上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

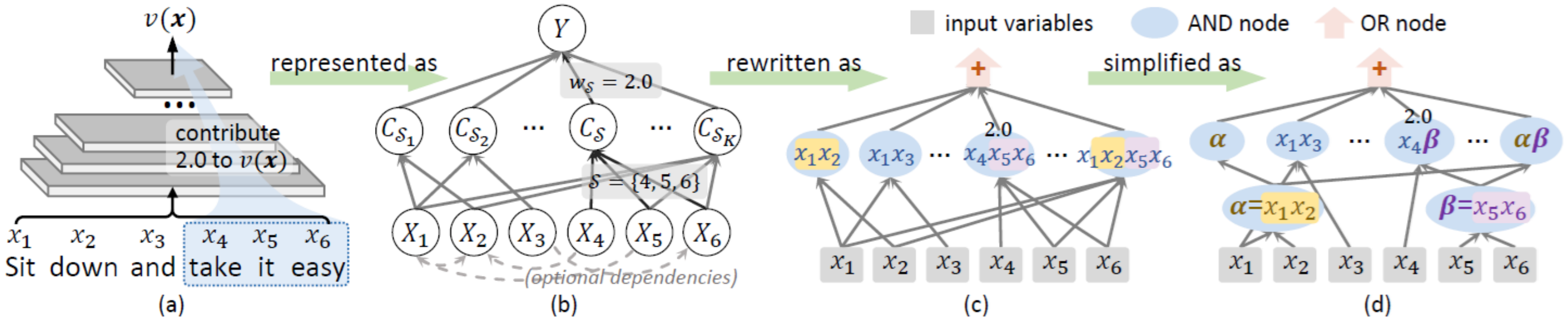
Defining and Quantifying the Emergence of Sparse Concepts in DNNs

Jie Ren*, Mingjie Li*, Qirui Chen, Huiqi Deng, Quanshi Zhang
Shanghai Jiao Tong University

THU-AM-361

Overview

Key discovery: sparse and symbolic interactive concepts between input variables emerge in various DNNs, when the DNN is sufficiently trained.



Overview

A theoretical definition for "concepts": the concept is precisely defined with a clear "boundary" that specifies the exact input variables involved in each concept.

$$v(x) \quad = \quad v(\emptyset) \quad + \quad I(S_1) \quad + \quad I(S_2) \quad + \quad I(S_3) \quad + \quad \cdots$$

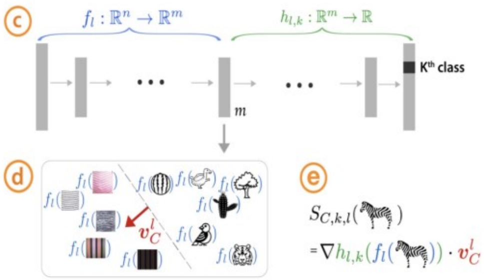
Network output **Constant bias** **Concept 1** **Concept 2** **Concept 3**

<u>Faithfulness:</u> Such concepts can exactly disentangle/explain the DNN output on any masked sample.

<u>Conciseness:</u> A sparse graph with a few of concepts can approximate the DNN's output.

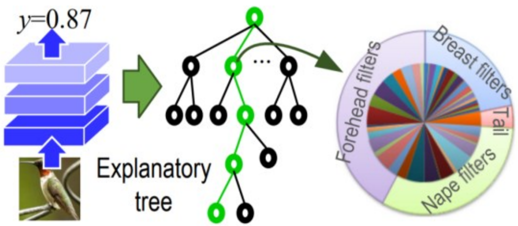
Background: symbolic explanations for DNNs

- Extracting concepts in DNNs.



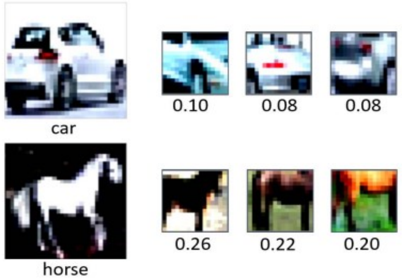
[Kim et al., 2018]

- Distilling the DNN into symbolic models (e.g., decision tree).



[Zhang et al., 2019]

- Decomposing prototype features from the data.



[Das et al., 2020]

Challenge: theoretically and objectively formulate “concepts” encoded by a DNN.

[Kim et al., 2018] Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). ICML 2018.
[Zhang et al., 2019] Interpreting CNNs vis decision trees. CVPR 2019.
[Das et al., 2020] Interpreting deep neural networks through prototype factorization. ICDMW 2020.

Interactive concepts

We define the interactive concept to decompose the DNN's output into effects of concepts.

- *E.g.*, in a vision task,

$$v(x) = v(\emptyset) + I(S_1) + I(S_2) + I(S_3) + \cdots$$

Network output | Constant bias | Torso concept | Tail concept | Chest concept

Interactive concepts

Mathematically, given

- **a DNN v:** $\mathbb{R}^n \rightarrow \mathbb{R}$
- **an input sample $\boldsymbol{x} \in \mathbb{R}^n$** with n input variables indexed by $N = \{1, 2, \dots, n\}$
- **an interactive concept $\boldsymbol{S} \subseteq \boldsymbol{N}$** is a subset of input variables in N, which has **an effect $\boldsymbol{I}(\boldsymbol{S})$** to DNN

The DNN's inference score $v(\boldsymbol{x})$ is **decomposed into the sum of effects of all potential interactive concepts.**

$$v(\boldsymbol{x}) = v(\emptyset) + \sum_{S \subseteq N, S \neq \emptyset} I(S)$$

Model output
The output score of sample $\boldsymbol{x}$

Constant bias
The output score when all input variable are masked.

Interaction effects
between variables in all potential subsets S

where $I(S) \triangleq \sum_{T \subseteq S} (-1)^{|S|-|T|} \cdot v(\boldsymbol{x}_T)$.[Harsanyi, 1963]

[Harsanyi, 1963] A simplified bargaining model for the n-person cooperative game. International Economic Review, 4(2):194–220, 1963.

Understanding the interactive concept

The interaction in each set S formulates the "AND" relationship between input variables in S.

$$v(N) = v(\emptyset) + \overset{0}{I(S_1)} + I(S_2) + I(S_3) + \cdots$$

Network output (bird) | **Constant bias** | **Torso of the bird** | **Tail of the bird** | **Chest of the bird**

Understanding the interactive concept

The interaction in each set S formulates the "AND" relationship between input variables in S.

input sentence	**activated** AND interactions	inactivated AND interactions
I think he is a green hand.	green hand, I think, I think he, he is, is green, a green hand	/
I think he is a _____ hand.	I think, I think he, he is	green hand, is green, a green hand
I think __ __ a green hand.	green hand, I think, a green hand	I think he, he is, is green

Faithfulness of the interactive concepts

Interactive concepts can exactly explain/fit the DNN output on any masked sample.

$$\forall S \subseteq N,\ v(S) = \sum_{S' \text{ activated}} I(S') = \sum_{S' \subseteq S} I(S')$$

input sentence	**activated** AND interactions	**output**
I think he is a green hand.	S_1 ={green hand} S_2 ={I think} S_3 ={I think he} S_4 ={he is} S_5 ={is green} S_6 ={a green hand}	$v(N) = I(S_1) + I(S_2) + I(S_3) + I(S_4) + I(S_5) + I(S_6)$
I think he is a _____ hand.	S_2 ={I think} S_3 ={I think he} S_4 ={he is}	$v(S) = I(S_2) + I(S_3) + I(S_4)$

Faithfulness of the interactive concept

1. **Efficiency property.** The output of a model can be decomposed into interactions of different subsets of variables, $v(N) = v(\emptyset) + \sum_{S \subseteq N, S \neq \emptyset} I(S)$.

2. **Linearity property.** If we merge outputs of two models, $u(S) = w(S) + v(S)$, then $\forall\ S \subseteq N$, the interaction $I_u(S)$ w.r.t. the new network u can be decomposed into $I_u(S) = I_w(S) + I_v(S)$

3. **Nullity property**: The dummy variable $i \in N$ satisfies $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$. It means that the variable i has no interactions with others, i.e. $\forall S \subseteq N \backslash \{i\}, I(S \cup \{i\}) = 0$.

4. **Symmetry property.** If input variables $i, j \in N$ have same cooperation with other variables $\forall S \subseteq N \backslash \{i, j\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then they have same interactions with other variables, $\forall S \subseteq N \backslash \{i, j\}, I(S \cup \{i\}) = I(S \cup \{j\})$.

5. **Anonymity property.** For any permutations π on N, we have $\forall S \subseteq N, I_v(S) = I_{\pi v}(\pi S)$, where $\pi S = \{\pi(i) | i \in S\}$, and πv is defined by $\pi v(\pi S) = v(S)$.

6. **Recursive property.** The interaction utility of $S \cup \{i\}$ is the difference of the interaction utility of S with and without the presence of i, i.e. $I(S \cup \{i\}) = I(S | i \text{ is always present}) - I(S)$.

7. **Interaction distribution property.** This axiom characterizes how interactions are distributed for “interaction functions”. An interaction function v_T parametrized by T satisfies $\forall S \subseteq N$, if $T \subseteq S$, $v_T(S) = c$; otherwise, $v_T(S) = 0$. Then, we have $I(T) = c$, and $\forall S \neq T, I(S) = 0$.

Faithfulness of the interactive concept

1. **Connection to the Shapley value [Shapley, 1953].** Let $\phi(i)$ denote the Shapley value of an input variable i. Then, the Shapley value can be represented as the weight sum of interaction utilities, i.e. $\phi(i) = \sum_{S\subseteq N\setminus\{i\}} \frac{1}{|S|+1} \cdot I(S \cup \{i\})$.

2. **Connection to the marginal benefit [Grabisch et al., 1999].** Let $\Delta v_T(S)$ denote the marginal benefit of variables in T given the environment S. Then it can be decomposed into the sum of interaction utilities inside T and sub-environments $S' \subseteq S$, i.e. $\Delta v_T(S) = \sum_{S'\subseteq S} I(T \cup S')$.

3. **Connection to the Shapley interaction index [Grabisch et al., 1999].** Let $I^{Shapley}(T)$ denote the Shapley interaction index of a subset of input variables $T \subseteq N$. The Shapley interaction index can be represented as the weighted sum of interaction utilities, i.e. $I^{Shapley}(T) = \sum_{S\subseteq N\setminus \mathrm{T}} \frac{1}{|S|+1} \cdot I(S \cup T)$.

4. **Connection to the Shapley Taylor interaction index [Sundararajan et al., 2020].** Let $I^{Shapley-Taylor}(T)$ denote the Shapley Taylor interaction index of order k. The Shapley Taylor can be represented as the weighted sum of interaction utilities if $|T| = k$, i.e. $I^{Shapley-Taylor}(T) = \sum_{S\subseteq N\setminus \mathrm{T}} \binom{|S|+k}{k}^{-1} \cdot I(S \cup T)$. Besides, $I^{Shapley-Taylor}(T) = I(T)$ if $|T| < k$, and $I^{Shapley-Taylor}(T) = 0$ if $|T| > k$.

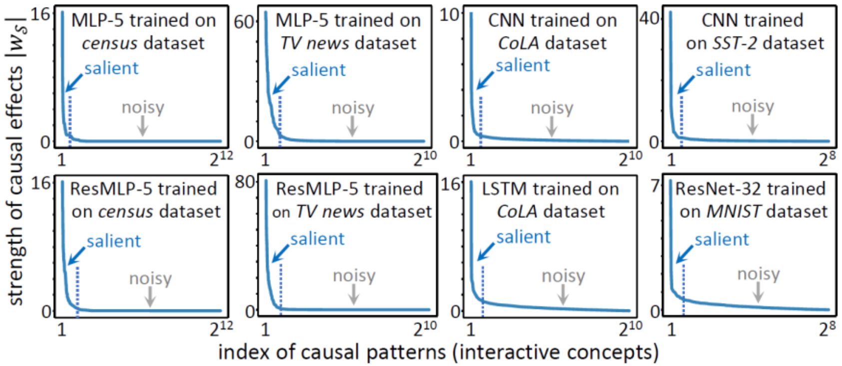
[Shapley, 1953] A value for n-person games. Contributions to the Theory of Games, 2(28):307-317, 1953.
[Grabisch et al., 1999] An axiomatic approach to the concept of interaction among players in cooperative games. International Journal of game theory, 28(4):547–565, 1999.
[Sundararajan et al., 2020] The shapley taylor interaction index. In ICML 2020.

Discovering and boosting the conciseness of the explanation

Technique 1: Only using a few salient concepts

The DNN output can be decomposed into effects of all potential interactive concepts.

- **Salient concepts**: have significant effects $|I(S)|$ on the DNN output;
- **Noisy patterns**: have negligible effects $|I(S)| \approx 0$ on the DNN output.

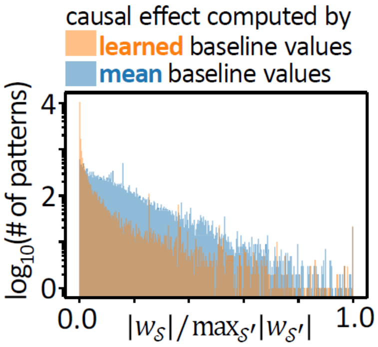


Discovering and boosting the conciseness of the explanation

Technique 2: Learning the optimal baseline value

The effects $I(S)$ of interactive concepts are computed based on baseline values, which are used to mask input variables to compute $v(S)$.

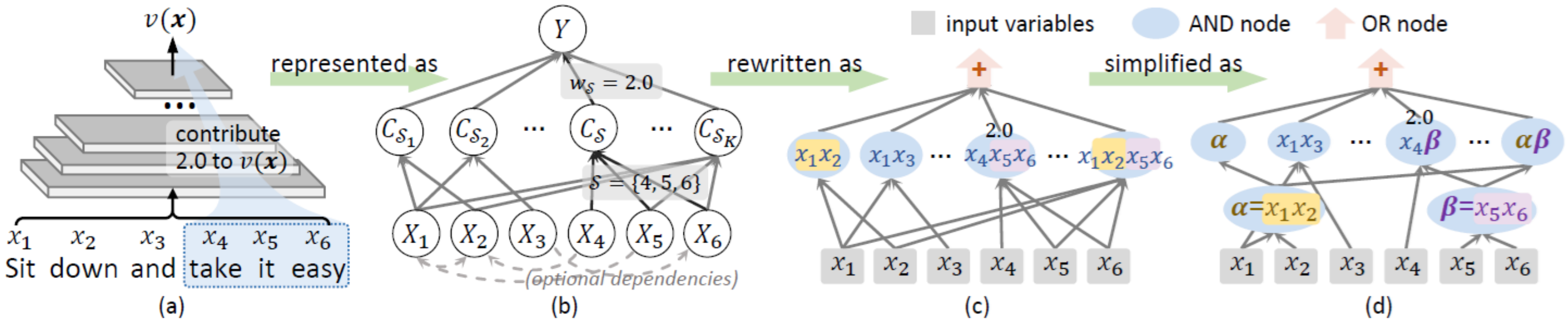
We learn the baseline values that **minimize the number of salient concepts while not destroying the faithfulness.**



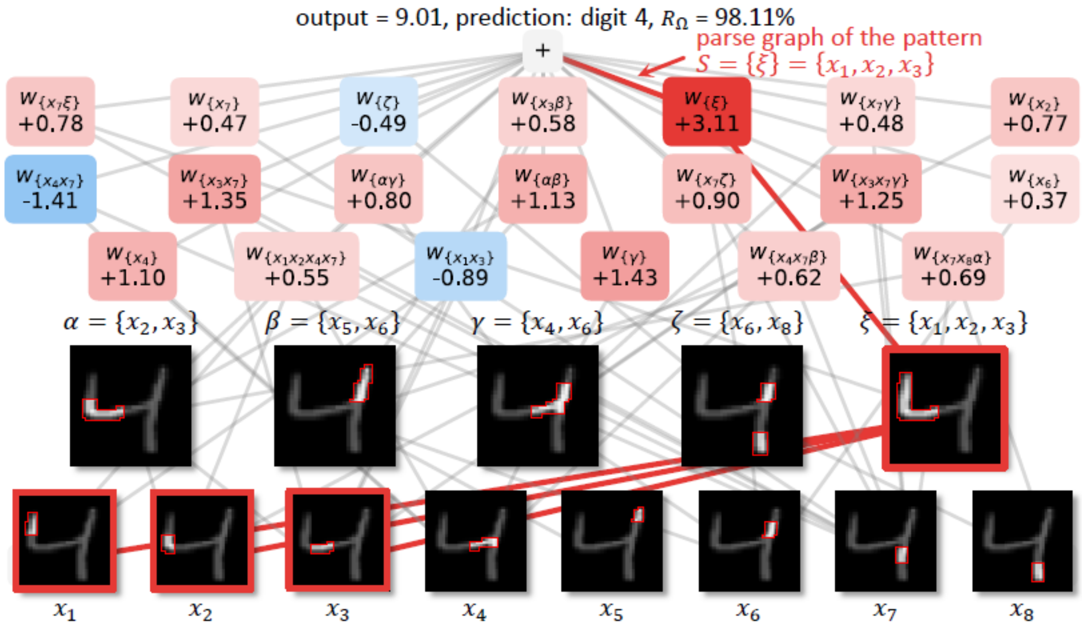
Using the interactive concepts to construct an And-Or graph

In an And-Or graph：

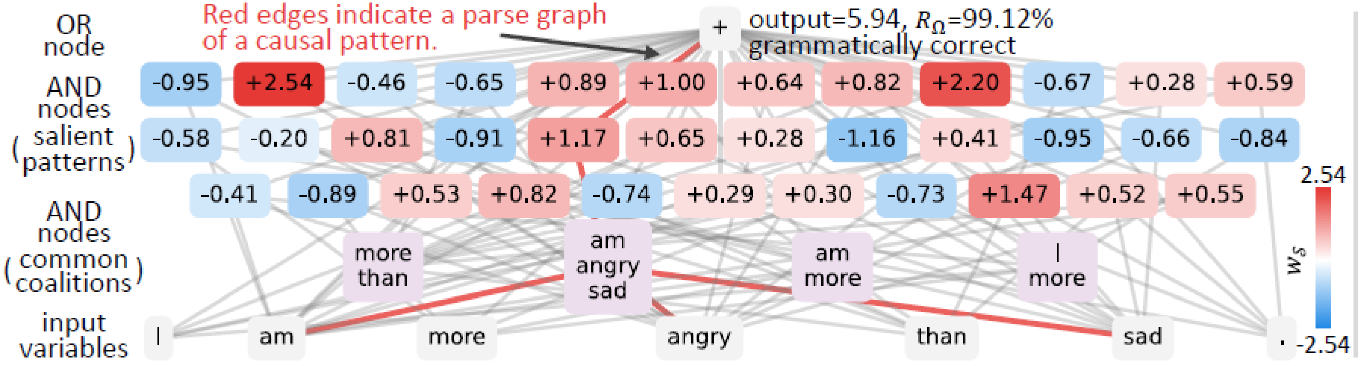
- **AND node:** a concept that represents the **AND relationship** between its child nodes
- **OR node**: the **sum** of interaction effects of all concepts

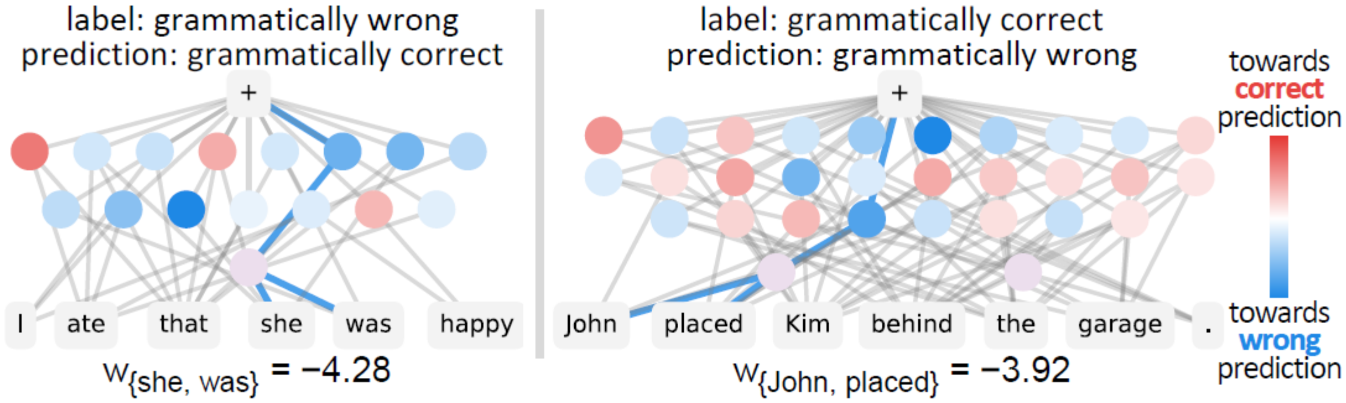


Using the interactive concepts to construct an And-Or graph



Using the interactive concepts to construct an And-Or graph

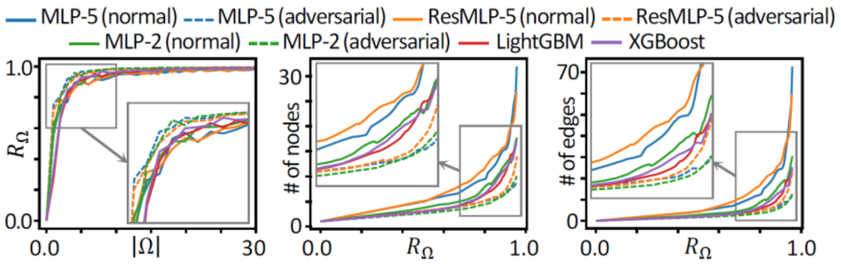




Explaining adversarial training based on interactive concepts

The interactive concepts provide **a new perspective** to understand the effects of different deep learning techniques, e.g. adversarial training.

- Adversarial training could boost the sparsity of interactive concepts.



- Adversarial training made different DNNs encode common interactive concepts for inference.

Jaccard similarity between two models. Two adversarially trained models were more similar than two normally trained ones.

		TV news	census	bike
MLP-2	normal	0.5965	0.4899	-
	adversarial	**0.6109**	**0.6292**	-
MLP-5	normal	0.3664	0.2482	0.3816
	adversarial	**0.6304**	**0.4971**	**0.4741**
ResMLP-5	normal	0.3480	0.2764	0.3992
	adversarial	**0.5731**	**0.4489**	**0.4491**

Conclusion

In this study,

- We define the interactive concept encoded in DNNs, and we prove such concepts can faithfully explain the DNN output.
- We discover and further boost the conciseness of interactive concepts.
- We build an And-Or Graph using a small number of interactive concepts to explain DNNs, which provides new insights for understanding the DNN.

Thank you!