# Towards the Difficulty for a Deep Neural Network to Learn Concepts of Different Complexities

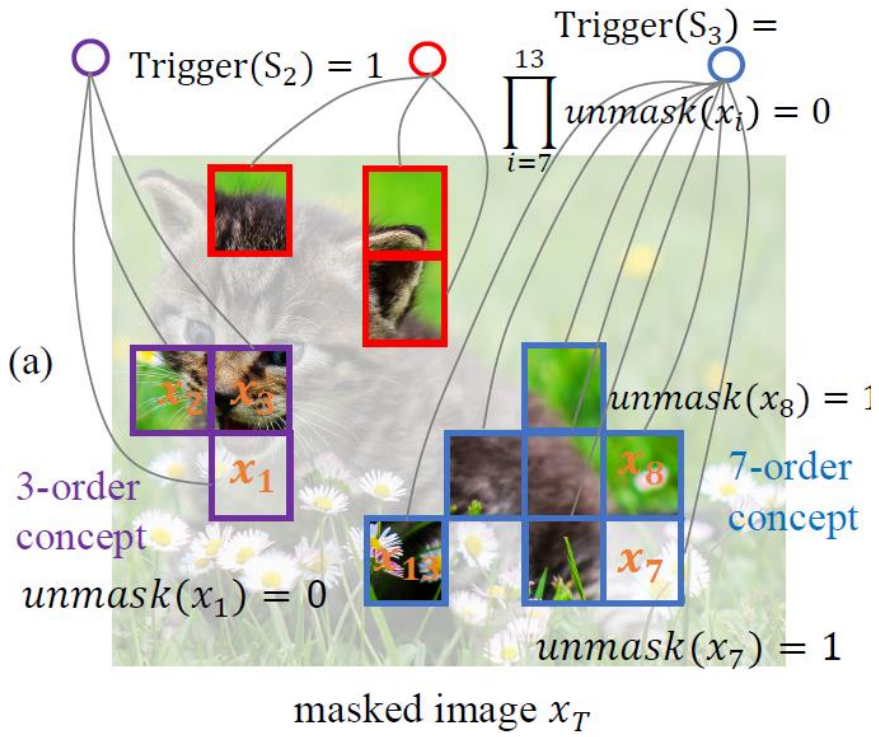
Dongrui Liu[1*], Huiqi Deng[1*], Xu Cheng[1], Qihan Ren[1], Kangrui Wang[2], Quanshi Zhang[1]

1. Shanghai Jiao Tong University  2. University of Chicago

# Preliminary: understanding interactions as AND relationships

[1-3] used the Harsanyi interaction $I(S)$ to study the emergence of concepts in neural networks

Each interaction represents an **AND relationship** between input variables in a set $S$




$I(S_2)$

Patches in $S_2$ are all present, so interaction $S_2$ is activated

Patches $x_1$ and $x_7$ are masked, so interactions $S_1$ and $S_3$ are deactivated

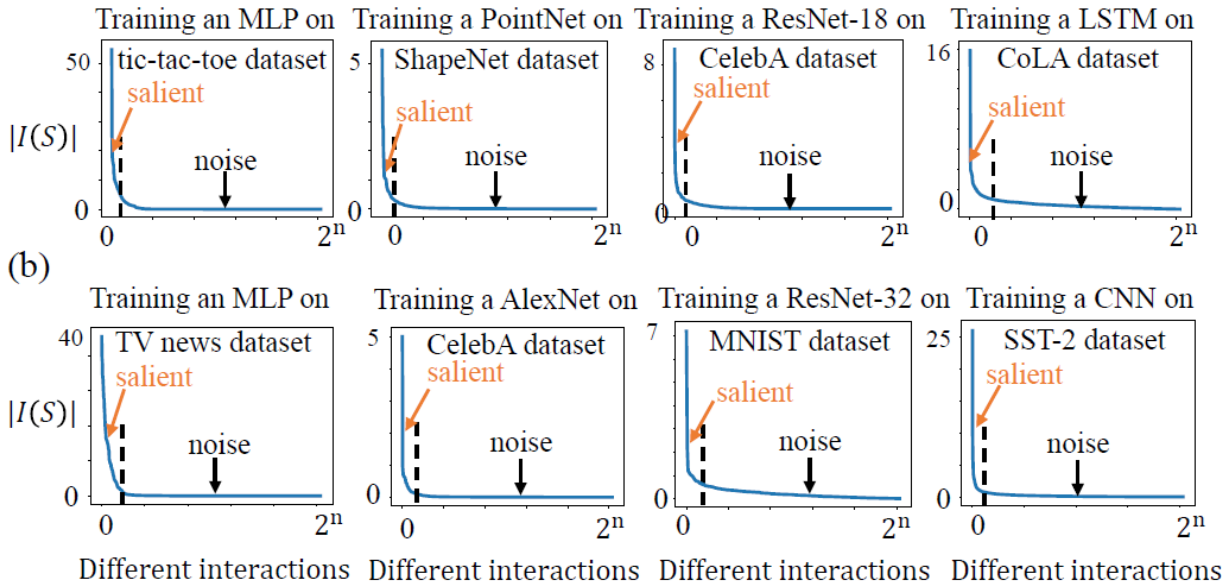

[1] Ren et al. Defining and Quantifying the Emergence of Sparse Concepts in DNNs. CVPR, 2023.
[2] Ren et al. Can we faithfully represent absence states to compute shapley values on a DNN?. ICLR, 2022.
[3] Li et al. Does a neural network really encode symbolic concept. ICML, 2023.

# Preliminary: neural networks usually only encode a small number of salient concepts

- **Faithfulness:** The output of a neural network on any arbitrarily masked sample $x_S$ can be disentangled into the sum of the effects of all interactions within set $S$

$$\forall S \subseteq N, \quad v(\boldsymbol{x}_S) = \sum_{T \subseteq S} I(T|\boldsymbol{x})$$

- **Sparsity:** **most** interactions have **near-zero effects** (noisy patterns), but only **a small number** of interactions have **significant effects** (salient interactive concepts).




[1] Ren et al. Defining and Quantifying the Emergence of Sparse Concepts in DNNs. CVPR, 2023.

# Overview

We theoretically explain the reason why it is easier for DNNs to learn simple concepts than complex concepts.

**Simple (low-order)** concepts
(composed of **a few** patches)

**Complex (high-order)** concepts
(composed of **massive** patches)

- We show that low-order interactive concepts in the data are much **more stable than** high-order concepts, which makes low-order interactive concepts more likely to be encoded.

- We provide new insights into several empirical findings *w.r.t.* the training of DNNs.

# Low-order interactive concepts in data are more stable

- Using input perturbations to roughly represent inevitable variations in data

  There are some inevitable variations in data. For example, image classification suffers from different variations, such as small shape deformation and small object rotations.

  In this paper, we analyze the impact of data variations on interactive concepts of different complexities. Such data variations are quite difficult to formulate, so we use a Gaussian perturbation as a rough representation.

# Low-order interactive concepts in data are more stable

---

The Harsanyi interaction effect of each concept can be rewritten.

**Theorem 2** (proven in the supplementary material). *Given a neural network $v$ and an arbitrary input sample $\boldsymbol{x}' \in \mathbb{R}^n$, the network output can be decomposed by using the Taylor expansion i.e., $v(\boldsymbol{x}') = \sum_{S \subseteq N} \sum_{\boldsymbol{\pi} \in Q_S} U_{S,\boldsymbol{\pi}} \cdot J(S, \boldsymbol{\pi} | \boldsymbol{x}')$. In this way, according to Equation (1), the Harsanyi interaction effect $I(S|\boldsymbol{x}')$ on the sample $\boldsymbol{x}'$ can be reformulated as follows.*

$$I(S|\boldsymbol{x}') = \sum\nolimits_{\boldsymbol{\pi} \in Q_S} U_{S,\boldsymbol{\pi}} \cdot J(S, \boldsymbol{\pi} | \boldsymbol{x}'). \tag{4}$$

*Here, $J(S, \boldsymbol{\pi}|\boldsymbol{x}') = \prod_{i \in S} \left( \text{sign}(x_i' - b_i) \cdot \frac{x_i' - b_i}{\tau} \right)^{\pi_i}$ denotes a Taylor expansion term of the degree $\boldsymbol{\pi}$, where the degree $\boldsymbol{\pi} \in Q_S = \{[\pi_1, \ldots, \pi_n] | \forall i \in S, \pi_i \in \mathbb{N}^+; \forall i \notin S, \pi_i = 0\}$ and $b_i$ is the baseline value to mask the input variable $x_i$. In addition, $U_{S,\boldsymbol{\pi}} = \frac{\tau^m}{\prod_{i=1}^n \pi_i!} \frac{\partial^m v(\boldsymbol{x}'_\emptyset)}{\partial x_1^{\pi_1} \cdots \partial x_n^{\pi_n}} \cdot \prod_{i \in S} [\text{sign}(x_i' - b_i)]^{\pi_i}$, where $\boldsymbol{x}'_\emptyset$ denotes the sample whose input variables are all masked. $m = \sum_{i=1}^n \pi_i$.*

The stability of interactive concepts decreases along with the order of interactive concepts.

**Theorem 3.** *Let us add a Gaussian perturbation $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ to the input sample $\boldsymbol{x}$. Let us first consider the case with the lowest degree $\hat{\boldsymbol{\pi}} = [\hat{\pi}_1, \ldots, \hat{\pi}_n] \in Q_S$, satisfying that $\forall i \in S, \hat{\pi}_i = 1; \forall i \notin S, \hat{\pi}_i = 0$. The mean and variance of $J(S, \hat{\boldsymbol{\pi}} | \boldsymbol{x} + \boldsymbol{\epsilon})$ over the Gaussian perturbation $\boldsymbol{\epsilon}$ are given as*

$$\mathbb{E}_{\boldsymbol{\epsilon}}[J(S, \hat{\boldsymbol{\pi}} | \boldsymbol{x} + \boldsymbol{\epsilon})] = 1, \quad \text{Var}_{\boldsymbol{\epsilon}}[J(S, \hat{\boldsymbol{\pi}} | \boldsymbol{x} + \boldsymbol{\epsilon})] = \left(1 + \left(\frac{\sigma}{\tau}\right)^2\right)^{|S|} - 1. \tag{5}$$

*Furthermore, for the more general case with an arbitrary degree $\boldsymbol{\pi} \in Q_S = \{[\pi_1, \cdots, \pi_n] | \forall i \in S, \pi_i \in \mathbb{N}^+; \forall i \notin S, \pi_i = 0\}$, the mean and variance of $J(S, \boldsymbol{\pi} | \boldsymbol{x} + \boldsymbol{\epsilon})$ are computed as*

$$\mathbb{E}_{\boldsymbol{\epsilon}}[J(S, \boldsymbol{\pi} | \boldsymbol{x} + \boldsymbol{\epsilon})] = \mathbb{E}_{\boldsymbol{\epsilon}}[\prod\nolimits_{i \in S} (1 + \frac{\epsilon_i}{\tau})^{\pi_i}], \ \text{Var}_{\boldsymbol{\epsilon}}[J(S, \boldsymbol{\pi} | \boldsymbol{x} + \boldsymbol{\epsilon})] = \text{Var}_{\boldsymbol{\epsilon}}[\prod\nolimits_{i \in S} (1 + \frac{\epsilon_i}{\tau})^{\pi_i}]. \tag{6}$$

# DNNs mainly learn low-order interactive concepts

The following equation enables us to understand a DNN for the classification task as a pseudo-linear function.

$$v(\boldsymbol{x}_T) = Y(\boldsymbol{x}_T) = \sum\nolimits_{S\in\Omega} U_S \cdot C_S(\boldsymbol{x}_T)$$

- If an interactive concept $S$ has a stable value (i.e., $C_S$ stably being present/absent) across all samples in the same category, then we consider this concept is discriminative and easy to learn.

- If the concept cannot be consistently present or consistently absent over samples in the same category, then this concept is hard to learn.

# DNNs mainly learn low-order interactive concepts

---

**Experiment 1: verifying the claim that high-order interactive concepts are more sensitive to data variations than low-order interactive concepts.** Therefore, we use the following two metrics to evaluate the discrimination power of each interactive concept *S*

Relative consistency of the interactive concept *S*

$$\beta(S) = \mathbb{E}_c\left[|\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}_c}[I(S|\boldsymbol{x})]|/\mathrm{Std}_{\boldsymbol{x}\in\boldsymbol{X}_c}[I(S|\boldsymbol{x})]\right]$$

Relative instability of the interactive concept *S*

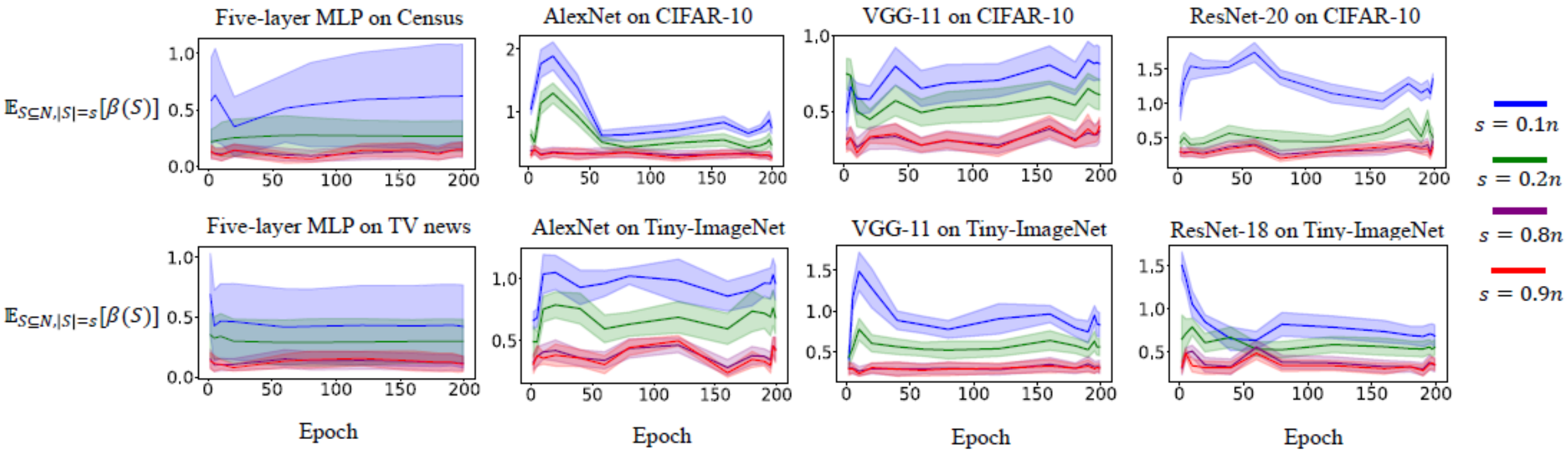
$$\kappa(S) = \mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}}[\mathbb{E}_{\boldsymbol{\epsilon}}[|I(S|\boldsymbol{x}+\boldsymbol{\epsilon}) - I(S|\boldsymbol{x})|]]/\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}}\left[|I(S|\boldsymbol{x})|\right]$$

# DNNs mainly learn low-order interactive concepts

---

**Experiment 1: verifying the claim that high-order interactive concepts are more sensitive to data variations than low-order interaction.**

Consistency of the interactive concept S $\beta(S) = \mathbb{E}_c\left[|\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}_c}[I(S|\boldsymbol{x})]|/\mathrm{Std}_{\boldsymbol{x}\in\boldsymbol{X}_c}[I(S|\boldsymbol{x})]\right]$

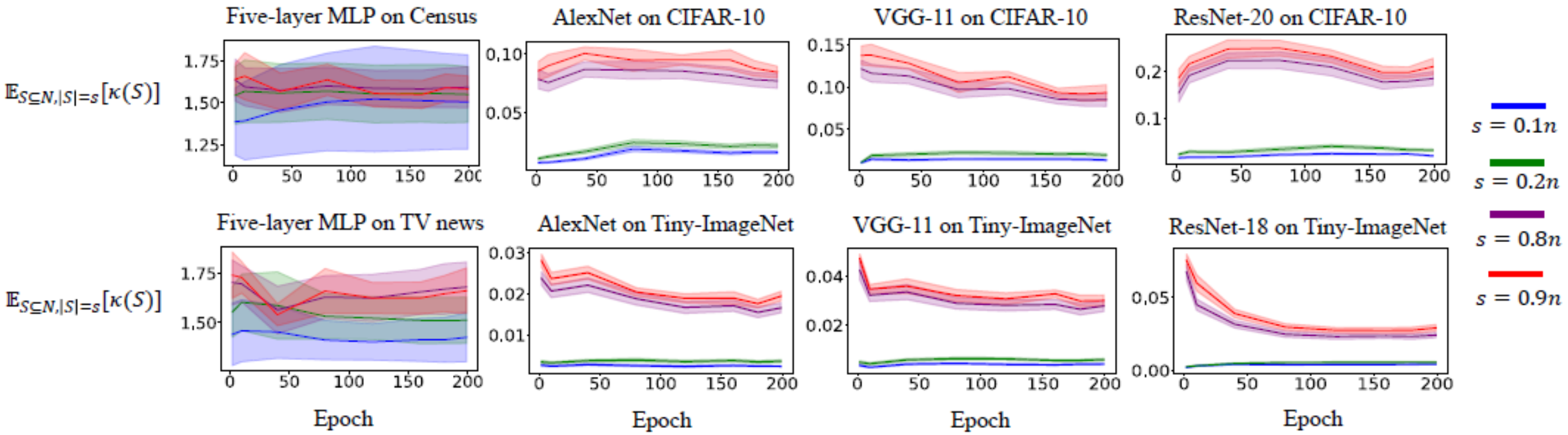


# DNNs mainly learn low-order interactive concepts

---

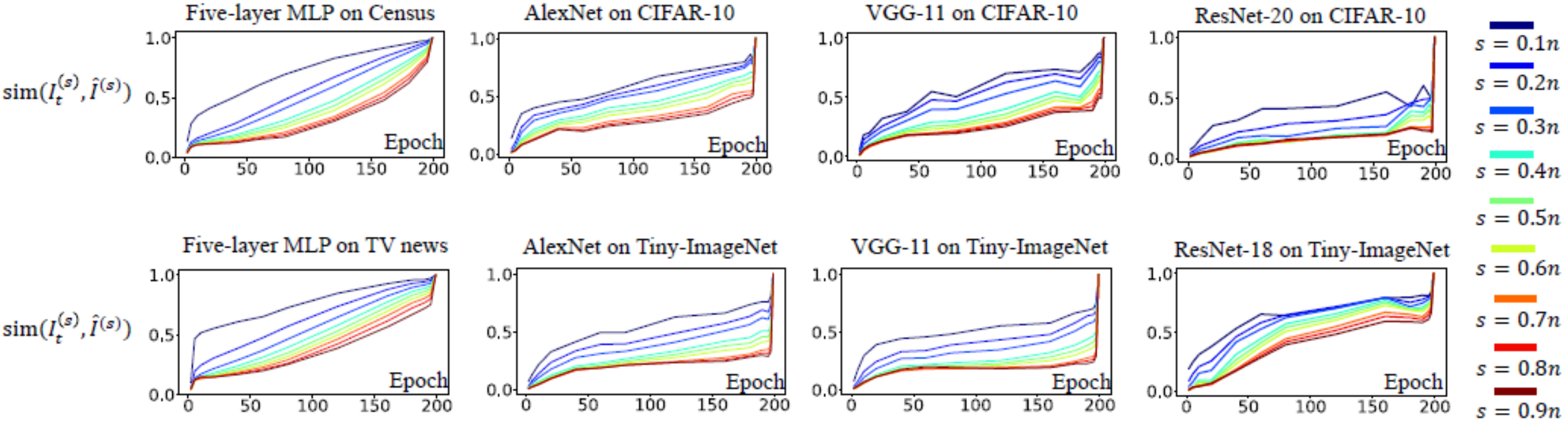
**Experiment 1: verifying the claim that high-order interactive concepts are more sensitive to data variations than low-order interaction.**

Instability of the interactive concept S $\quad \kappa(S) = \mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}}[\mathbb{E}_{\boldsymbol{\epsilon}}[|I(S|\boldsymbol{x}+\boldsymbol{\epsilon}) - I(S|\boldsymbol{x})|]]/\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}}\left[|I(S|\boldsymbol{x})|\right]$



# DNNs mainly learn low-order interactive concepts

**Experiment 2: verifying the phenomenon that low-order interactive concepts are usually learned faster than high-order concepts.** Specifically, we examine the similarity between interactive concepts in the network $v_t$ and the interactive concepts in the finally-learned DNN $\hat{v}$.

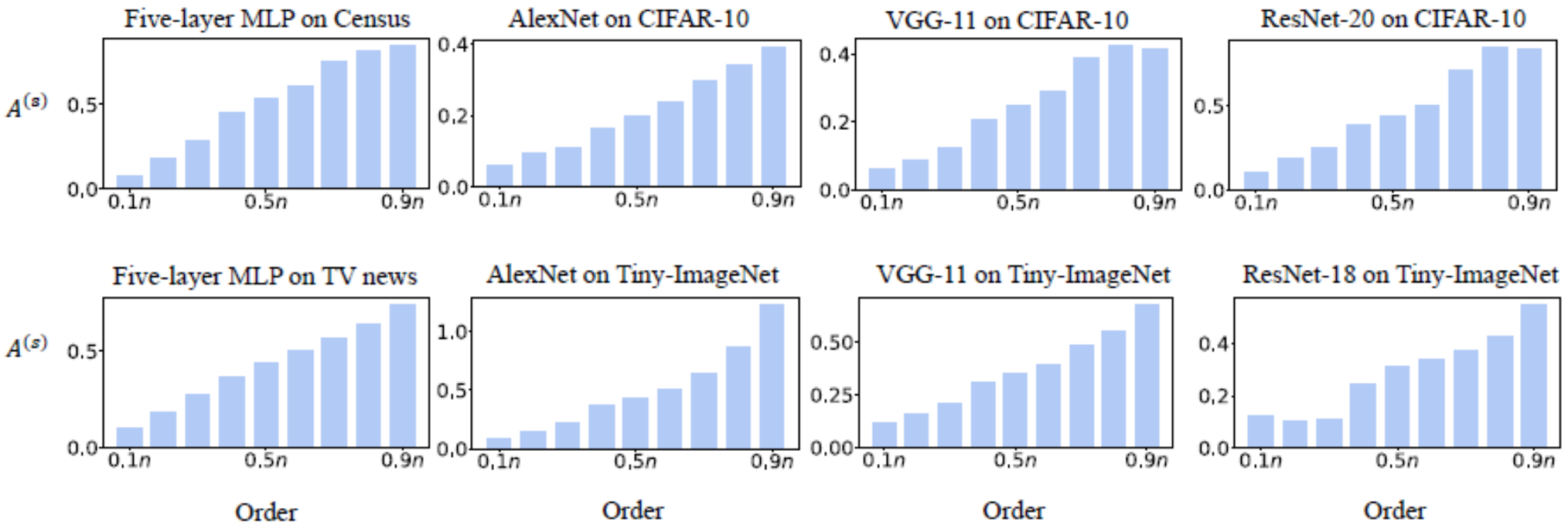


# Explaining generalization power and adversarial robustness

## Low-order interactive concepts are less sensitive to adversarial perturbations

- Sensitivity of the interactive concept S to adversarial perturbations $\delta$

$$\alpha(S) = \mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}}[|I(S|\boldsymbol{x}+\boldsymbol{\delta}) - I(S|\boldsymbol{x})|] \,/\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}}[|I(S|\boldsymbol{x})|] \qquad A^{(s)} = \mathbb{E}_{S\subseteq N,|S|=s}[\alpha(S)]$$



# Explaining existing findings about what are learned first

We discuss some related studies on which kind of knowledge is usually first learned by a DNN. We find that our theorems can partially explain mechanisms behind some previous findings[1-4].

[1] Devansh Arpit, Stanisław Jastrz˛ebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In International conference on machine learning, pp. 233–242. PMLR, 2017..
[2] Karttikeya Mangalam and Vinay Uday Prabhu. Do deep neural networks learn shallow learnable examples first? 2019..
[3] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. arXiv preprint arXiv:1901.06523, 2019..
[4] Chen Liu, Zhichao Huang, Mathieu Salzmann, Tong Zhang, and Sabine Süsstrunk. On the impact of hard adversarial instances on overfitting in adversarial training. arXiv preprint arXiv:2112.07324, 2021.
[5] 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.

# Conclusion

- We prove that low-order interactive concepts in the data are much **more stable than** high order concepts, which makes low-order interactive concepts more likely to be encoded.

- **We provide new insights into several empirical findings** *w.r.t.* the conceptual representation of DNNs.