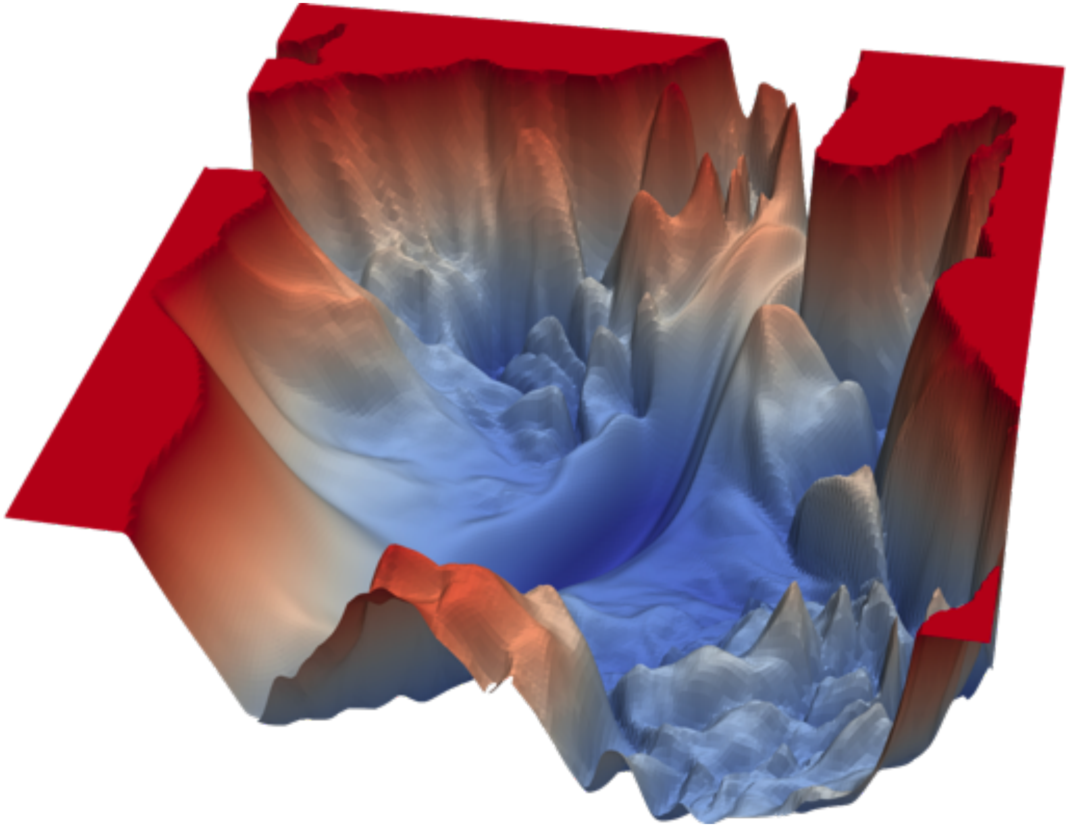
Functional Transparency for Structured Data: a Game-Theoretic Approach

Guang-He Lee, Wengong Jin, David Alvarez Melis, and Tommi S. Jaakkola

Massachusetts Institute of Technology

CSAIL

Goal: understand the (complex) network

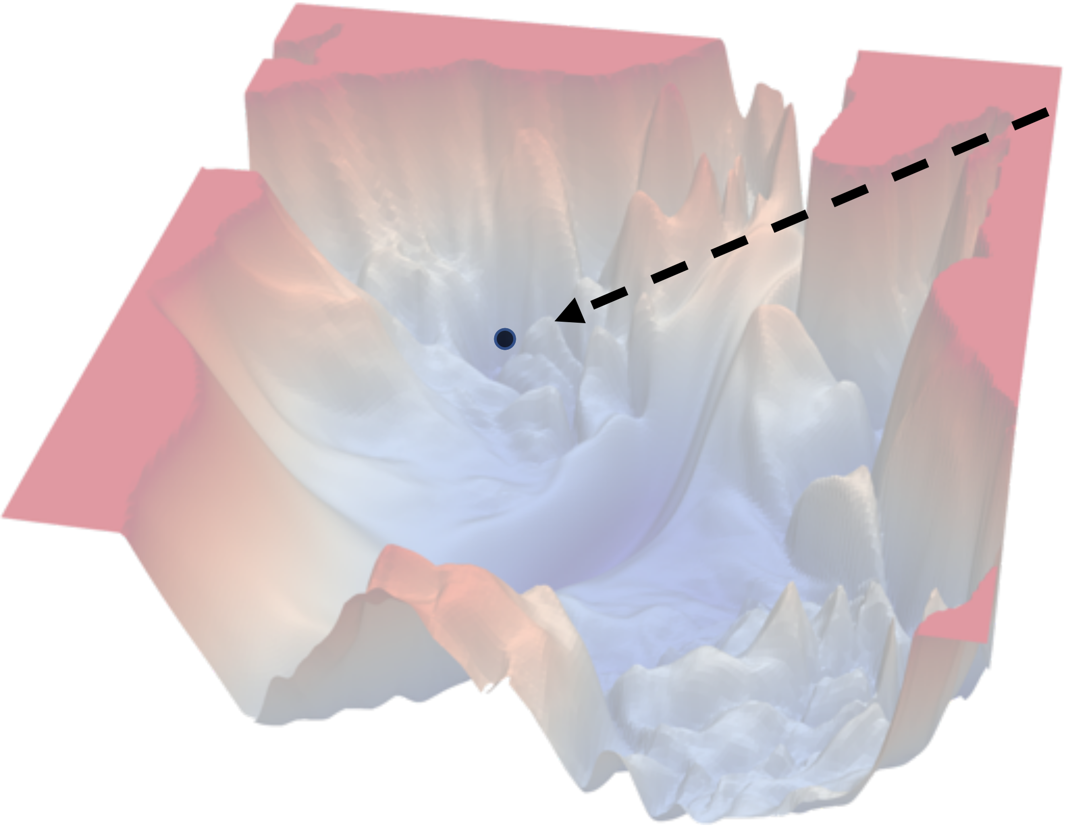


dog

deep nets

Img: https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

Typical method: post-hoc explanation

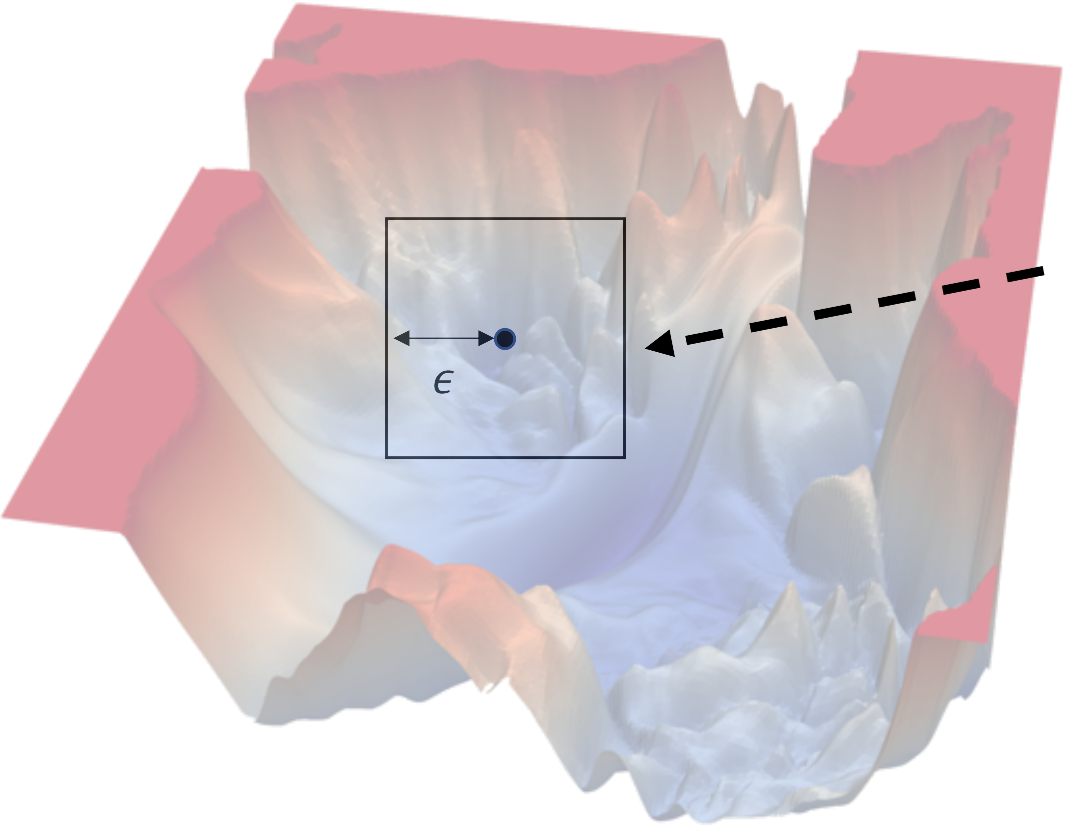


1. Given an example x_i

Deep nets

Img: https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

Typical method: post-hoc explanation



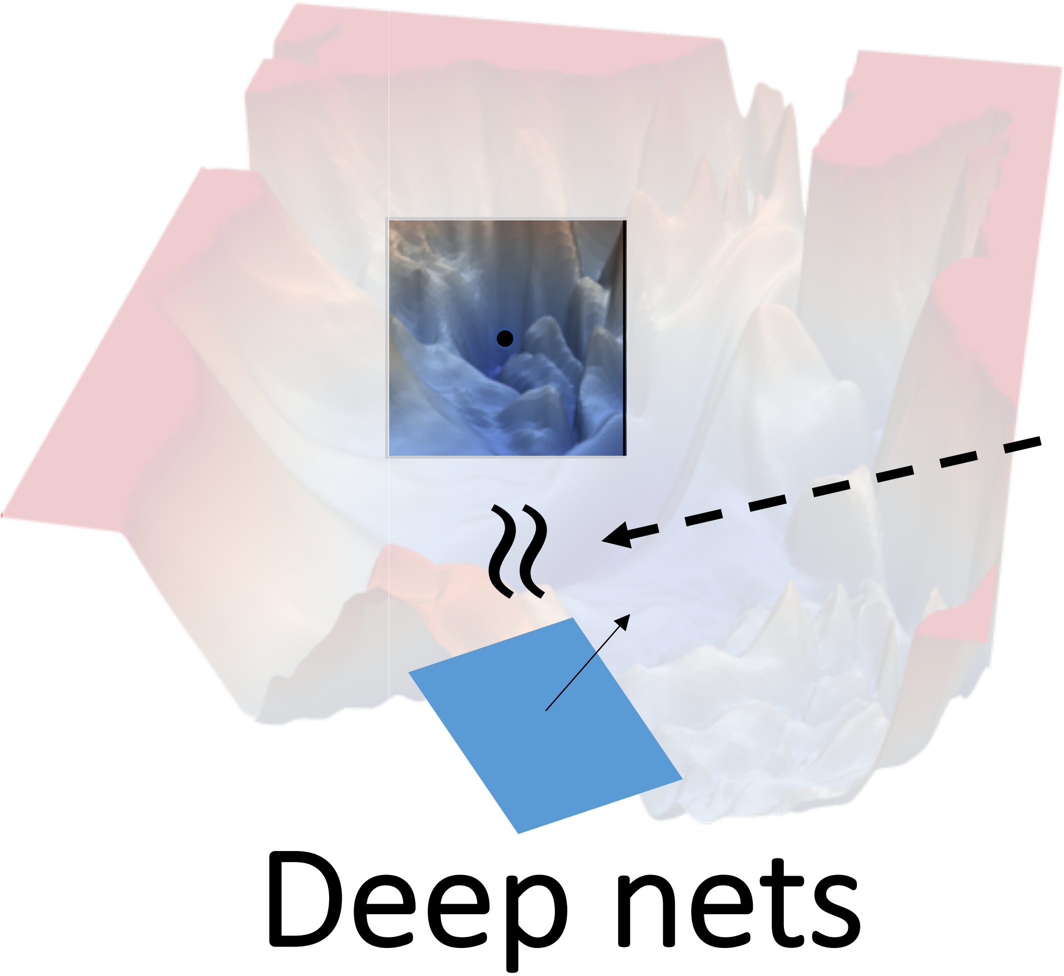
Deep nets

1. Given an example x_i

2. choose a neighborhood $\mathcal{B}(x_i)$

Img: https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

Typical method: post-hoc explanation



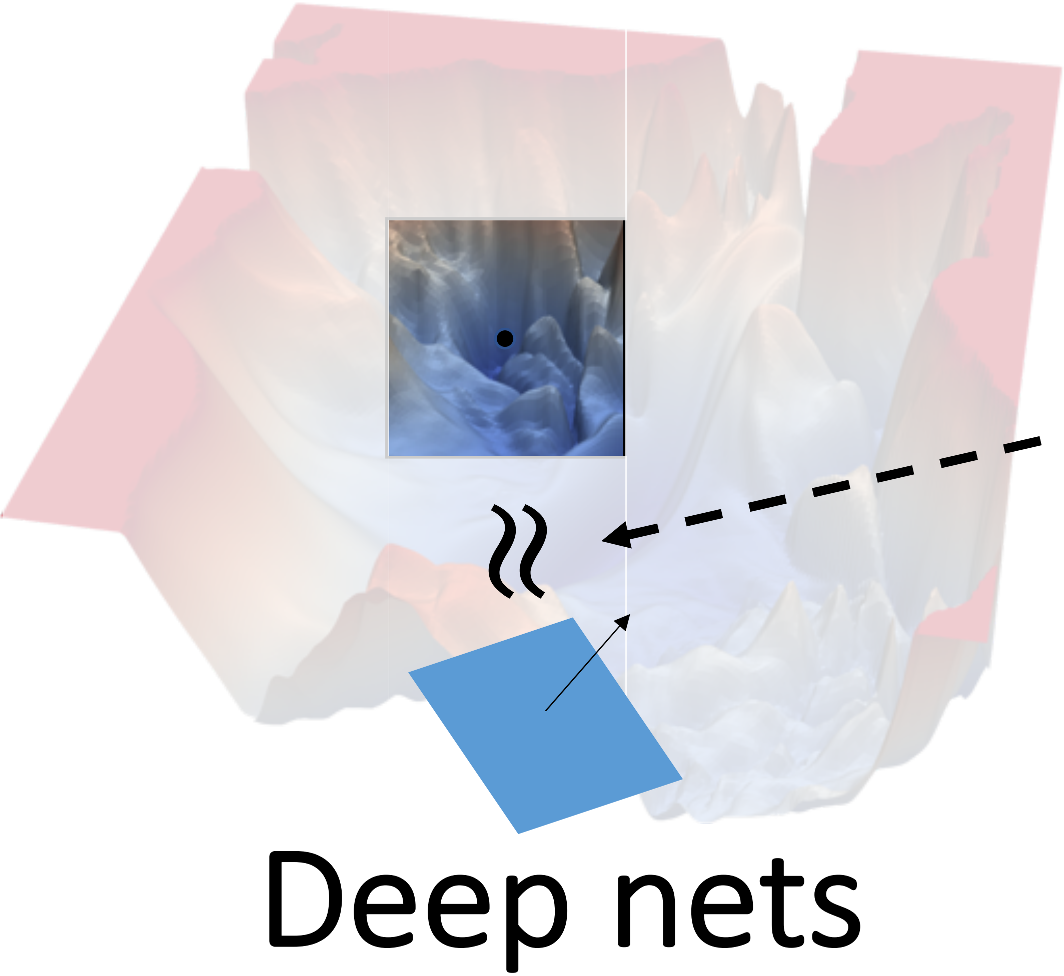
1. Given an example x_i 

2. choose a neighborhood $\mathcal{B}(x_i)$

3. Find a simple approximation
   - e.g., linear model, decision tree.

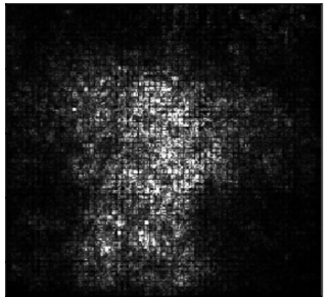
Img: https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

Typical method: post-hoc explanation



1. Given an example x_i

2. choose a neighborhood $\mathcal{B}(x_i)$
3. Find a simple approximation
   - e.g., linear model, decision tree.

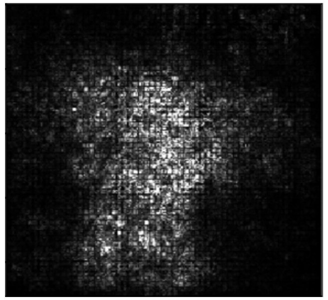


Img: https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

Post-hoc explanations are not stable

Input 1

Explanation 1

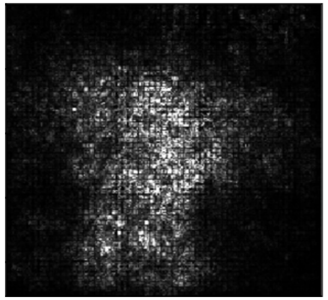


(Alvarez-Melis & Jaakkola, 18’, Ghorbani et al., 19’)

Post-hoc explanations are not stable

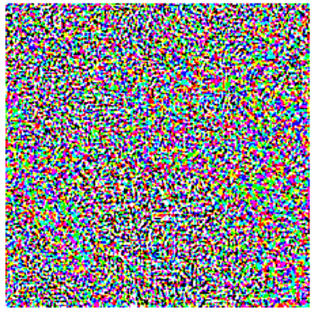
Input 1

Explanation 1

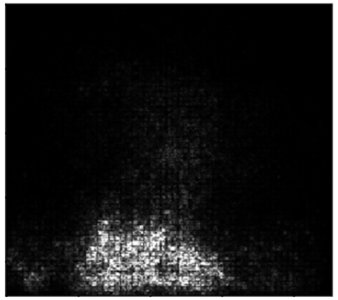


Input 2

$+ \epsilon$

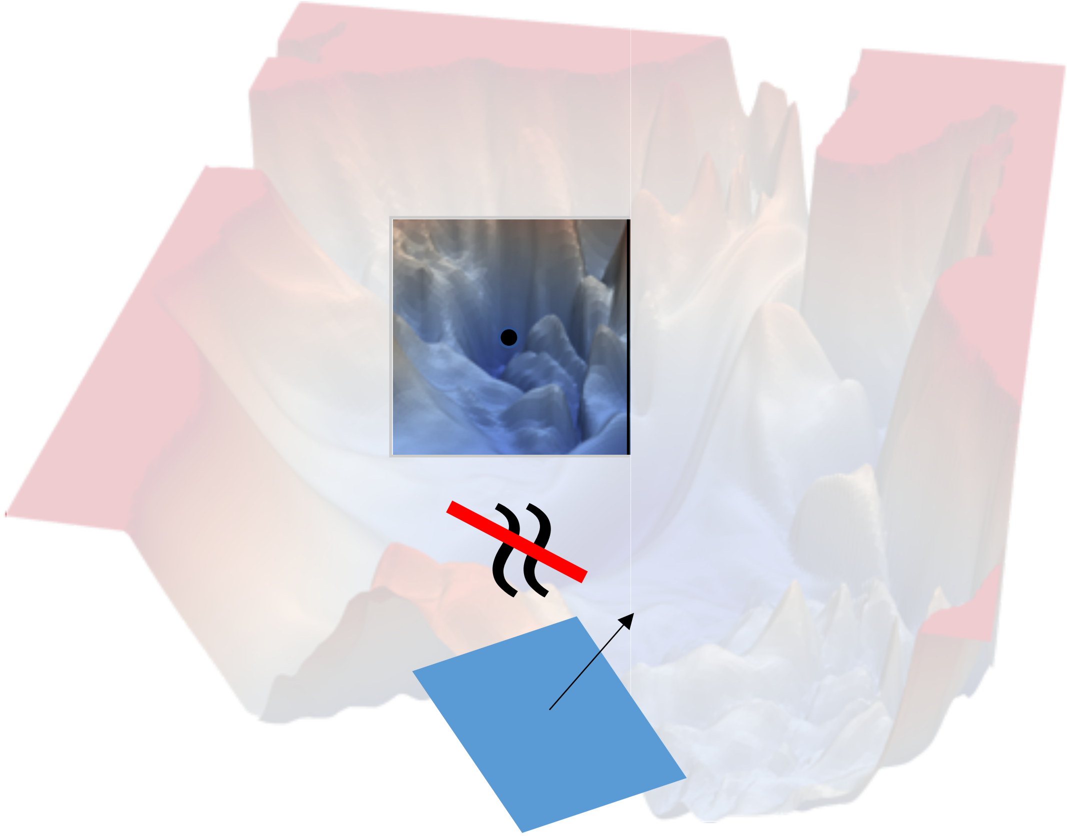


Explanation 2



(Alvarez-Melis & Jaakkola, 18', Ghorbani et al., 19')

Reason: the network does not operate as the desired explanation



Training complex models to exhibit meaningful properties locally

stability, **transparency**, …

Training complex models to exhibit meaningful properties locally

stability, **transparency**, …

$\Rightarrow$ $\mathcal{G}$:the set of functions with desired property

Training complex models to exhibit meaningful properties locally

stability, **transparency**, …

⇨ $\mathcal{G}$:the set of functions with desired property

- example for **transparency**: linear model, decision tree

Training complex models to exhibit meaningful properties locally

stability, **transparency**, ...

⇨ $\mathcal{G}$:the set of functions with desired property

⇨ $$\min_{g \in \mathcal{G}} \frac{1}{|\mathcal{B}(x_i)|} \sum_{x_j \in \mathcal{B}(x_i)} d(f(x_j), g(x_j))$$

Degree to which the property is enforced on f around x_i.

Training complex models to exhibit meaningful properties locally

stability, **transparency**, ...

$\mathcal{G}$:the set of functions with desired property

$$\min_{g \in \mathcal{G}} \frac{1}{|\mathcal{B}(x_i)|} \sum_{x_j \in \mathcal{B}(x_i)} d(f(x_j), g(x_j))$$

Degree to which the property is enforced on f around x_i.

Regularize the model f towards the property

Functional property enforcement

For each x_i, the witness $\hat{g}_{x_i}$ measures the enforcement

$$\hat{g}_{x_i} = \arg\min_{g \in \mathcal{G}} \sum_{x_j \in \mathcal{B}(x_i)} d(\hat{f}(x_j), g(x_j))$$

Functional property enforcement

For each x_i, the witness $\hat{g}_{x_i}$ measures the enforcement

$$\hat{g}_{x_i} = \arg\min_{g \in \mathcal{G}} \sum_{x_j \in \mathcal{B}(x_i)} d(\hat{f}(x_j), g(x_j))$$

We regularize the predictor $\hat{f}$ towards agreement

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{(x_i, y_i) \in \mathcal{D}} \left[\mathcal{L}(f(x_i), y_i) + \lambda\, d(f(x_i), \hat{g}_{x_i}(x_i)) \right]$$

Functional property enforcement

A co-operative game:

$$\begin{cases} \hat{f} = \underset{f \in \mathcal{F}}{\arg\min} \sum_{(x_i, y_i) \in \mathcal{D}} \left[\mathcal{L}(f(x_i), y_i) + \lambda \, d(f(x_i), \hat{g}_{x_i}(x_i)) \right] \\ \hat{g}_{x_i} = \underset{g \in \mathcal{G}}{\arg\min} \sum_{x_j \in \mathcal{B}(x_i)} d(\hat{f}(x_j), g(x_j)) \end{cases}$$

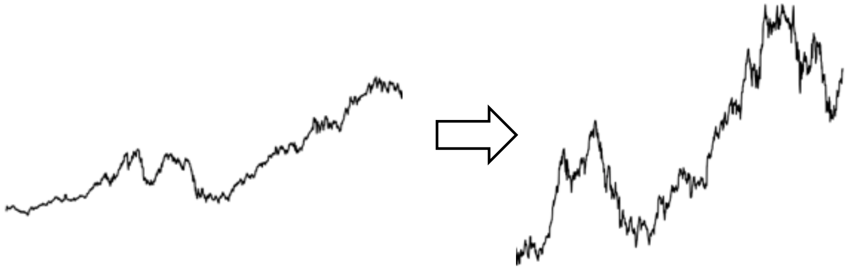
The asymmetry leads to efficiency in optimization.
(see the paper for more details)

Examples

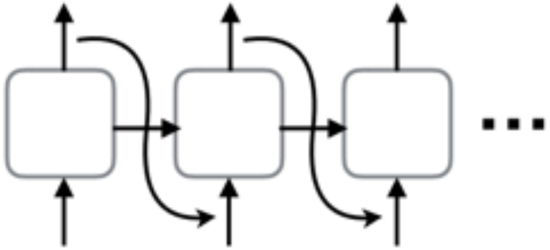
Task	Predictor	Witness

Examples

Task



Predictor

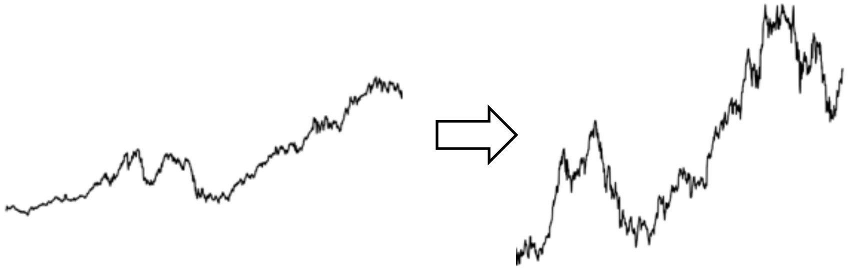


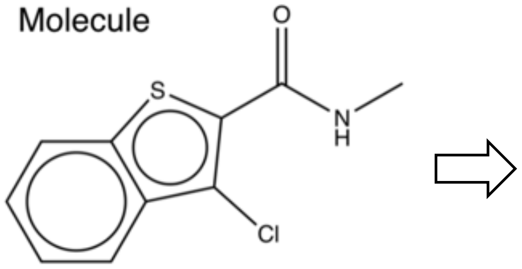
Witness

$$\sum_{k=0}^{K-1} \theta_{k+1} \cdot x_{t-k} + \theta_0$$

Examples

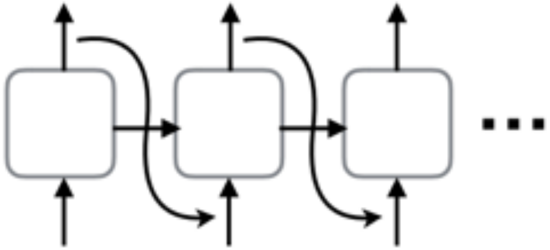
Task

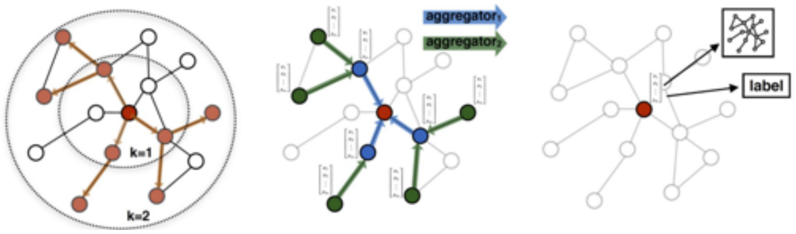




toxic

Predictor

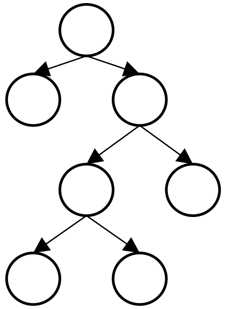




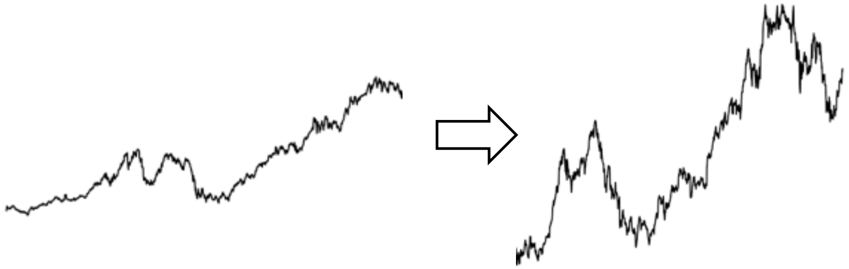
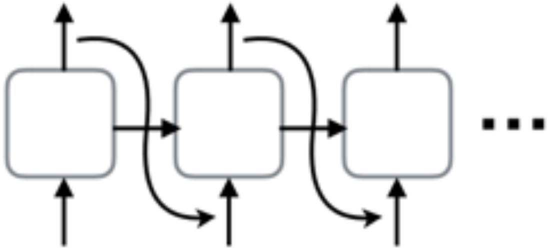
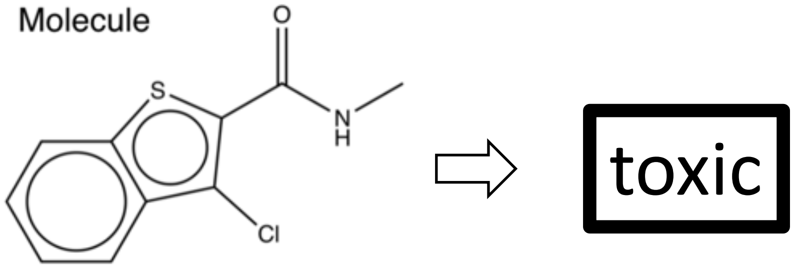
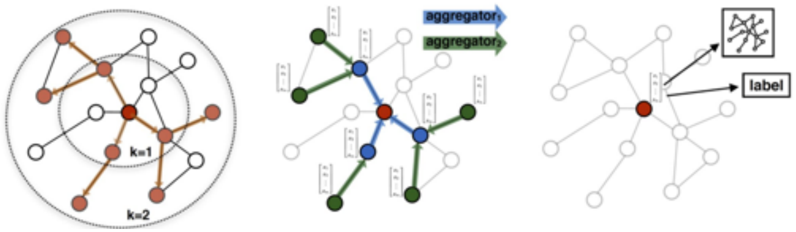
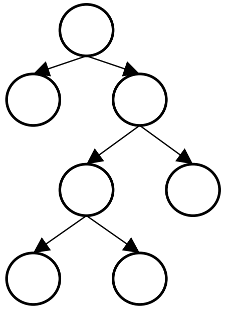
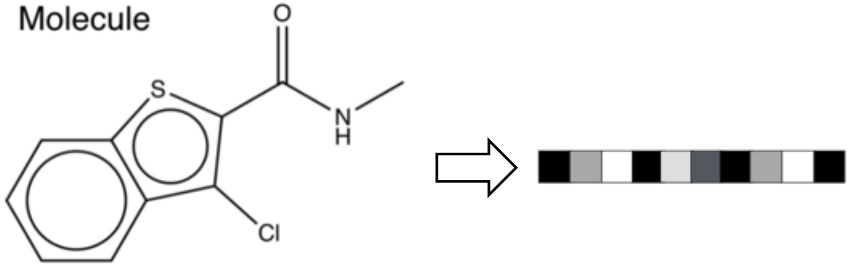
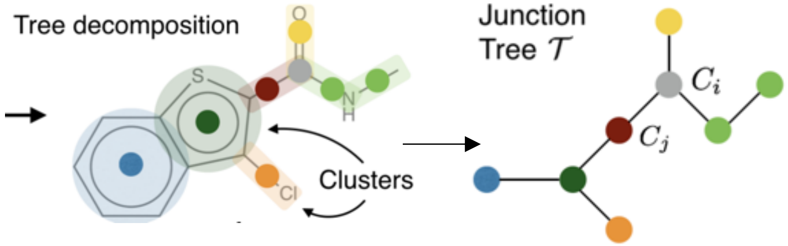
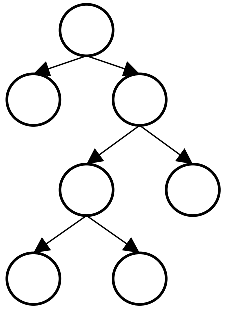
(Hamilton et al., 17')

Witness

$$\sum_{k=0}^{K-1} \theta_{k+1} \cdot x_{t-k} + \theta_0$$



Examples

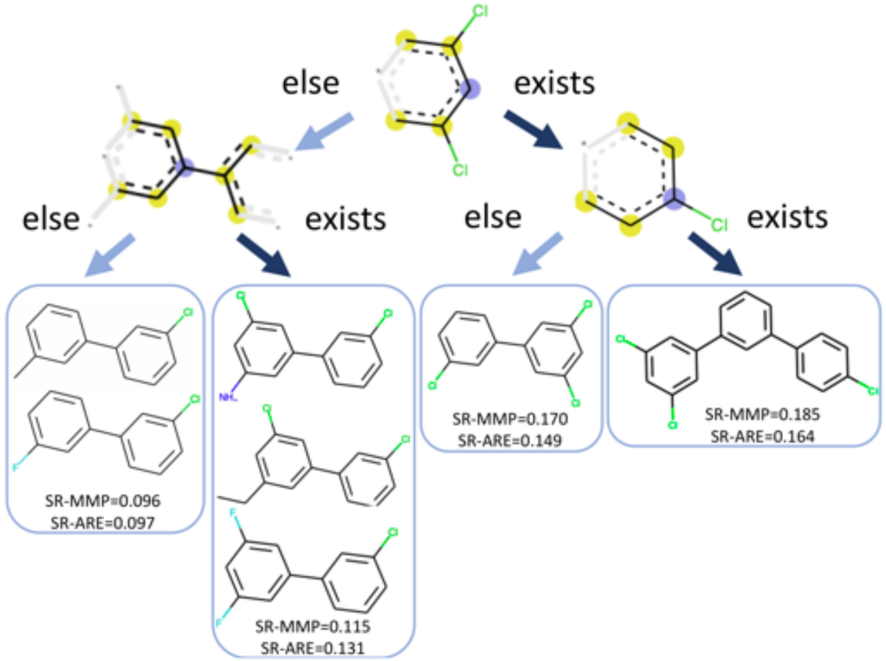
Task	Predictor	Witness
		$$\sum_{k=0}^{K-1} \theta_{k+1} \cdot x_{t-k} + \theta_0$$
	 (Hamilton et al., 17')	
	 (Jin et al., 18')	

Empirical study

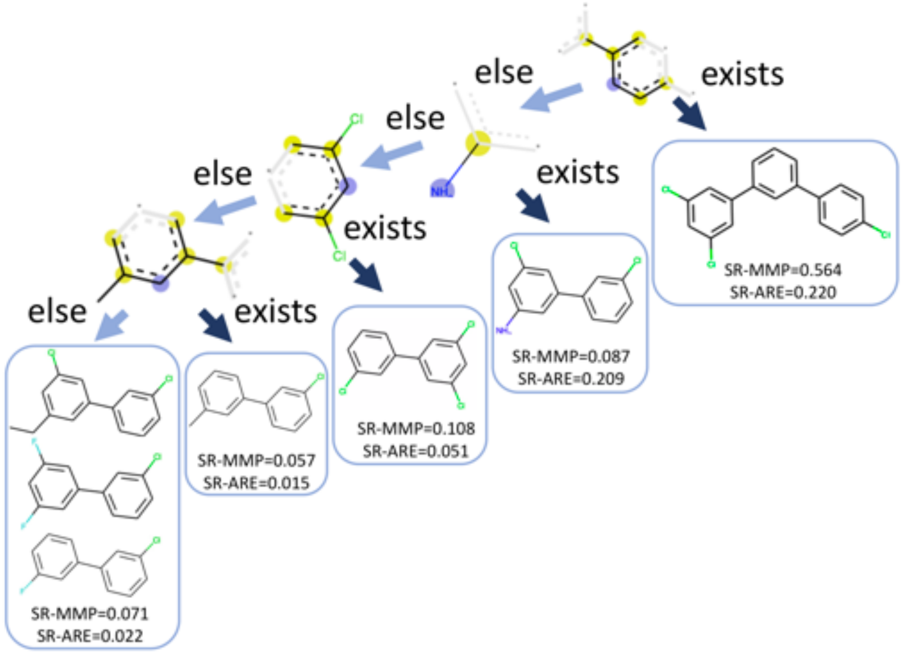
- we can measure transparency based on deviation between predictor and explainer.

Aspect	Measure	GAME	DEEP
Performance	$\mathrm{AUC}(f, y)$	**0.826**	0.815
Transparency	$\mathrm{AUC}(\hat{g}_{\mathcal{M}}, f)$	**0.967**	0.922

Models trained w/ this approach yield more compact explanations



The explanation from our model



The explanation from a normal model

Poster:

06:30 -- 09:00 PM @ Pacific Ballroom #64

- Details and analysis about the framework

Related work on functional transparency:

Towards Robust, Locally Linear Deep Networks, ICLR 19’