Explainable Deep Learning

and its potential use in TSF

Tirtharaj Dash

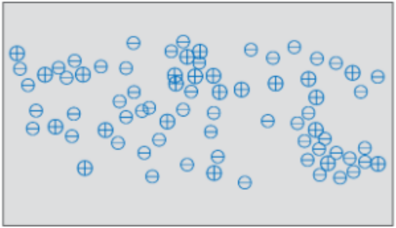
University of Cambridge, UK

September 28, 2024

Today

- Explainability and Deep Learning
- “Logically” explainable Deep Neural Nets
- Application in TSF, esp. Healthcare

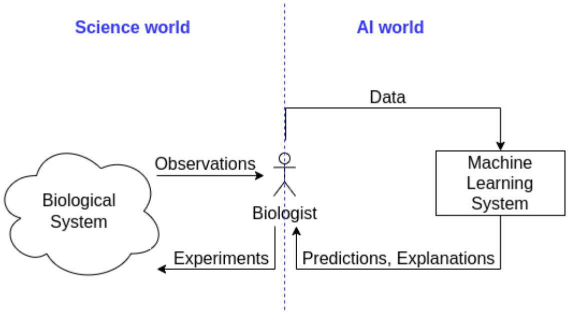
Explainability and Deep Neural Nets



A classical ML problem: classification.

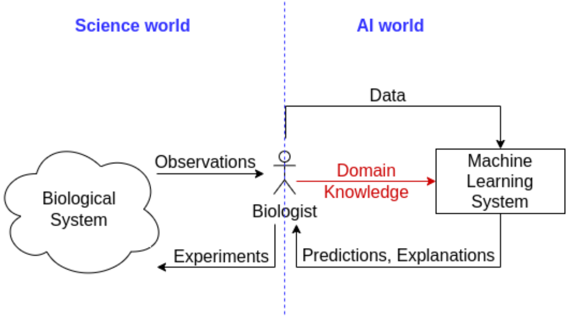
Explainability and Deep Neural Nets

Classical ML:



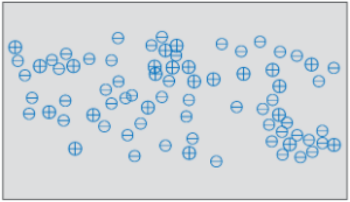
Explainability and Deep Neural Nets

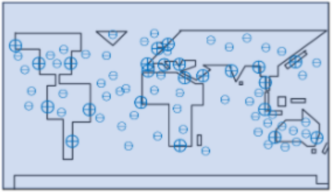
Human-in-the-Loop ML:



Background knowledge comes from domain-expert(s).

Explainability and Deep Neural Nets





The $\oplus$ points are the port cities, inferrable from boundaries.

Muggleton, Srinivasan, Bain: Compression, significance and accuracy, *ML Proceedings*, 1992.

Explainability and Deep Neural Nets

Inclusion of domain-knowledge into deep neural networks significantly improves its predictive performance.

DNN Type	Comparative Performance (with domain-knowledge)		
	Better	Same	Worse
MLP	71	0	2
GNN	63	9	1

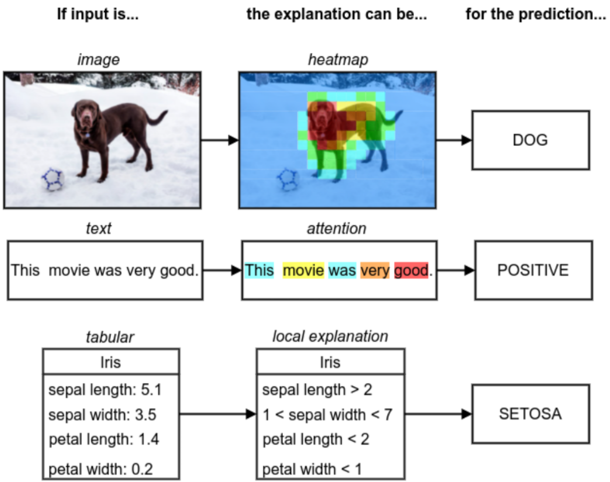
Dash et al.: *MLJ* (2021, 2022, 2023), *ILP* (2018, 2021), *Sci.Rep.* (2022).

Dash, *PhD Thesis*, BITS Pilani, 2022.

Explainability and Deep Neural Nets

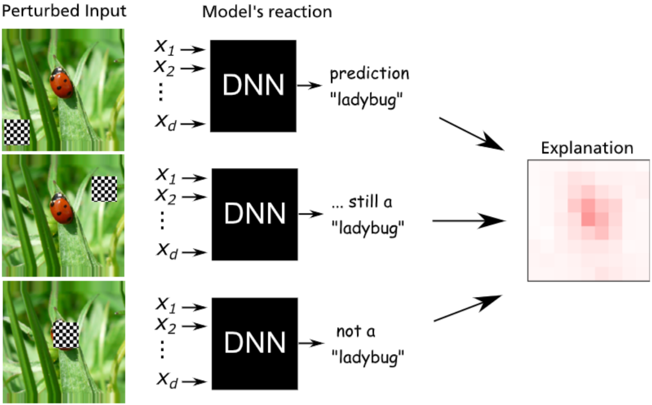
Explainability or Interpretability is the concept that a machine learning model and its output can be explained in a way that "makes sense" to a human being at an acceptable level.

Explainability and Deep Neural Nets



Ras et al.: Explainable deep learning: A field guide for the uninitiated, *JAIR*, 2022.

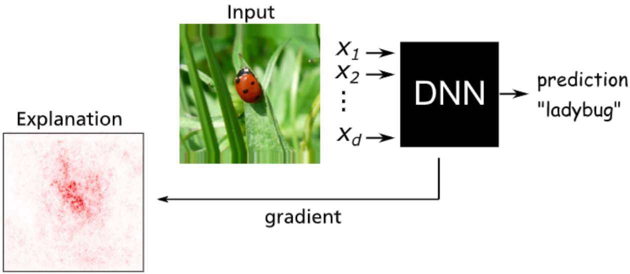
Explainability and Deep Neural Nets



Perturbation-based explanation generation.

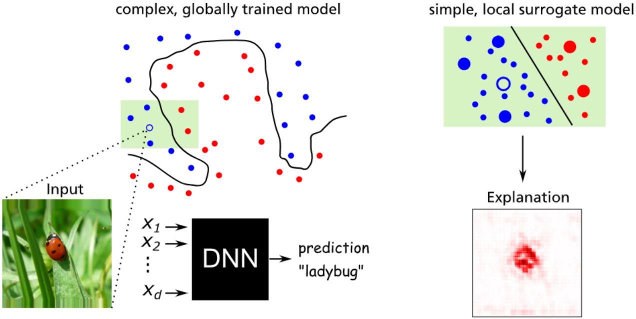
Samek et al.: Explainable deep learning: concepts, methods, and new developments. In Explainable Deep Learning AI (pp. 7-33). Academic Press, 2023.

Explainability and Deep Neural Nets



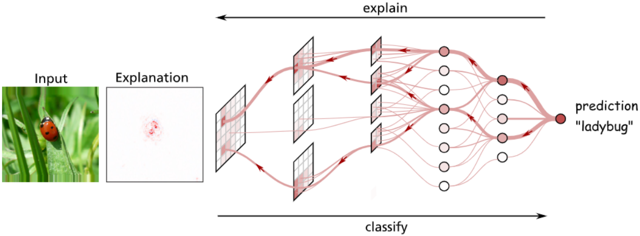
Gradient-based explanation generation.

Explainability and Deep Neural Nets



Surrogate-based explanation generation.

Explainability and Deep Neural Nets



Relevance propagation based explanation generation.

Explainability and Deep Neural Nets

So, what did we see?

Explainability and Deep Neural Nets

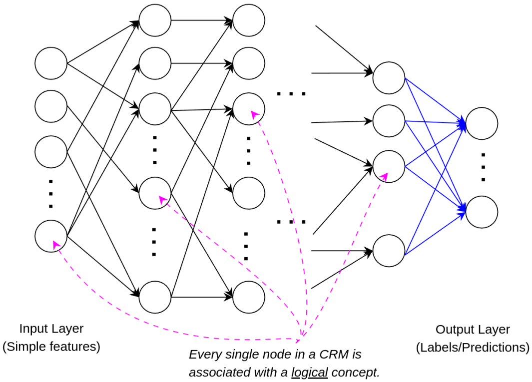
Methods to explain "black-box" deep networks *post hoc*.

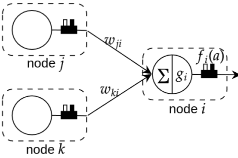
Explainability and Deep Neural Nets

Q1. But, is a deep neural network "truly" explainable?

Q2. Can we build deep networks that are "explainable by design"?

Logically explainable DNNs





$$h_i(a) = \begin{cases} f_i(a) & \text{if } v_i \in I \\ f_i(a) g_i\left(\sum_{(v_j,v_i)\in E} w_{ji} h_j(a)\right) & \text{if } v_i \notin I \end{cases}$$

CRM nodes as Gated nodes

Logically explainable DNNs

Relational Features:

A relational feature takes a clausal form:

$$C : \forall X \; (p(X) \leftarrow \exists \boldsymbol{Y} \; Body(X, \boldsymbol{Y}))$$

or,

$$C : (p(X) \leftarrow Body(X, \boldsymbol{Y}))$$

Here,

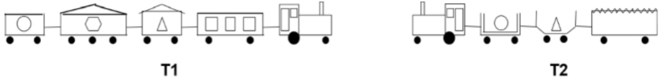
$$\begin{aligned} p(X) &: \text{Head literal} \\ Body(X, \boldsymbol{Y}) &: \text{Conjunction of body literals} \end{aligned}$$

Assumption: C is not self-recursive. We call C a "feature-clause".

Logically explainable DNNs

Feature clauses:

Let's look at the classic trains problem:



Some feature-clauses for trains:

$$C_1 : p(X) \leftarrow (has\_car(X, Y), short(Y))$$
$$C_2 : p(X) \leftarrow (has\_car(X, Y), short(Y), closed(Y))$$
$$C_3 : p(X) \leftarrow (has\_car(X, Y), has\_car(X, Z), short(Y), closed(Z))$$

The predicates $has\_car/2$, $short/1$, $closed/1$, etc. are defined as part of the background knowledge (B) about trains.

Logically explainable DNNs

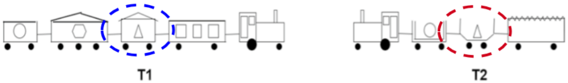
Feature functions:

A feature function is defined, for $X = a$ as:

$$f_{C,B}(a) = \begin{cases} 1 & \text{if } B \cup (C\{X/a\}) \models p(a) \\ 0 & \text{otherwise} \end{cases}$$

Simply, for a feature-clause C_i, we refer to the corresponding feature-function as $f_i(X)$.

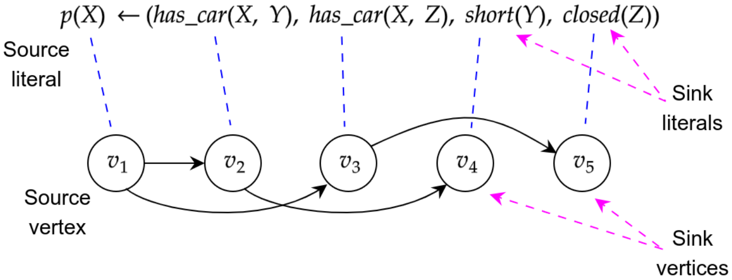
Example:



Some feature functions are: $f_1(t_1) = 1, f_2(t_1) = 1, f_2(t_2) = 0$.

Logically explainable DNNs

Ordered Clause:
We impose an ordering of the literals in a clause. If C is a clause of the form $\lambda_1 \leftarrow \lambda_2, \ldots, \lambda_k$, then the ordered clause is:
$\langle C \rangle = \langle \lambda_1, \lambda_2, \ldots, \lambda_k \rangle$.



Logically explainable DNNs

ρ-derivation of feature-clauses (Composition):

Example 1:

$$
\begin{aligned}
p(X) \leftarrow\ & has\_car(X,\ Y), short(Y),\\
& has\_car(X,\ Z), closed(Z),\\
& Y = Z
\end{aligned}
$$

$$\uparrow$$

$$\boxed{\rho_1}$$

$$\uparrow$$

$$
\begin{aligned}
p(X) \leftarrow\ & has\_car(X,\ Y), short(Y),\\
& has\_car(X,\ Z), closed(Z)
\end{aligned}
$$

$$\uparrow$$

$$\boxed{\rho_2}$$

$$\nearrow \qquad \nwarrow$$

$$\mathrm{C1}:\ p(X) \leftarrow has\_car(X, Y), short(Y) \qquad \mathrm{C2}:\ p(X) \leftarrow has\_car(X, Z),\ closed(Z)$$

Logically explainable DNNs

Example 2:

$$p(X) \leftarrow has\_car(X,\ U),\ has\_car(X,\ V),\ smaller(U,\ V),\ has\_car(X,\ Y), short(Y),\ U = V,\ U = Y$$

$$\uparrow$$

$$\boxed{\rho_1}$$

$$\uparrow$$

$$p(X) \leftarrow has\_car(X,\ U),\ has\_car(X,\ V),\ smaller(U,\ V),\ has\_car(X,\ Y), short(Y),\ U = V$$

$$\uparrow$$

$$\boxed{\rho_1}$$

$$\uparrow$$

$$p(X) \leftarrow has\_car(X,\ U),\ has\_car(X,\ V),\ smaller(U,\ V),\ has\_car(X,\ Y), short(Y)$$

$$\uparrow$$

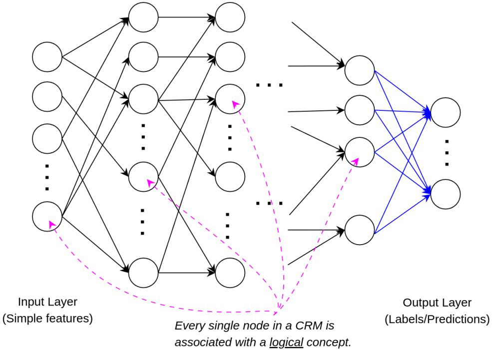
$$\boxed{\rho_2}$$

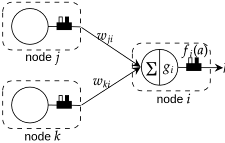
$$\nearrow \qquad \nwarrow$$

$$\mathrm{C1}:\ p(X) \leftarrow has\_car(X, Y), short(Y) \qquad \mathrm{C2}:\ p(X) \leftarrow has\_car(X, U),\ has\_car(X, V),\ smaller(U, V)$$

Logically explainable DNNs

Compositional Relational Machines (CRMs):





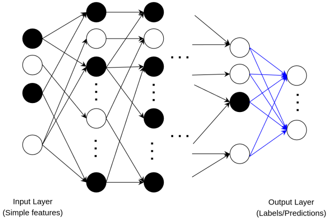
$$h_i(a) = \begin{cases} f_i(a) & \text{if } v_i \in I \\ f_i(a) g_i\left(\sum_{(v_j, v_i) \in E} w_{ji} h_j(a)\right) & \text{if } v_i \notin I \end{cases}$$

CRM nodes as Gated nodes

Logically explainable DNNs

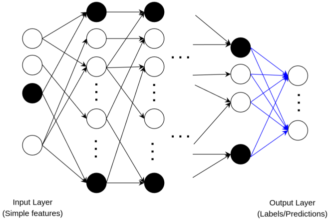
Relational instance 1:

T1



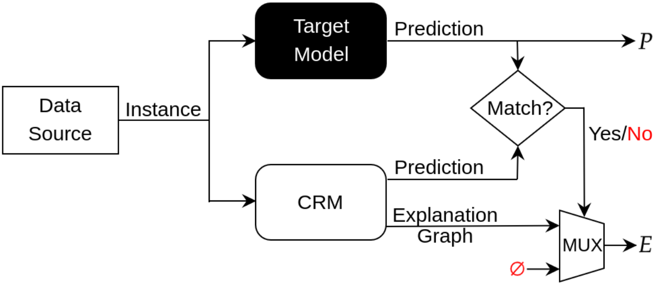
Relational instance 2:

T2



Logically explainable DNNs

Evaluation: (a) Predictive fidelity, (b) Explanatory fidelity



Explanation: Constructed by back-tracing the top activations in each layer of the deep neural network.

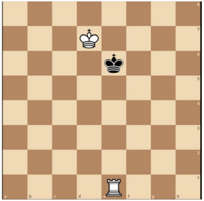
Logically explainable DNNs

(A) Synthetic datasets (Trains and Chess)

- Target theory (model) known

Trains

Chess

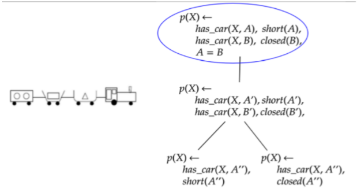


- Results

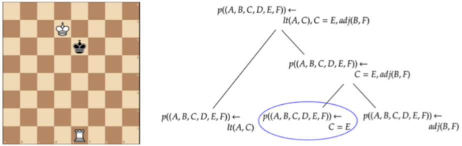
Dataset	Fidelity			
	CRM		Baseline	
	Pred.	Expl.	Pred.	Expl.
Trains	1.0	1.0	0.5	0.4
Chess	1.0	0.9	0.7	0.7

Logically explainable DNNs

- Some explanations generated by the CRM:



Train t_1 (With the substitution $\{X/t1\}$)



Board $(d, 7, e, 1, e, 6)$ (With the substitution $\{A/d, B/7, \ldots, F/6\}$)

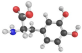
<u>Target theory:</u> Train X has a car Y and Y is short and closed.

<u>Target theory:</u> White Rook and Black King are on the same file (column).

Logically explainable DNNs

(B) Real datasets (drug design: NCI GI50; $n = 10$)

- Target theory is not known. BotGNNs are used as the target models [Dash et al., *MLJ*, 2022].

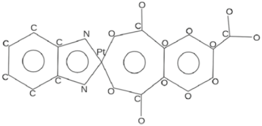
( , class)

- Results: Predictive fidelity

Dataset	CRM	Baseline
786_0	0.77	0.53
A498	0.79	0.59
A549_ATCC	0.85	0.63
ACHN	0.73	0.58
BT_549	0.78	0.51
CAKI_1	0.81	0.69
CCRF_CEM	0.82	0.68
COLO_205	0.77	0.53
DLD_1	0.90	1.00
DMS_114	0.89	0.91
Avg.	0.81 (0.05)	0.66 (0.17)

Logically explainable DNNs

- An explanation generated by CRM:



$p(X) \leftarrow$
has_struc$(X, A, B,$ *hetero_aromatic*$)$,
lteq$(B, 7)$,
has_struc$(X, C, D,$ *amine*$)$,
lteq$(D, 2)$,
has_struc$(X, E, F,$ *oxide*$)$,
lteq$(F, 7)$

$p(X) \leftarrow$
has_struc$(X, A, B,$ *amine*$)$,
lteq$(B, 2)$,
has_struc$(X, C, D,$ *oxide*$)$,
lteq$(D, 7)$

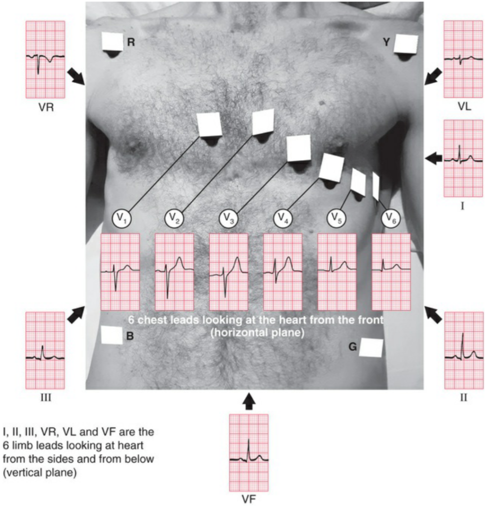
$p(X) \leftarrow$
has_struc$(X, A, B,$ *hetero_aromatic*$)$,
lteq$(B, 7)$

$p(X) \leftarrow$
has_struc$(X, A, B,$ *amine*$)$,
lteq$(B, 2)$

$p(X) \leftarrow$
has_struc$(X, A, B,$ *oxide*$)$,
lteq$(B, 7)$

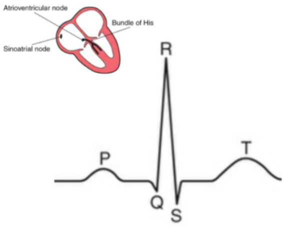
Srinivasan, Baskar, Dash, Shah: *MLJ*, 2023.

XDL in Healthcare TSF



Lead positions for a 12-lead ECG with 12 views of the heart

XDL in Healthcare TSF

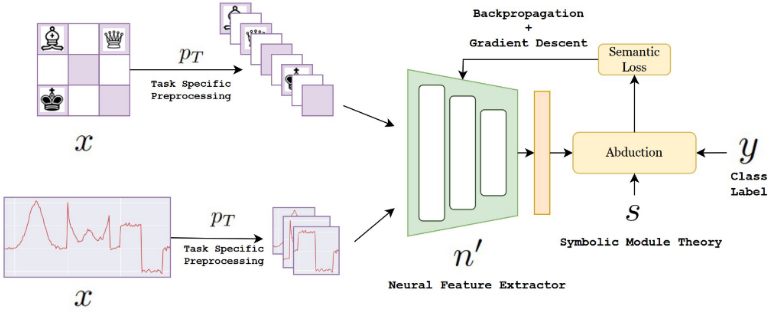


R	Rate	What is the rate (measured in beats per minute [bpm])?
R	Rhythm	What is the rhythm?
P	P wave	Is there one P wave before every QRS complex?
W	Width	Is the width of the QRS complex normal (< 3 small squares)?
Q	Q wave	Are there any deep Q waves present?
S	ST segment	Is there ST segment depression or elevation?
T	T wave	Are there any abnormal inverted (upside down) T waves?

ECG complex - 1 heartbeat

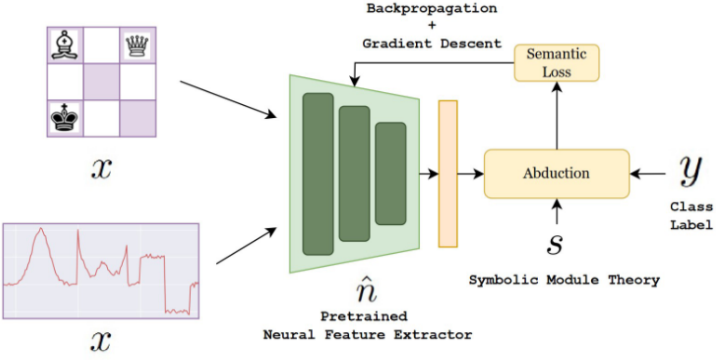
How doctors see it?

XDL in Healthcare TSF



Shah et al.: NEUROLOG, *arXiv*, 2022.

XDL in Healthcare TSF



Thank you!

tirtharajdash.github.io