“Logically” Explainable Deep Networks

and an application in drug discovery

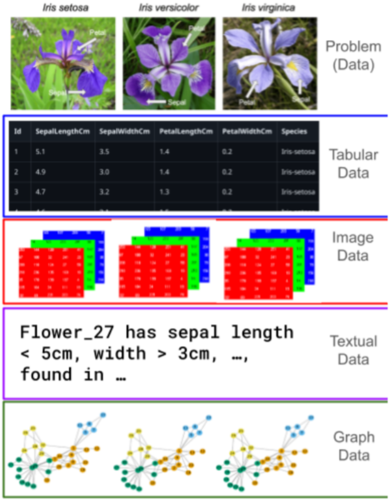
Tirtharaj Dash

Susanne Bornelöv Lab, CRUK Cambridge Institute
University of Cambridge, UK

December 5, 2024

Problem
(Data)

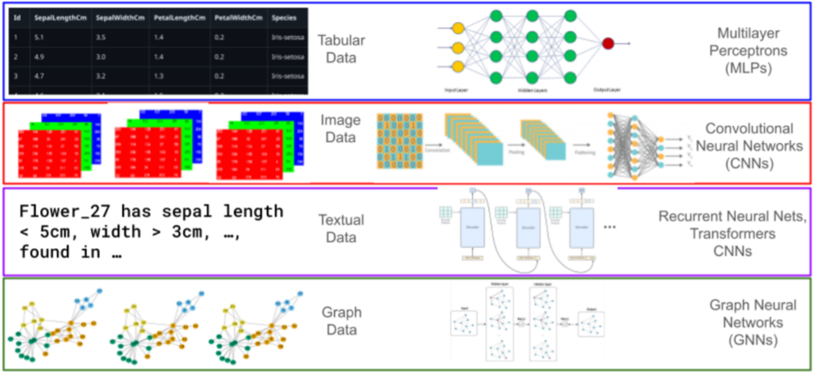
Data (observations)



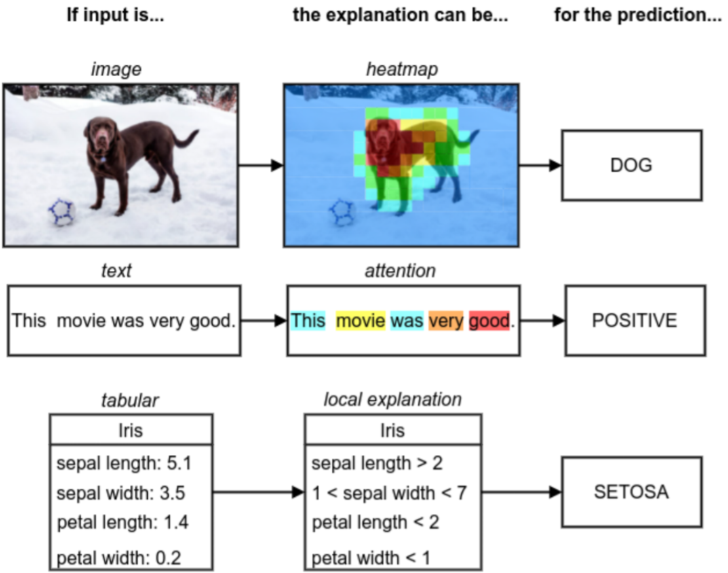
Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa

How we represent them

and, machines that learn from them

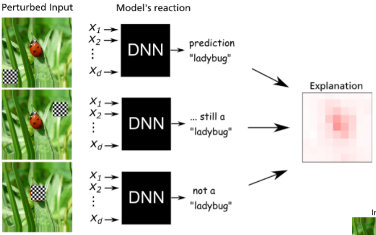


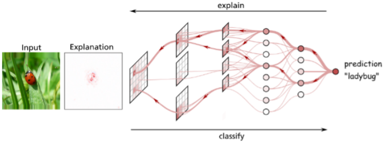
How do we explain these machines

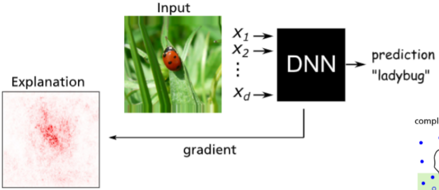


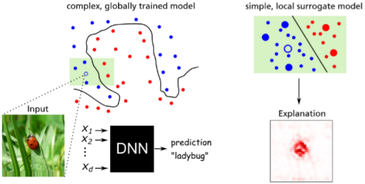
Ras et al.: Explainable deep learning: A field guide for the uninitiated, *JAIR*, 2022.

using several *post hoc* methods



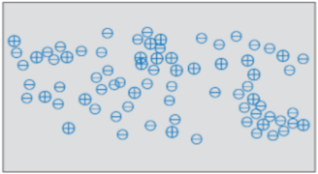


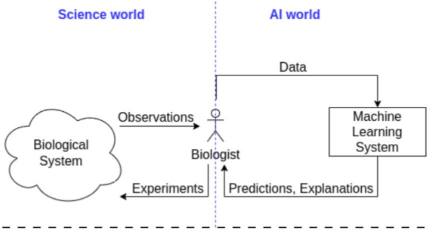




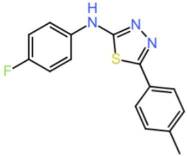
Samek et al.: Explainable deep learning: concepts, methods, and new developments. Academic Press, 2023.

ML



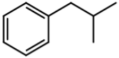


m1: CC1=CC=C(C=C1)C2=NN=C(S2)NC3=CC=C(C=C3)F



class: Positive

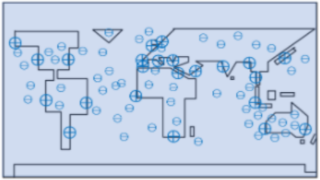
m2: CC(C)CC1=CC=CC=C1

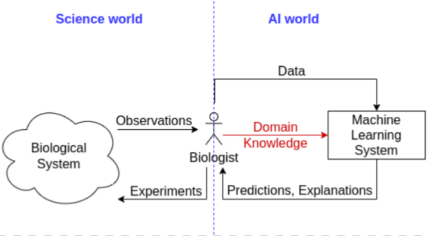


class: Negative

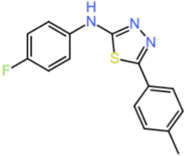
ML with Domain-Knowledge

⊕ points are port cities.





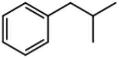
m1: CC1=CC=C(C=C1)C2=NN=C(S2)NC3=CC=C(C=C3)F



class: Positive

Tofacitinib:
contains aromatic and heterocyclic rings with functional groups (amine, thiol, fluorine) and a higher degree of molecular complexity, potentially allowing interaction with JAK2.

m2: CC(C)CC1=CC=CC=C1



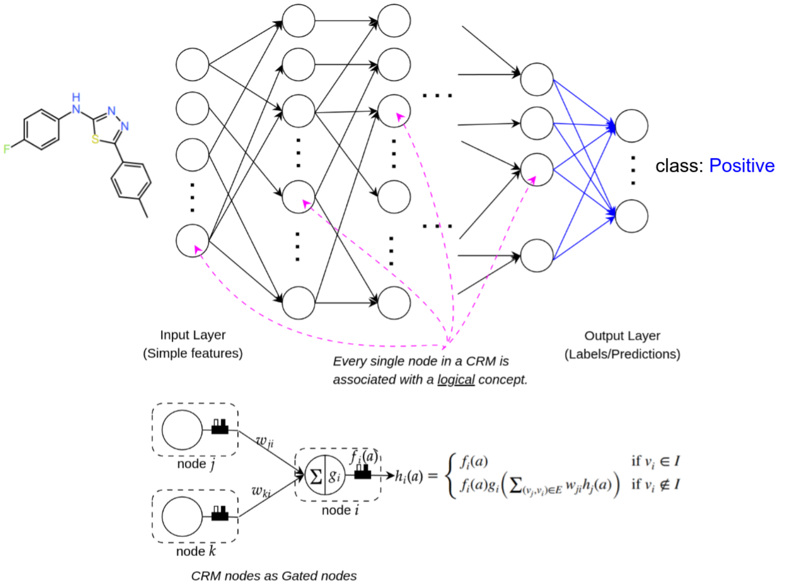
class: Negative

Isobutylbenzene:
Contains only a benzene ring and an alkyl chain, simple structure, less likely to engage in specific interactions with JAK2.

ML with Domain-Knowledge

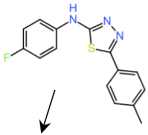
- A deep model's decision should be explained in a manner that domain-experts can understand.
- Constructing deep models using data and domain-knowledge can help us achieve that.

This talk: Compositional Relational Machine (CRM)



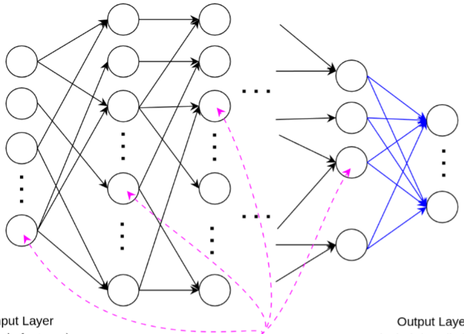
This talk: Compositional Relational Machine (CRM)

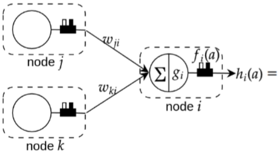
Data and BK:
Relational




```
atom(m1,1,c).
atom(m1,2,c).
...
bond(m1,1,2,1).
bond(m1,2,3,ar).
...
```

```
amine(m1, ...).
benzene(m1,...).
pyrazole(m1,...).
...
conn(m1,...).
fuzed(m1,...).
...
```





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

CRM nodes as Gated nodes

CRM

1. Data and background knowledge are uniformly represented in a relational representation (e.g. Prolog).
2. Using some language restrictions, a set of simple features are constructed. (a "template library of features")
3. These simple features can be composed to produce "complex" features. We propose two ρ-operations (ρ_1 and ρ_2).
4. A d-depth composition results in a composition graph.

CRM

Simple features:

```
p(X) :- q(X,Y), r(Y).
p(X) :- q(X,Y), r(Y), s(Y).
```

```
p(X) :- q(X,Y), r(X,Z), s(Y), t(Z).
```

The last feature is not a simple feature.

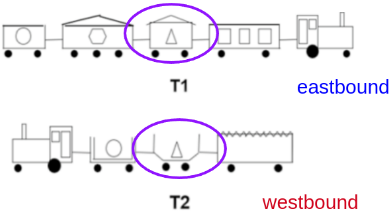
We read `p() :- q(), r().` as "if q and r then p."

Or, if q is TRUE and r is TRUE then p is TRUE.

p() :- q(). is also denoted as $p() \leftarrow q()$.

CRM

Train classification problem:



Simple features library:

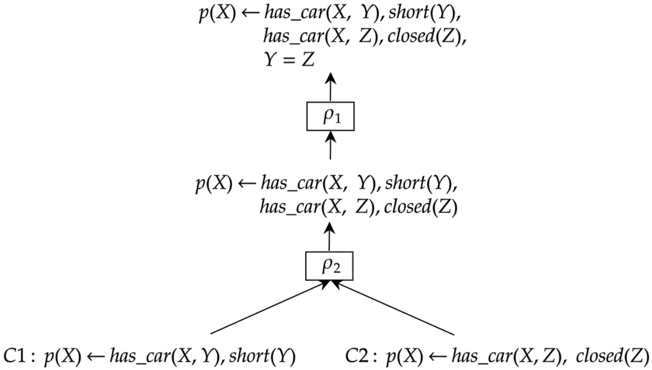
```
C1: p(X) :- has_car(X,Y),
            short(Y).

C2: p(X) :- has_car(X,Y),
            short(Y),
            closed(Y).

C3: p(X) :- has_car(X,Y),
            has_car(X,Z),
            short(Y),
            closed(Z).
```

$f_1(t_1) = 1$. Train T1 contains a short car. Clause C_1 evaluates to TRUE.
$f_2(t_2) = 0$. Train T2 contains a short car, but it is not closed. Clause C_2 evaluates to FALSE.
...

CRM

ρ-derivation of feature-clauses (Composition):

Example 1:



CRM

Example 2:

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

$$\uparrow \boxed{\rho_1} \uparrow$$

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

$$\uparrow \boxed{\rho_1} \uparrow$$

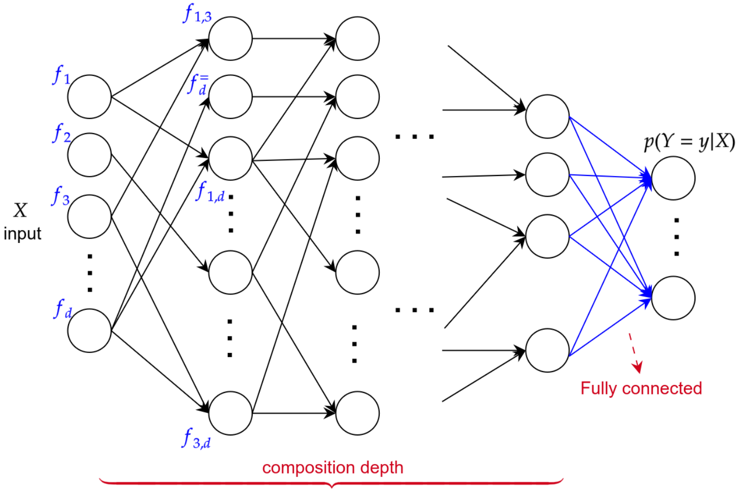
$$p(X) \leftarrow has\_car(X,\ U),\ has\_car(X,\ V),\ smaller(U,\ V),$$
$$has\_car(X,\ Y), short(Y)$$

$$\uparrow \boxed{\rho_2}$$

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

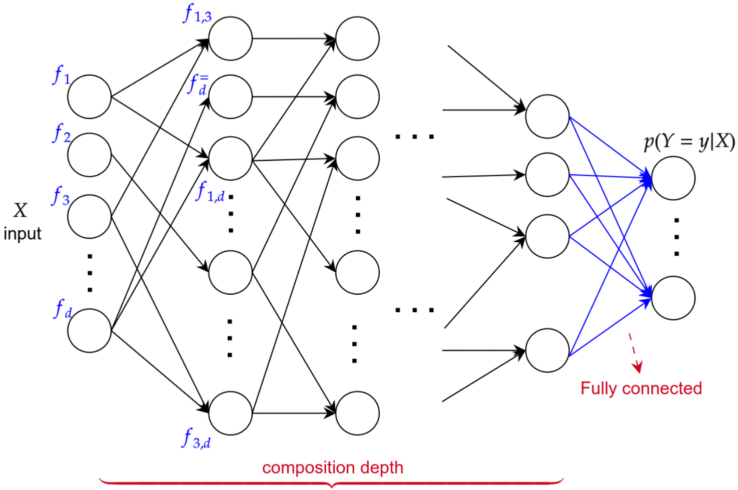
CRM

d-depth composition results in a <u>template</u> for a DNN structure.



Maximum in-degree is 2. $\rightarrow$ A CRM is a compressed network.

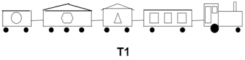
CRM

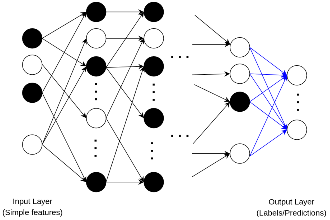


For each data instance, we can now *ground* this structure template.

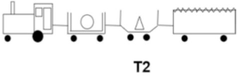
CRM

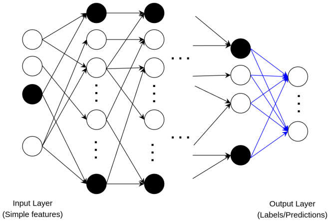
Relational instance 1:





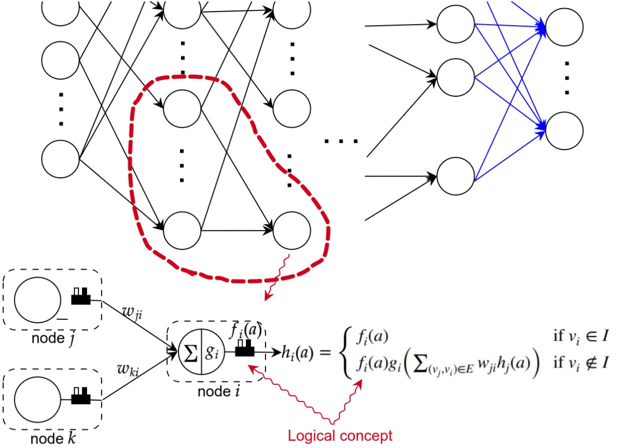
Relational instance 2:





CRM

Computation in a CRM:



CRM nodes as Gated nodes

Inspired from: Alan Turing's idea of B-type networks and unorganised machines (1948)

CRM

For a mini-batch of data instances:

1. Perform forward pass to compute the class-conditional probabilities, $p(Y = y_i|X)$
2. Compute loss (e.g. cross-entropy for classification)
3. Perform gradient descent to update model parameters ($\boldsymbol{w_{ij}}$s)

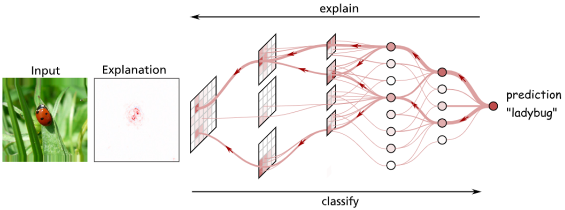
CRM

Model Explanation:

For any data instance, X:

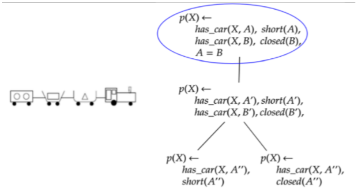
1. Compute the prediction, $\hat{y} = m^*(X)$
2. Perform Layerwise Relevance Propagation (Bach et al., PloS one, 2015.)

CRM's explanation is a structured tree.

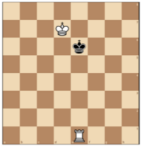


Evaluation

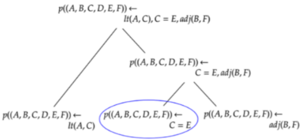
(A) Synthetic datasets: Target theory (model) is known.



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



Board $(d, 7, e, 1, e, 6)$

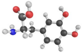


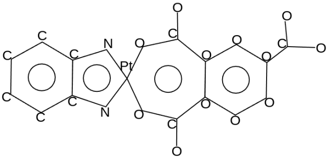
(With the substitution $\{A/d, B/7, \ldots, F/6\}$)

Target theory: Train X has a car Y and Y is short and closed.

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

Evaluation

(B) Real datasets (NCI GI-50): Activity of chemical compounds in cancer cell-line experiments; a database of ( , class) pairs.




```
Domain-knowledge used:
1. Functional groups
2. Ring structures
3. Fused structures
4. Connected structures
```
3, 4 are inferred from 1, 2.

$p(X) \leftarrow$
$\quad has\_struc(X, A, B, hetero\_aromatic),$
$\quad lteq(B, 7),$
$\quad has\_struc(X, C, D, amine),$
$\quad lteq(D, 2),$
$\quad has\_struc(X, E, F, oxide),$
$\quad lteq(F, 7)$

$p(X) \leftarrow$
$\quad has\_struc(X, A, B, amine),$
$\quad lteq(B, 2),$
$\quad has\_struc(X, C, D, oxide),$
$\quad lteq(D, 7)$

$p(X) \leftarrow$
$\quad has\_struc(X, A, B, hetero\_aromatic),$
$\quad lteq(B, 7)$

$p(X) \leftarrow$
$\quad has\_struc(X, A, B, amine),$
$\quad lteq(B, 2)$

$p(X) \leftarrow$
$\quad has\_struc(X, A, B, oxide),$
$\quad lteq(B, 7)$

Target theory is not known.

Thank you.

Joint work with: Ashwin Srinivasan, A. Baskar, Devanshu Shah

Machine Learning Journal, 113, 1091–1132, (2024).
https://rdcu.be/d2jGF

Code: https://github.com/tirtharajdash/CRM

Appendix

$$\underbrace{p(\Theta|D,B)}_{} \propto \overset{\text{Prior}}{p(\Theta|B)} \times \overset{\text{Likelihood}}{p(D|\Theta,B)}$$

If data D is limited, the prior becomes very important.