Defining and Quantifying the Emergence of Sparse Concepts in DNNs

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

Abstract

This paper aims to illustrate the concept-emerging phenomenon in a trained DNN. Specifically, we find that the inference score of a DNN can be disentangled into the effects of a few interactive concepts. These concepts can be understood as causal patterns in a sparse, symbolic causal graph, which explains the DNN. The faithfulness of using such a causal graph to explain the DNN is theoretically guaranteed, because we prove that the causal graph can well mimic the DNN's outputs on an exponential number of different masked samples. Besides, such a causal graph can be further simplified and re-written as an And-Or graph (AOG), without losing much explanation accuracy. The code is released at https://github.com/sjtu-xai-lab/aog.

1. Introduction

It is widely believed that the essence of deep neural networks (DNNs) is a fitting problem, instead of explicitly formulating causality or modeling symbolic concepts like how graphical models do. However, in this study, we surprisingly discover that *sparse and symbolic interactive relationships between input variables emerge in various DNNs trained for many tasks*, when the DNN is sufficiently trained. In other words, the inference score of a DNN can be faithfully disentangled into effects of only a few interactive concepts.

In fact, the concept-emerging phenomenon does exist and is even quite common for various DNNs, though somewhat counter-intuitive and seeming conflicting with the DNN's layerwise inference. To clarify this phenomenon, let us first define interactive concepts that emerge in the DNN. Let a DNN have n input variables (*e.g.* a sentence with n words). As Fig. 1(a) shows, given the sentence "*sit down and take it easy*," the co-appearance of a set of words $\mathcal{S} = \{$*take, it, easy*$\}$ causes the meaning of "*calm down*," which makes a considerable numerical contribution $w_\mathcal{S}$ to the network output. Such a combination of words is termed an *interactive concept*. Each interactive concept $\mathcal{S}$ represents an AND relationship between the set of words in $\mathcal{S}$. In other words, only their co-appearance will trigger this interactive concept. The absence (masking) of any words in $\{$*take, it, easy*$\}$ will remove the effect $w_\mathcal{S}$ towards "*calm down*" from the network output.

Causal graph based on interactive concepts. Given an input sample, we introduce how to extract a set of interactive concepts Ω from a trained DNN, and how to organize all such concepts $\mathcal{S} \in \Omega$ into a three-layer causal graph in Fig. 1(b). We also prove that such a causal graph can mimic the inference score of the DNN. Specifically, each source node X_i $(i = 1, ..., n)$ in the bottom layer represents the binary state of whether the i-th input variable is masked $(X_i = 0)$ or not $(X_i = 1)$. Each intermediate node $C_\mathcal{S}$ $(\mathcal{S} \in \Omega)$ in the causal graph represents an interactive concept $\mathcal{S}$ that encodes the AND relationship between input variables in $\mathcal{S}$. In fact, $\mathcal{S}$ can also be interpreted as a causal pattern for the DNN's inference, as follows. If the interactive concept appears in the sample, then the causal pattern $\mathcal{S}$ is triggered $C_\mathcal{S} = 1$; otherwise, $C_\mathcal{S} = 0$. Each triggered pattern $\mathcal{S}$ contributes a causal effect $w_\mathcal{S}$ to the causal graph's output Y in the top layer. Therefore, the output Y of the causal graph can be specified by a structural causal model (SCM) [25], which sums up all triggered causal effects, *i.e.* $Y = \sum_\mathcal{S} w_\mathcal{S} \cdot C_\mathcal{S}$. **Note that we study the mathematical causality between the input and the output of the DNN, instead of the natural true causality potentially hidden in data.**

In this study, we discover that we can always construct a causal graph with a relatively small number of causal patterns (interactive concepts) to *faithfully* and *concisely* explain a DNN's inference on an input sample.

- *Faithfulness.* Given an input sample with n variables, there are 2^n different ways to randomly mask input variables. Given any one of all the 2^n masked input samples, we prove that the output Y of the causal graph can always mimic the DNN's output. This guarantees that the causal graph encodes the same logic (*i.e.* the same set of interactive concepts) as the DNN. Thus, we can consider such a causal graph as a faithful explanation for the inference logic of the DNN.

- *Conciseness.* Theoretically, we may extract at most 2^n causal patterns (interactive concepts) from a DNN with n input variables. However, we discover that most causal

[*] These authors contributed equally to this work.

[†] Quanshi Zhang is the corresponding author. He is with the Department of Computer Science and Engineering, the John Hopcroft Center, at the Shanghai Jiao Tong University, China. zqs1022@sjtu.edu.cn.

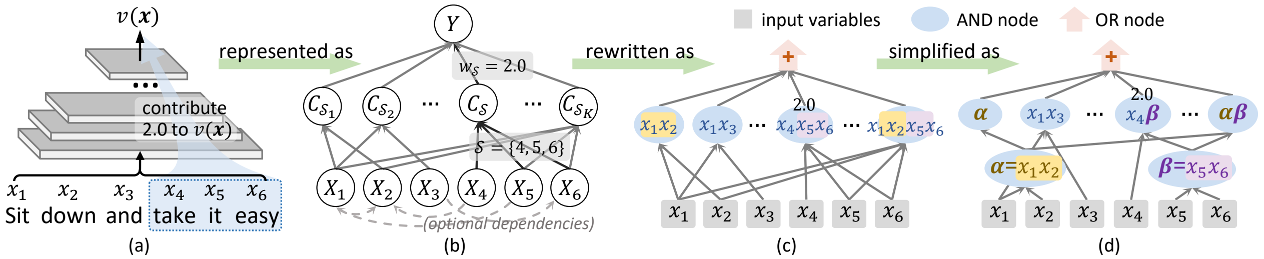
$v(\boldsymbol{x})$
contribute 2.0 to $v(\boldsymbol{x})$
x_1 x_2 x_3 x_4 x_5 x_6
Sit down and take it easy
(a)
represented as
Y
$w_S = 2.0$
C_{S_1} C_{S_2} ··· C_S ··· C_{S_K}
$S = \{4,5,6\}$
X_1 X_2 X_3 X_4 X_5 X_6
(optional dependencies)
(b)
rewritten as
input variables
AND node
OR node
+
x_1x_2 x_1x_3 ··· $x_4x_5x_6$ ··· $x_1x_2x_5x_6$
2.0
x_1 x_2 x_3 x_4 x_5 x_6
(c)
simplified as
+
$\boldsymbol{\alpha}$ x_1x_3 ··· $x_4\boldsymbol{\beta}$ ··· $\boldsymbol{\alpha\beta}$
2.0
$\boldsymbol{\alpha}=x_1x_2$
$\boldsymbol{\beta}=x_5x_6$
x_1 x_2 x_3 x_4 x_5 x_6
(d)

Figure 1. Emergence of symbolic interactive concepts in a sufficiently trained DNN (a), which make considerable numerical effects on the network output. (b) All interactive concepts can be faithfully organized into a causal graph, which reflects the DNN's inference logic. (c,d) Besides, the causal graph can be further simplified as an And-Or graph (AOG), which extracts common coalitions.

patterns have almost zero effects on the output Y, so we can use a sparse graph with a small number of salient causal patterns to approximate the DNN's output in real applications. Furthermore, as Fig. 1(c,d) shows, we propose to summarize common coalitions shared by salient causal patterns to simplify the causal graph to a deep And-Or graph (AOG).

Note that since the DNN encodes complex inference logic, different samples may activate different sets of salient causal patterns and generate different causal graphs.

• *Universality.* As Fig. 2 shows, given DNNs with various architectures trained on different tasks, we find that the inference of each DNN can all be faithfully and concisely explained by a few salient causal patterns.

In addition, we prove that causal patterns extracted from the DNN have broad theoretical connections with classical interaction/attribution metrics for explaining DNNs. Specifically, the causal effects can explain the elementary mechanism of the Shapley value [31], the Shapley interaction index [13], and the Shapley-Taylor interaction index [36].

Contributions of this paper can be summarized as follows: (1) We discover and prove that the inference logic of a complex DNN on a certain sample can be represented as a relatively simple causal graph. (2) Furthermore, such a causal graph can be further simplified as an AOG. (3) The trustworthiness of using the AOG to explain a DNN is verified in experiments.

2. Explainable AI (XAI) theories based on game-theoretic interactions

This study provides a solid foundation for XAI theories based on game-theoretic interactions. Our research group led by Dr. Quanshi Zhang in Shanghai Jiao Tong University has developed a theory system based on game-theoretic interactions to address two challenges in XAI, *i.e.*, (1) extracting explicit and countable concepts from implicit knowledge encoded by a DNN, and (2) using explicit concepts to explain the representation power of DNNs. More crucially, this interaction also enables us to unify the common mechanisms shared by various empirical findings on DNNs.

• *Extracting concepts encoded by DNNs.* Defining the interactions between input variables is a typical approach in XAI [36, 38]. Based on game theory, we defined the multivariate interaction [44, 46] and the multi-order interaction [45] to investigate interactions from different perspectives. In this study, we first demonstrate that game-theoretic interactions are faithful (Theorem 1) and very sparse (Remard 1). [20] further found that salient interactions were usually discriminative and shared by different samples and different DNNs. These findings enabled us to consider salient interactions as concepts encoded by a DNN. Based on this, [28] formulated the optimal baseline values in game-theoretic explanations for DNNs. Furthermore, [6] investigated the different behaviors of the DNN when encoding shapes and textures. [7] further found that salient interactions usually represented the prototypical concepts encoded by a DNN.

• *Game-theoretic interactions enable us to explain the representation power of DNNs.* We used interactions to explain the various capacities of a DNN, including its adversarial robustness [27,39], adversarial transferability [40], and generalization power [45, 50]. [9] proved that a DNN is less likely to encode interactions of the intermediate complexity. In comparison, [29] proved that a Bayesian neural network is less likely to encode complex interactions, thereby avoiding over-fitting.

• *Game-theoretic interactions also reveal the common mechanism underlying many empirical findings.* [10] discovered that the interactions could be considered as elementary components of fourteen attribution methods. [49] proved that the reduction of interactions is the common utility of twelve previous methods of boosting adversarial transferability.

3. Method

3.1. Causal graph based on interactive concepts

In this paper, we discover and prove a concept-emerging phenomenon that the inference logic of a DNN on an input sample can be represented as a causal graph, in which each causal pattern can be considered as an interactive concept[1]. Thus, in order to clarify this phenomenon, let us

[1] Note that unlike previous studies [12], the concept in this paper is defined based on interactions between input variables.

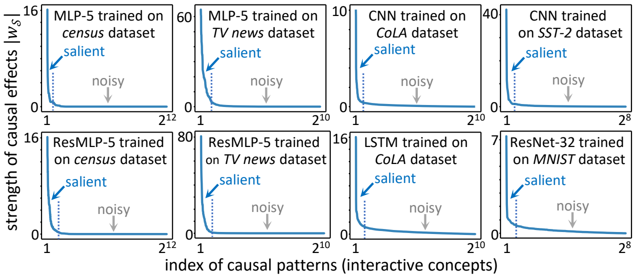

Figure 2. Strength of causal effects of different causal patterns shown in descending order. It shows that sparse causality (sparse interactive concepts) is universal for various DNNs.

first introduce how to build the causal graph. Given a pre-trained DNN $v(\cdot)$ and an input sample $\boldsymbol{x}$ with n variables $\mathcal{N} = \{1, 2, \ldots, n\}$ (*e.g.*, a sentence with n words), let $v(\boldsymbol{x}) \in \mathbb{R}$ denote the DNN's output[2] on the sample $\boldsymbol{x}$. Then, the causal graph corresponding to the inference logic on $\boldsymbol{x}$ is shown in Fig. 1(b). As Fig. 1(b) shows, each source node X_i $(i = 1, ..., n)$ in the bottom layer represents the binary state of whether the i-th input variable is masked ($X_i = 0$) or not ($X_i = 1$). The second layer consists of a set Ω of all causal patterns. Each causal pattern $\mathcal{S} \in \Omega$ represents the AND relationship between a subset of input variables $\mathcal{S} \subseteq \mathcal{N}$. For example, in Fig. 1(b), the co-appearance of the three words in $\mathcal{S} = \{$*take, it, easy*$\}$ forms a phrase meaning "calm down". In other words, only when all three words are present, the causal pattern $\mathcal{S}$ will be triggered, denoted by $C_\mathcal{S} = 1$; otherwise, $C_\mathcal{S} = 0$. As the output of the causal graph, the single sink node Y depends on triggering states $C_\mathcal{S}$ of all causal patterns in Ω. Thus, the transition probability in this causal graph is given as follows.

$$
\begin{aligned}
P(C_\mathcal{S} = 1 | X_1, X_2, ..., X_n) &= \prod\nolimits_{i \in \mathcal{S}} X_i, \\
P(Y | \{C_\mathcal{S} | \mathcal{S} \in \Omega\}) &= \mathbb{1}\left(Y = \sum\nolimits_{\mathcal{S} \in \Omega} w_\mathcal{S} \cdot C_\mathcal{S}\right),
\end{aligned}
\tag{1}
$$

where $Y \in \{v(x_S) | S \subseteq N\}$. $P(C_\mathcal{S} = 0 | X_1, X_2, ..., X_n) = 1 - P(C_\mathcal{S} = 1 | X_1, X_2, ..., X_n)$. $\mathbb{1}(\cdot)$ refers to the indicator function.

$w_\mathcal{S}$ can be understood as the ***causal effect*** of the pattern $\mathcal{S}$ to the output Y. Specifically, each triggered causal pattern $C_\mathcal{S}$ will contribute a certain causal effect $w_\mathcal{S}$ to the DNN's output. For example, the triggered causal pattern *"take it easy"* would contribute a considerable additional effect $w_\mathcal{S} > 0$ that pushes the DNN's output towards the positive meaning "calm down." The quantification of the causal effect $w_\mathcal{S}$ will be introduced later.

According to Eq. (1), the causal relationship between $C_\mathcal{S}$ ($\mathcal{S} \in \Omega$) and the output Y in the causal graph can be specified by the following structural causal model (SCM) [25].

$$
Y(X) = \sum\nolimits_{\mathcal{S} \in \Omega} w_\mathcal{S} \cdot C_\mathcal{S}(X) \tag{2}
$$

• Faithfulness of the causal graph. In this paragraph, we prove that there exists at least one causal graph parameterized by $\{w_\mathcal{S}\}$ in Eq. (1) that can faithfully mimic the inference logic of a DNN on the sample $\boldsymbol{x}$. Specifically, given an input sample $\boldsymbol{x}$ with n variables, we have 2^n ways to mask input variables in $\boldsymbol{x}$, and generate 2^n different masked samples. If the output Y of a causal graph can always mimic the DNN's output[2] on all the 2^n input samples, we can consider that the causal graph is faithful. To this end, given a subset of input variables $\mathcal{S} \subseteq \mathcal{N}$, let $\boldsymbol{x}_\mathcal{S}$ denote the masked sample, where variables in $\mathcal{N} \backslash \mathcal{S}$ are masked, and other variables in $\mathcal{S}$ keep unchanged. Let $v(\boldsymbol{x}_\mathcal{S})$ and $Y(\boldsymbol{x}_\mathcal{S})$ denote the DNN's output[2] and the causal graph's output on this sample $\boldsymbol{x}_\mathcal{S}$, respectively.

Theorem 1 (Proof in Appendix C)**.** *Given a certain input $\boldsymbol{x}$, let the causal graph in Fig. 1 encode 2^n causal patterns,* i.e., $\Omega = 2^\mathcal{N} = \{\mathcal{S} : \mathcal{S} \subseteq \mathcal{N}\}$. *If the causal effect $w_\mathcal{S}$ of each causal pattern $\mathcal{S} \in \Omega$ is measured by the Harsanyi dividend [15],* i.e. $w_\mathcal{S} \triangleq \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} \cdot v(\boldsymbol{x}_{\mathcal{S}'})$*, then the causal graph faithfully encodes the inference logic of the DNN, as follows.*

$$
\forall \mathcal{S} \subseteq \mathcal{N}, \;\; Y(\boldsymbol{x}_\mathcal{S}) = v(\boldsymbol{x}_\mathcal{S}) \tag{3}
$$

In fact, the Harsanyi dividend $w_\mathcal{S}$ was first proposed in game theory to measure the interaction between players. Here, we first use it in the SCM to explain the causal effect of each causal pattern $\mathcal{S} \subseteq \mathcal{N}$ for the DNN's inference.

Theorem 1 proves the faithfulness of using such a causal graph to represent the inference logic of the DNN on a certain sample $\boldsymbol{x}$. In other words, we can exactly disentangle/explain the DNN output on any masked sample into the causal effects. It ensures that *we can use the causal graph to predict DNN outputs on randomly masked samples, thereby showing the trustworthiness of the causal graph.* In comparison, previous explanation methods [1,4,23,30,42] cannot mimic inferences on the masked samples (*i.e.,* not satisfying the faithfulness in Theorem 1). Note that no matter whether input variables are dependent or not, the faithfulness will not be affected, *i.e.*, the causal graph can always accurately mimic the DNN's output on all 2^n possible masked input samples.

However, different original samples $\boldsymbol{x}$ mainly trigger different sets of causal patterns and generate different causal graphs. For example, given a cat image, pixels on the head (in $\mathcal{S}$) may form a head pattern, and the DNN may assign a significant effect $w_\mathcal{S}$ on the pattern. Whereas, we cannot find the head pattern in a bus image, so the same set of pixels $\mathcal{S}$ in the bus image probably do not form any meaningful pattern and have ignorable effect $w_\mathcal{S} \approx 0$.

Specifically, given the sample $\boldsymbol{x}$, each masked sample $\boldsymbol{x}_\mathcal{S}$ is implemented by masking all variables in $\mathcal{N} \backslash \mathcal{S}$ using baseline values just like in [2, 8], as follows.

$$
(\boldsymbol{x}_\mathcal{S})_i = \begin{cases} x_i, & i \in \mathcal{S} \\ r_i, & i \in \mathcal{N} \backslash \mathcal{S} \end{cases}, \tag{4}
$$

[2] Note that people can apply different settings for the DNN's output $v(\boldsymbol{x})$. In particular, in the multi-category classification task, we set $v(\boldsymbol{x}) = \log \frac{p(y = y^{\text{truth}} | \boldsymbol{x})}{1 - p(y = y^{\text{truth}} | \boldsymbol{x})} \in \mathbb{R}$ by following [9].

where $\boldsymbol{r} = [r_1, r_2, \ldots, r_n]$ denotes the baseline values of the n input variables. The DNN's output $v(\boldsymbol{x}_\mathcal{S})$[2] is computed by taking the masked sample $\boldsymbol{x}_\mathcal{S}$ as the input. According to the SCM in Eq. (2), the output $Y(\boldsymbol{x}_\mathcal{S})$ of the causal graph is computed as $Y(\boldsymbol{x}_\mathcal{S}) = \sum_{\mathcal{T}\in\Omega} w_\mathcal{T} \cdot C_\mathcal{T}(\boldsymbol{x}_\mathcal{S}) = \sum_{\mathcal{T}\subseteq\mathcal{S},\mathcal{T}\in\Omega} w_\mathcal{T}$. In particular, $Y(\boldsymbol{x} = \boldsymbol{x}_\mathcal{N}) = \sum_{\mathcal{S}\in\Omega} w_\mathcal{S}$. In Section 3.2, we will introduce how to learn optimal baseline values r_i that further enhance the conciseness of the causal graph.

• **Generality of causal patterns.** Besides, we also prove that the above causal effects $w_\mathcal{S}$ based on Harsanyi dividends satisfy *the efficiency, linearity, dummy, symmetry, anonymity, recursive, and interaction distribution axioms* in game theory (see Appendix B and D.1), which further demonstrates the trustworthiness of the causal effects. More crucially, we also prove that causal effects $w_\mathcal{S}$ can explain the elementary mechanism of existing game-theoretic metrics. Please see Appendix D.2 for the proof.

Theorem 2 (Connection to the Shapley value, proved by [15])**.** *Let $\phi(i)$ denote the Shapley value [31] of an input variable i. Then, the Shapley value $\phi(i)$ can be explained as the result of uniformly assigning causal effects to each involving variable i,* i.e., $\phi(i) = \sum_{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}} \frac{1}{|\mathcal{S}|+1} w_{\mathcal{S}\cup\{i\}}$.

The Shapley value [31] was first proposed in game theory and has been used by previous studies [23] to estimate attributions of input variables in the DNN. The Shapley value satisfies four satisfactory axioms and is widely considered as a relatively fair estimation of attributions. Theorem 2 proves that the Shapley value can be considered as a re-allocation of causal effects to input variables.

In Appendix D.2, we further prove that the Shapley interaction index [13] and the Shapley Taylor interaction index [36] can also be understood as the assignment of causal effects $w_\mathcal{S}$ to different coalitions.

3.2. Discovering and boosting the conciseness of the causal graph

Remark 1. *Given a DNN $v(\cdot)$ and an input sample $\boldsymbol{x}$ with n variables, we can find a small set of causal patterns Ω subject to $|\Omega| \ll 2^n$, such that the DNN's output can be approximated by the causal graph's output,* i.e. $\forall \mathcal{S} \subseteq \mathcal{N},\ Y(\boldsymbol{x}_\mathcal{S}) \approx v(\boldsymbol{x}_\mathcal{S})$.

• **Discovering the conciseness.** We have discovered that lots of DNNs with various architectures trained for different tasks can all be explained using sparse causal patterns. Although Theorem 1 indicates that the causal graph needs to encode 2^n causal patterns to precisely fit the DNN's output on all the 2^n masked samples, Remark 1 shows a common phenomenon that the causal effects $w_\mathcal{S}$ extracted from the DNN are usually very sparse. To this end, we trained various DNNs for different tasks, and Fig. 2 shows the strength of causal effects $|w_\mathcal{S}|$ in descending order for various DNNs. We found that most causal patterns had little influence on the output with negligible values $|w_\mathcal{S}| \approx 0$, and they were termed ***noisy causal patterns***. Only a few causal patterns had considerable effects $|w_\mathcal{S}|$, and they were termed ***salient causal patterns***. Furthermore, we also conducted experiments in Section 4.2, and Figs. 3, 4, and 6 show that we could use a small number of causal patterns (empirically 10 to 100 causal patterns for most DNNs) in Ω to approximate the DNN's output, as stated in Remark 1.

• **Boosting the conciseness.** Inspired by Remark 1, we aim to learn a more concise causal graph. To this end, we propose the following objective of learning faithful and sparse causal effects $w_\mathcal{S}$.

$$\begin{aligned} &\min\nolimits_{\boldsymbol{w},\Omega}\ \text{unfaith}(\boldsymbol{w}_\Omega)\ \ s.t.\ \ |\Omega| \le M \\ \Leftrightarrow\ &\min\nolimits_{\boldsymbol{w},\Omega}\ \text{unfaith}(\boldsymbol{w}_\Omega)\ \ s.t.\ \ \|\boldsymbol{w}_\Omega\|_0 \le M, \\ &\text{unfaith}(\boldsymbol{w}_\Omega) = \sum\nolimits_{\mathcal{S}\subseteq\mathcal{N}} \Big[v(\boldsymbol{x}_\mathcal{S}) - Y_{\boldsymbol{w}_\Omega}(\boldsymbol{x}_\mathcal{S})\Big]^2 \end{aligned} \tag{5}$$

where $\boldsymbol{w}_\Omega \overset{\text{def}}{=} [w'_{\mathcal{S}_1}, w'_{\mathcal{S}_2}, ..., w'_{\mathcal{S}_{2^n}}]$. If $\mathcal{S} \in \Omega$, then $w'_\mathcal{S} = w_\mathcal{S}$; otherwise, $w'_\mathcal{S} = 0$. The L_0-norm $\|\boldsymbol{w}_\Omega\|_0$ refers to the number of non-zero elements in $\boldsymbol{w}_\Omega$, thereby $\|\boldsymbol{w}_\Omega\|_0 = |\Omega|$. In this way, the above objective function enables people to use a small number of causal patterns to explain the DNN.

However, direct optimization of Eq. (5) is difficult. Therefore, we propose several techniques to learn sparse causal effects based on Eq. (5) to faithfully mimic the DNN's outputs on numerous masked samples. The following paragraphs will introduce how to relax the Harsanyi dividend in Theorem 1 by removing noisy causal patterns and learning the optimal baseline value, so as to boost the sparsity of causal effects. Besides, we also discovered that adversarial training [24] can make the DNN encode much more sparse causal effects.

First, boosting conciseness by learning the optimal baseline value. In fact, the sparsity of causal patterns does not only depend on the DNN itself, but it is also determined by the choice of baseline values in Eq. (4). Specifically, input variables are masked by their baseline values $\boldsymbol{r} = [r_1, r_2, \ldots, r_n]$ to represent their absence states in the computation of causal effects. Thus, $\boldsymbol{w}_\Omega$ can be represented as a function of $\boldsymbol{r}$, *i.e.*, $\boldsymbol{w}_\Omega(\boldsymbol{r})$. To this end, some recent studies [2, 8, 28] defined baseline values from a heuristic perspective, *e.g.* simply using mean/zero baseline values [8, 37]. However, it still remains an open problem to define optimal baseline values.

Thus, we further boost the sparsity of causal patterns by learning the optimal baseline values that enhance the conciseness of the causal graph. However, it is difficult to learn optimal baseline values by directly optimizing Eq. (5). To this end, we relax the optimization problem in Eq. (5) (L_0 regression) as a Lasso regression (L_1 regression) as follows.

$$\begin{aligned} &\min\nolimits_{\Omega,\boldsymbol{r}}\ \text{unfaith}(\boldsymbol{w}_\Omega)\ \ s.t.\ \ \|\boldsymbol{w}_\Omega\|_0 \le M \\ \Leftrightarrow\ &\min\nolimits_{\Omega,\boldsymbol{r}}\ \text{unfaith}(\boldsymbol{w}_\Omega) + \lambda\|\boldsymbol{w}_\Omega\|_0 \\ \xRightarrow{\text{relax}}\ &\min\nolimits_{\Omega,\boldsymbol{r}}\ \text{unfaith}(\boldsymbol{w}_\Omega) + \lambda\|\boldsymbol{w}_\Omega\|_1 \end{aligned} \tag{6}$$

We learn optimal baseline values by minimizing the loss $\mathcal{L}(\boldsymbol{r},\Omega) = \mathrm{unfaith}(\boldsymbol{w}_\Omega) + \lambda \cdot \|\boldsymbol{w}_\Omega\|_1$. More crucially, the learning of baseline values is the *safest* way of optimizing $\mathcal{L}(\boldsymbol{r},\Omega)$, because the change of baseline values always ensures $\mathrm{unfaith}(\boldsymbol{w}) = 0$ and just affects $\|\boldsymbol{w}_\Omega\|_1$. In this way, learning baseline values significantly boosts the conciseness of causal effects. In practice, we usually initialize the baseline value r_i as the mean value of the variable i over all samples, and then we constrain r_i within a relatively small range, *i.e.*, $\|r_i - r_i^{\text{initial}}\|^2 \leq \tau$, to represent the absence state[3].

Second, boosting conciseness by neglecting noisy causal patterns. Considering the optimization problem, we use a greedy strategy to remove the noisy causal patterns from $2^{\mathcal{N}} = \{\mathcal{S} : \mathcal{S} \subseteq \mathcal{N}\}$ and keep the salient causal patterns to construct the set $\Omega \subseteq 2^{\mathcal{N}}$ that minimizes the loss $\mathcal{L}(\boldsymbol{r},\Omega)$ in Eq. (6). It is worth noting that we do not directly learn causal effects by blindly optimizing Eq. (6), because automatically optimized causal effects usually lack sufficient support for their physical meanings, while the setting of Harsanyi dividends is a meaningful interaction metric in game theory [15]. The Harsanyi dividend satisfies *the efficiency, linearity, dummy, symmetry axioms* axioms, which ensures the trustworthiness of this metric. In other words, although automatically optimized causal effects can minimize $\mathrm{unfaith}(\boldsymbol{w})$, they still cannot be considered as reliable explanations from the perspective of game theory. Thus, we only recursively remove noisy causal patterns from Ω to update Ω, *i.e.*, $\Omega \leftarrow \Omega \backslash \{\mathcal{S}\}$, without creating any new causal effect outside the paradigm of the Harsanyi dividends in Theorem 1. Specifically, we remove noisy causal patterns by following a greedy strategy, *i.e.*, iteratively removing the noisy causal pattern such that $\mathrm{unfaith}(\boldsymbol{w}_\Omega)$ is minimized in each step. In this way, we just use the set of retained causal patterns, denoted by Ω, to approximate the output, *i.e.*, $v(\boldsymbol{x}) \approx Y(\boldsymbol{x}) = \sum_{\mathcal{S}\in\Omega} w_\mathcal{S} \cdot C_\mathcal{S}(\boldsymbol{x}) = \sum_{\mathcal{S}\in\Omega} w_\mathcal{S}$.

Ratio of the explained causal effects R_Ω. We propose a metric R_Ω to quantify the ratio of the explained salient causal effects in Ω to the overall network output.

$$R_\Omega = \frac{\sum_{\mathcal{S}\in\Omega} |w_\mathcal{S}|}{\sum_{\mathcal{S}\in\Omega} |w_\mathcal{S}| + |\Delta|} \qquad (7)$$

where $\Delta = v(\boldsymbol{x}) - \sum_{\mathcal{S}\in\Omega} w_\mathcal{S}$ denotes effects of the unexplained causal patterns.

Third, discovering that adversarial training boosts the conciseness. As discussed in Section 4.3, we also discover that adversarial training [24] makes the DNN encode more sparse causal patterns than standard training, thus boosting the conciseness of the causal graph.

3.3. Rewriting the causal graph as an AOG

The AOG is a hierarchical graphical model that encodes how semantic patterns are formed for inference, which has been widely used for interpretable knowledge representation [21, 48], object detection [35], *etc.* In this section, we show that the above causal graph can be rewritten into an And-Or graph (AOG), which summarizes common coalitions shared by different causal patterns to further simplify the explanation. According to the SCM in Eq. (2), the causal graph in Section 3.1 actually represents the And-Sum representation encoded by the DNN, *i.e.*, $v(\boldsymbol{x}) \approx \sum_{\mathcal{S}\in\Omega} w_\mathcal{S} \cdot C_\mathcal{S}(\boldsymbol{x}) = \sum_{\mathcal{S}\in\Omega} w_\mathcal{S}$. In fact, such And-Sum representation can be equivalently transformed into an AOG.

The structure of a simple three-layer AOG is shown in Fig. 1(c). Just like the causal graph in Fig. 1(b), at the bottom layer of the AOG in Fig. 1(c), there are n leaf nodes representing n variables of the input sample. The second layer of the AOG has multiple AND nodes, each representing the AND relationship between its child nodes. For example, the AND node $x_4x_5x_6$ indicates the causal pattern $\mathcal{S} = \{x_4, x_5, x_6\}$ with the causal effect $w_\mathcal{S} = 2.0$. The root node is a *noisy OR* node (as discussed in [21]), which sums up effects of all its child AND nodes to mimic the network output, *i.e.*, *output* $= \sum_{\mathcal{S}\in\Omega} w_\mathcal{S} \cdot C_\mathcal{S}$.

Furthermore, in order to simplify the AOG, we extract common coalitions shared by different causal patterns as new nodes to construct a deeper AOG. For example, in Fig. 1(c), input variables x_5 and x_6 frequently co-appear in different causal patterns. Thus, we consider x_5, x_6 as a coalition and add an AND node $\beta = \{x_5, x_6\}$ to represent their co-appearance. Accordingly, the pattern $\{x_4, x_5, x_6\}$ is simplified as $\{x_4, \beta\}$ (see Fig. 1(d)). Therefore, for each coalition / causal pattern $\mathcal{S}$ in an intermediate layer, its triggering state $C_\mathcal{S} = \prod_{\mathcal{S}'\in\text{Child}(\mathcal{S})} C_{\mathcal{S}'}$, where Child($\mathcal{S}$) denotes all input variables or coalitions composing $\mathcal{S}$. *I.e.*, each coalition / causal pattern $\mathcal{S}$ is triggered if and only if all its child nodes in Child($\mathcal{S}$) are triggered.

In order to extract common coalitions, we use the minimum description length (MDL) principle [14] to learn the AOG g as the simplest description of causal patterns. The MDL is a classic way of summarizing patterns from data for decades, which has solid foundations in information theory. Given an AOG g and input variables $\mathcal{N}$, let $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}}$ denote the set of all leaf nodes and AND nodes in the bottom two layers, *e.g.* $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}} = \{x_1, x_2, ..., x_6\} \cup \{\alpha, \beta\}$ in Fig. 1(d). The objective of minimizing the description length $L(g, \mathcal{M})$ is given as follows.

$$\min_{\mathcal{M}} \; L(g,\mathcal{M}) \;\; s.t. \;\; L(g,\mathcal{M}) = L(\mathcal{M}) + L_\mathcal{M}(g), \qquad (8)$$

where $L(\mathcal{M})$ denotes the complexity of describing the set of nodes $\mathcal{M}$, and $L_\mathcal{M}(g)$ denotes the complexity of

[3] The setting of τ is introduced in Section 4.2. Please see Appendix E for more discussions

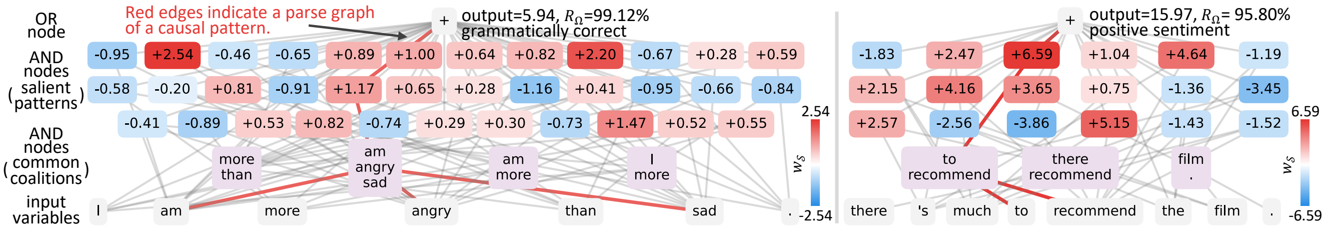

Figure 3. AOGs that explained correct predictions made by the neural network. The networks were trained on (left) the CoLA dataset and (right) the SST-2 dataset, respectively. The red color of nodes in the second layer indicates causal patterns with positive effects, while the blue color represents patterns with negative effects. Red edges indicate the parse graph of a causal pattern.

using nodes in $\mathcal{M}$ to describe patterns in g. The MDL principle usually formulates the complexity (description length) of the set of nodes $\mathcal{M}$ as the entropy $L(\mathcal{M}) = -\kappa \sum_{m\in\mathcal{M}} p(m)\log p(m)$. We set the occurring probability $p(m)$ of the node $m \in \mathcal{M}$ proportional to the overall strength of causal effects of the node m's all parent nodes $\mathcal{S}$, Child$(\mathcal{S}) \ni m$. $\forall m\in\mathcal{M}$, $p(m)=count(m)/\sum_{m'\in\mathcal{M}} count(m')$ s.t. $count(m)=\sum_{\mathcal{S}\in\Omega:\text{Child}(\mathcal{S})\ni m} |w_{\mathcal{S}}|$. $\kappa=10/Z$ is a scalar weight, where $Z=\sum_{\mathcal{S}\in\Omega} |w_{\mathcal{S}}|$. The second term $L_{\mathcal{M}}(g) = -\mathbb{E}_{\mathcal{S}\sim p(\mathcal{S}|g)} \sum_{m\in\mathcal{S}} \log p(m)$ represents the complexity (description length) of using nodes in $\mathcal{M}$ to describe all causal patterns in g. The appearing probability of the causal pattern $\mathcal{S}$ in the AOG g is sampled as $p(\mathcal{S}|g) \propto |w_{\mathcal{S}}|$. The time cost of the MDL method is $O(|\Omega|^2)$. The loss $L(g, \mathcal{M})$ can be minimized by recursively adding common coalitions into $\mathcal{M}$ via the greedy strategy by following [14]. Please see Appendix F for more discussions.

Limitations of the AOG explainer. Although we prove that the AOG explainer is the unique faithful explanation, it is still far from a computationally efficient explanation. Thus, extending the theoretical solution to the practical one is our future work, *e.g.* developing approximated methods or accelerating techniques for computation. In Appendix H, we have discussed some techniques to reduce the time cost on image datasets.

4. Experiments

Datasets and models. We focused on classification/regression tasks based on NLP datasets, image datasets, and tabular datasets. For NLP tasks, we explained LSTMs [17] and CNNs used in [26]. Each model was trained for sentiment classification on the SST-2 dataset [34] or for linguistic acceptability classification on the CoLA dataset [41], respectively. For vision tasks, we explained ResNets [16] and VGG-16 [33] trained on the MNIST dataset [19] and the CelebA dataset [22] (please see Appendix G.2 for results on the CelebA dataset). The tabular datasets included the UCI census income dataset [11], the UCI bike sharing dataset [11], and the UCI TV news channel commercial detection dataset [11]. These datasets were termed *census, bike,* and *TV news* for simplicity. Each tabular dataset was used to train MLPs, LightGBM [18], and XGBoost [5]. For MLPs, we used two-layer MLPs (namely *MLP-2*) and five-layer MLPs (namely *MLP-5*), where each layer contained 100 neurons. Besides, we added a skip-connection [16] to each layer of MLP-5 to build *ResMLP-5*. Please see Appendix G.1 for more details.

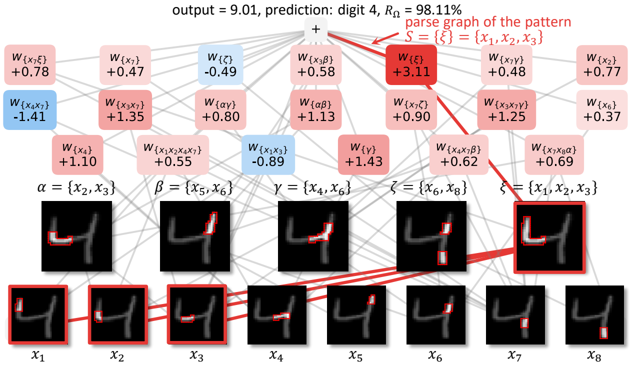

Figure 4. An AOG that explained the prediction made by ResNet-20 trained on the MNIST dataset. Red edges indicate the parse graph of a causal pattern.

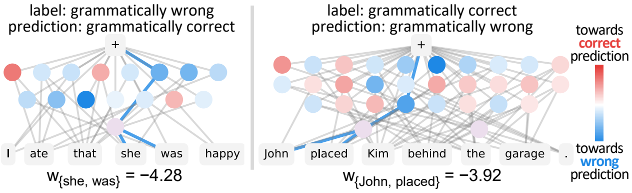
label: grammatically wrong
prediction: grammatically correct
+
I
ate
that
she
was
happy
$w_{\{she, was\}} = -4.28$
label: grammatically correct
prediction: grammatically wrong
+
John
placed
Kim
behind
the
garage
.
$w_{\{John, placed\}} = -3.92$
towards correct prediction
towards wrong prediction

Figure 5. AOGs for a network trained on the CoLA dataset. We randomly highlight a parse graph (blue) in the AOG.

Explaining network inferences and discovering representation flaws of DNNs. Figs. 3 and 4 show AOG explanations for correct predictions in NLP tasks and the image classification task, respectively. The highlighted parse graph in each figure corresponds to a single causal pattern. We only visualized a single parse graph in each AOG for clarity. We found that AOGs extracted meaningful word collocations and typical digit shapes used by the DNN for inference. Besides, Fig. 5 shows AOG explanations for incorrect predictions in the NLP task. Results show that the AOG explainer could reveal the representation flaws that were responsible for incorrect predictions. For example, local correct grammar "she was" in Fig. 5(left) was mistakenly learned to make negative impacts on the linguistic acceptability of the whole

Dataset	Model	Average IoU			
		SI	STI (k=2)	STI (k=3)	ours
Add-Mul dataset [47]	functions in the dataset	0.61	0.27	0.55	**1.00**
Dataset in [28]		0.99	0.50	0.59	**1.00**
Manually labeled And-Or dataset	MLP-5	0.87	0.35	0.69	**0.97**
	ResMLP-5	0.90	0.35	0.69	**0.98**

Table 1. IoU (↑) on synthesized datasets. The AOG explainer correctly extracted causal patterns.

sentence. The phrase "John placed" in Fig. 5(right) directly hurt the linguistic acceptability without considering the complex structure of the sentence. Please see Appendix G.4 for more results.

4.1. Examining whether the AOG explainer reflects faithful causality

In this section, we proposed two metrics to examine whether the AOG explainer faithfully reflected the inference logic encoded by DNNs.

Metric 1: intersection over union (IoU) between causal patterns in the AOG explainer and ground-truth causal patterns. This metric evaluated whether causal patterns (nodes) in the AOG explainer correctly reflected the interactive concepts encoded by the model. Given a model and an input sample, let m denote the number of ground-truth causal patterns $m = |\Omega^{\text{truth}}|$ in the input. Then, for fair comparisons, we also used m causal patterns $\Omega^{\text{top-}m}$ in the AOG explainer with the top-m causal effects $|w_{\mathcal{S}}|$. We measured the IoU between Ω^{truth} and $\Omega^{\text{top-}m}$ as $\textit{IoU} = |\Omega^{\text{top-}m} \cap \Omega^{\text{truth}}| / |\Omega^{\text{top-}m} \cup \Omega^{\text{truth}}|$ to evaluate the correctness of the extracted causal patterns in the AOG explainer. A higher IoU value means a larger overlap between the ground-truth causal patterns and the extracted causal patterns, which indicates higher correctness of the extracted causal patterns.

However, for most realistic datasets and models, people could not annotate the ground-truth patterns, as discussed in [47]. Therefore, we used the off-the-shelf functions with ground-truth causal patterns in the Addition-Multiplication (Add-Mul) dataset [47] and the dataset proposed in [28], to test whether the learned AOGs could faithfully explain these functions. The ground-truth causal patterns of functions in both datasets can be easily determined. For example, for the function $y = x_1x_3 + x_3x_4x_5 + x_4x_6$, $x_i \in \{0, 1\}$ in the Add-Mul dataset, the ground-truth causal patterns are $\Omega^{\text{truth}} = \{\{x_1, x_3\}, \{x_3, x_4, x_5\}, \{x_4, x_6\}\}$ given the input sample $\boldsymbol{x} = [1, 1, ..., 1]$. It was because the multiplication between binary input variables could be considered as the AND relationship, thereby forming explicit ground-truth causal patterns. In other words, the co-appearance of variables in each causal pattern would contribute 1 to the output score y.

Similarly, we also constructed the third dataset containing pre-defined And-Or functions with ground-truth causal patterns, namely the *manually labeled And-Or dataset* (see Appendix G.3). Then, we learned the aforementioned MLP-5 and ResMLP-5 networks to regress each And-Or function. We considered causal patterns in such And-Or functions as ground-truth causal patterns in the DNN.

Explanation methods		TV news		census		bike	
		MLP-5	ResMLP-5	MLP-5	ResMLP-5	MLP-5	ResMLP-5
Attribution -based explanations	Shapley	125.5	130.8	55.6	51.4	1.1E+4	7953.9
	I×G	738.7	2586.1	408.1	1325.1	1.4E+5	1.1E+5
	LRP	317.6	9.4E+4	155.1	1.4E+04	1.4E+5	5.8E+8
	OCC	1386.2	1117.5	638.7	287.4	6.2E+4	3.7E+4
Interaction -based explanations	SI	6231.2	5598.6	2726.1	2719.0	1.2E+5	1.2E+5
	STI (k=2)	182.0	236.0	34.7	38.8	7685.0	5219.8
	STI (k=3)	177.7	252.4	41.0	60.5	1.0E+4	5045.8
ours		**9.4E-12**	**1.1E-11**	**8.5E-12**	**8.5E-12**	**2.6E-9**	**1.9E-9**

Table 2. Unfaithfulness ρ^{unfaith} (↓) of different explanation methods. Our AOG exhibited the lowest unfaithfulness.

As for baseline methods, previous studies usually did not directly extract causal patterns from a trained DNN at a low level as input units. To this end, interaction metrics (such as the Shapley interaction (SI) index [13] and the Shapley-Taylor interaction (STI) index [36]) were widely used to quantify numerical effects of different interactive patterns between input variables on the network output. Thus, we computed interactive patterns with top-ranked SI values, or patterns with top-ranked STI values of orders $k = 2$ and $k = 3$, as competing causal patterns for comparison. Based on the IoU score defined above, Table 1 shows that our AOG explainer successfully explained much more causal patterns than other interaction metrics.

Metric 2: evaluating faithfulness of the AOG explainer. We also proposed a metric ρ^{unfaith} to evaluate whether an explanation method faithfully extracted causal effects encoded by DNNs. As discussed in Section 3.2, if the quantified causal effects $\boldsymbol{w}$ are faithful, then they are supposed to minimize $\text{unfaith}(\boldsymbol{w})$. Therefore, according to the SCM in Eq. (2), we defined $\rho^{\text{unfaith}} = \mathbb{E}_{\mathcal{S} \subseteq \mathcal{N}}[v(\boldsymbol{x}_{\mathcal{S}}) - \sum_{S' \subseteq S} w_{\mathcal{S}'})]^2$ to measure the unfaithfulness. As mentioned above, we considered the SI values and STI values as numerical effects $w_{\mathcal{S}}$ of different interactive patterns $\mathcal{S}$ on a DNN's inference. Besides, we could also consider that attribution-based explanations quantified the causal effect $w_{\{i\}}$ of each variable i. Therefore, Table 2 compares the extracted causal effects in the AOG with SI values, STI values, and attribution-based explanations (including the Shapley value [31], Input×Gradient [32], LRP [3], and Occlusion [43]). Our AOG explainer exhibited much lower ρ^{unfaith} values than baseline methods.

4.2. Conciseness of the AOG explainer

The conciseness of an AOG depends on a trade-off between the ratio of the explained causal effects R_{Ω} and the simplicity of the explanation. In this section, we evaluated the effects of baseline values on the simplicity of the AOG explainer, and examined the relationship between the ratio of causal effects being explained and the simplicity of the AOG explainer.

Effects of baseline values on the conciseness of ex-

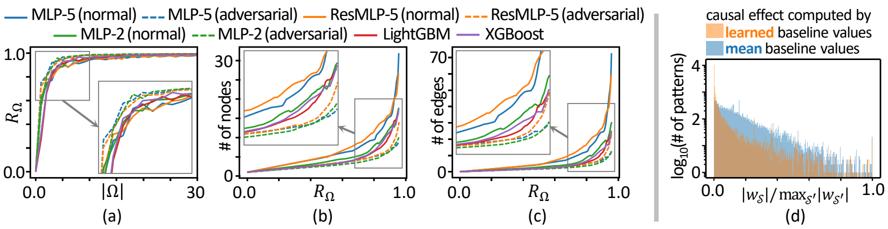

Figure 6. (a) The change of R_Ω along with the number of causal patterns $|\Omega|$ in AOGs. (b,c) The change of the node/edge number in AOGs along with R_Ω. (d) The histogram of re-scaled causal effects. The learned baseline values boosted the sparsity of causal patterns in the AOG explainer. Please see Appendix G.6 and G.7 for results on other datasets.

planations. In this experiment, we explored whether the learning of baseline values in Section 3.2 could boost the sparsity of causal patterns. To this end, we followed [8] to initialize baseline values of input variables as their mean values over different samples. Then, we learned baseline values via Eq. (6). The baseline value r_i of each input variable i was constrained within a certain range around the data average, *i.e.*, $\|r_i - \mathbb{E}_x[x_i]\|^2 \leq \tau$. In experiments, we set $\tau = 0.01 \cdot \text{Var}_x[x_i]$, where $\text{Var}_x[x_i]$ denotes the variance of the i-th input variable over different samples. Fig. 6(d) shows the histogram of the relative strength of causal effects $\frac{|w_S|}{\max_{S' \subseteq N} |w_{S'}|}$, which was re-scaled to the range of $[0, 1]$. *Compared with mean baseline values, the learned baseline values usually generated fewer causal patterns with significant strengths, which boosted the sparsity of causal effects and enhanced the conciseness of explanations.* In this experiment, we used MLP-5 and computed relative strengths of causal effects in 20 randomly selected samples in the TV news dataset. Please see Appendix G.7 for more results.

Ratio of the explained causal effects R_Ω. There was a trade-off between faithfulness (the ratio of explained causal effects) and conciseness of the AOG. A good explanation was supposed to improve the simplicity while keeping a large ratio of causal effects being explained. As discussed in Section 3.2, we just used causal patterns in Ω to approximate the DNN's output. Fig. 6(a) shows the relationship between $|\Omega|$ and the ratio of the explained causal effects R_Ω in different models based on the TV news dataset. *When we used a few causal patterns, we could explain most effects of causal patterns to the DNN's output.* Fig. 6(b,c) shows that the node and edge number of the AOG increased along with the increase of R_Ω.

4.3. Effects of adversarial training

In this experiment, we learned MLP-2, MLP-5, and ResMLP-5 on the TV news dataset via adversarial training [24]. Fig. 6(a) shows that compared with normally trained models, we could use less causal patterns (smaller $|\Omega|$) to explain the same ratio of causal effects R_Ω in adversarially trained models. Moreover, Fig. 6(b,c) also shows that AOGs for adversarially trained models contained fewer nodes and edges than AOGs for normally trained models. This indicated that *adversarial training made models encode more sparse causal patterns than normal training*.

		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**

Table 3. Jaccard similarity between two models. Two adversarially trained models were more similar than two normally trained ones.

Besides, *adversarial training also made different models encode common patterns*. To this end, we trained different pairs of models with the same architecture but with different initial parameters. Given the same input, we measured the Jaccard similarity coefficient between causal effects of each pair of models, in order to examine whether the two models encoded similar causal patterns. Let w_S and w'_S denote causal effects in the two models. The Jaccard similarity coefficient was computed as $J = \frac{\sum_{S \subseteq N} \min(|w_S|, |w'_S|)}{\sum_{S \subseteq N} \max(|w_S|, |w'_S|)}$. A high Jaccard similarity indicated that the two models encoded similar causal patterns for inference. Table 3 shows that the similarity between two adversarially trained models was significantly higher than that between two normally trained models. This indicated adversarial training made different models encode common causal patterns for inference.

5. Conclusion

In this paper, we discover and study the concept-emerging phenomenon in a DNN. Specifically, we show that the inference logic of a DNN can usually be mimicked by a sparse causal graph. To this end, we theoretically prove and experimentally verify the faithfulness of using a sparse causal graph to represent interactive concepts encoded in a DNN. We also propose several techniques to boost the conciseness of such causal representation. Furthermore, we show that such a causal graph can be rewritten as an AOG, which further simplifies the explanation. The AOG explainer provides new insights for understanding the inference logic of DNNs.

Acknowledgements. This work is partially supported by the National Nature Science Foundation of China (62276165), National Key R&D Program of China (2021ZD0111602), Shanghai Natural Science Foundation (21JC1403800,21ZR1434600), National Nature Science Foundation of China (U19B2043).

References

[1] Kamil Adamczewski, Frederik Harder, and Mijung Park. Bayesian importance of features (bif). *arXiv preprint arXiv:2010.13872*, 2010. 3

[2] Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International Conference on Machine Learning*, pages 272–281. PMLR, 2019. 3, 4

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 7

[4] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 882–891. PMLR, 2018. 3

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 6

[6] Xu Cheng, Chuntung Chu, Yi Zheng, Jie Ren, and Quanshi Zhang. A game-theoretic taxonomy of visual concepts in dnns. *arXiv preprint arXiv:2106.10938*, 2021. 2

[7] Xu Cheng, Xin Wang, Haotian Xue, Zhengyang Liang, and Quanshi Zhang. A hypothesis for the aesthetic appreciation in neural networks. *arXiv preprint arXiv::2108.02646*, 2021. 2

[8] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*, 2017. 3, 4, 8

[9] Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang. Discovering and explaining the representation bottleneck of DNNS. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 2, 3

[10] Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, Ziwei Yang, Zheyang Li, and Quanshi Zhang. Understanding and unifying fourteen attribution methods with taylor interactions. *arXiv preprint*, 2022. 2

[11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. 6

[12] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9273–9282, 2019. 2

[13] Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999. 2, 4, 7

[14] Mark H Hansen and Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001. 5, 6

[15] John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. 3, 4, 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[18] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017. 6

[19] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 6

[20] Mingjie Li and Quanshi Zhang. Does a neural network really encode symbolic concept? *arXiv preprint arXiv:2302.13080*, 2023. 2

[21] Xilai Li, Xi Song, and Tianfu Wu. Aognets: Compositional grammatical architectures for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6220–6230, 2019. 5

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 6

[23] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017. 3, 4

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 4, 5, 8

[25] Judea Pearl. *Causality*. Cambridge university press, 2009. 1, 3

[26] A Rakhlin. Convolutional neural networks for sentence classification. *GitHub*, 2016. 6

[27] Jie Ren, Die Zhang, Yisen Wang, Lu Chen, Zhanpeng Zhou, Yiting Chen, Xu Cheng, Xin Wang, Meng Zhou, Jie Shi, and Quanshi Zhang. Towards a unified game-theoretic view of adversarial perturbations and robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3797–3810. Curran Associates, Inc., 2021. 2

[28] Jie Ren, Zhanpeng Zhou, Qirui Chen, and Quanshi Zhang. Can we faithfully represent masked states to compute shapley values on a dnn? In *The eleventh International Conference on Learning Representations, ICLR 2023, Kigali Rwanda, May 1-5, 2023*, 2023. 2, 4, 7

[29] Qihan Ren, Huiqi Deng, Yunuo Chen, Siyu Lou, and Quanshi Zhang. Bayesian neural networks tend to ignore complex and sensitive concepts. *arXiv preprint arXiv:2302.13095*, 2023. 2

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 3

[31] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953. 2, 4, 7

[32] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016. 7

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[34] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013. 6

[35] Xi Song, Tianfu Wu, Yunde Jia, and Song-Chun Zhu. Discriminatively trained and-or tree models for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3278–3285, 2013. 5

[36] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International Conference on Machine Learning*, pages 9259–9268. PMLR, 2020. 2, 4, 7

[37] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328, 2017. 4

[38] Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *arXiv preprint arXiv:2203.00870*, 2022. 2

[39] Xin Wang, Shuyun Lin, Hao Zhang, Yufei Zhu, and Quanshi Zhang. Interpreting attributions and interactions of adversarial attacks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1075–1084. IEEE, 2021. 2

[40] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. A unified approach to interpreting and boosting adversarial transferability. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2

[41] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. 6

[42] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. INVASE: instance-wise variable selection using neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 3

[43] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 7

[44] Die Zhang, Hao Zhang, Huilin Zhou, Xiaoyi Bao, Da Huo, Ruizhao Chen, Xu Cheng, Mengyue Wu, and Quanshi Zhang. Building interpretable interaction trees for deep NLP models. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14328–14337. AAAI Press, 2021. 2

[45] Hao Zhang, Sen Li, Yinchao Ma, Mingjie Li, Yichen Xie, and Quanshi Zhang. Interpreting and boosting dropout from a game-theoretic view. In *International Conference on Learning Representations*, 2021. 2

[46] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10877–10886. AAAI Press, 2021. 2

[47] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In *AAAI*, 2021. 7

[48] Quanshi Zhang, Jie Ren, Ge Huang, Ruiming Cao, Ying Nian Wu, and Song-Chun Zhu. Mining interpretable aog representations from convolutional networks via active question answering. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 5

[49] Quanshi Zhang, Xin Wang, Jie Ren, Xu Cheng, Shuyun Lin, Yisen Wang, and Xiangming Zhu. Proving common mechanisms shared by twelve methods of boosting adversarial transferability. *arXiv preprint arXiv:2207.11694*, 2022. 2

[50] Huilin Zhou, Hao Zhang, Huiqi Deng, Dongrui Liu, Wen Shen, Shih-Han Chan, and Quanshi Zhang. Concept-level explanation for the generalization of a dnn. *arXiv preprint arXiv:2302.13091*, 2023. 2