A Unified Interpretation of Training-Time Out-of-Distribution Detection: Supplementary material

Xu Cheng, Xin Jiang, Zechao Li*

School of Computer Science and Engineering, Nanjing University of Science and Technology

{xcheng8, xinjiang, zechao.li}@njust.edu.cn

A. Common conditions for proving the sparsity property of interactions

Ren et al. [5] have proven that, under three common conditions, a well-trained DNN typically encodes a small set of salient interactions, denoted as Ω_{salient}, for inference, where $|\Omega_{\text{salient}}| \ll 2^n$.
(1) The DNN is assumed not to encode interactions of very high orders, *i.e.*, high-order derivatives of the DNN output with respect to input variables are assumed to be zero.
(2) The classification confidence of the DNN on partially masked input samples is assumed to increase monotonically as the number of unmasked input variables increases, *i.e.,* $\forall i \in N\backslash S$ and $\forall S \subseteq N\backslash i, v(\boldsymbol{x}_{S\cup\{i\}}) > v(\boldsymbol{x}_S)$.
(3) The network output for masked input samples is assumed to be neither excessively high nor excessively low.

B. Proof of Theorem 2

Theorem 2. *The change of the inference score $\Delta v^{(m_1,m_2)}$ is proven to be represented as the sum of interaction effects of different orders.*

$$\Delta v^{(m_1,m_2)} = \sum\nolimits_{m=0}^{n} w^{(m)} \cdot \mathbb{E}_{S\subseteq N,|S|=m}[I(S|\boldsymbol{x})],$$
$$w^{(m)} = \begin{cases} C_{m_2 n}^{m} - C_{m_1 n}^{m}, & m \le m_1 n, \\ C_{m_2 n}^{m}, & m_1 n < m \le m_2 n, \\ 0, & m_2 n < m \le n, \end{cases} \tag{1}$$

Proof. The difference in inference scores between different randomly masked samples is represented as:

$$\begin{aligned} \Delta v^{(m_1,m_2)} &= \mathbb{E}_{\substack{T_1,T_2\subseteq N \& T_1\subsetneq T_2 \\ |T_2|=m_2 n, |T_1|=m_1 n}} [v(\boldsymbol{x}_{T_2}) - v(\boldsymbol{x}_{T_1})] \\ &= \mathbb{E}_{\substack{T_2\subseteq N, \\ |T_2|=m_2 n}} [v(\boldsymbol{x}_{T_2})] - \mathbb{E}_{\substack{T_1\subseteq N, \\ |T_1|=m_1 n}} [v(\boldsymbol{x}_{T_1})], \end{aligned} \tag{2}$$

where subsets T_1 and T_2 are randomly sampled from the universal set N, $0 \le m_1 \le m_2 < 1$.

Then, according to the Theorem 1, the first term in Eq.(2) can be re-written as:

$$\begin{aligned} \mathbb{E}[v(T_1)] &= \mathbb{E}_{T_1}[\sum\nolimits_{S\subseteq T_1} I(S|\boldsymbol{x})] \\ &= \mathbb{E}_{T_1}[\sum\nolimits_{m=0}^{m_1 n} \sum\nolimits_{S\subseteq T_1,|S|=m} I(S|\boldsymbol{x})] \\ &= \sum\nolimits_{m=0}^{m_1 n} \mathbb{E}_{T_1}[C_{m_1 n}^{m} \mathbb{E}_{S\subseteq T_1,|S|=m}[I(S|\boldsymbol{x})]] \\ &= \sum\nolimits_{m=0}^{m_1 n} C_{m_1 n}^{m} \mathbb{E}_{T_1}[\mathbb{E}_{S\subseteq T_1,|S|=m}[I(S|\boldsymbol{x})]], \end{aligned} \tag{3}$$

Similarly, we can obtain,

$$\mathbb{E}[v(T_2)] = \sum\nolimits_{m=0}^{m_2 n} C_{m_2 n}^{m} \mathbb{E}_{T_2}[\mathbb{E}_{S\subseteq T_2,|S|=m}[I(S|\boldsymbol{x})]], \tag{4}$$

*Corresponding author.

Note that $\mathbb{E}_{S\subseteq T_1,|S|=m}[I(S|\boldsymbol{x})]$ is averaged over subsets T_1 and $\mathbb{E}_{S\subseteq T_2,|S|=m}[I(S|\boldsymbol{x})]$ is averaged over subsets T_2. Then, we can obtain,

$$\mathbb{E}_{T_1}[\mathbb{E}_{S\subseteq T_1,|S|=m}[I(S|\boldsymbol{x})]] = \mathbb{E}_{T_2}[\mathbb{E}_{S\subseteq T_2,|S|=m}[I(S|\boldsymbol{x})]] = \mathbb{E}_{S\subseteq N,|S|=m}[I(S|\boldsymbol{x})], \tag{5}$$

Then, we substitute Eq. (3) and Eq. (4) into Eq. (1), and the output change $\Delta v(m_1, m_2)$ can be rewritten as follows:

$$\begin{aligned}
\Delta v(m_1,m_2) &= \mathbb{E}_{T_1,T_2:\emptyset\subseteq T_1\subset T_2\subseteq N}[v(T_2)-v(T_1)] \\
&= \mathbb{E}_{\substack{T_2\subseteq N,\\ |T_2|=m_2 n}}[v(\boldsymbol{x}_{T_2})] - \mathbb{E}_{\substack{T_1\subseteq N,\\ |T_1|=m_1 n}}[v(\boldsymbol{x}_{T_1})] \\
&= \sum\nolimits_{m=0}^{m_2 n} C^m_{m_2 n}\mathbb{E}_{T_2}[\mathbb{E}_{S\subseteq T_2,|S|=m}[I(S|\boldsymbol{x})]] - \sum\nolimits_{m=0}^{m_1 n} C^m_{m_1 n}\mathbb{E}_{T_1}[\mathbb{E}_{S\subseteq T_1,|S|=m}[I(S|\boldsymbol{x})]] \\
&= \sum\nolimits_{m=0}^{m_2 n} C^m_{m_2 n}\mathbb{E}_{S\subseteq N,|S|=m}[I(S|\boldsymbol{x})] - \sum\nolimits_{m=0}^{m_1 n} C^m_{m_1 n}\mathbb{E}_{S\subseteq N,|S|=m}[I(S|\boldsymbol{x})] \\
&= \sum\nolimits_{m=0}^{n} w^{(m)}\mathbb{E}_{S\subseteq N,|S|=m}[I(S|\boldsymbol{x})], \\
w^{(m)} &= \begin{cases} C^m_{m_2 n} - C^m_{m_1 n}, & m \le m_1 n, \\ C^m_{m_2 n}, & m_1 n < m \le m_2 n, \\ 0, & m_2 n < m \le n. \end{cases}
\end{aligned} \tag{6}$$

Then, Theorem 2 is proven. □

C. OOD Detection performance of baseline models and enhancement models

Table 1. OOD detection performance of baseline model and the model trained with different training-time OOD detection enhancement methods.

Dataset	Method	ResNet-18		ResNet-34		WideResNet-40-2	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIAFR-10	Baseline	62.03	88.48	50.09	89.12	56.99	89.02
	LogitNorm [9]	41.70	92.63	37.53	91.74	46.55	90.72
	CSI [7]	31.78	94.96	19.24	96.56	24.27	95.66
	T2FNorm [3]	37.54	93.15	27.87	94.57	23.76	95.54
	DAL [8]	12.31	96.94	12.50	96.70	12.93	96.05
CIAFR-100	Baseline	79.70	78.15	78.95	78.30	78.01	76.90
	LogitsNorm	75.61	81.86	66.82	82.30	79.10	79.61
	CSI	59.37	85.93	54.36	86.11	73.45	83.44
	T2FNorm	77.54	82.55	72.71	81.38	71.36	82.49
	DAL	54.48	87.82	55.14	87.15	50.54	87.87

Table 1 reports the OOD detection performance of baseline models and the models trained with training-time OOD detection enhancement methods, which shows enhancement models achieve superior OOD detection performance to the baseline model.

D. Experiment Details

D.1. Annotating Semantic Parts

We follow [2, 4, 6, 10] to annotate semantic parts, because given an input sample $\boldsymbol{x} \in \mathbb{R}^n$, the DNN theoretically encodes 2^n interactions. Thus, the computational costs to compute interactions are very high, when n is sufficiently large. To address this issue, [2, 4, 6, 10] annotated 12 semantic parts in each input sample, such that the annotated semantic parts were aligned over different samples. In this way, [2, 4, 6, 10] treated each semantic part of each input sample as a "single" input variable for the DNN. In experiments, given an image in the CIFAR-10 dataset or CIFAR-100 dataset, we first resized it to 32×32 before feeding it into the DNN. Then, we follow [2, 4, 6, 10] to divide the resized image into small patches of size 4×4, resulting in a total of 8×8 image patches. We randomly selected $n = 12$ patches from 6×6 image patches located in the center of the image to reduce computational costs, because [2, 4, 6, 10] considered the DNN mainly used foreground information to make inference.

Nevertheless, we conducted experiments to examine whether the distribution of interactions of different orders was stable, when we sampled different sets of input variables to compute interactions. If the stability of the distribution of the interaction order was successfully examined, it means that we could follow [2, 4, 6, 10] to compute interactions on a small set of input variables to reduce the time cost, instead of computing interactions on all input variables. Specifically, we extracted interactions from the ResNet-18 model, and we followed the settings in [2, 4] to divide each image in the CIFAR-10 dataset into 8×8 image patches. Then, we randomly sampled two different sets of $n = 12$ patches as two sets of input variables, denoted by N_1, and N_2. For the set N_1, we calculated the mean interaction strength $\mathbb{E}_{S\subseteq N_1,|S|=m}[|I(S|\boldsymbol{x})|]$ of each order m. We computed the Jaccard similarity between two distributions of interactions computed based on N_1 and N_2.

$$\textit{Jaccard Similarity} = \frac{\sum_m \min(\mathbb{E}_{S\subseteq N_1,|S|=m}[|I(S|\boldsymbol{x})|], \mathbb{E}_{S\subseteq N_2,|S|=m}[|I(S|\boldsymbol{x})|])}{\sum_m \max(\mathbb{E}_{S\subseteq N_1,|S|=m}[|I(S|\boldsymbol{x})|], \mathbb{E}_{S\subseteq N_2,|S|=m}[|I(S|\boldsymbol{x})|])} \tag{7}$$

Besides, we also whether the distribution of interactions of different orders was stable under different settings of image patch sizes. To this end, we divided each image in the CIFAR-10 dataset into 4×4 image patches and sampled $n = 12$ as input variables, denoted by N_3. Then, we calculated the mean interaction strength $\mathbb{E}_{S\subseteq N_3,|S|=m}[|I(S|\boldsymbol{x})|]$ of each order m. We computed the Jaccard similarity between two distributions of interactions computed based on N_1 and N_3.

$$\textit{Jaccard Similarity} = \frac{\sum_m \min(\mathbb{E}_{S\subseteq N_1,|S|=m}[|I(S|\boldsymbol{x})|], \mathbb{E}_{S\subseteq N_3,|S|=m}[|I(S|\boldsymbol{x})|])}{\sum_m \max(\mathbb{E}_{S\subseteq N_1,|S|=m}[|I(S|\boldsymbol{x})|], \mathbb{E}_{S\subseteq N_3,|S|=m}[|I(S|\boldsymbol{x})|])} \tag{8}$$

Fig. 1 and Fig. 2 show that under different settings of patch sizes and sampled sets N, the distribution of interactions over different orders was similar. Thus, the above experiments verified that we could follow [2, 4, 6] to simply sample a small set of input variables to reduce the computational cost of interactions, which did not affect the analysis of OOD detection.

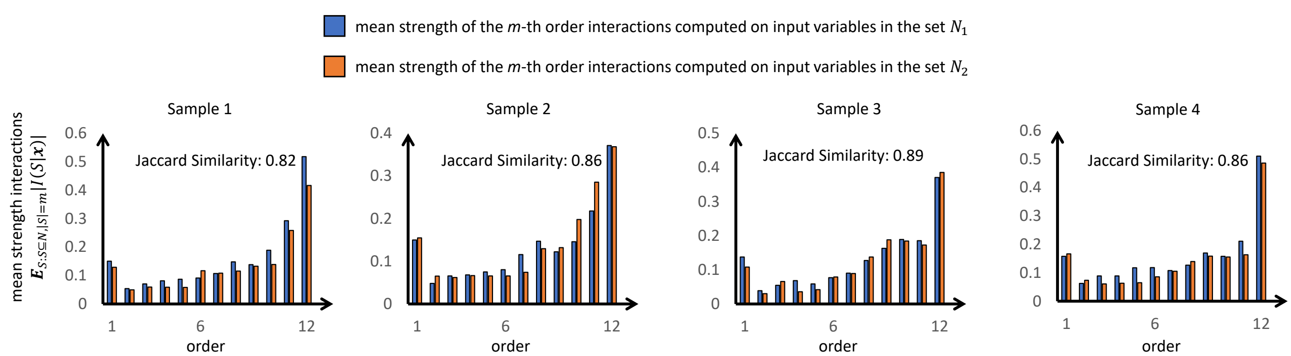

Figure 1. Comparison between the averaged interaction strength calculated over different N.

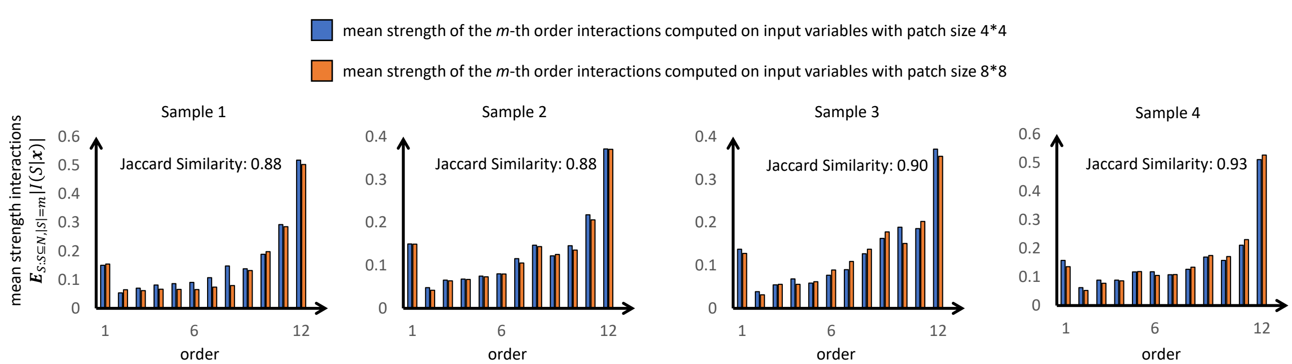

Figure 2. Comparison between the averaged interaction strength calculated with different patch sizes.

D.2. Training Details

We follow [1] to fine-tune ResNet-18, ResNet-34 and WideResNet-40-2 on the CIFAR-10 and CIFAR-100 datasets. For each DNN, we trained five versions, including a baseline model trained with the cross-entropy loss, and four enhancements trained with training-time OOD methods, respectively. We set the training hyper-parameters of the baseline models and enhancement models to be consistent to ensure fair comparisons. Each DNN was trained for 100 epochs using SGD with the momentum 0.9, weight decay $5e{-}4$, and learning rate 0.01.

References

[1] Xuefeng Du, Yiyou Sun, and Yixuan Li. When and how does in-distribution label help out-of-distribution detection? *arXiv preprint arXiv:2405.18635*, 2024. 4

[2] Mingjie Li and Quanshi Zhang. Does a neural network really encode symbolic concepts? In *International conference on machine learning*, pages 20452–20469, 2023. 2, 3

[3] Sudarshan Regmi, Bibek Panthi, Sakar Dotel, Prashnna K Gyawali, Danail Stoyanov, and Binod Bhattarai. T2fnorm: Extremely simple scaled train-time feature normalization for ood detection. *arXiv preprint arXiv:2305.17797*, 2023. 2

[4] Jie Ren, Mingjie Li, Qirui Chen, Huiqi Deng, and Quanshi Zhang. Defining and quantifying the emergence of sparse concepts in dnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20280–20289, 2023. 2, 3

[5] Qihan Ren, Jiayang Gao, Wen Shen, and Quanshi Zhang. Where we have arrived in proving the emergence of sparse interaction primitives in dnns. In *The Twelfth International Conference on Learning Representations*. 1

[6] Qihan Ren, Junpeng Zhang, Yang Xu, Yue Xin, Dongrui Liu, and Quanshi Zhang. Towards the dynamics of a DNN learning symbolic interactions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 3

[7] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020. 2

[8] Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment distributions for out-of-distribution detection. *Advances in neural information processing systems*, 36:73274–73286, 2023. 2

[9] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644, 2022. 2

[10] Huilin Zhou, Hao Zhang, Huiqi Deng, Dongrui Liu, Wen Shen, Shih-Han Chan, and Quanshi Zhang. Explaining generalization power of a dnn using interactive concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17105–17113, 2024. 2, 3