WBS and Rothko

Explainable, Interpretable AI:

The Future of Investment Management

19 Nov 2021

On the need for knowledge extraction from deep networks

Artur d'Avila Garcez

City, University of London

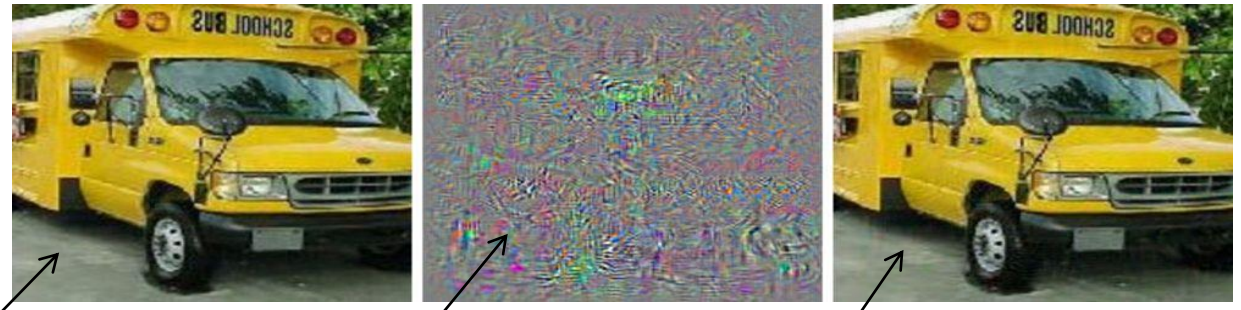
a.garcez@city.ac.uk

Outline

1. Why knowledge extraction is needed?
2. Measuring XAI
3. Global and local extraction
4. CLEAR method
5. Next Steps

Deep Networks' success: image, text, audio, video, etc.

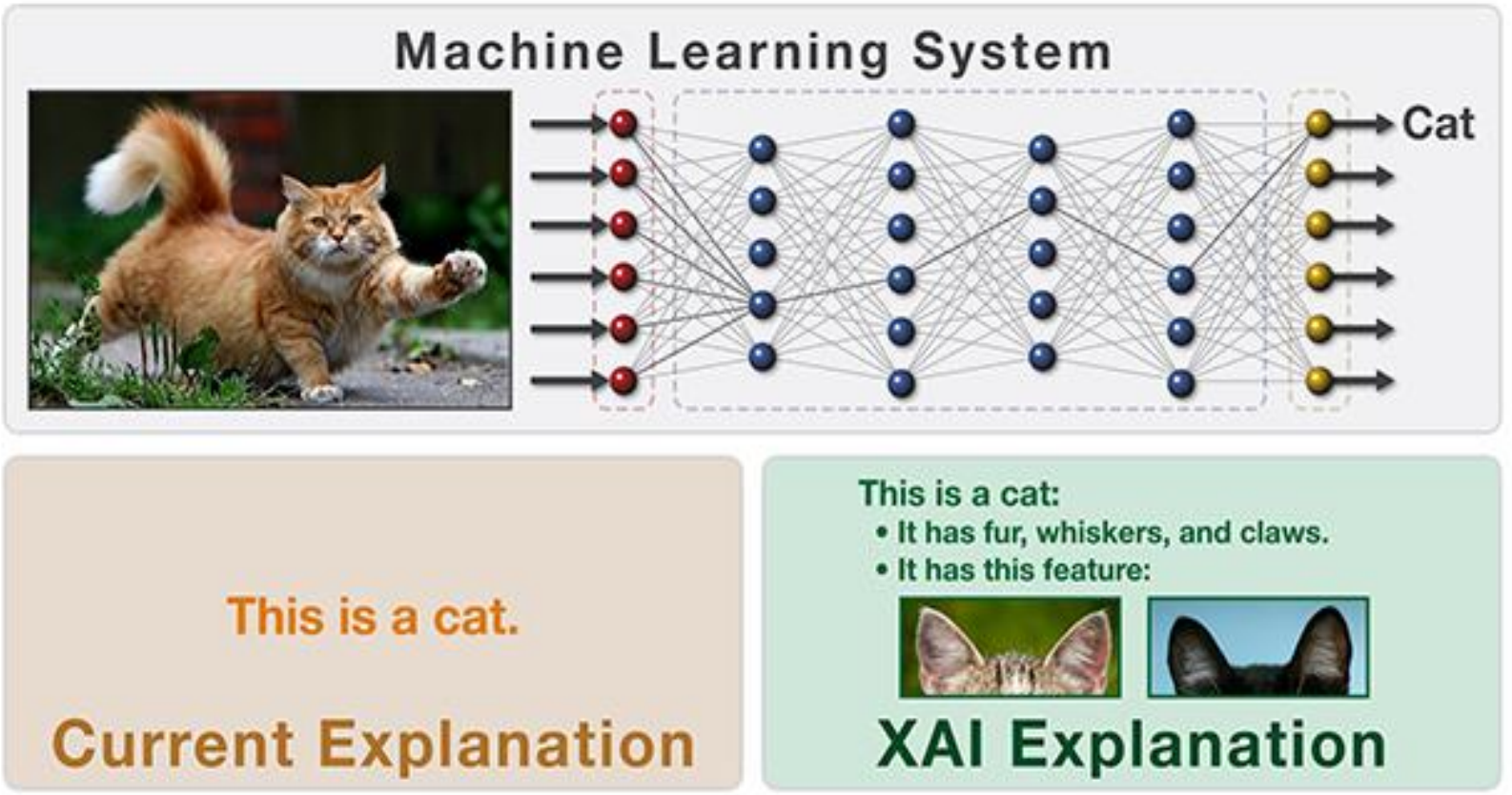
Deep Networks' problems: robustness, energy consumption, fairness, trust!



school bus + adversarial perturbation = ostrich

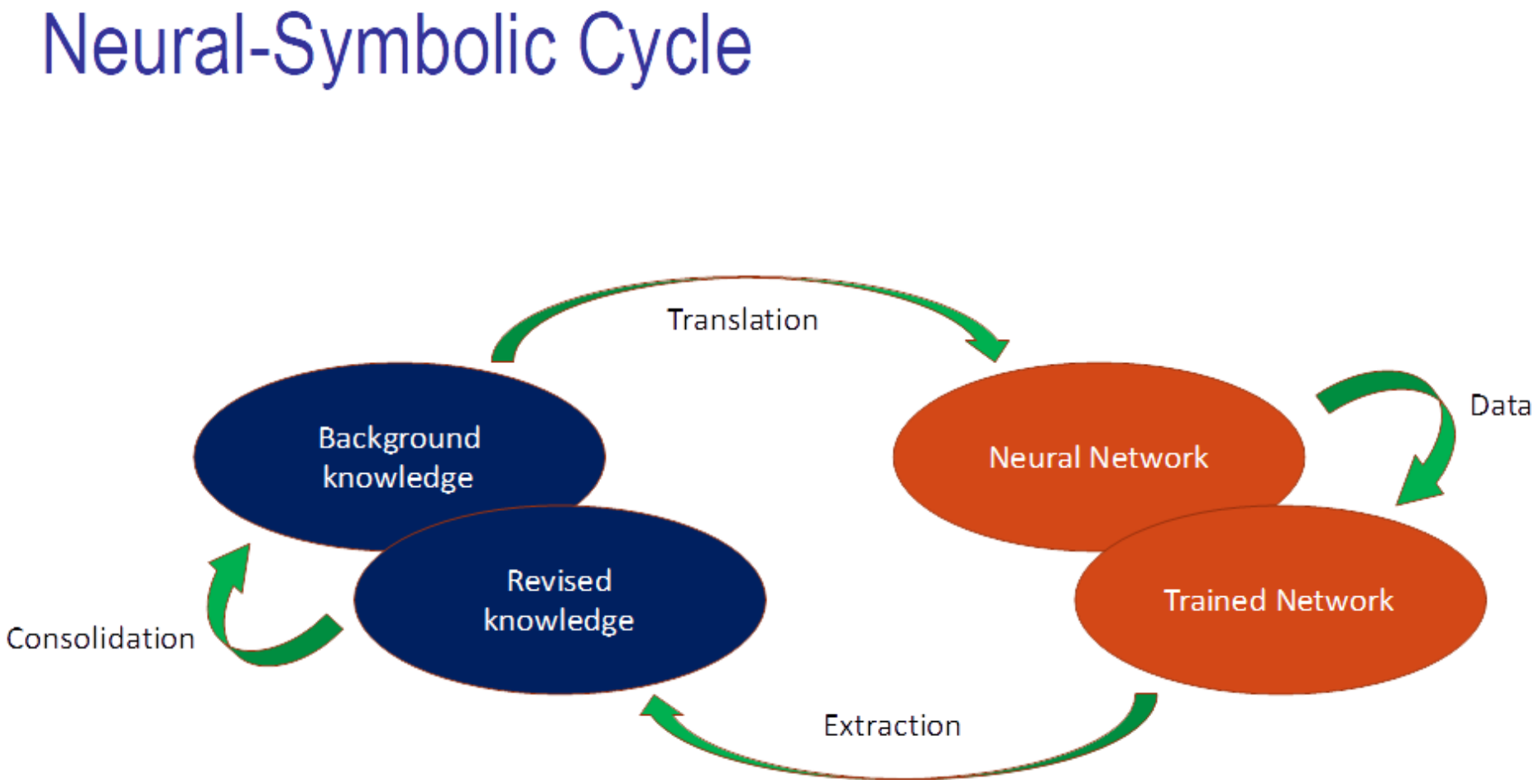
https://arxiv.org/abs/1312.6199

DARPA's Explainable AI (XAI)



- XAI = Interpretable ML
- Explanation = Knowledge extraction, not XAI

Knowledge Extraction: an integral part of Neurosymbolic AI cycle



A. d'Avila Garcez and L. Lamb, Neurosymbolic AI: The 3rd Wave, Dec 2020
https://arxiv.org/abs/2012.05876

Knowledge Extraction (benefits)

Proofs of soundness

Proof history (goal-directed reasoning)

Levels of abstraction (modularity)

Transfer learning (analogy)

System maintenance/improvement

Knowledge Extraction techniques

- Soundness is important
- Pedagogical, Decompositional, Eclectic
- Early methods: MofN, CILP
- Decision tree extraction – TREPAN, ERIC
- Recurrent networks - Finite Automata extraction
- Current work: layerwise extraction, soft decision trees, probabilistic MofN, distillation, network querying (Logic Tensor Networks)

Fidelity

For every complex problem there is an answer that is clear, simple, and wrong

H. L. Mencken

Fidelity = accuracy of symbolic knowledge extracted w.r.t. trained model, not data

Measure of fidelity should precede user studies!

Recent application in industry: Reducing harm from gambling

- Playtech plc system is required to provide explanations to the regulator, gambling operators and to the player.
- Neural nets and Random Forests performed considerably better than logistic regression and Bayesian nets.
- Extracted decision trees and risk curves have been shown to help debug the system and improve predictions.

Knowledge Extraction (challenges)

Computational complexity and Comprehensibility

Global knowledge extraction is a hard problem; networks now have millions or billions of parameters…

Thus, **local** XAI methods, which seek to explain each case rather than the entire model (LIME, SHAP), have an important role to play in practice.

But local knowledge extraction also require a measure of fidelity and an equation to explain alternative outcomes!

CLEAR: Counterfactual Local Explanations via Regression

Measurable (fidelity)

Counterfactual (equation)

Contrastive (when applied to images)

Measurable counterfactual local explanations for any classifier. A White, A d'Avila Garcez, In Proc ECAI, Aug 2020. preprint arXiv:1908.03020.

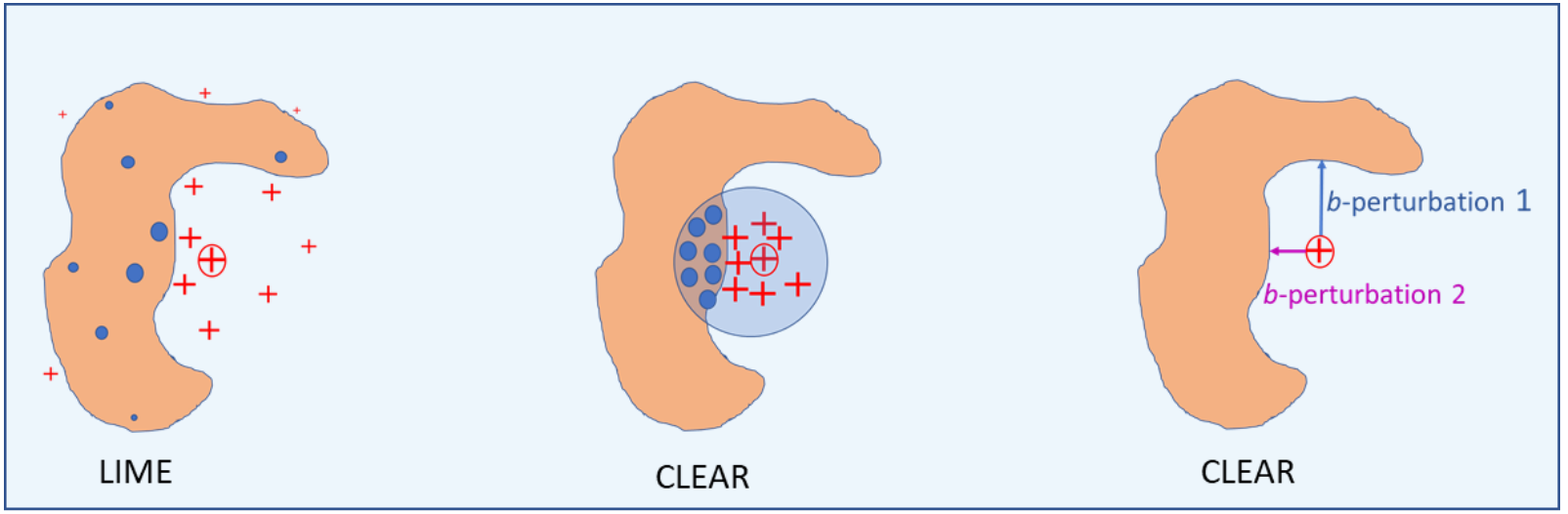
Counterfactual Instances Explain Little. A White and A d'Avila Garcez. arXiv:2109.09809, Sep 2021.

Contrastive Counterfactual Visual Explanations With Overdetermination. A. White et al., Sep 2021. https://arxiv.org/abs/2106.14556

The CLEAR Project (1)

The *CLEAR* Method:

1. Determine $\boldsymbol{x}$'s actual perturbations.
2. Generate synthetic observations that are then labelled by the AI system.
3. Create a balanced neighbourhood data set (including counterfactuals from step 1).
4. Perform a step-wise regression including 2nd degree terms and interaction terms.



Joint work with Adam White

The CLEAR Project (2)

5. Estimate perturbation:

MLP on toy diabetes dataset:

$\boldsymbol{x}$ = {Glucose = 0.537, BloodPressure = 3.1}, $P_{\text{MLP}}(\boldsymbol{x} \in class1)$= 0.69

CLEAR generates a logistic regression equation (step 4):

$$(1 + e^{-\boldsymbol{w}^T \boldsymbol{x}})^{-1} = 0.69$$

$\boldsymbol{w}^T \boldsymbol{x}$ = -0.8 + 1.73 Glucose + 0.25 BloodPressure + 0.31 Glucose^2

For $\boldsymbol{x}$ to be on the decision boundary ($\boldsymbol{w}^T \boldsymbol{x}$ = 0), the estimated perturbation is obtained by substituting the value of BloodPressure in $\boldsymbol{x}$:

$$0.31 \text{ Glucose}^2 + 1.73 \text{ Glucose} - 0.025 = 0$$

Glucose = 0.014

Estimated perturbation = 0.014 – 0.537 = -0.523

The CLEAR Project (3)

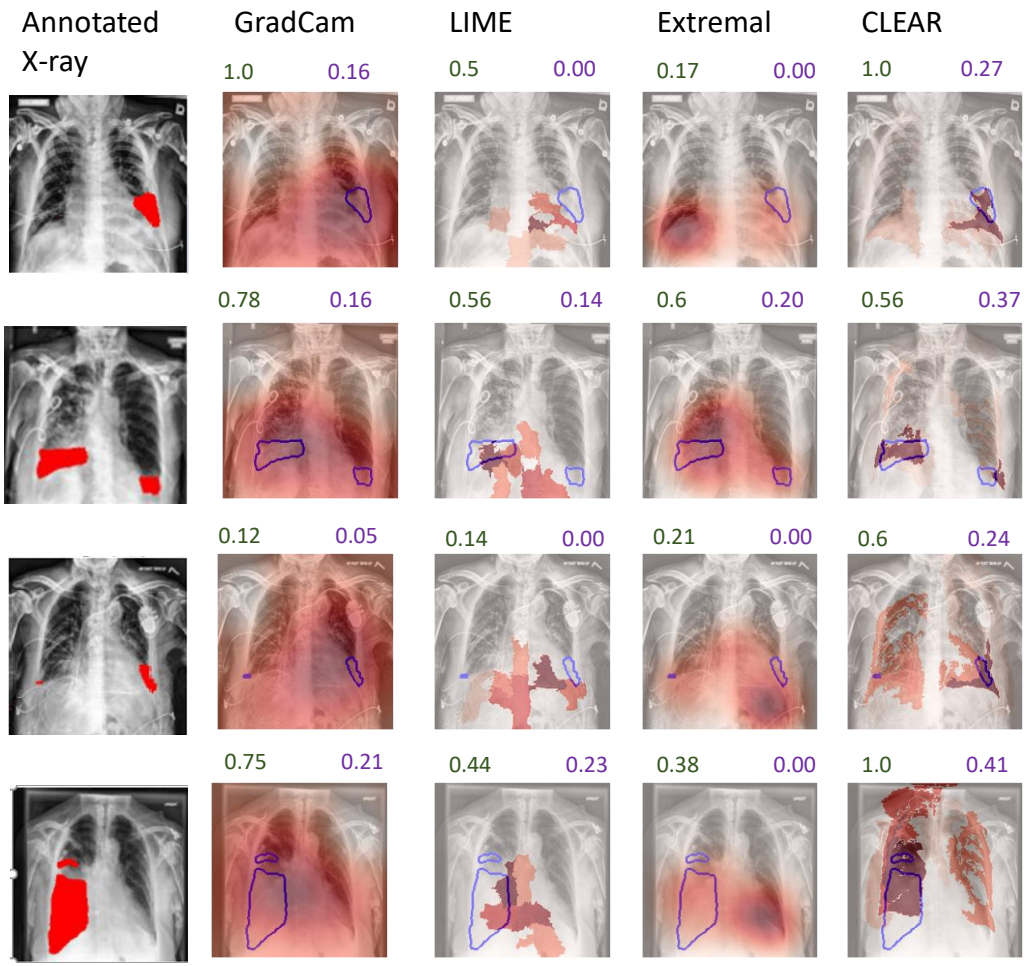
Measure the fidelity of the regression coefficients:

estimated perturbation (step 5) = -0.523

actual perturbation (step 1) = -0.557

fidelity error = | estimated – actual | = 0.034

CLEAR Application (chest x-rays)

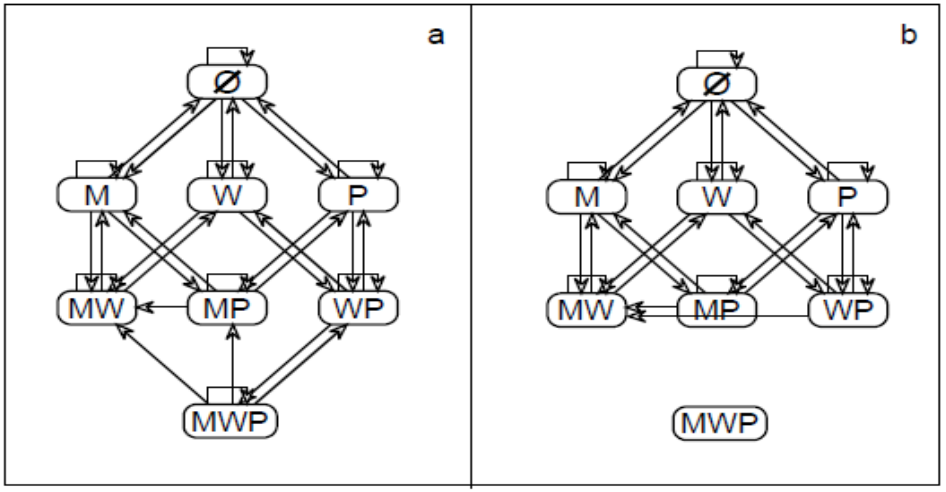


Code available from:
https://github.com/ClearExplanationsAI/CLEAR

Recurrent networks

- Extraction of state transition diagrams...

  CrMeth = M (level of methane is critical)
  HiWat = W (level of water is high)
  PumpOn = P (pump is turned on)

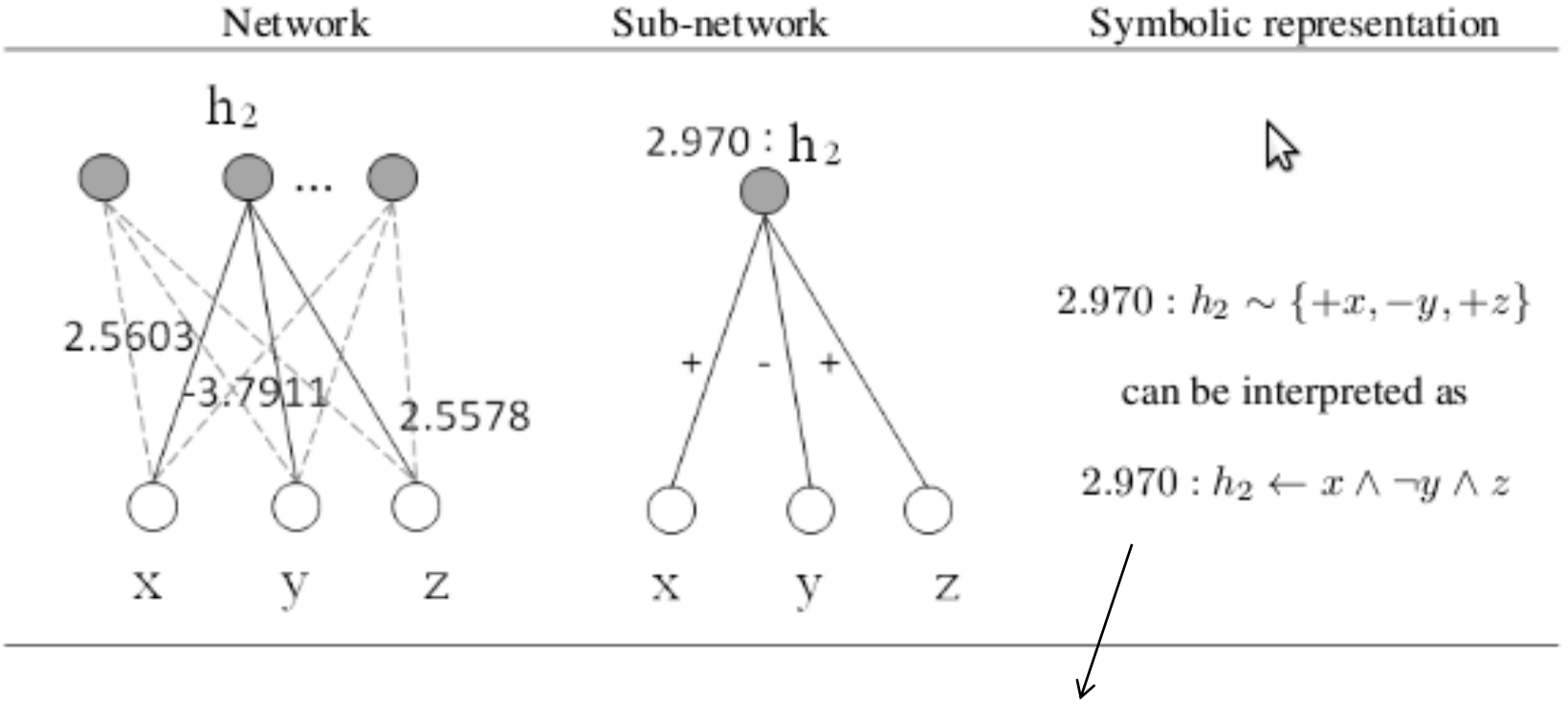


Extracting Automata from Recurrent Neural Networks Using Queries and Counterexamples, Gail Weiss, Yoav Goldberg, Eran Yahav, 2017
https://arxiv.org/abs/1711.09576

Learning and Representing Temporal Knowledge in Recurrent Networks, Rafael V. Borges, Artur d'Avila Garcez, Luis C. Lamb, IEEE TNNLS, 2011

Extraction from RBMs and DBN

Knowledge extraction from RBMs (originally the building block of (modular) deep nets, c.f. Hinton's Deep Belief Nets)



Each rule has a confidence value ∑||w||/n

S. Tran and A. S. d'Avila Garcez. Deep Logic Networks: Inserting and Extracting Knowledge from Deep Belief Networks. IEEE TNNLS, Nov, 2016

Layerwise Extraction and Probabilistic MofN

- We can improve the accuracy of rules extracted from RBMs by extracting MofN rules
- Search values for M given extracted rules, e.g. M=0,1,2,3 in

$$2.970 : h_2 \leftarrow \text{M of } \{x, \sim y, z\}$$

Extracting M of N Rules from Restricted Boltzmann Machines. Simon Odense and Artur d'Avila Garcez, ICANN 2017.

Layerwise Knowledge Extraction from Deep Convolutional Networks. Simon Odense and Artur d'Avila Garcez, arXiv:2003.09000 [cs.AI], March 2020.

ERIC: Extracting Relations Inferred from Convolutions. Joe Townsend, Theodoros Kasioumis, Hiroya Inakoshi, arXiv:2010.09452 [cs.LG], Oct 2020.

Logic Tensor Networks (LTN)

- Neural nets with rich structure can represent more than classical propositional logic
- But neural nets are essentially propositional (i.e. do not use variables explicitly)
- To take advantage of full FOL, a hybrid approach is needed whereby the logical statements act as soft **constraints** on the network...

Neural-symbolic integration for fairness in AI. B Wagner, A. d'Avila Garcez. AAAI Spring Symposium, March 2021. CEUR Proceedings 2846.

LTN: Semantic Image Interpretation

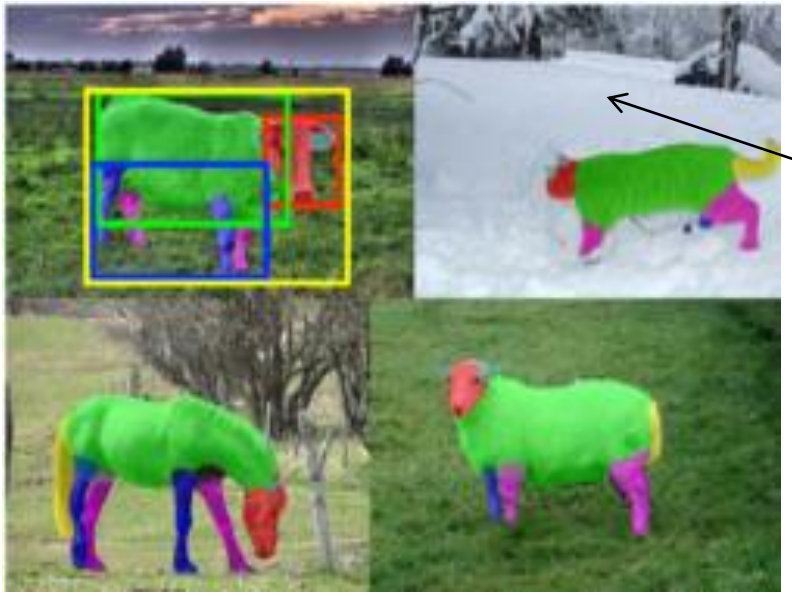
Given a picture extract a graph that describes its semantic content

Normally, every cat has a tail

Q. Get me the red thing next to the sheep

A. The horse's muzzle? Yes.

$$\forall xy(\mathsf{partOf}(x,y) \rightarrow \neg \mathsf{partOf}(y,x))$$



Make sure your system does not distinguish cats from wolves 99% correctly just because of the snow in the background...

I. Donadello, L. Serafini and A. S. d'Avila Garcez. Logic Tensor Networks for Semantic Image Interpretation. In Proc. IJCAI'17, Melbourne, Aug 2017.

Current Challenges/Opportunities

- Extraction of meaningful concepts and concept relations from large networks.
- Reuse of extracted concepts and relations in OOD transfer problems.
- Extraction of FOL from neural nets at different levels of abstraction (e.g. normative rules)
- Explaining time series, CNNs, LSTMs.