NEURAL INFORMATION PROCESSING SYSTEMS

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

A Unified Game-Theoretic Interpretation of Adversarial Robustness

Jie Ren[1*], Die Zhang[1*], Yisen Wang[2*], Lu Chen[1], Zhanpeng Zhou[1], Xu Cheng[1], Xin Wang[1], Yiting Chen[1], Jie Shi[3], Quanshi Zhang[1]

1. Shanghai Jiao Tong University 2. Peking University 3. Huawei Technologies Inc.

Previous studies of explaining adversarial robustness

Previous explanations lack an essential and unified explanation.

What is the essence of adversarial attacks and defense?

- Explanations for **adversarial examples**
  - Linearity of feature representations
  - Non-robust yet discriminative features
- Understandings of **adversarial training**
  - Learning more shape-biased features
  - Enumeration of potential adversarial perturbations

How to explain adversarial robustness from the perspective of feature representation?

- Understanding of the **robustness**
  - Proving the theoretical bounds

Contributions of this paper

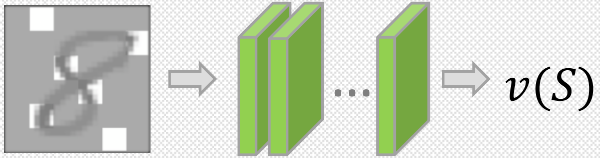
- We discover that adversarial attacks mainly affect high-order interactions between input variables.
- Adversarial **training** boosts the robustness of DNNs by learning more discriminative low-order interactions.
- We proposed a unified explanation for several adversarial defense methods.

Shapley values: the importance of input variables

Game

- Input variables $N = \{1,2,\dots,n\}$ -> players
- Scalar network output $v(N)$ -> total reward

Given input variables $S \subseteq N$,



- Shapley value is considered as a method that fairly allocates the reward to players[1,2].

$$\phi(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{(n - |S| - 1)!\,|S|!}{n!} [v(S \cup \{i\}) - v(S)]$$

[1] Lloyd S Shapley. “A value for n-person games”. In: Contributions to the Theory of Games 2.28 (1953), pp. 307–317.
[2] Scott M. Lundberg, and Su-In Lee, “A unified approach to interpreting model predictions” in NeurIPS 2017.

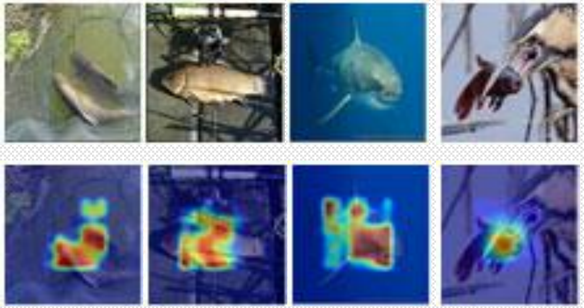
Game-theoretic interactions

- Different pixels cooperate with each other for inference, rather than work individually.
- Shapley Interaction index[3] between two input variables (i, j): the change of the importance (Shapley value) of i when j is present, w.r.t. the importance when j is absent.

$$I(i,j) = \phi_{w/\ j}(i) - \phi_{w/o\ j}(i) = \mathrm{E}_{S \subseteq N \backslash \{i,j\}}[\Delta v(i,j,S)]$$

Shapley value of i when j is present

Shapley value of i when j is absent



$$\text{where } \Delta v(i,j,S) = v(S \cup \{i,j\}) - v(S \cup \{i\}) - v(S \cup \{j\}) + v(S)$$

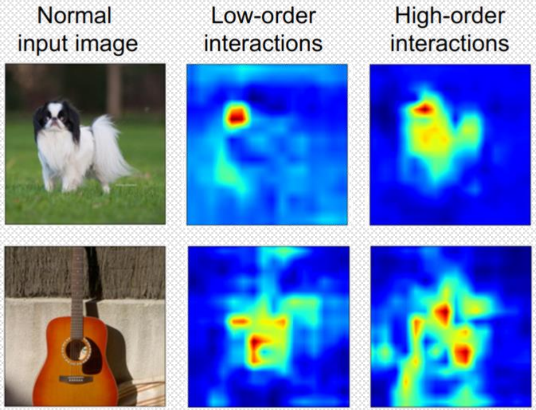
[3] Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. International Journal of game theory, 28(4):547–565, 1999.

Game-theoretic multi-order interactions to represent the complexity of representations

- Our team further define interactions of different orders as follows[4].

$$I_{ij}^{(m)} = \mathrm{E}_{S\subseteq N\setminus\{i,j\},|S|=m}[\Delta v(i,j,S)], \qquad I(i,j) = \frac{1}{n-1}\sum_{m=0}^{n-2} I_{ij}^{(m)}$$

$I_{ij}^{(m)}$ measures the average interaction between variables (i,j) under all contexts consisting of m variables.



Low order m: simple contextual collaborations with a few variable → represent simple concepts

High order m: complex contextual collaborations with massive variables → represent complex concepts

[4] Hao Zhang, Sen Li, Yinchao Ma, Mingjie Li, Yichen Xie, and Quanshi Zhang. Interpreting and boosting dropout from a game-theoretic view. In ICLR, 2021.

Game-theoretic multi-order interactions: properties

Properties of multi-order interactions

- **Linearity property:** If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $I_u^{(m)}(i,j) = I_{ij,v}^{(m)} + I_{ij,w}^{(m)}$
- **Dummy property:** If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, then $\forall j \in N \backslash \{i\}, I_{ij}^{(m)} = 0$
- **Symmetry property:** If $\forall S \subseteq N \backslash \{i,j\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\forall k \in N \backslash \{i,j\}, I_{ik}^{(m)} = I_{jk}^{(m)}$
- **Commutativity property:** $\forall i \neq j \in N, I_{ij}^{(m)} = I_{ji}^{(m)}$
- **Efficiency property:** $v(N) - v(\emptyset) = \sum_{i \in N}[v(\{i\}) - v(\emptyset)] + \sum_{i,j \in N, i \neq j}[\sum_{m=0}^{n-2} \frac{n-1-m}{n(n-1)} I_{ij}^{(m)}]$
- **Accumulation property:** $\phi(i|N) = E_m E_{j \in N \backslash \{i\}}\left[I_{ij}^{(m)}\right] + v(\{i\}) - v(\emptyset)$
- **Marginal contribution property**: $\forall i \neq j \in N, \phi^{(m+1)}(i) - \phi^{(m)}(i) = E_{j \in N \backslash \{i\}}\left[I_{ij}^{(m)}\right]$

Game-theoretic multi-order interactions: efficiency property

- **Efficiency property** of the multi-order interaction:

$$v(N) = v(\emptyset) + \sum_{i \in N} \phi^{(0)}(i|) + \sum_{i,j \in N, i \neq j} \sum_{m=0}^{n-2} J_{ij}^{(m)}, \qquad J_{ij}^{(m)} = \frac{n-1-m}{n(n-1)} I_{ij}^{(m)}$$

$$\phi^{(0)}(i) = v(\{i\}) - v(\emptyset)$$

Effects of a single variable

utility of multi-order interactions to the model output

Contributions of this paper

- We discover that adversarial attacks mainly affect high-order interactions between input variables.
- Adversarial **training** boosts the robustness of DNNs by learning more discriminative low-order interactions.
- We proposed a unified explanation for several adversarial defense methods.

Attacks mainly affect high-order interactions

Given the normal sample x, let $\tilde{x} = x + \delta$ denote the adversarial example.

Decompose the total adversarial utility of perturbations into attacking utilities on different interactions of different orders:

$$\Delta v(N|x) = v(N|x) - v(N|\tilde{x}) = \sum_{i \in N} \Delta\phi^{(0)}(i|N, x) + \sum_{i,j \in N, i \neq j} \sum_{m=0}^{n-2} \Delta J_{ij}^{(m)},$$

Small and can be ignored

$$\Delta J_{ij}^{(m)} = \frac{n-1-m}{n(n-1)} \Delta I_{ij}^{(m)}, \quad \Delta I_{ij}^{(m)} = I_{ij}^{(m)}(x) - I_{ij}^{(m)}(\tilde{x})$$

Attacks mainly affect high-order interactions

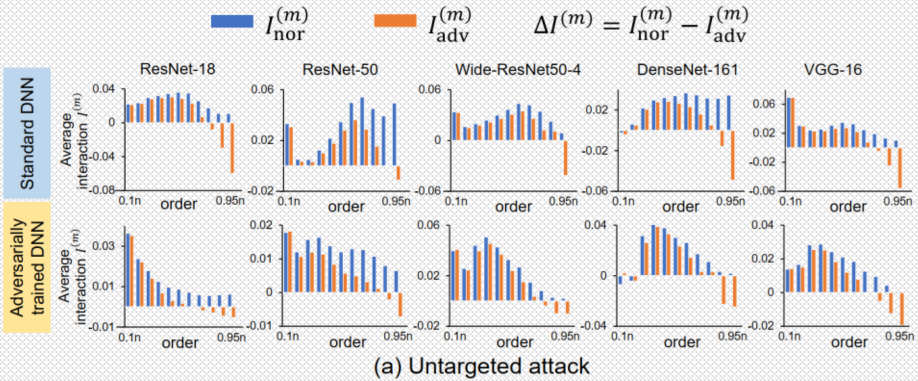


Figure: The multi-order interaction in normal samples and that in adversarial examples of standard DNNs and adversarially trained DNNs.

We discover that adversarial attacks mainly affect high-order interactions between input variables.

Attacks mainly affect high-order interactions

Theoretic explanation of the sensitivity of high-order interactions:

Proposition 1 (equivalence between the multi-order interaction and the mutual information):

$$I_{ij}^{(m)} = \mathbb{E}_{S\subseteq N\setminus\{i,j\},|S|=m} MI(X_i;X_j;Y|X_S)$$

high-order interactions → conditioned on larger contexts S → suffering more from adversarial perturbations.

Contributions of this paper

- We discover that adversarial attacks mainly affect high-order interactions between input variables.
- Adversarial **training** boosts the robustness of DNNs by learning more discriminative low-order interactions.
- We proposed a unified explanation for several adversarial defense methods.

AT boosts the robustness of high-order interactions

Attacking utility of m-order interactions: $\Delta J_{ij}^{(m)} = \frac{n-1-m}{n(n-1)} \Delta I_{ij}^{(m)}$

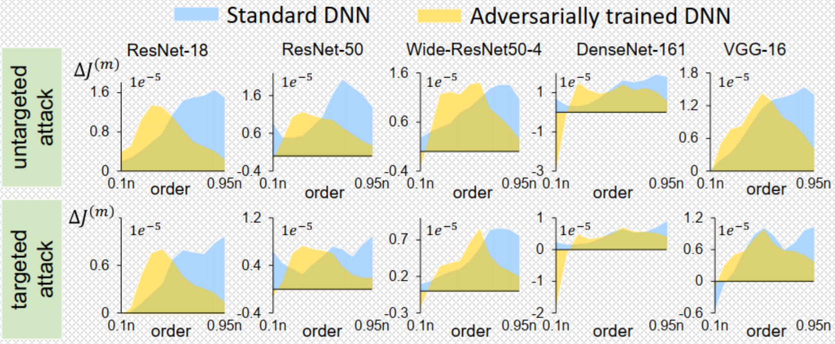


Figure: Distribution of compositional attacking utilities caused by interactions of different orders in standard DNNs and adversarially trained DNNs.

In adversarially learned DNNs, attacking utilities of high-order interactions significantly decreased.

AT learns more reliable low-order interactions to boost the robustness of high-order interactions

Disentanglement:

$$D^{(m)} = \mathbb{E}_{x\in\Omega}\mathbb{E}_{\substack{i,j\in N\\ i\neq j}} \frac{|I_{ij}^{(m)}(x)|}{\sum_{S\subseteq N\setminus\{i,j\},|S|=m}|\Delta v(i,j,S|x)|} = \mathbb{E}_{x\in\Omega}\mathbb{E}_{\substack{i,j\in N\\ i\neq j}} \frac{|\sum_{S\subseteq N\setminus\{i,j\},|S|=m}\Delta v(i,j,S|x)|}{\sum_{S\subseteq N\setminus\{i,j\},|S|=m}|\Delta v(i,j,S|x)|}$$

whether the m-order interactions represent discriminative information of a specific category.

In adversarially trained DNNs, low-order interactions exhibited higher disentanglement
-> more category-specific information
-> strengthen the robustness of high-order interactions.

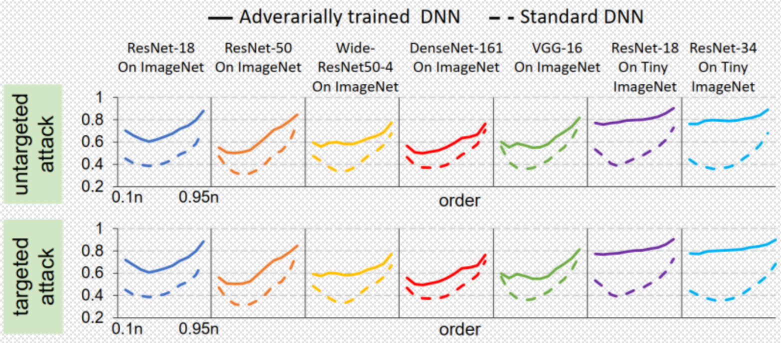


Figure: The interaction disentanglement.

Contributions of this paper

- We discover that adversarial attacks mainly affect high-order interactions between input variables.
- Adversarial **training** boosts the robustness of DNNs by learning more discriminative low-order interactions.
- We proposed a unified explanation for several adversarial defense methods.

The unified explanation for previous adversarial defenses

- Attribution-based method for detecting adversarial examples: ML-LOO[5]
- Rank-based method for detecting adversarial examples[6]

Detecting the **highest-order interaction** (the most sensitive component).

- Cutout method[7]
- High recoverability of adversarial examples in adversarially trained DNNs

Utilizing discriminative low-order interactions and **removing sensitive high-order interactions** to boost the robustness.

[5] Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. ML-LOO: detecting adversarial examples with feature attribution. CoRR, abs/1906.03499, 2019.
[6] Malhar Jere, Maghav Kumar, and Farinaz Koushanfar. A singular value perspective on model robustness. arXiv preprint arXiv:2012.03516, 2020.
[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

Contributions of this paper

- We discover that adversarial attacks mainly affect high-order interactions between input variables.
- Adversarial **training** boosts the robustness of DNNs by learning more discriminative low-order interactions.
- We proposed a unified explanation for several adversarial defense methods.

THANK YOU !