PAPER REVIEW PRESENTATION

Where We Have Arrived in Proving the Emergence of Sparse Interaction Primitives in DNNs

Qihan Ren, Jiayang Gao, Wen Shen, and Quanshi Zhang
ICLR, 2024

2026.4.30

SNU IDEA Lab. 정유립

01 Interaction-based works (prof. Quanshi Zhang)

[Foundation (Mathematical Methodology)]

1. AAAI 2021 Interpreting Multivariate Shapley Interactions in DNNs

[Interaction-based Understanding of DNNs]

2. CVPR 2023 Defining and Quantifying the Emergence of Sparse Concepts in DNNs
3. ICML 2023 Does a Neural Network Really Encode Symbolic Concepts?
4. ICLR 2024 Where We Have Arrived in Proving the Emergence of Sparse Interaction Primitives in DNNs
5. NeurIPS 2024 Towards the Dynamics of a DNN Learning Symbolic Interactions

[Application]

6. ICML 2023 Bayesian Neural Networks Avoid Encoding Complex and Perturbation-Sensitive Concepts
7. AAAI 2024 Explaining Generalization Power of a DNN Using Interactive Concepts

[Architectural Innovation]

8. ICML 2023 HarsanyiNet: Computing Accurate Shapley Values in a Single Forward Propagation

01 Interaction-based works (prof. Quanshi Zhang)

1. AAAI 2021 Interpreting Multivariate Shapley Interactions in DNNs
2. CVPR 2023 Defining and Quantifying the Emergence of Sparse Concepts in DNNs
3. ICML 2023 Does a Neural Network Really Encode Symbolic Concepts?
4. ICLR 2024 Where We Have Arrived in Proving the Emergence of Sparse Interaction Primitives in DNNs

1. AAAI 2021

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

3. ICLR 2024

$$v(x) \approx \sum_{S \in \Omega_{salient}} I(S) + v(x_\emptyset)$$

$|\Omega_{salient}| \ll 2^n$

Notation

Harsanyi Interaction $I(S) \overset{\text{def}}{=} \sum_{T \subseteq S} (-1)^{|S|-|T|} \cdot u(T)$

masked value function $u(T) \overset{\text{def}}{=} v(x_T) - v(x_\emptyset)$

Trained DNN v

Output of the DNN on the sample x $v(x) \in \mathbb{R}$

Input sample $x = [x_1, \dots, xn]^T$

Baseline value $b = [b_1, \dots, bn]^T = x_\emptyset$

Masked input sample $x_T = [z_1, \dots, z_n]^T$, Where $z_i = \begin{cases} x_i, & i \in T \\ b_i, & i \in N \setminus T \end{cases}$

Set $N = \{1, \dots, n\}$

Salient Subset $\Omega_{salient} \subseteq N$

02 INTERACTION AS A SYMBOLIC CONCEPTS

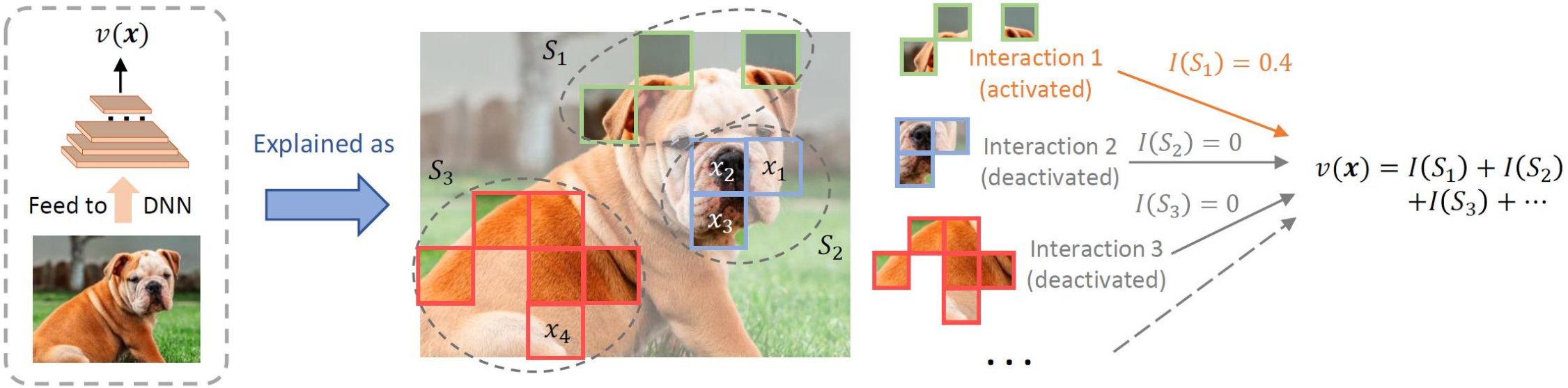


Figure 1 : Illustration of interactions encoded by a DNN. Each interaction S corresponds to an AND relationship among a specific set S of input variables (image patches). The patches x_1 and x_4 are masked, so that interactions S_2 and S_3 are deactivated

02 INTERACTION AS A SYMBOLIC CONCEPTS

Axiomatic properties

to define interactions as representations of faithful inference patterns (or concepts) encoded by a DNN

Ren et al. (ICML 2023)

(1) Sparsity property.

A DNN is supposed to encode few salient interactions on a specific sample.

(2) Universal matching property.

The network output on any arbitrarily masked sample is supposed to be well matched by the effects of specific interactions.

Theorem 1 (Proven in Ren et al. (2023a) and Appendix B.1). *Let the input sample $\boldsymbol{x}$ be arbitrarily masked to obtain a masked sample $\boldsymbol{x}_S$. The output of the DNN on masked sample $\boldsymbol{x}_S$ can be disentangled into the sum of all interaction effects within S: $\forall S \subseteq N$, $v(\boldsymbol{x}_S) = \sum_{T \subseteq S} I(T) + v(\boldsymbol{x}_\emptyset)$.*

※

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

(3) Sample-wise transferability property.

Salient interactions are supposed to be shared across different samples in the same category

03 PROVING THE SPARSITY OF INTERACTIONS

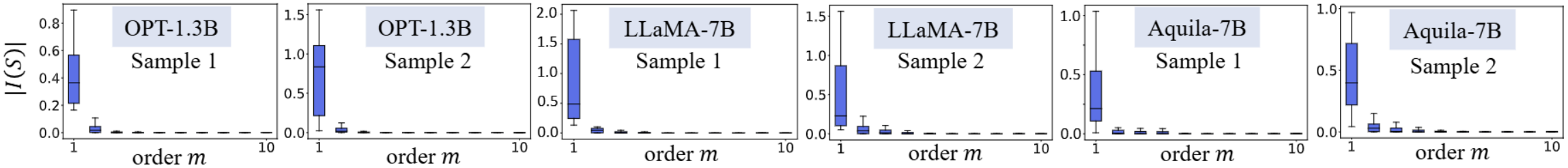
Assumption 1

The DNN has at most M-th order non-zero derivatives, where M < n, and n is the number of input variables to the DNN

$$\forall S \subseteq N, \quad |S| \geq M+1, I(S) = 0$$

Justification

(a) empirical evidence



(b) heuristic reasoning

(c) literature precedent

: 기존 game-theoretic interaction 연구들도 고차항 truncate 함.

03 PROVING THE SPARSITY OF INTERACTIONS

Assumption 2

The DNN can be used on occluded samples (e.g., an image with some patches being masked), and yields a higher classification confidence when the sample is less occluded.

$$\forall\, m' \leq m \Rightarrow \bar{u}^{(m')} \leq \bar{u}^{(m)}$$

$$\bar{u}(m) \overset{def}{=} \mathbb{E}_{|S|=m}[u(S)]$$
$$u(S) = v(x_S) - v(x_\emptyset)$$

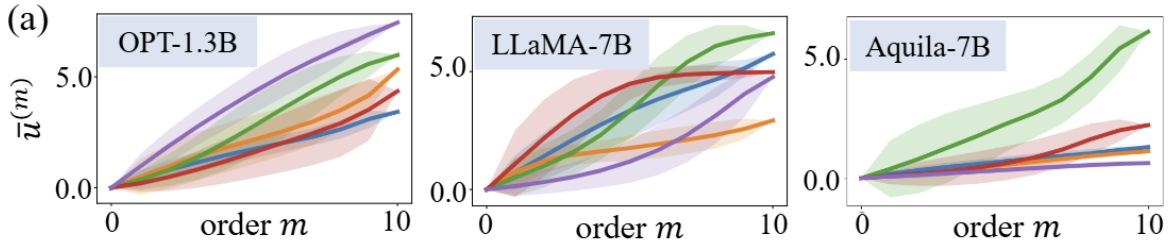
Assumption 3

The classification confidence of the DNN does not significantly degrade on occluded samples

$$\bar{u}^{(m')} \geq \left(\frac{m'}{m}\right)^p \bar{u}^{(m)} \quad \text{for } m' \leq m \text{ , } p > 0 \text{ (positive constatn)}$$

Justification

(a) empirical evidence



(b)

	OPT-1.3B	LLaMA-7B	Aquila-7B
Average p	1.18±0.22	1.18±0.42	1.11±0.39

(b) heuristic reasoning

03 PROVING THE SPARSITY OF INTERACTIONS

Theorem 2 (Proven in Appendix B.3). *There exists* $m_0 \in \{n, n-1, \cdots, n-M\}$, *such that for all* $1 \leq k \leq M$, *the sum of the effects of all k-order interactions can be written as*

$$A^{(k)} = (\lambda^{(k)} n^{p+\delta} + a^{(k)}_{\lfloor p \rfloor - 1} n^{\lfloor p \rfloor - 1} + \cdots + a^{(k)}_1 n + a^{(k)}_0)\, \bar{u}^{(1)}, \tag{3}$$

where $|\lambda^{(k)}| \leq 1$, $|a^{(k)}_0| < n$, $|a^{(k)}_i| \in \{0, 1, \cdots, n-1\}$ *for* $i = 1, \cdots, \lfloor p \rfloor - 1$, *and*

$$\delta \leq \log_n \left(\frac{1}{\lambda} \left(1 - \frac{a_{\lfloor p \rfloor - 1}}{n^{p - \lfloor p \rfloor + 1}} - \cdots - \frac{a_0}{n^p} \right) \right), \quad \text{if } \lambda > 0, \tag{4}$$

$$\delta \leq \log_n \left(\frac{1}{-\lambda} \left(\frac{a_{\lfloor p \rfloor - 1}}{n^{p - \lfloor p \rfloor + 1}} + \cdots + \frac{a_0}{n^p} \right) \right), \quad \text{if } \lambda < 0. \tag{5}$$

Here, $\lambda \overset{\text{def}}{=} \sum_{k=1}^{M} \frac{\binom{m_0}{k}}{\binom{n}{k}} \lambda^{(k)} \neq 0$, $a_i \overset{\text{def}}{=} \sum_{k=1}^{M} \frac{\binom{m_0}{k}}{\binom{n}{k}} a^{(k)}_i$ *for* $i = 0, 1, \cdots, \lfloor p \rfloor - 1$, *and* $\lfloor p \rfloor$ *denotes the greatest integer that is less than or equal to* p.

The above theorem indicates that the sum of effects of all k-order interactions is $O(n^{p+\delta})$.

03 PROVING THE SPARSITY OF INTERACTIONS

Theorem 3 (Proven in Appendix B.4). *$R^{(k)}$ has the following upper bound:*

$$R^{(k)} \leq \frac{\bar{u}^{(1)}}{\tau|\eta^{(k)}|}|\lambda^{(k)}n^{p+\delta} + a^{(k)}_{\lfloor p \rfloor -1}n^{\lfloor p \rfloor -1} + \cdots + a^{(k)}_0|. \tag{6}$$

$$\eta^{(k)} \overset{\text{def}}{=} \frac{\sum_{|S|=k} I(S)}{\sum_{|S|=k} |I(S)|} \qquad R^{(k)} \overset{\text{def}}{=} |\{S \subseteq N \mid |S| = k, |I(S)| \geq \tau\}|$$

The above theorem indicates that if positive interactions do not fully cancel with negative interactions (i.e., $|\eta(k)|$ is not extremely small), then the number of valid interactions R(k) of the k-th order has an upper bound of $O(np+\delta /|\tau\, \eta(k)|)$, which is much less than the total number of nk potential interactions of the k-th order in most cases

Table 2: Comparison between the number of valid interactions and the derived upper bound.

	OPT-1.3B	LLaMA-7B	Aquila-7B	MLP (tabular dataset)
Real # of valid interactions	28.73±52.37	50.53±40.37	30.13±26.20	54.42±36.81
Upper bound	197.84±188.87	293.20±287.28	184.23±124.71	229.11±139.52

감사합니다!