensor Networks: Theory and Applications <i>Luciano Serafini, Artur d'Avila Garcez, Samy Badreddine, Ivan Donadello, Michael Spranger and Federico Bianchi</i> 370 |

New Book for 2023

Compendium of Neuro-Symbolic Artificial Intelligence (tentative)



approx. 30 chapters and 800 pages

Each chapter based on 2 or more related published papers.

Book will provide an even more comprehensive overview of the state of the art.

New Journal



- **Neurosymbolic Artificial Intelligence journal, IOS Press**
- **Open and Transparent reviewing (like Semantic Web journal)**
- **Will open for submissions early 2023.**

- **Preliminary announcement:**
<https://www.iospress.com/catalog/journals/neurosymbolic-artificial-intelligence>

- **EiCs:**
 - **Tarek Besold**
 - **Artur Garcez**
 - **Pascal Hitzler**

NeSy Workshop

- **Annual Workshop on Neural-Symbolic Learning and Reasoning (NeSy)**
- **17th Installation, July 2023, Siena, Italy**
- **Announcement this week or so**



Community slack

- **Community slack for Neurosymbolic AI**
- **neurosymbolic-group.slack.com**
- **email me to receive an invite (hitzler@ksu.edu)**



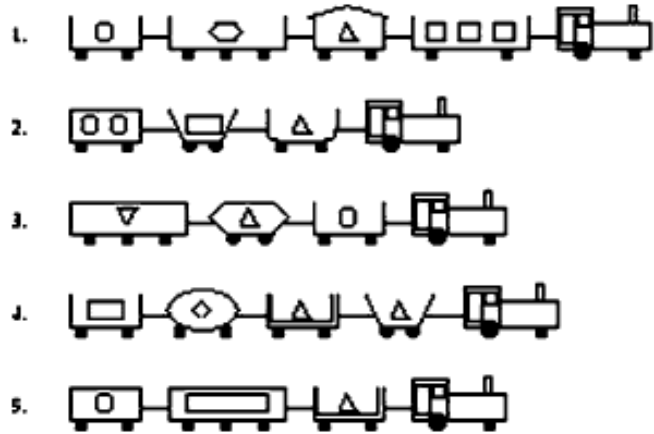
- **Neurosymbolic Artificial Intelligence**
- **Concept Induction**
- **Explainability Framework**
- **Explaining Hidden Neuron Activations**
- **Are Concept Induction Explanations Meaningful To Humans?**
- **Improving Deep Learning Through Concept Induction**

DL-Learner [Lehmann, Hitzler]

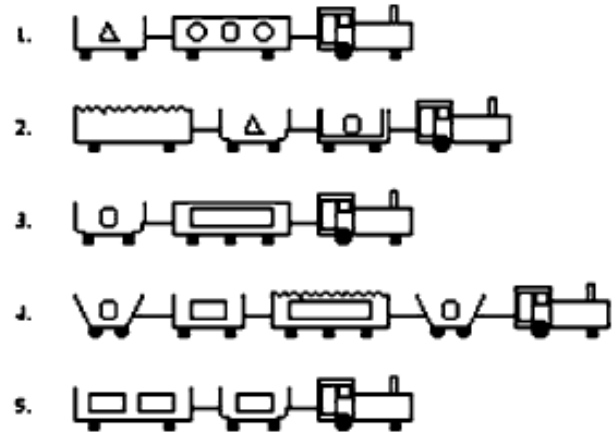


Approach similar to inductive logic programming, but using Description Logics (the logic underlying OWL).

Positive examples:



negative examples:



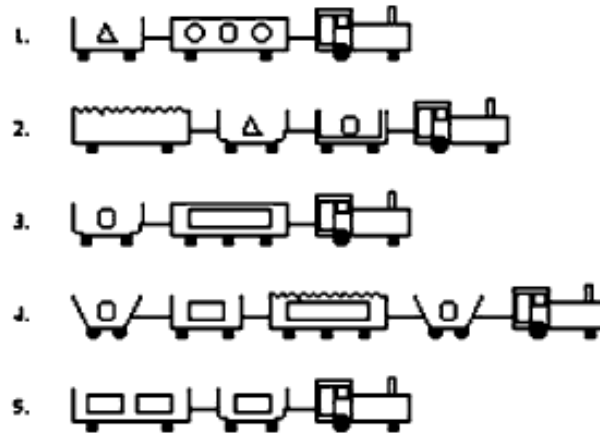
Task: find a class description (logical formula) which separates positive and negative examples.

Jens Lehmann, Pascal Hitzler, Concept Learning in Description Logics Using Refinement Operators. Machine Learning 78 (1-2), 203-250, 2010.

Positive examples:



negative examples:



DL-Learner result:

$\exists \text{hasCar} . (\text{Closed} \sqcap \text{Short})$

In FOL:

$$\{x \mid \exists y (\text{hasCar}(x, y) \wedge \text{Closed}(y) \wedge \text{Short}(y))\}$$

Scalability Issues with DL-Learner



- For large-scale experiments, DL-Learner took 2 hours or more for one run.
- We knew we needed at least thousands of runs.
- So we needed a more scalable solution.
- The provably correct algorithms have very high complexity.
- Hence we had to develop a heuristic which trades (some) correctness for speed.
- It is also currently restricted to using a class hierarchy as underlying knowledge base.

ECII algorithm and system



- We thus implemented our own system, ECII (Efficient Concept Induction from Instances) which trades some correctness for speed. [Sarker, Hitzler, AAI-19]

| 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: heuristic Concept Induction system



- For scalability, we developed ECII (Efficient Concept Induction from Instances) which trades some correctness for speed. [Sarker, Hitzler, AAI-19]

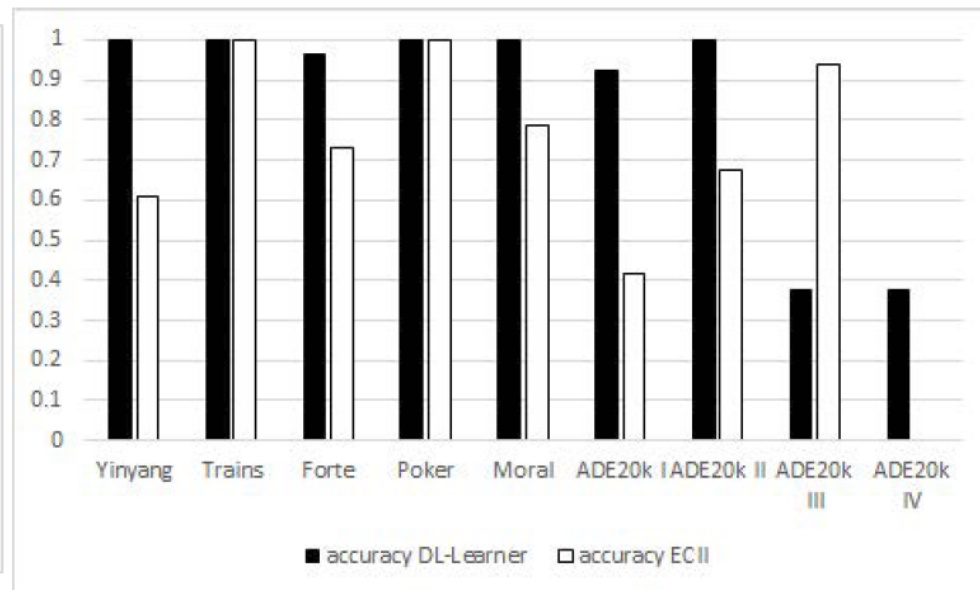
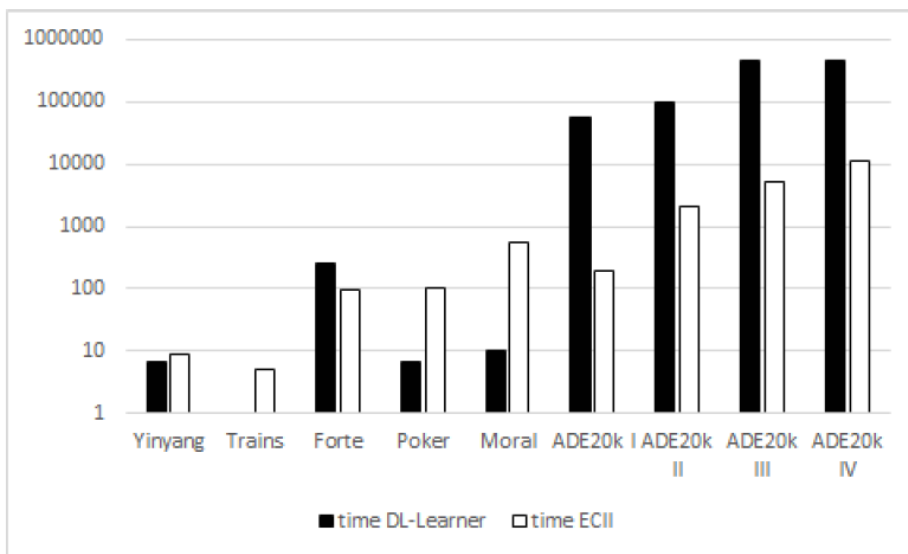
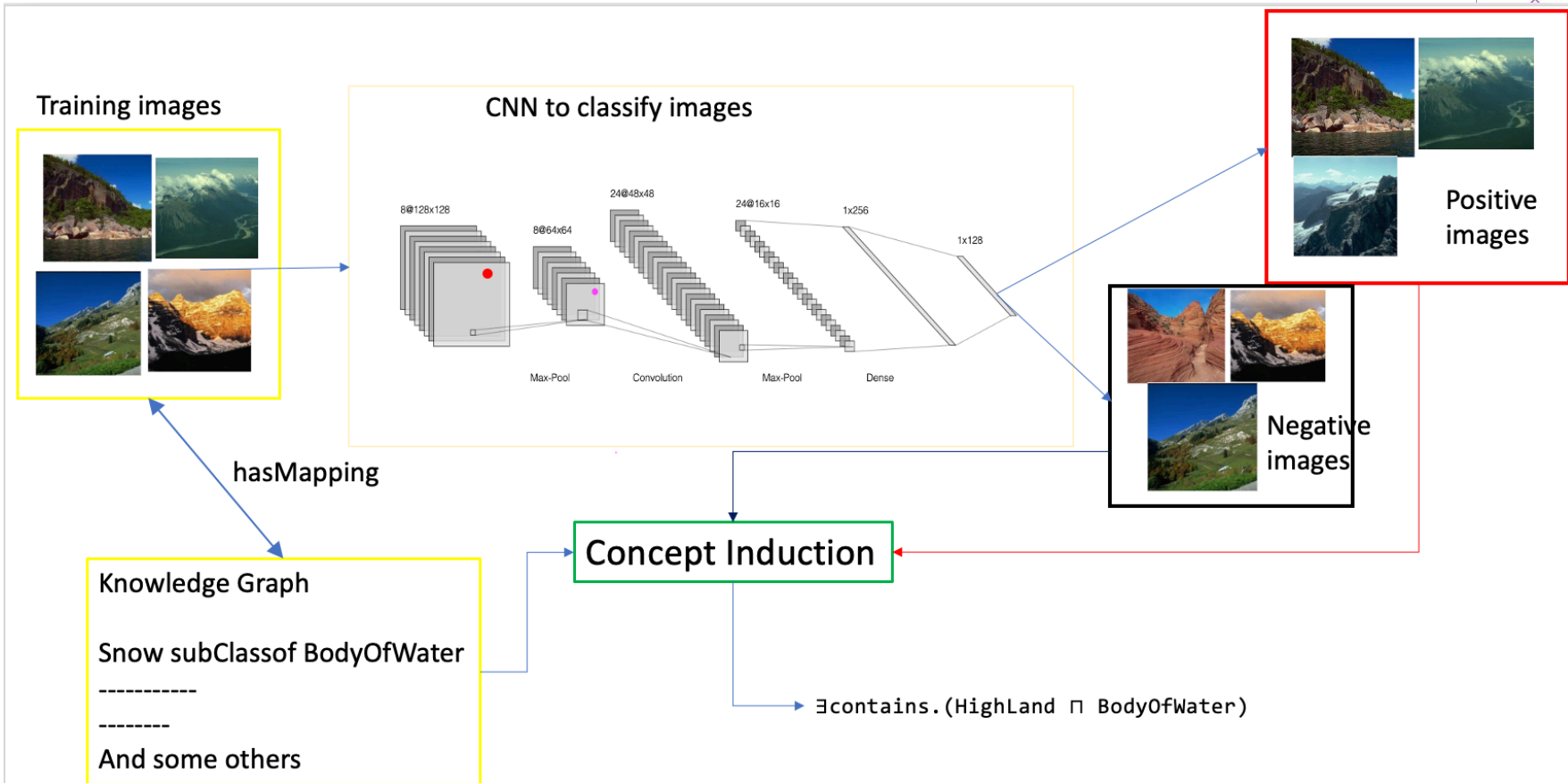


Figure 1: Runtime comparison between DL-Learner and ECII. The vertical scale is logarithmic in hundredths of seconds, and note that DL-Learner runtime has been capped at 4,500 seconds for ADE20k III and IV. For ADE20k I it was capped at each run at 600 seconds.

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.

- **Neurosymbolic Artificial Intelligence**
- **Concept Induction**
- **Explainability Framework**
- **Explaining Hidden Neuron Activations**
- **Are Concept Induction Explanations Meaningful To Humans?**
- **Improving Deep Learning Through Concept Induction**

Concept



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 d'Avila Garcez, Isaac Noble, Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning, NeSy 2017, London, UK, July 17-18 2017. CEUR Workshop Proceedings Vol. 2003, 2017.

Proof of Concept Experiment



Positive:



Negative:



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 Background Knowledge



- **Wikipedia category hierarchy (curated)**
[Sarker et al, **KGSWC 2020**]
- **approx. 2M concepts**
- **For each known object in image, create an individual for the ontology which is in the appropriate class.**

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



Proof of Concept Experiment



Positive:



Negative:



Contains Transitway

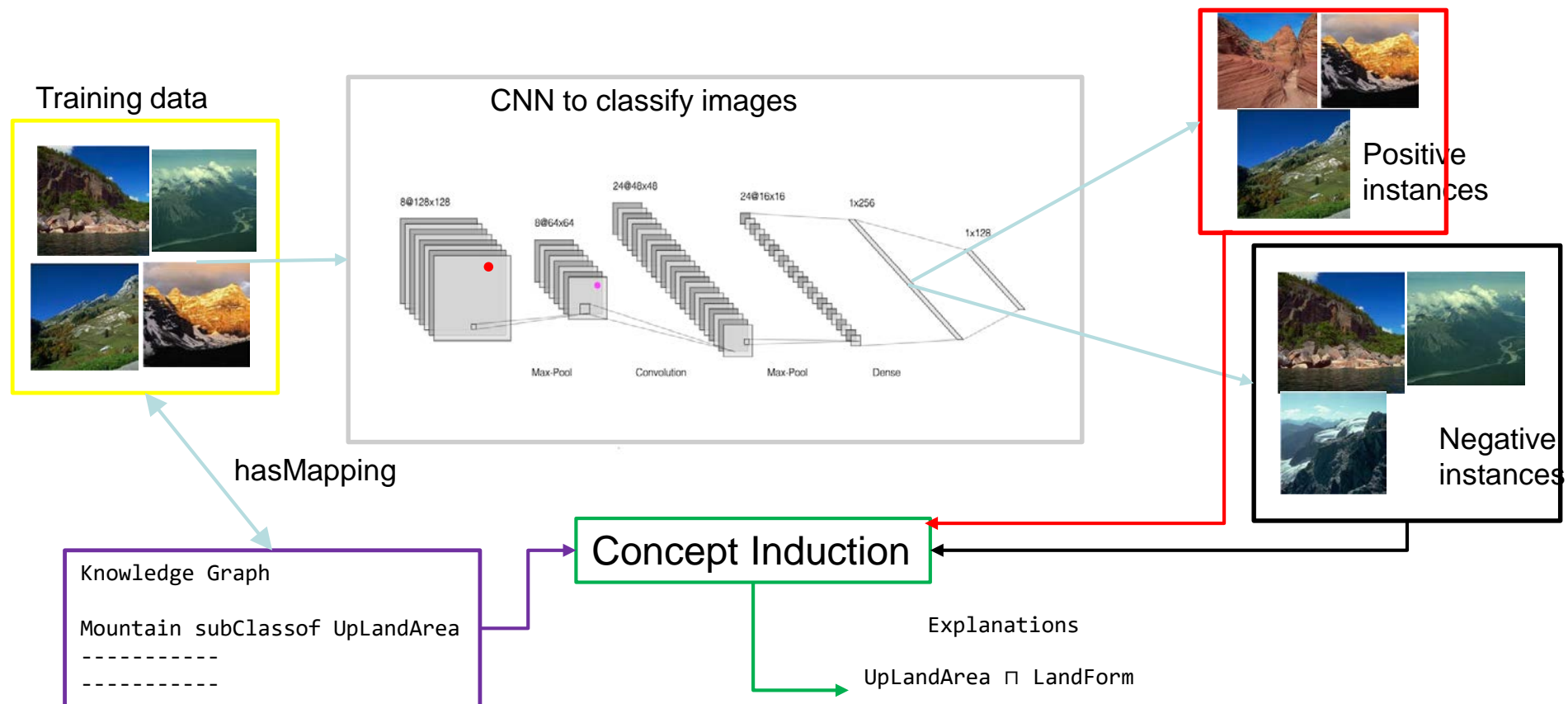
Contains Land Area

- **Neurosymbolic Artificial Intelligence**
- **Concept Induction**
- **Explainability Framework**
- **Explaining Hidden Neuron Activations**
- **Are Concept Induction Explanations Meaningful To Humans?**
- **Improving Deep Learning Through Concept Induction**

Idea Recap



- Generate explanation of the whole model
- Global explanation



Results (communicated by Abhilekha Dalal)



Neuron number 04 (dense layer, i.e. before output layer):

- Total number of images that got activated = **612/1370** (1370= test_dataset)
- Highest activation = **12.627778**
- Total number of positives = **149 (images that has value ≥ 6)**
- Total number of negatives = **150 (images that has value < 6)**

Solution given by ECI analysis for neuron 04

solution 1: (:Bed)
 solution 2: (:WN_Bed)
 solution 3: (:WN_Table)
 solution 4: (:WN_Lamp)
 solution 5: ((:WN_Table) \sqcap (:Bed))
 solution 6: (:Night_table)
 solution 7: (:Cushion)
 solution 8: ((:Cushion) \sqcap (:WN_Cushion))
 solution 9: (:WN_Shade)
 solution 10: ((:Pillow) \sqcap (:WN_Bed))
 solution 14: (:WN_Pillow)
 solution 17: ((:WN_Cushion) \sqcap (:WN_Lamp))
 solution 19: (:WN_Headboard)
 solution 24: ((:WN_Lamp) \sqcap (:Pillow))
 solution 25: (:WN_Table)



Distinct Concepts from the solution

Bed
 Table
 Night Table
 Lamp
 Pillow
 Cushion
 Headboard

Results

Google analysis for Neuron number 04 :

- Take each concept from distinct concept list for eg: Bed, Table and collect images from Google.
- First set analysis, all images activate (853 images)
- Second set analysis, all images activate (900 images)



ADE20K Dataset

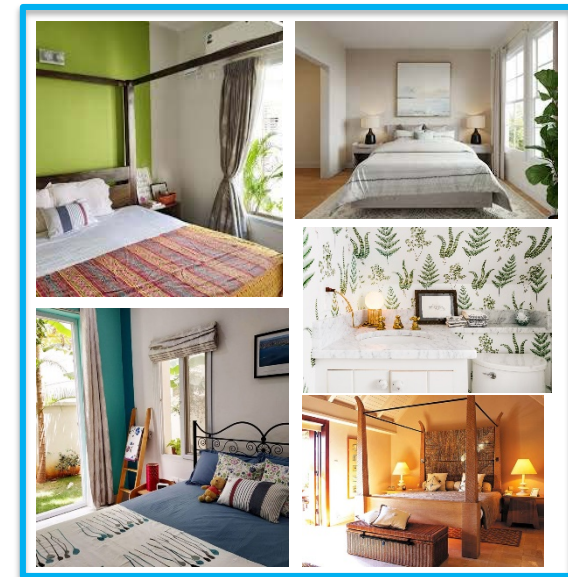


Positive Images



Negative Images

Google Images



Neuron number 05 :

- Total number of images that got activated =
- Highest activation =
- Total number of positives =
- Total number of negatives =

787/1370 (1370= test_dataset)

10.196102

116 (images that has value ≥ 5)

150 (images that has value < 5)

Solution given by ECII analysis for neuron 04

solution 1: (:WN_Table)
 solution 2: (:Floor)
 solution 4: (:WN_Flooring)
 solution 5: (:Window)
 solution 7: ((:WN_Flooring) \sqcap (:Window))
 solution 10: ((:Ceiling) \sqcap (:WN_Table))
 solution 15: (:Picture)
 solution 17: (:WN_Picture)
 solution 22: (:Chair)
 solution 24: (:WN_Lamp)
 solution 26: ((:WN_Windowpane) \sqcap (:WN_Painting))

Distinct Concepts from the solution

Table
 Floor
 Window
 Ceiling
 Picture
 Chair
 Lamp
 Painting



Results

Google analysis for Neuron number 05 :

- Take each concept from distinct concept list for eg: Window, Chair, Picture and collect images from google.
- First set analysis, all images activate (1500 images)
- Second set analysis, all images activate (508 images)



ADE20K Dataset



Positive Images



Negative Images

Google Images



Neuron number 11 :

- Total number of images that got activated =
- Highest activation =
- Total number of positives =
- Total number of negatives =

794/1370 (1370= test_dataset)
17.6951
262 (images that has value ≥ 9)
250 (images that has value < 9)

Solution given by ECII analysis for neuron 11

solution 1: (:WN_Edifice)
 solution 2: (:WN_Building)
 solution 3: (:Building)
 solution 4: (:WN_Sky)
 solution 5: (:Sky)
 solution 6: (:WN_Road)
 solution 7: (:WN_Route)
 solution 8: (:Road)
 solution 9: (:WN_Tree)
 solution 10: ((:WN_Motorcar) \sqcap (:WN_Machine))
 solution 14: (:WN_Automobile)
 solution 17: ((:WN_Route) \sqcap (:WN_Building))
 solution 19: ((:WN_Automobile) \sqcap (:WN_Route))
 solution 24: (:Sidewalk)
 solution 25: (:WN_Pavement)



Distinct Concepts from the solution

Edifice(Building)
 Building
 Sky
 Road
 Route
 Tree
 Motorcar
 Machine
 Automobile
 Sidewalk
 Pavement

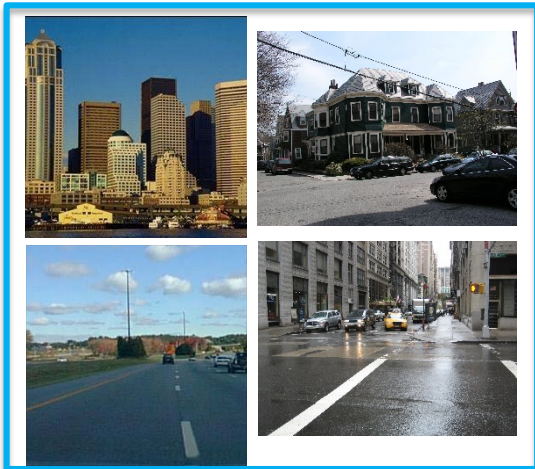
Results

Google analysis for Neuron number 11 :

- Take each concept from distinct concept list for eg: Building, Sky and collect images from google.
- First set analysis, all images activate (183 images)
- Second set analysis, all images activate (454 images)



ADE20K Dataset



Positive Images



Negative Images

Google Images



- **Neurosymbolic Artificial Intelligence**
- **Concept Induction**
- **Explainability Framework**
- **Explaining Hidden Neuron Activations**
- **Are Concept Induction Explanations Meaningful To Humans?**
- **Improving Deep Learning Through Concept Induction**

Reference for this section:

Cara Widmer, Md Kamruzzaman Sarker, Srikanth Nadella, Joshua Fiechter, Ion Juvina, Brandon Minnery, Pascal Hitzler, Joshua Schwartz, Michael Raymer
Towards Human-Compatible XAI: Explaining Data Differentials with Concept Induction over Background Knowledge, arXiv:2209.13710

Are the results human-compatible? Part I



- Hypothesis:
 - ECII explanations are better than semi-random explanations, but worse than human-generated explanations.
- Experimental setting as before.
- 300 Amazon Mechanical Turk participants
- Seven concepts taken from top ECII results.
- 45 image set pairs, each set corresponding to a category.



Which of these better represents what the images in group A have that the images in group B do not?

Bake, Bakery, Bread, Indoor, Product, Store, Woman

Basket, Bread, Cake, Ceiling, Floor, Person, Wall

Are the results human-compatible? Part I



A



B

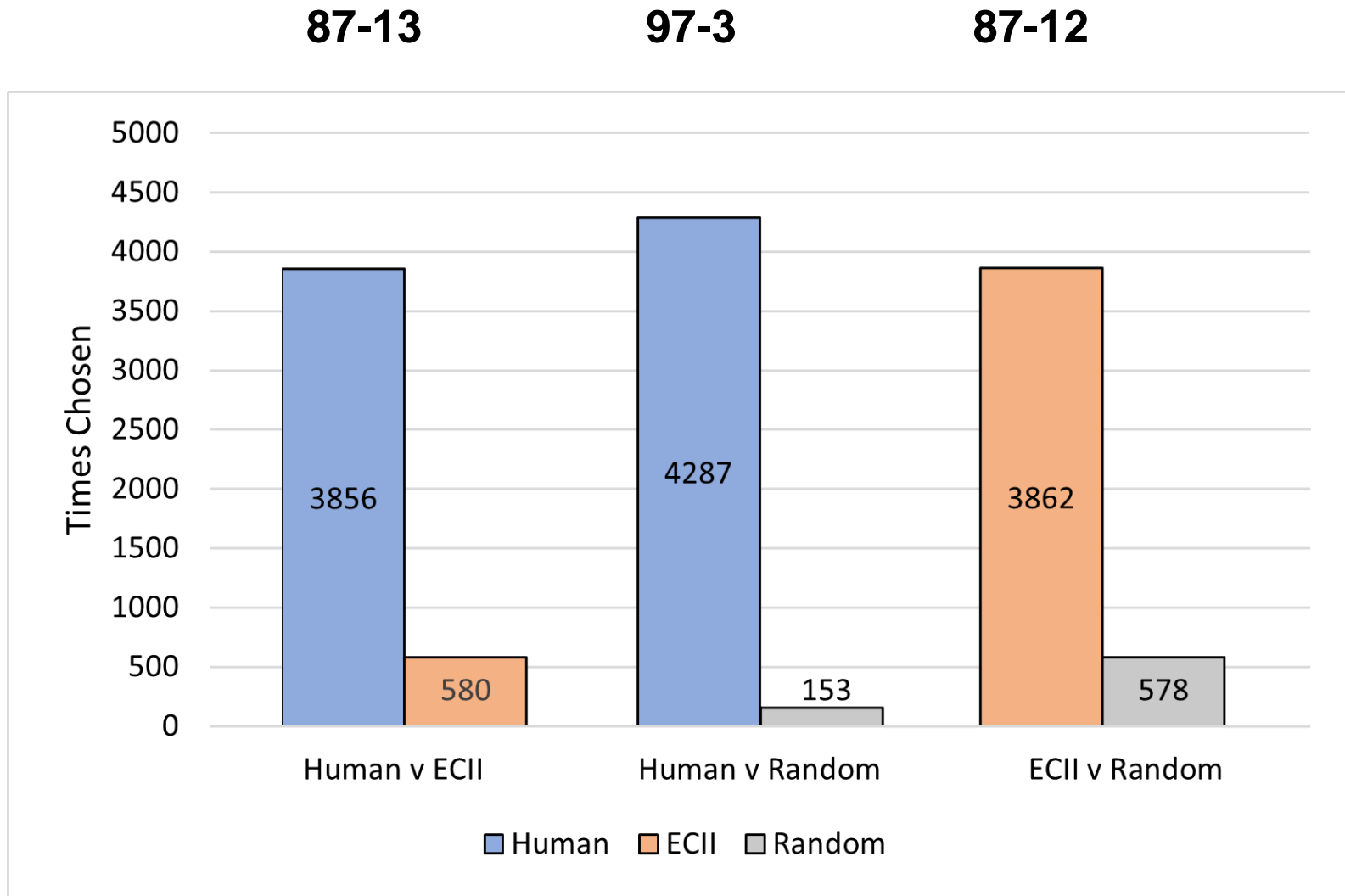


Which of these better represents what the images in group A have that the images in group B do not?

Bake, Bakery, Bread, Indoor, Product, Store, Woman

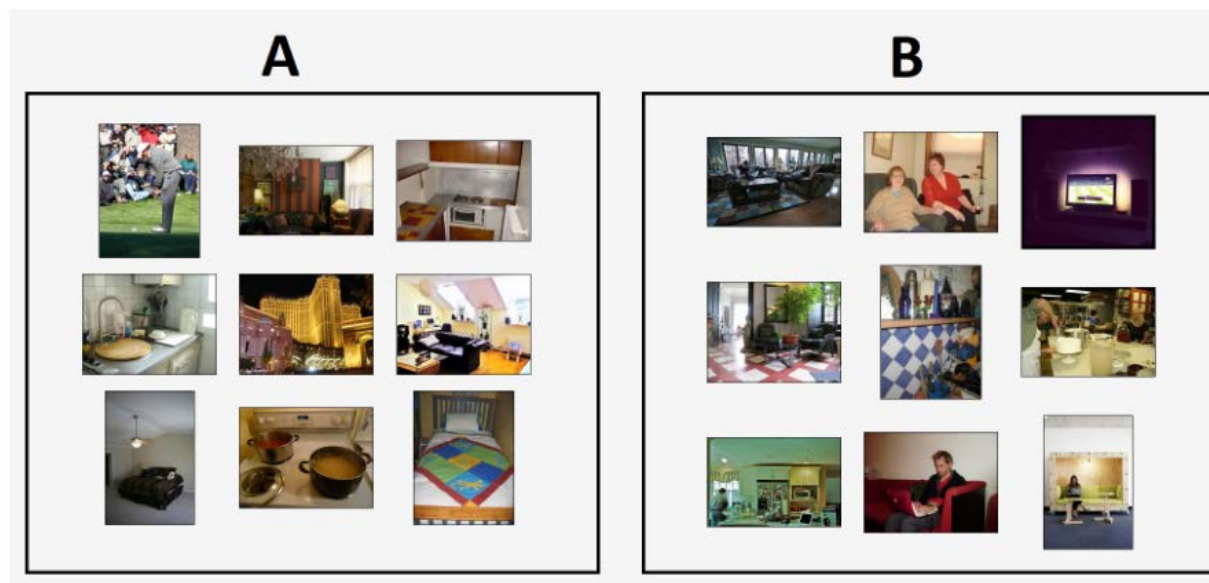
Basket, Bread, Cake, Ceiling, Floor, Person, Wall

Are the results human-compatible? Part I



Are the results human-compatible? Part II

- Hypothesis:
 - ECII explanations matched to correct images better than chance, but not as frequently as human generated explanations
- Experimental setting as before.
- 100 Amazon Mechanical Turk participants
- 16 image sets, from ML decision errors (logistic regression classifier)



Explanation: Home, Manufacturing, Clothing, Clothing Manufacturers, People, Chairs, Tableware

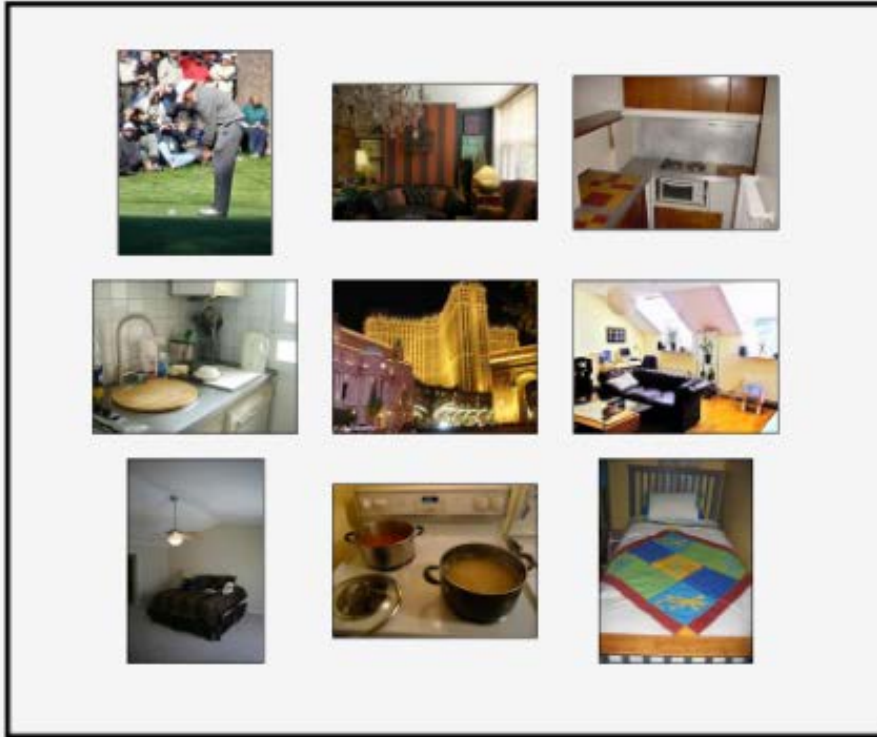
Which group of images do you think this explanation refers to?

Image Group A

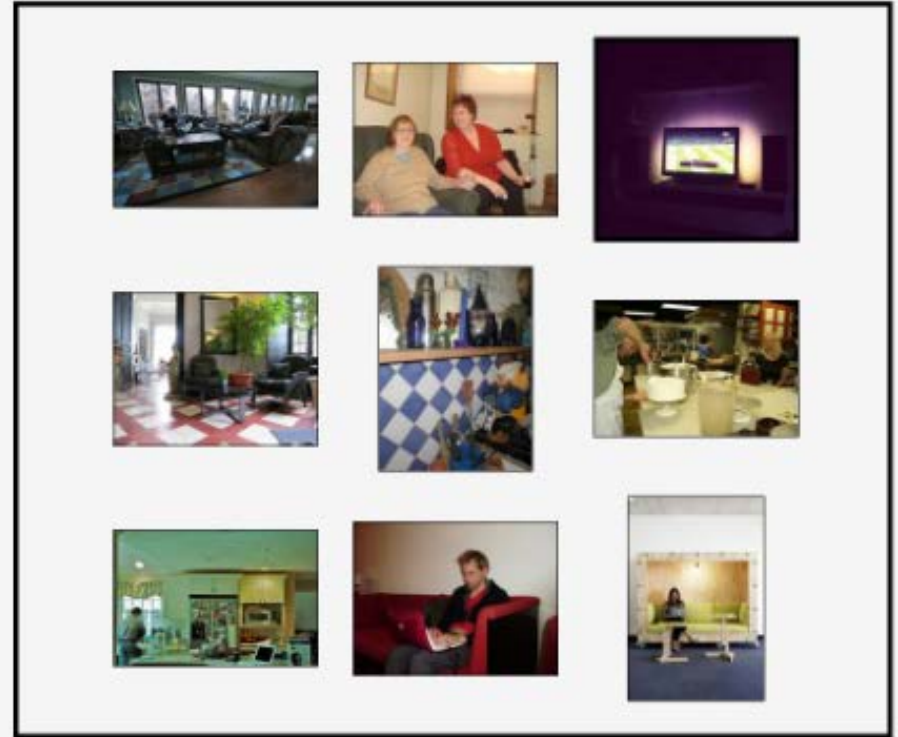
Image Group B

Are the results human-compatible? Part II

A



B



Explanation: Home, Manufacturing, Clothing, Clothing Manufacturers, People, Chairs, Tableware

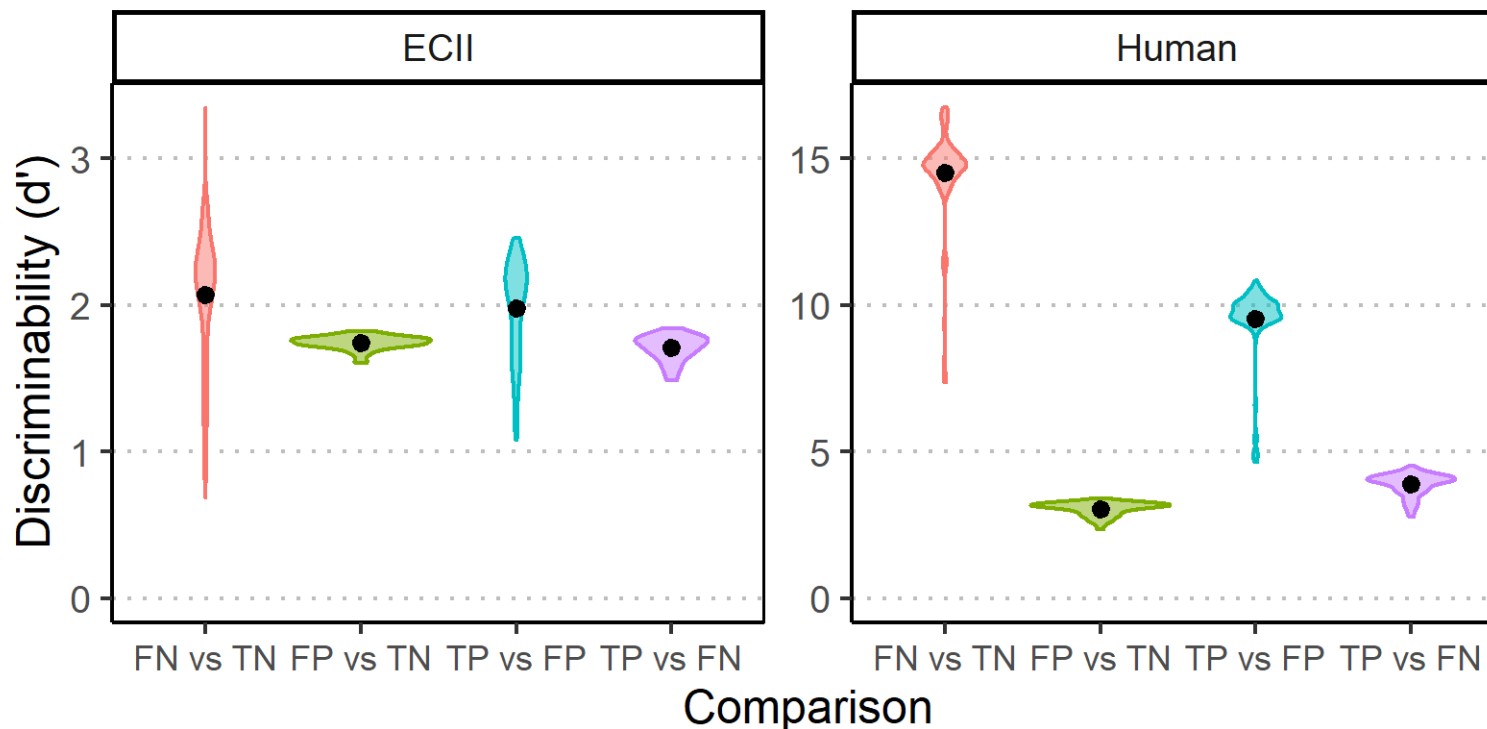
Which group of images do you think this explanation refers to?

Image Group A

Image Group B

Are the results human-compatible? Part II

- **Bayesian hierarchical signal-detection model (SDT)**
 - yields discriminability measure



- **Neurosymbolic Artificial Intelligence**
- **Concept Induction**
- **Explainability Framework**
- **Explaining Hidden Neuron Activations**
- **Are Concept Induction Explanations Meaningful To Humans?**
- **Improving Deep Learning Through Concept Induction**

Improving deep learning



Experimental set-up

- **Dataset : Twitter Dataset for toxicity analysis**
 - <https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data>
 - **Classes like “Lie, Dangerous, Insult”**
- **Language Model Used: Bert Base Model**
 - **12 layers**
 - **768 hidden layer neurons**
 - **110M parameters**

**Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova:
BERT: Pre-training of Deep Bidirectional Transformers for Language
Understanding. NAACL-HLT (1) 2019: 4171-4186**

Data examples – “Insult” class



- "Fiore, an occupation sympathizer..." This article makes me feel sick. An insult to Oregonians who have tolerated 41 days and more from this unwanted intrusion. An insult to the LE that put their lives and reputations at risk to resolve this. The mutual admiration between her and Bundy's counsel is to be expected.
correctly classified
- I'm not sure what you're trying to say, or what the source is of you're information you've implied is somehow not relevant to this article. Forget about mainstream media and the tired and over used commentary that dismiss all mainstream media and politicians making up canned rhetoric repeating it so often that easily manipulated people actually believe them. We all need to worry about individuals that have an ax to grind and make statements out of thin air, try to shock and change the subject on issues. There is racism in our country and it has been passed down from one generation to another but all good people with moral compasses will continue to work within the process by joining together for the rights of all human beings, we will all benefit and it has nothing to do with political sides blather or insults directed at media. We have options, as a society, our sources for information from credible research is unlimited. You may be looking for truth in all the wrong places.
incorrectly classified

Concept Induction Analysis



- **Run ECII on false positives vs. true positives**
- **Take first 20 results from ECII**
- **Get new examples that fall under all of the ECII classes**
- **Retrain with the additional examples**
 - **initial training set size: 10,000**
 - **retraining set size: 11,800**
 - **i.e. 18% added**

Does retraining improve classification?

Results before and after training



| Class | Accuracy (before) | Accuracy (after) | Precision (before) | Precision (after) | F-Measure (before) | F-Measure (after) | Recall (before) | Recall (after) |
|------------|-------------------|-------------------------|--------------------|--------------------------|--------------------|--------------------------|-----------------|-----------------------|
| Lie | 0.9483 | 0.9721 | 0.9333 | 0.9464 | 0.9589 | 0.9789 | 0.9859 | 0.9897 |
| Dangerous | 0.8731 | 0.8947 | 0.8485 | 0.8711 | 0.8682 | 0.8890 | 0.8889 | 0.9120 |
| Crazy | 0.8911 | 0.9105 | 0.8511 | 0.8784 | 0.8791 | 0.8962 | 0.9090 | 0.9465 |
| Corruption | 0.9455 | 0.9788 | 0.9167 | 0.9533 | 0.8800 | 0.9125 | 0.8462 | 0.8782 |
| Fool | 0.8983 | 0.9427 | 0.9483 | 0.9788 | 0.9016 | 0.9433 | 0.8594 | 0.9652 |
| Insult | 0.7813 | 0.8333 | 0.7885 | 0.8123 | 0.7961 | 0.8211 | 0.8039 | 0.8349 |

Results as communicated by Sulogna Chowdhury

Conclusions

- **While I presented only first data, it seems clear that concept induction for explainable deep learning can be made to work.**
- **explaining hidden neurons**
- **improving deep learning**
- **explanations are meaningful to humans**
- **We're looking into**
 - **consolidating the results**
 - **refining the approach**
 - **improving the concept induction approach**
 - **other application scenarios**





Thanks!

References

Pascal Hitzler, Md Kamruzzaman Sarker (eds.), **Neuro-Symbolic Artificial Intelligence – The State of the Art. Frontiers in Artificial Intelligence and Applications Vol. 342, IOS Press, Amsterdam, 2022.**



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. In: Pascal Hitzler, Md Kamruzzaman Sarker (eds.), Neuro-Symbolic Artificial Intelligence: The State of the Art. Frontiers in Artificial Intelligence and Applications Vol. 342, IOS Press, Amsterdam, 2022.**

Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, Pascal Hitzler
Neuro-Symbolic Artificial Integration: Current Trends
AI Communications 34 (3), 197-209, 2022.

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 d'Avila Garcez, Isaac Noble, Proceedings of the Twelfth International Workshop on Neural-Symbolic Learning and Reasoning, NeSy 2017, London, UK, July 17-18 2017. CEUR Workshop Proceedings Vol. 2003, 2017.**

References

Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, Md Kamruzzaman Sarker, Neural-Symbolic Integration and the Semantic Web. Semantic Web 11 (1), 2020, 3-11.



Cara Widmer, Md Kamruzzaman Sarker, Srikanth Nadella, Joshua Fiechter, Ion Juvina, Brandon Minnery, Pascal Hitzler, Joshua Schwartz, Michael Raymer, Towards Human-Compatible XAI: Explaining Data Differentials with Concept Induction over Background Knowledge. arXiv:2209.13710

Md Kamruzzaman Sarker, Pascal Hitzler, Efficient Concept Induction for Description Logics. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press 2019 , pp. 3036-3043.

References



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.

Pascal Hitzler, Semantic Web: A Review of the Field. Communications of the ACM 64 (2), 76-82, 2021.

Md Kamruzzaman Sarker, Joshua Schwartz, Pascal Hitzler, Lu Zhou, Srikanth Nadella, Brandon Minnery, Ion Juvina, Michael L. Raymer, William R. Aue, Wikipedia Knowledge Graph for Explainable AI. In: Boris Villazón-Terrazas, Fernando Ortiz-Rodríguez Sanju M. Tiwari, Shishir K. Shandilya (eds.), Knowledge Graphs and Semantic Web. Second Iberoamerican Conference and First Indo-American Conference, KGSWC 2020, Mérida, Mexico, November 26-27, 2020, Proceedings. Communications in Computer and Information Science, vol. 1232, Springer, Heidelberg, 2020, pp. 72-87.