Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A survey towards the first principle of explaining DNNs: interactions explain the learning dynamics*

Huilin ZHOU, Qihan REN, Junpeng ZHANG, Quanshi ZHANG‡

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

E-mail: zhouhuilin116@sjtu.edu.cn; renqihan@sjtu.edu.cn; zhangjp63@sjtu.edu.cn; zqs1022@sjtu.edu.cn

Received Nov. 25, 2024; Revision accepted Jan. 26, 2025; Crosschecked

Unedited

Abstract: Most explanation methods are designed in an empirical manner, so whether there exists a first-principal explanation of deep neural networks (DNNs) becomes the next core scientific problem in explainable artificial intelligence (XAI). Although it is still an open problem, in this paper, we discuss whether the interaction-based explanation can serve as the first-principal explanation of a DNN. The strong explanatory power of interaction theory comes from the following aspects: (1) it establishes a new axiomatic system to quantify the decision-making logic of a DNN into a set of symbolic interaction concepts; (2) it simultaneously explains various deep learning phenomena, such as generalization power, adversarial sensitivity, representation bottleneck, and learning dynamics; (3) it provides mathematical tools that uniformly explain the mechanisms of various empirical attribution methods and empirical adversarial-transferability-boosting methods; and (4) it explains the extremely complex learning dynamics of a DNN by analyzing the two-phase dynamics of interaction complexity, which further reveals the internal mechanism of why and how the generalization power/adversarial sensitivity of a DNN changes during the learning process.

Key words: The first-principal explanation; Theory of equivalent interactions; Two-phase dynamics of interactions; Dynamics of the generalization power of a DNN

https://doi.org/10.1631/FITEE.2401025 **CLC number:**

1 Introduction

Although deep neural networks (DNNs) have exhibited superior performance in various tasks, their decision-making logic is not transparent. The lack of interpretability is especially critical in high-risk tasks, such as medical diagnosis and autonomous driving. Many studies (Simonyan et al., 2014; Yosinski et al., 2015; Bau et al., 2017; Selvaraju et al., 2017; Kim et al., 2018) on explainable AI aim to explain the complex learning behaviors of a DNN. However, the explainable AI community has not reached a consensus on "the first-principal explanation" that can faithfully explain both the knowledge and the generalization power of neural networks. How to define the first-principal explanation of a DNN remains an open problem.

As a result, many explanation methods are mainly designed in an empirical manner. To this end, the first-principal explanation is supposed to be a unified theory system with a set of axioms and theorems that can comprehensively explain the internal mathematical mechanisms of various deep-learning phenomena, including knowledge representation, generalization power, adversarial robustness, and learning dynamics.

Therefore, in this paper, we aim to discuss the potential of the theory system of interactions serving as the first-principal explanation of a DNN, including both the theoretical achievements of interaction theory and its limitations. Specifically, we will an-

‡ Corresponding author

* Project supported by the National Science and Technology Major Project (No. 2021ZD0111602), the National Natural Science Foundation of China (Nos. 62276165 and 92370115), and the Shanghai Natural Science Foundation (No. 24ZR1491700)

ORCID: Huilin ZHOU, https://orcid.org/0000-0001-8834-4665

© Zhejiang University Press 2025

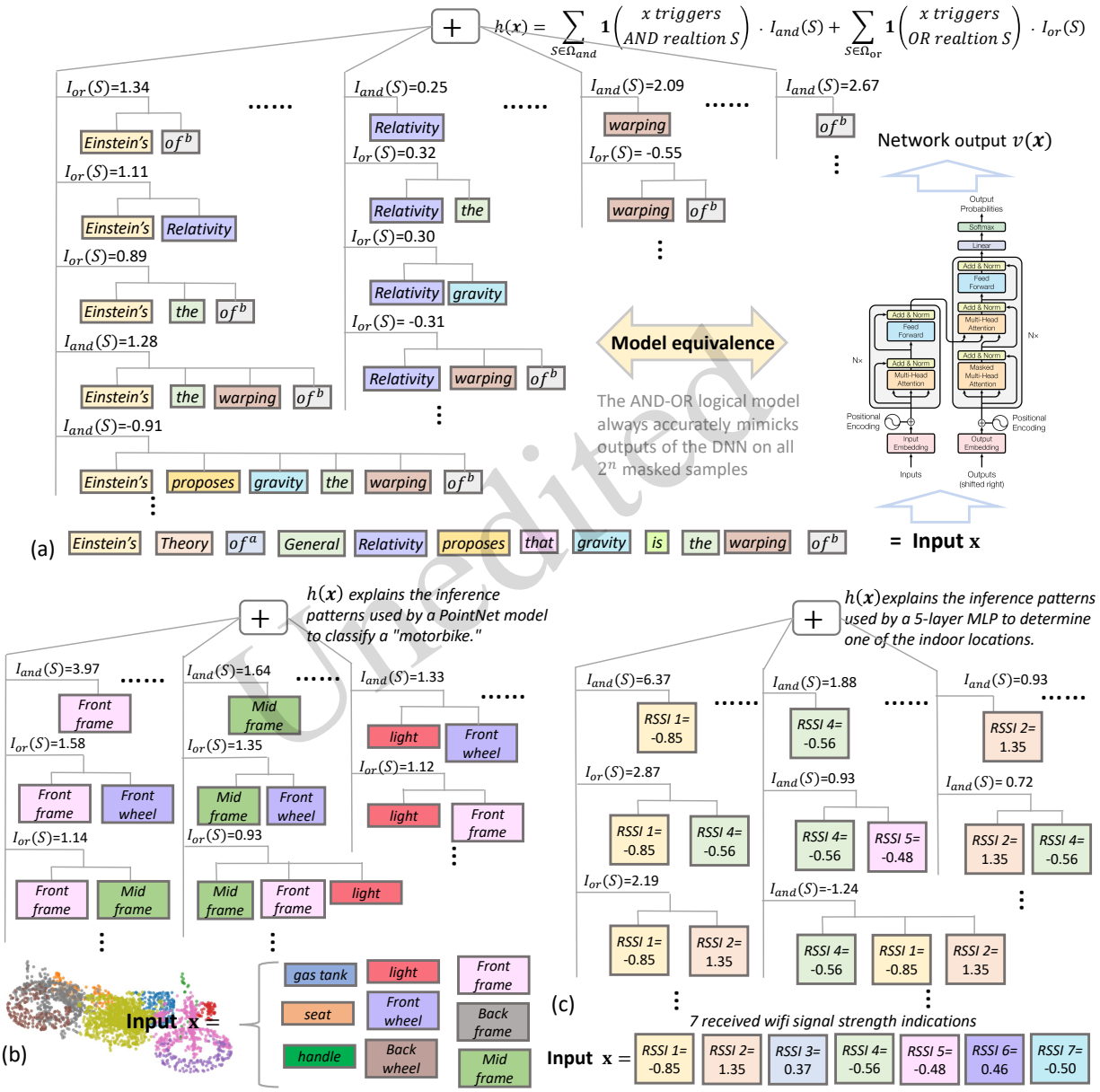

Fig. 1 AND-OR logical model that explains the inference patterns used by a DNN. The inference of a DNN on a certain input sample is equivalent to a surrogate logical model that uses a small number of AND-OR interactions for inference. Each interaction corresponds to an AND relationship or an OR relationship between a set of input variables. We visualize (a) interactions between different tokens in inputs from the sentence dataset used by an LLM (Shen et al., 2023); (b) interactions between different point cloud clusters in inputs from the ShapeNet dataset (Yi et al., 2016) used by a PointNet model (Qi et al., 2017) to classify a "motorbike;" and (c) interactions between different RSSIs in inputs from the UCI wireless indoor localization dataset (Dua and Graff, 2017) used by a five-layer MLP for indoor localization. DNN, deep neural networks; RSSIs, received WiFi signal strength indications.

Unedited

alyze the representation power of interaction theory from various perspectives, *e.g.*, explaining the knowledge representation of a DNN, explaining the mechanism behind the performance, using interactions to unify empirical deep-learning methods, and, more importantly, explaining the extremely complex learning dynamics of a DNN. All these perspectives are necessary issues *w.r.t.* the first-principal explanation of a DNN.

2 Interaction-based explanation

2.1 Background

Before the discussion of the first-principal explanation, let us first revisit the long-lasting disappointing view of the faithfulness of the post-hoc explanations of a DNN. One of the most fundamental and challenging issues for an interpretability theory is whether the internal decision-making logic of neural networks can be accurately and rigorously explained. However, many well-known scholars have considered that faithfully explaining the decision-making logic of a DNN as symbolic inference patterns was an impossible task. Specifically, Rudin (2019) called to stop using inaccurate post-hoc explanation methods to explain black-box machine learning models for high-stakes applications, and use interpretable models instead. Ghassemi et al. (2021) claimed that using current explanation methods to aid medical decisions was a false hope. Adebayo et al. (2018) found that some existing saliency methods were misleading, because they were independent of both the model and the data generating process.

Therefore, faithfully explaining the output score of a DNN has become one of the biggest challenges in the explainable AI community. The faithfulness of explaining the output score of a DNN is the first challenge.

2.2 Preliminaries: definition of interactions

As a direct response to the above disappointment of post-hoc explanations, Ren et al. (2023a) proposed the use of interactions between different input variables in the input sample to represent the knowledge (or concepts) encoded by the DNN. Given a trained DNN $v : \mathbb{R}^n \to \mathbb{R}$, and a specific input sample $\boldsymbol{x} = [x_1, ..., x_n]^\top \in \mathbb{R}^n$, indexed by $N = \{1, 2, ..., n\}$. The DNN usually only encodes a small set of salient interactions $\Omega_{\text{and}}, \Omega_{\text{or}}$ for inference on the input $\boldsymbol{x}$. As Figure 1 shows, a surrogate logical model based on AND-OR interactions $h(\boldsymbol{x}_T) = \sum_{S\in\Omega_{\text{and}}} \mathbb{1}\left(\substack{\boldsymbol{x}_T \text{ triggers} \\ \text{AND relation } S}\right) \cdot I_{\text{and}}(S) + \sum_{S\in\Omega_{\text{or}}} \mathbb{1}\left(\substack{\boldsymbol{x}_T \text{ triggers} \\ \text{OR relation } S}\right) \cdot I_{\text{or}}(S)$ can accurately mimic network outputs $v(\boldsymbol{x}_T)$ on all randomly masked input samples $\boldsymbol{x}_T$, where $\boldsymbol{x}_T$ denotes *w.r.t.* $T \subseteq N$.

$I_{\text{and}}(S)$ represents the numerical effect of an AND interaction encoded by the DNN v. For example, in Figure 1 (a), given the input prompt of "*Einstein's Theory of General Relativity proposes that gravity is the warping of*", let the DNN v predict the next token "*space*" within the word "*space-time.*" The interaction $I_{\text{or}}(S = \{Einstein's, of\})$ is encoded by the DNN. This is because the token "*Einstein*" is statistically related to the predicted token "*space*," and the preposition "*of*" is typically followed by a noun, in this case, "*space.*" Furthermore, more complex interaction like $I_{\text{and}}(S = \{Einstein's, the, warping, of\})$ is also used by the DNN to generate "*space.*" When all words in $S = \{$*Einstein's, the, warping,of*$\}$ appear in the prompt, the function $\mathbb{1}\left(\substack{\boldsymbol{x}_T \text{ triggers} \\ \text{AND relation } S}\right)$ is triggered and returns 1. If any word in S is masked, the function $\mathbb{1}\left(\substack{\boldsymbol{x}_T \text{ triggers} \\ \text{AND relation } S}\right)$ is salient and returns 0.

The following three properties guarantee that the interactions can be faithfully considered as primitive inference patterns encoded by the DNN. (1) **Sparsity property**: Li and Zhang (2023) discovered and Ren et al. (2023d) proved that the DNNs only encode a small number of salient interactions for inference under three common conditions. (2) **Universal-matching property**: Chen et al. (2024) proved that we can use interactions $S \subseteq N$ to accurately mimic the network outputs on all randomly masked samples $\boldsymbol{x}_T$. (3) **Transferability property**: Li and Zhang (2023) observed that many salient interactions extracted from one input sample can also be extracted from other input samples in the same category, and many salient concepts encoded by a DNN are usually also encoded by other DNNs trained for the same task. Chen et al. (2024) further proposed a method to extract generalizable interactions shared by different DNNs.

Unedited

2.3 Explaining the representation capacity of DNNs using interactions

In this paper, how to define the first-principal explanation of a DNN is the core issue. To this end, we believe that if the theory system of interactions can well explain various deep-learning phenomena, then the theory system of interactions is more likely to be considered as the first-principal explanation of the DNN.

Therefore, in this subsection, we briefly review how interactions can be used to explain the internal mechanisms that determine the adversarial robustness, the generalization power, and the representation bottleneck of the DNN.

(1) The complexity of interactions directly determines the adversarial robustness of the DNNs. Considering that Ren et al. (2023a) demonstrated that the decision-making logic of the DNN can be explained as the sum of a small number of salient interactions, the adversarial sensitivity of the entire AI model can be represented as the sum of the adversarial sensitivity of these interactions. To this end, Ren et al. (2021) discovered, and Liu et al. (2023) and Ren et al. (2023c) further proved that the adversarial sensitivity of an interaction has an exponential relation with the order of the interaction. Here, the complexity (order) of an interaction S is defined as the number of input variables in S. Ren et al. (2021) discovered that adversarial training enhances the discrimination power of low-order interactions, thereby boosting the adversarial robustness of the model. Liu et al. (2023) mathematically proved and empirically demonstrated that DNNs tend to learn simple interactions more easily than complex interactions. Ren et al. (2023c) proved that, compared to a standard DNN, a Bayesian neural network (BNN) is more likely to avoid encoding complex interactions.

(2) The complexity of interactions directly determines the generalization power of the DNNs. Zhou et al. (2024) found that complex (high-order) interactions are less generalizable than simple (low-order) interactions. If an interaction is frequently triggered by different training samples, and this interaction is also triggered by testing samples at a similar frequency, then this interaction is considered to be well-generalized to testing samples; otherwise, it is not. Zhang et al. (2020) discovered that the dropout operation improves the generalization power of a DNN by decreasing the interaction strength. Cheng et al. (2024) explained how DNNs gradually learned and forgot inference patterns during forward propagation, which provided new insights into how newly merged interactions and forgotten old interactions were related to the generalization power of a DNN.

(3) The complexity of interactions directly determines the representation bottleneck of the DNNs. Deng et al. (2021) discovered and proved that the representation bottleneck of the DNN, *i.e.*, a DNN usually tends to encode interactions of both extremely high orders and extremely low orders, but a DNN is less likely to encode interactions of medium orders. The cognition gap between the DNNs and humans is considered as the representation bottleneck of the DNN.

2.4 Summarize common mechanisms for empirical deep learning methods

Another requirement for the first-principal explanation of a DNN is the capacity to explain the internal mechanisms of different deep-learning methods. Currently, most deep-learning methods are developed in an empirical manner without a solid theoretical foundation. To this end, we find that the theory system of interactions well explains the deep-learning methods in the following two directions.

First, unifying different attribution methods. Deng et al. (2024) used interactions to unify the mechanisms of different attribution methods, which are developed to estimate the importance/attribution of a DNN's input variables to the output score. They proposed the Taylor interaction and proved that 14 attribution methods can all be reformulated as a reallocation of interaction effects encoded by the DNN. For instance, the Shapley value (Shapley, 1953) can be considered to be computed by evenly assigning the interaction effect to each variable involved in the interaction.

Second, unifying different adversarial-transferability-boosting methods. Wang et al. (2021) further used interactions to explain a mathematical mechanism shared by different adversarial-transferability-boosting methods. These methods are developed to generate adversarial examples with enhanced adversarial transferability. Various methods are developed empirically without attempting to

Unedited

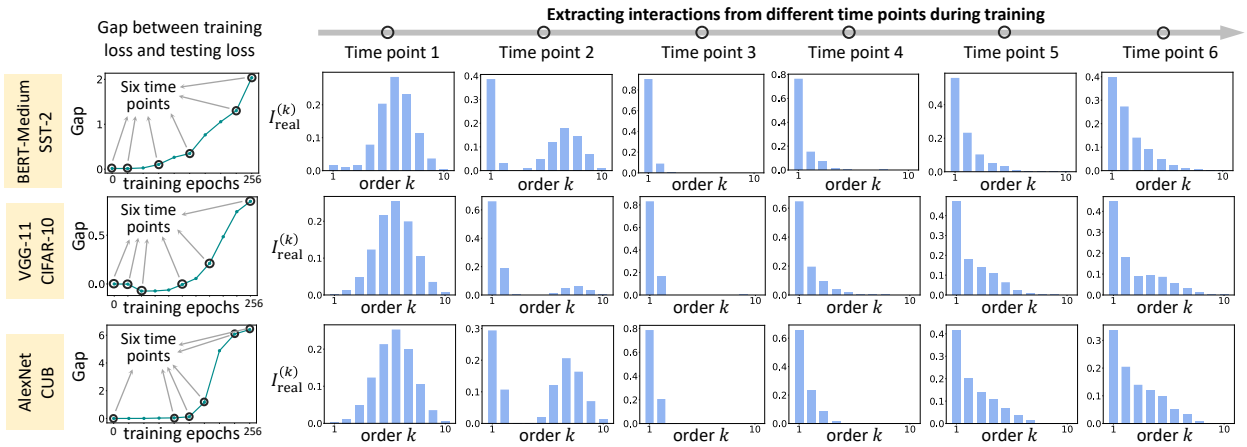

Fig. 2 The two-phase phenomenon of the change of interaction's complexity (Ren et al., 2024).

formulate the essence of the transferability. To this end, Wang et al. (2021) found that adversarial transferability and multi-order interactions are negatively correlated, and different adversarial-transferability-boosting methods all decrease interactions between adversarial perturbations.

3 Explaining the dynamics of the generalization power of a DNN

Although we have briefly revisited the achievements of the interaction theory in explaining DNNs in the previous section, we still believe that the core evidence for the first-principal explanation is the capacity to explain the extremely complex learning dynamics of a DNN. It is well known that predicting the learning dynamics of a model is much more challenging than explaining a static model. Therefore, in this section, let us introduce the latest progress in explaining a DNN's learning dynamics.

3.1 Discovering and proving two-phase dynamics of learning interactions

Zhang et al. (2024) discovered a two-phase phenomenon in the training process of DNNs as shown in Figure 2. **(1) Before the training process:** A DNN with initialized parameters tends to encode interactions of moderate orders, while encoding very few interactions of extremely high orders or extremely low orders. **(2) In the first phase:** Almost all initial interactions are quickly removed from the DNN. Only interactions of very low orders (*e.g.*, interactions of the first and second orders) are gradually learned. **(3) In the second phase:** The DNN begins to learn interactions of increasing orders. This can be explained as the hypothesis that the DNN begins to learn over-fitted features.

The above two-phase phenomenon has been widely observed on various DNNs trained for different tasks. Specifically, Zhang et al. (2024) and Ren et al. (2024) trained LeNet (LeCun et al., 1998), AlexNet (Krizhevsky et al., 2012), and VGG (Simonyan and Zisserman, 2014) on image datasets, including the MNIST dataset (LeCun et al., 1998), the CIFAR-10 dataset (Krizhevsky and Hinton, 2009), the CUB-200-2011 dataset (Wah et al., 2011), and the Tiny-ImageNet dataset (Le and Yang, 2015). They also trained the BERT model (Devlin et al., 2019) on the SST-2 dataset (Socher et al., 2013) for sentiment classification, and trained DGCNN (Wang et al., 2019) on the ShapeNet dataset (Yi et al., 2016) for 3D point cloud classification.

Figure 2 displays how the distribution of normalized interaction strength across different orders evolves during training. The normalized strength of m-order interactions is defined as $I^{(m)} = \frac{\mathbb{E}_{\boldsymbol{x}}[\sum_{S\in\Omega_{\text{salient}}:|S|=m}|I(S)|]}{\mathbb{E}_{1\le m'\le n}\mathbb{E}_{\boldsymbol{x}}[\sum_{S\in\Omega_{\text{salient}}:|S|=m'}|I(S)|]}$, where $\Omega_{\text{salient}} = \{S\subseteq N : |I(S)| > \tau\}$ denotes all salient interactions, the threshold is set to $\tau = 0.03\cdot\mathbb{E}_{\boldsymbol{x}}[|v(\boldsymbol{x}) - v(\emptyset)|]$, and $v(\boldsymbol{x}) = \log(p(y^{\text{true}}|\boldsymbol{x})/[1-p(y^{\text{true}}|\boldsymbol{x})])$ denotes the scalar confidence of the ground-truth output y^{true}.

Figure 2 illustrates the widespread existence of two-phase dynamics across various DNNs trained on diverse datasets. Before training (shown in the 1st column), the DNN mainly encodes interactions of medium orders. During the first phase (from the 2nd

Unedited

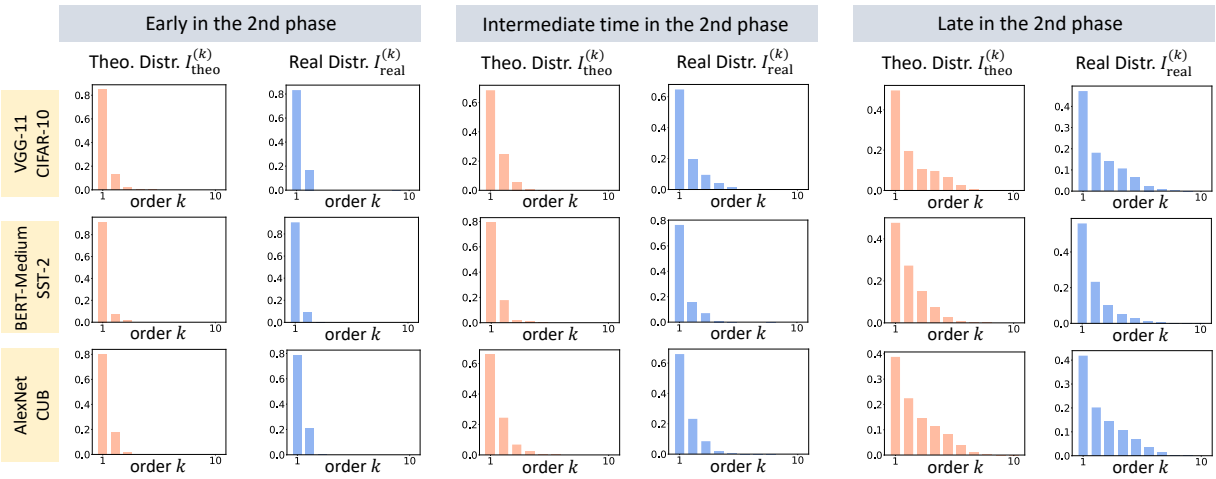

Fig. 3 Comparison between the theoretically predicted distribution of interactions across different orders and the real distribution in the second phase (Ren et al., 2024)

Unedited

to the 3rd column), the strength of medium-order and high-order interactions gradually diminishes to zero, while low-order interactions grow stronger. This progress can be considered as the removal of initial interactions for noise patterns. In the second phase (from the 3rd column to the 6th column in the figure), the DNN begins to learn interactions of increasing orders. This can be explained as the over-fitting process, which learns complex features.

Ren et al. (2024) followed the universal matching property of interactions (Ren et al., 2023a) to reformulate the inference score of the DNN $v(\boldsymbol{x}=\hat{\boldsymbol{x}}_S)$ as the weighted sum of interaction-triggering function.

$$\begin{aligned} &\forall S \subseteq N, v(\boldsymbol{x}=\hat{\boldsymbol{x}}_S) = g(\boldsymbol{x}=\hat{\boldsymbol{x}}_S) \\ &s.t. \quad g(\boldsymbol{x}) = \sum\nolimits_{T\subseteq N} w_T J_T(x), \end{aligned} \tag{1}$$

where the function $J_T(\boldsymbol{x}) \in \{0,1\}$ is a binary interaction-triggering function that approximates the triggering state of an interaction *w.r.t.* T on the given input sample $\boldsymbol{x}$. w_T is a scalar weight. Thus, the learning of a DNN can be considered as the learning of the scalar weight w_T for each interaction-triggering function $J_T(\boldsymbol{x})$. Ren et al. (2024) considered the learning problem as a regression to a set of potentially true interactions.

Specifically, given a training sample $\boldsymbol{x}$, the output score of the finally converged DNN on all 2^n randomly masked samples $\{\boldsymbol{x}_S : S \subseteq N\}$ can be written as $y_S = v(\boldsymbol{x}_\emptyset) + \sum_{\emptyset \neq T \subseteq S} w^*_T$, according to the universal matching property of interactions. $\{w^*_T : T \subseteq N\}$ can be taken as a set of true interactions that the DNN needs to learn. Ren et al. (2024) assumed that if the training of the DNN is subject to parameter noises, then the DNN's training process can be viewed as a process of gradually reducing the noise on network parameters. Then, the learning of the converged interactions on the training sample $\boldsymbol{x}$ can be represented as the following regression problem.

$$L(\boldsymbol{w}) = \mathbb{E}_{\epsilon}\mathbb{E}_{S\subseteq N}[(y_S - \boldsymbol{w}^\top(\boldsymbol{J}(\boldsymbol{x}_S) + \boldsymbol{\epsilon})^2] \tag{2}$$

where $\boldsymbol{w} = \text{vec}(\{w_T : T \subseteq N\}) \in \mathbb{R}^{2^n}$ denotes the weight vector of 2^n different interactions. $\boldsymbol{J}(\boldsymbol{x}_S) = \text{vec}(\{J_T(\boldsymbol{x}_S) : T \subseteq N\}) \in \mathbb{R}^{2^n}$ denotes the binary triggering states of 2^n interactions on the masked sample $\boldsymbol{x}_S$. $\boldsymbol{\epsilon} = \text{vec}(\{\epsilon_T : T \subseteq N\}) \in \mathbb{R}^{2^n}$ denotes the vector of the noises.

The optimal solution to $\min_{\boldsymbol{w}} L(\boldsymbol{w})$ is $\hat{\boldsymbol{w}} = (\boldsymbol{J}^\top\boldsymbol{J} + 2^n \text{diag}(\boldsymbol{c}))^{-1}\boldsymbol{J}^\top\boldsymbol{w}^* = \hat{\boldsymbol{M}}\boldsymbol{w}^*$. Here, $\boldsymbol{J} = [\boldsymbol{J}(\boldsymbol{x}_{S_1}), \boldsymbol{J}(\boldsymbol{x}_{S_2}), ..., \boldsymbol{J}(\boldsymbol{x}_{S_{2^n}})]^\top \in \mathbb{R}^{2^n\times 2^n}$ represents the binary triggering states of 2^n interactions on 2^n masked samples, and $\boldsymbol{c} = \text{vec}(\{\text{Var}[\epsilon_T] : T \subseteq N\}) = \text{vec}(\{2^{|T|}\sigma^2 : T \subseteq N\}) \in \mathbb{R}^{2^n}$ represents the vector of variances of the triggering strength of 2^n interactions. The analytic solution to $\min_{\boldsymbol{w}} L(\boldsymbol{w})$ shows that under parameter noises, high-order interactions are significantly suppressed.

Explaining the dynamics in the second phase. We can consider the second phase as the change of the optimal solution to $\min_{\boldsymbol{w}} L(\boldsymbol{w})$ when parameter noises are gradually removed during the training process. In this way, if the parameter noises are weakened, then the suppression of high-order interactions

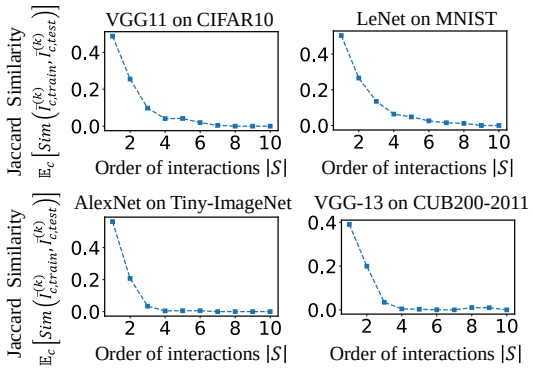

Fig. 4 The Jaccard similarity between interactions extracted from training samples and interactions extracted from testing samples (Zhang et al., 2024).

is also weakened. This explains the learning of interactions of increasing orders in the second phase.

Experimental verification. Furthermore, Ren et al. (2024) conducted experiments to examine the fitness between the theoretically predicted distribution of interactions over different orders and the real distribution on these DNNs. Ren et al. (2024) trained AlexNet and VGG on the MNIST dataset, the CIFAR-10 dataset, the CUB-200-2011 dataset, and the Tiny-ImageNet dataset. They also trained the BERT-Tiny model and the BERT-Medium model on the SST-2 dataset, and trained the DGCNN model on the ShapeNet dataset. Figure 3 shows that the theoretically predicted distribution can match well with the real distribution.

Explaining the dynamics in the first phase. Zhang et al. (2024) proved that the randomly initialized DNN usually encodes interactions of medium and high orders. In the first phase, the DNN removes initial interactions of medium orders and high orders, and mainly learns low-order interactions. Thus, the first phase can be viewed as the DNN gradually converging toward the optimal solution under large parameter noises.

3.2 Using the knowledge/concept in a DNN to explain its generalization power

When we discuss the first-principal explanation of a DNN, the analysis of the generalization power of a DNN is a key perspective. The first-principal explanation of a DNN is supposed to reflect the internal mechanism of why and how the generalization power of a DNN changes during the learning process.

To this end, we find that the two-phase dynamics of interaction complexity are also a key factor determining the dynamics of a DNN's generalization power. In previous studies, the over-fitting level is usually measured by the gap between the training loss and the testing loss. However, there has not been a mathematical connection between the loss-gap metric and the knowledge representation of a DNN. In other words, the "knowledge" encoded by the DNN cannot be used to explain its performance, even though in human cognition, the correctness of knowledge must serve as the first explanation of the performance.

Alignment between the two-phase training process and the loss gap. Zhang et al. (2024) and Zhou et al. (2024) both indicated that high-order interactions are the key reason for over-fitting of a DNN. For example, Zhang et al. (2024) discovered that the two-phase dynamics of the interaction complexity are temporally aligned with the dynamics of the loss gap (*i.e.*, the over-fitting level). Specifically, let the DNN's over-fitting level be measured by the gap between training and testing losses. Throughout the entire first phase, the DNN gradually removes the initial interactions of medium orders and high orders and mainly encodes the generalizable low-order interactions. Thus, the loss gap is relatively small. In comparison, in the second phase, the DNN learns interactions of increasing orders, and the loss gap begins to grow.

How to understand the over-fitting of a DNN from the interaction complexity. Since Ren et al. (2023a) proved that the output score of a DNN can usually be decomposed into the effects of interactions, the generalization power of the entire DNN can be attributed to the generalization power of all interactions that the DNN activates. To this end, although the generalization power of a DNN is usually defined by the testing accuracy, the generalization power of each individual interaction can be defined in a more straightforward manner. Specifically, let us consider a valid interaction that is frequently triggered across different training samples. If this interaction is also triggered by testing samples at a similar frequency, then this interaction is considered well generalized to testing samples. Otherwise, if this interaction is less frequently or not at all triggered on testing samples, it is considered non-generalizable.

In this way, Zhou et al. (2024) defined the gen-

Unedited

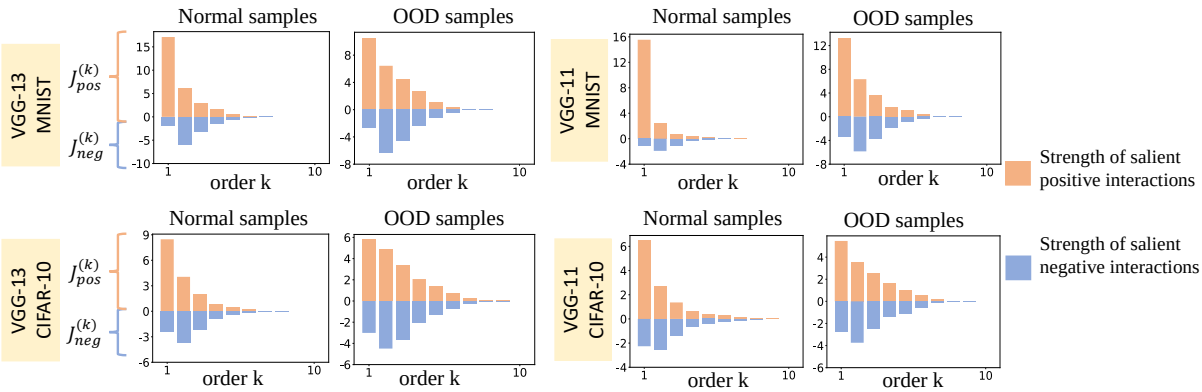

Fig. 5 The distribution of interactions extracted from original samples and the distribution of interactions extracted from incorrectly labeled samples (Zhang et al., 2024). OOD, out-of-distribution.

eralization power of m-order interactions as the Jaccard similarity between the distribution of m-order interactions extracted from training samples and the distribution of m-order interactions extracted from testing samples. Zhang et al. (2024) conducted experiments to compare the generalization power of interactions of different orders. They trained VGG-11 (Simonyan and Zisserman, 2014) on the CIFAR-10 dataset (Krizhevsky and Hinton, 2009), LeNet (LeCun et al., 1998) on the MNIST dataset (LeCun et al., 1998), AlexNet (Krizhevsky et al., 2012) on the Tiny-ImageNet dataset (Le and Yang, 2015), and VGG-13 (Simonyan and Zisserman, 2014) on the CUB200-2011 dataset (Wah et al., 2011). Figure 4 shows that the average Jaccard similarity between the interaction distribution on training samples and that on testing samples decreases as the order m of interactions increases, indicating that high-order interactions are the key reason for over-fitting of a DNN.

In another experiment, Zhang et al. (2024) compared the interaction complexity of interactions extracted from original samples with those extracted from incorrectly labeled samples. Figure 5 visualizes the distribution of interactions over different orders encoded by the DNN. Compared to the classification of original samples, these DNNs usually use interactions of higher orders for the classification of incorrectly labeled samples.

The two-phase dynamics of interaction complexity reflect the two-phase dynamics of a DNN's generalization power. In the first phase, the DNN gradually eliminates non-generalizable high-order interactions. By the end of this phase, it primarily encodes generalizable low-order interactions. In the second phase, the DNN begins to learn interactions of gradually increasing orders, which are typically non-generalizable, indicating that the DNN becomes increasingly over-fitted.

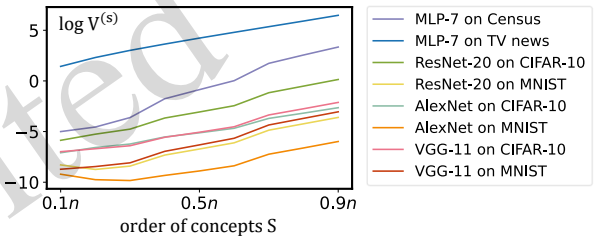

Fig. 6 The variance (adversarial sensitivity) of the interactions increases exponentially along with the order of interactions (Zhou et al., 2024).

High-order interactions are more adversarially sensitive. The adversarial sensitivity of a DNN can also be understood through interaction complexity. Specifically, Liu et al. (2023) mathematically proved that when a Gaussian perturbation $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \delta^2 \boldsymbol{I})$ is added to the input sample $\boldsymbol{x}$, the variance (instability) of the interaction effect $I(S|\boldsymbol{x}+\boldsymbol{\epsilon})$ increases exponentially with the order of interactions $|S|$. In experimental verification, let $V^{(s)} = \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{S:|S|=s}[\mathrm{Var}_{\boldsymbol{\epsilon}}(I(S|\boldsymbol{x}+\boldsymbol{\epsilon}))]]$ measure the average variance of s-order interactions *w.r.t.* the Gaussian perturbation $\boldsymbol{\epsilon}$. Figure 6 shows that the variance of the interactive effect $V^{(s)}$ increases exponentially with the order.

4 Conclusions

In this study, we have explored the potential of the interaction theory serving as a first-principle

Unedited

explanation for DNNs. Unlike empirical methods, the interaction theory offers a new axiomatic system that explains the decision-making logic of DNNs through symbolic interaction concepts. These symbolic interaction concepts simultaneously clarify the internal mathematical mechanisms underlying various deep-learning phenomena, including generalization power, adversarial robustness, representation bottleneck, and learning dynamics. Additionally, the interaction theory unifies diverse empirical methods, revealing shared mechanisms across 14 attribution methods and 12 transferability-boosting methods.

However, the interaction theory still has some limitations. First, the computation of interactions is highly complex. As the number of input variables increases, the computational burden grows exponentially. Kang et al. (2024) mathematically proved that although computing the Shapley value or the Banzhaf value of an input with n input variables involves 2^n coefficients, it becomes tractable when the Shapley value is sparse and of low degree. However, most real applications do not satisfy these assumed conditions, indicating that the computational bottleneck cannot be efficiently overcome.

Second, the explanatory power of interactions partially relies on the setting of the appropriate baseline value for masking input variables. The baseline value is supposed to effectively remove all signals of the input variable without introducing any out-of-distribution (OOD) features. From a theoretical perspective, this raises a deeper question, *i.e.*, how can we define reliable baseline values that effectively represent a "no-information" state? Ren et al. (2023b) proposed a method to learn baseline values, but this approach is fundamentally an engineering approximation, rather than a theoretically guaranteed optimal solution.

Third, the interaction theory has not been used to explain some classical phenomena in deep learning, such as catastrophic forgetting in continual learning.

Fourth, although Ren et al. (2023d) theoretically proved that based on the universal matching property and the sparsity property, a small-size AND-OR logical model can approximate the outputs of a DNN under 2^n masked states of the input sample. However, this does not mean that the DNN itself physically encodes such AND-OR logic in specific neurons. The logical model merely provides an equivalent explanation for the neural network's inference score on masked states. This equivalence does not imply that the learning/optimization of a DNN is conducted toward the exact AND-OR logic, but the DNN encodes a much more complex function.

Fifth, while the interaction theory can explain the learning dynamics of a DNN, how to use it to improve a DNN's performance remains an open problem. Directly penalizing high-order interactions is the most straightforward approach, but the optimization of high-order interactions would suffer from a serious gradient-oscillation problem in theory.

Contributors

Huilin ZHOU, Qihan REN, Junpeng ZHANG and Quanshi ZHANG designed the research. Huilin ZHOU and Junpeng ZHANG processed the data and conducted the experiments. Qihan REN proved the related theory. Huilin ZHOU drafted the paper. Quanshi ZHANG helped organize the paper. Huilin ZHOU and Quanshi ZHANG revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

References

Adebayo J, Gilmer J, Muelly M, et al., 2018. Sanity checks for saliency maps. Proc 32nd Int Conf on Neural Information Processing Systems, p.9525-9536. https://doi.org/10.5555/3327546.3327621

Bau D, Zhou BL, Khosla A, et al., 2017. Network dissection: Quantifying interpretability of deep visual representations. Proc IEEE conference on computer vision and pattern recognition, p.6541-6549. https://doi.org/10.1109/CVPR.2017.354

Chen L, Lou SY, Huang BH, et al., 2024. Defining and extracting generalizable interaction primitives from dnns. The 12nd Int Conference on Learning Representations.

Cheng X, Cheng L, Peng ZR, et al., 2024. Layerwise change of knowledge in neural networks. Proc 41st Int Conf on Machine Learning. https://doi.org/10.5555/3692070.3692386

Deng HQ, Ren QH, Zhang H, et al., 2021. Discovering and explaining the representation bottleneck of dnns. The 9th Int Conf on Learning Representations.

Deng HQ, Zou N, Du MN, et al., 2024. Unifying fourteen post-hoc attribution methods with taylor interactions. *IEEE Transactions on Pattern Anal Mach Intell*, 46(7):4625-4640. https://doi.org/10.1109/TPAMI.2024.3358410

Devlin J, Chang MW, Lee K, et al., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. Proc Conf North American Chapter of the Association for Computational Linguistics: Human

Unedited

Language Technologies, Volume 1 (Long and Short Papers), p.4171-4186.
https://doi.org/10.18653/v1/N19-1423

Dua D, Graff C, 2017. UCI machine learning repository, http://archive.ics.uci.edu/ml

Ghassemi M, Oakden-Rayner L, Beam AL, 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745-e750.
https://doi.org/10.1016/S2589-7500(21)00208-9

Kang J, Erginbas YE, Butler L, et al., 2024. Learning to understand: Identifying interactions via the möbius transform. *Proc* 38th *Int Conf on Neural Information Processing Systems*, 37:46160-46202.

Kim B, Wattenberg M, Gilmer J, et al., 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). Proc 35th Int Conf on Machine Learning, p.2668-2677.

Krizhevsky A, Hinton G, 2009. Learning multiple layers of features from tiny images, .

Krizhevsky A, Sutskever I, Hinton GE, 2012. Imagenet classification with deep convolutional neural networks. *Proc* 26th *Int Conf on Neural Information Processing Systems*, 25:1097-1105.

Le Y, Yang X, 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3.

LeCun Y, Bottou L, Bengio Y, et al., 1998. Gradient-based learning applied to document recognition. *Proc IEEE*, 86(11):2278-2324.
https://doi.org/10.1109/5.726791

Li MJ, Zhang QS, 2023. Does a neural network really encode symbolic concepts? Proc 40th Int Conf on Machine Learning, p.20452-20469.

Liu DR, Deng HQ, Cheng X, et al., 2023. Towards the difficulty for a deep neural network to learn concepts of different complexities. *Advances in Neural Information Processing Systems*, 36:41283-41304.

Qi CR, Su H, Mo KC, et al., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.652-660.
https://doi.org/10.1109/CVPR.2017.16

Ren J, Zhang D, Wang YS, et al., 2021. Towards a unified game-theoretic view of adversarial perturbations and robustness. *Proc* 35th *Int Conf on Neural Information Processing Systems*, 34:3797-3810.

Ren J, Li MJ, Chen QR, et al., 2023a. Defining and quantifying the emergence of sparse concepts in dnns. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, p.20280-20289.
https://doi.org/10.1109/CVPR52729.2023.01942

Ren J, Zhou ZP, Chen QR, et al., 2023b. Can we faithfully represent absence states to compute shapley values on a dnn? The 11st Int Conf on Learning Representations.

Ren QH, Deng HQ, Chen YN, et al., 2023c. Bayesian neural networks avoid encoding complex and perturbation-sensitive concepts. Proc 40th Int Conf on Machine Learning, p.28889-28913.

Ren QH, Gao JY, Shen W, et al., 2023d. Where we have arrived in proving the emergence of sparse interaction primitives in dnns. The 12th Int Conf on Learning Representations.

Ren QH, Zhang JP, Xu Y, et al., 2024. Towards the dynamics of a dnn learning symbolic interactions. *Proc* 38th *Int Conf on Neural Information Processing Systems*, 37:50653-50688.

Rudin C, 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206-215.
https://doi.org/10.1038/s42256-019-0048-x

Selvaraju RR, Cogswell M, Das A, et al., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proc IEEE international conference on computer vision, p.618-626.
https://doi.org/10.1109/ICCV.2017.74

Shapley LS, 1953. A value for n-person games. *Contribution to the Theory of Games Princeton University Press, Princeton, USA*, 2:307-317.
https://doi.org/10.1515/9781400881970-018

Shen W, Cheng L, Yang YX, et al., 2023. Can the inference logic of large language models be disentangled into symbolic concepts?,
https://arxiv.org/abs/2304.01083

Simonyan K, Zisserman A, 2014. Very deep convolutional networks for large-scale image recognition,
https://arxiv.org/abs/1409.1556

Simonyan K, Vedaldi A, Zisserman A, 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps,
https://arxiv.org/abs/1312.6034

Socher R, Perelygin A, Wu J, et al., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. Proc 2013 Conf on empirical methods in natural language processing, p.1631-1642.

Wah C, Branson S, Welinder P, et al., 2011. The caltech-ucsd birds-200-2011 dataset. *Technical Report No CNS-TR-2011-001, California Institute of Technology, Pasadena, USA*, .

Wang X, Ren J, Lin SY, et al., 2021. A unified approach to interpreting and boosting adversarial transferability. Proc 9th Int Conf on Learning Representations.

Wang Y, Sun Y, Liu Z, et al., 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1-12.
https://doi.org/10.1145/3326362

Yi L, Kim VG, Ceylan D, et al., 2016. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1-12.
https://doi.org/10.1145/2980179.2980238

Yosinski J, Clune J, Nguyen A, et al., 2015. Understanding neural networks through deep visualization,
https://arxiv.org/abs/1506.06579

Zhang H, Li S, Ma YC, et al., 2020. Interpreting and boosting dropout from a game-theoretic view. Proc 8th Int Conf on Learning Representations.

Zhang JP, Li Q, Lin L, et al., 2024. Two-phase dynamics of interactions explains the starting point of a dnn learning over-fitted features,
https://arxiv.org/abs/2405.10262

Unedited

Zhou HL, Zhang H, Deng HQ, et al., 2024. Explaining generalization power of a dnn using interactive concepts. Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, p.17105-17113. https://doi.org/10.1609/aaai.v38i15.29655

Unedited