上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

XAI: Theoretically Unifying Conceptual Explanation and Generalization of DNNs

Game-theoretic interactions to unify
1. attribution explanations
2. encoding of visual concepts
3. generalization power
4. adversarial transferability and robustness

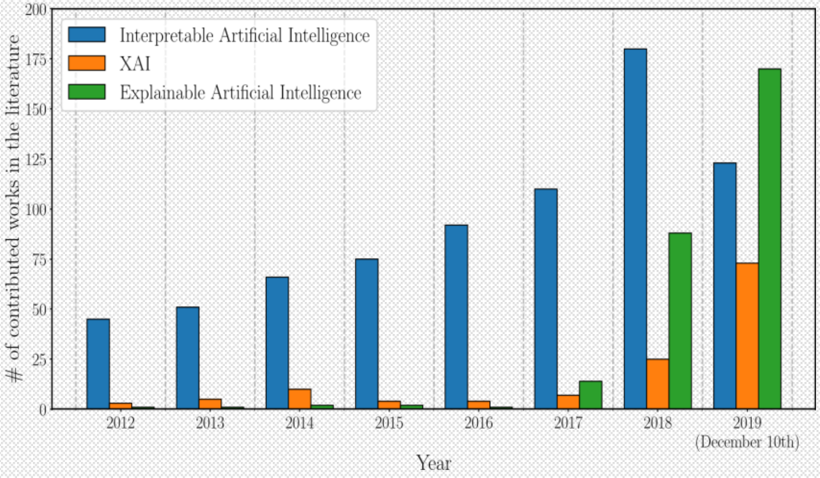
Quanshi Zhang

Associate Professor
John Hopcroft Center,
Shanghai Jiao Tong University, China

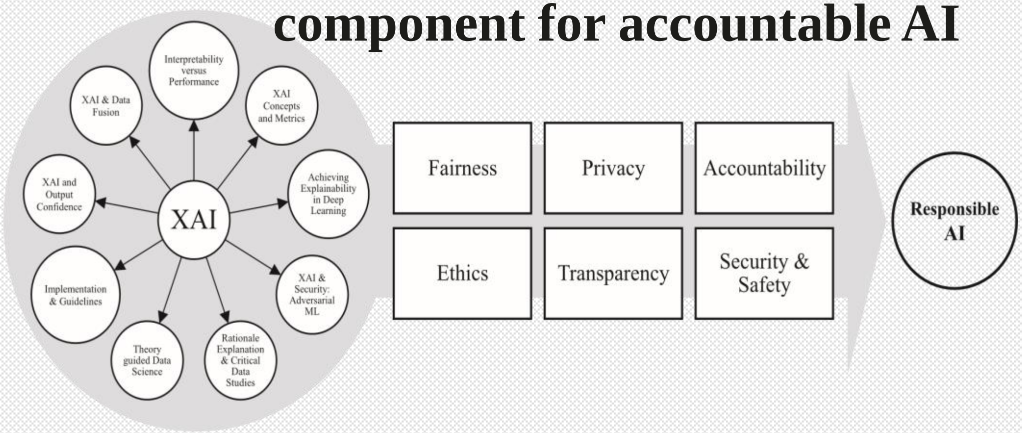
Why XAI is important ?

- **Key applications**
  - Finance, autonomous driving, medical diagnosis, military
- Set standards for the AI safety and interpretability

The growth of papers in XAI[1]



Interpretability is a necessary component for accountable AI



[1] Gonzalo Recio Dom`enech "Analysis of Explainability of Deep Learning Models for Medical Applicability" Minds Brains and Machines (MBM) — Master in Artificial Intelligence.

Topics of explaining DNNs

Semantic explanation

- Which semantic concepts are modeled and used for prediction
- How to quantify and improve the trustworthiness of a DNN
- End-to-end learn interpretable features
- Communicative learning at the semantic level
- How to evaluate the explanation

Mathematical explanation

- Model and explain the representation capacity of a DNN
- How to bridge the architecture with the knowledge representation
- Explain classical deep-learning techniques (e.g., distillation, adversarial learning, compression)
- How to debug DNNs using mathematical diagnosis of DNN features

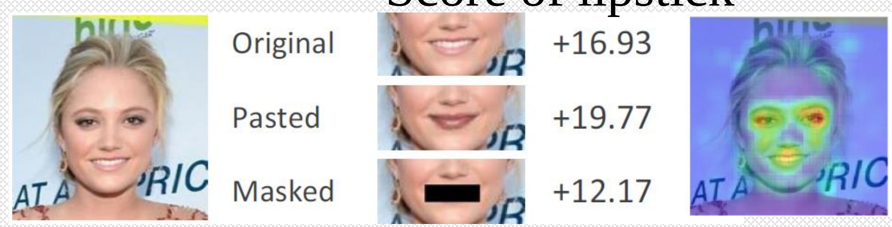
XAI topics

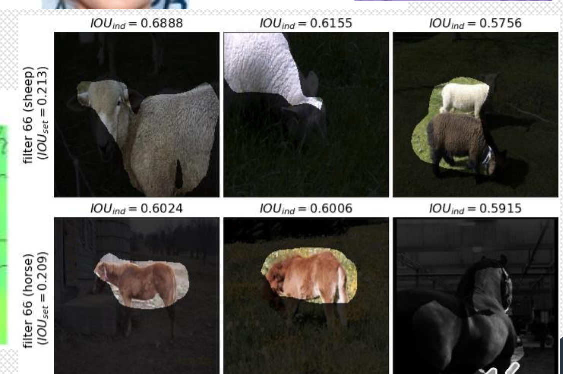
Semantic explanation

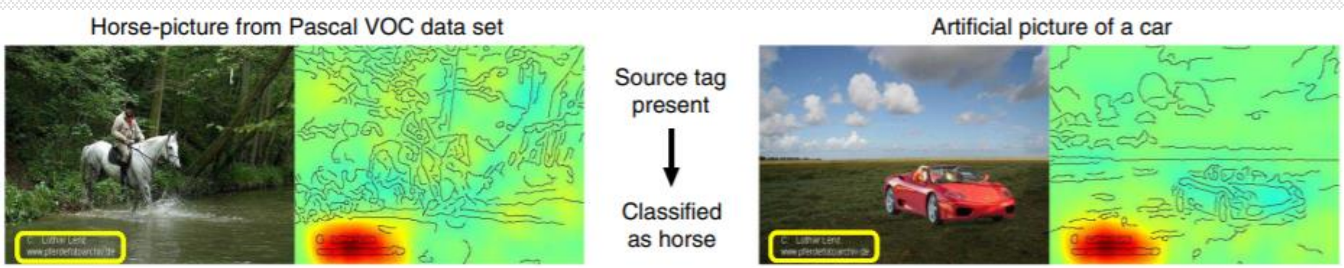
- Which semantic concepts are modeled and used for prediction
- How to quantify and improve the trustworthiness of a DNN
- End-to-end learn interpretable features
- Communicative learning at the semantic level
- How to evaluate the explanation

Make a surgery. Score=0.9
It is because
1) From Organ A. Score=0.2
2) From Organ B. Score=0.1
…

Score of lipstick







Lapuschkin et al. “unmasking clever hans predictors and assessing what machines really learn“ in Nat Commun 10 1096, 2019
Fong et al. "Net2Vec: Quantifying and Explaining how Concepts are encoded by filters in deep neural networks" in CVPR 2018
Zhang et al. “Examining CNN Representations with respect to Dataset Bias” in AAAI 2018

XAI topics

Mathematical explanation

- Model and explain the representation capacity of a DNN
- How to bridge the architecture with the knowledge representation
- Explain classical deep-learning techniques (e.g., distillation, adversarial learning, compression)
- How to debug DNNs using mathematical diagnosis of DNN features

- **How does an accident happen?**
- **What is the accident frequency if the car has run safely for a year?**
  - Once per year?
  - Once per ten years?
- **How to further boost the safety even without accident records?**

XAI topics

Mathematical explanation

- Model and explain the representation capacity of a DNN
- How to bridge the architecture with the knowledge representation
- Explain classical deep-learning techniques (e.g., distillation, adversarial learning, compression)
- How to debug DNNs using mathematical diagnosis of DNN features

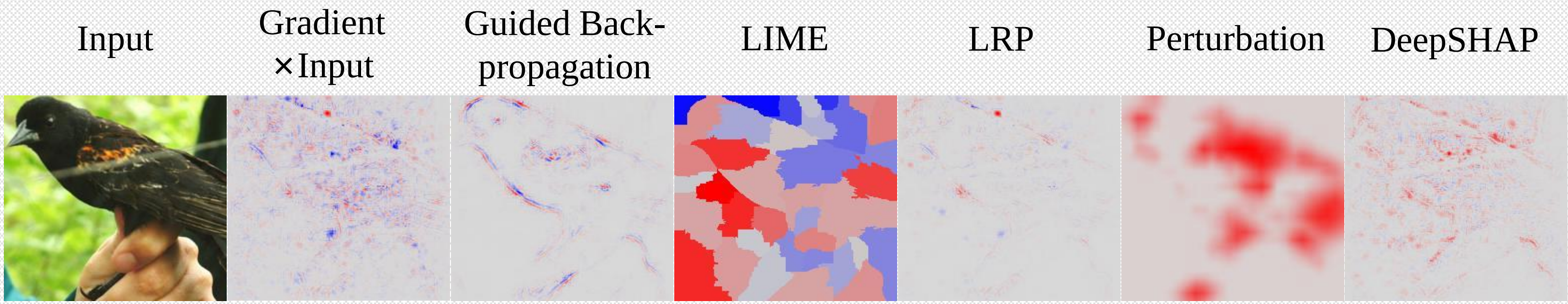
- **How to evaluate the generalization power of a DNN?**
- **Why does a specific DNN architecture outperform another architecture in a specific task?**
- **What is the relationship between the architecture and the knowledge.**
- **What is the common essence of existing DL methods? How to further improve these methods?**

Problems with semantic explanations

Many semantic explanations are still heuristic technologies, rather than science

- **Only self-consistency, no mutuality** between XAI methods
- **Very few theoretic foundations**
- **Difficult to improve DNNs**
- **Lack of convincing enough evaluation metrics**

Explanation results conflict with each other.



Problems with explaining the representation power

Analysis of the representation capacity of a DNN

- **Limited to certain assumptions (shallow nets or infinite width)**
- **Cannot provide semantic explanations**
- **Cannot explain the emergence of semantics in deep layers.**

"Mathematic proof" is not equivalent to "understanding."

Theorem 3 (Pitas et al. (2017)) *Let B an upper bound on the ℓ_2 norm of any point in the input domain. For any $B, \gamma, \delta > 0$, the following bound holds with probability $1 - \delta$ over the training set:*

$$L \leq \hat{L}_\gamma + \sqrt{\frac{\left(84B\sum_{i=1}^{d} k_i\sqrt{c_i} + \sqrt{\ln(4n^2 d)}\right)^2 \prod_{i=1}^{d} \|\mathbf{W}_i\|_2^2 \sum_{j=1}^{d} \frac{\|\mathbf{W}_j - \mathbf{W}_j^0\|_F^2}{\|\mathbf{W}_j\|_2^2} + \ln(\frac{m}{\delta})}{\gamma^2 m}} \tag{24}$$

Pitas, K., Davies, M., and Vandergheynst, P. (2017). Pac-bayesian margin bounds for convolutional neural networks. arXiv preprint arXiv:1801.00171

Vision for XAI science

Although still far from science

Regional explanation with strict meanings

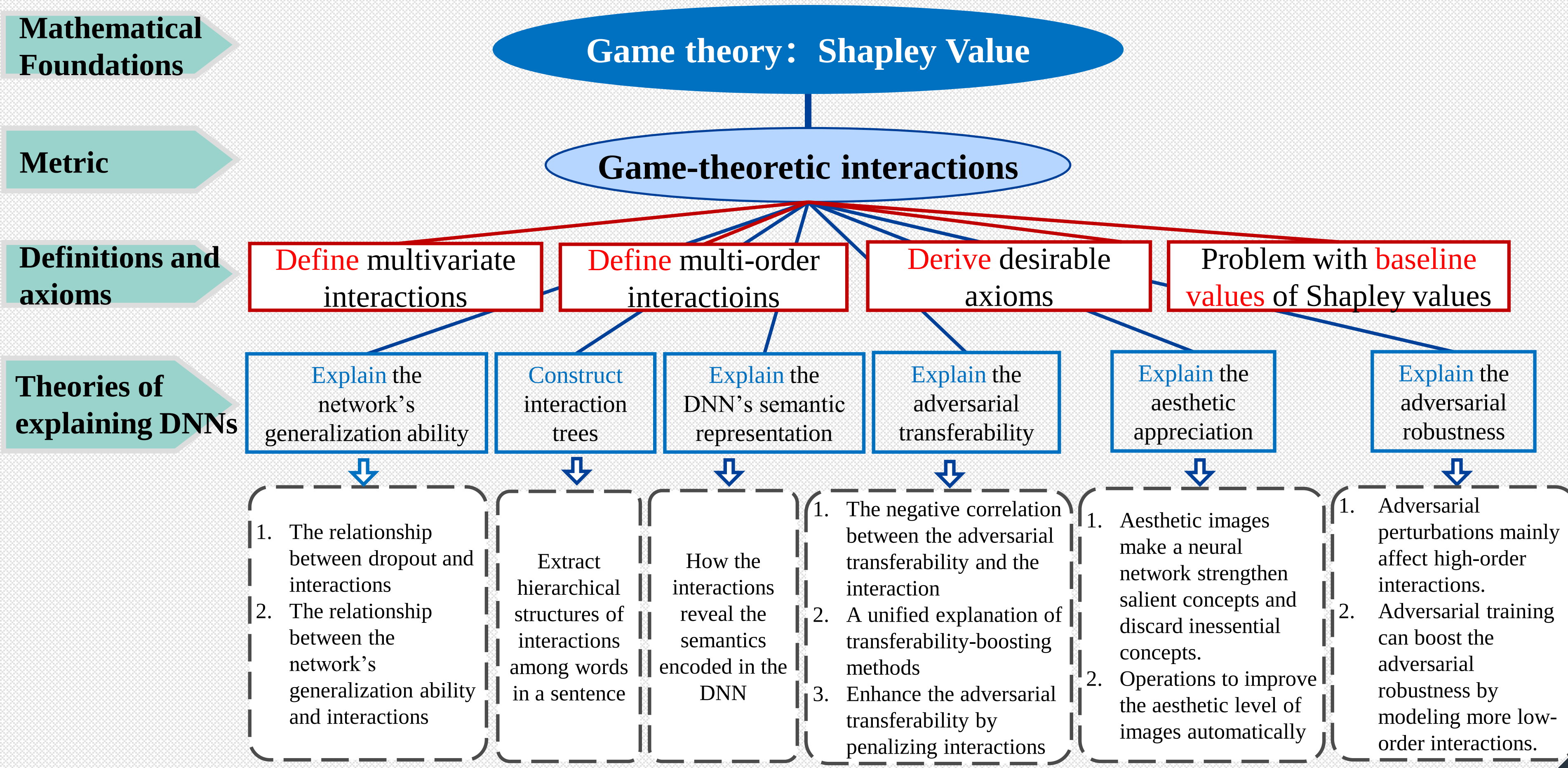
- Strict meanings of visual concepts
- Accurate attributions

XAI metrics for representation power of DNNs

- Mutuality between different metrics
  - Feature transferability
  - Adversarial robustness/transferability
  - Transformation complexity
  - Generalization
  - Disentanglement
  - Feature information
  - Interactions
- Essence of existing deep-learning methods
  - Summarize effective factors
  - Improve existing methods
- Guide deep learning
  - Guide the design of network architecture
  - Guide the learning process

Well-proved theoretic foundation

Game-theoretic interactions

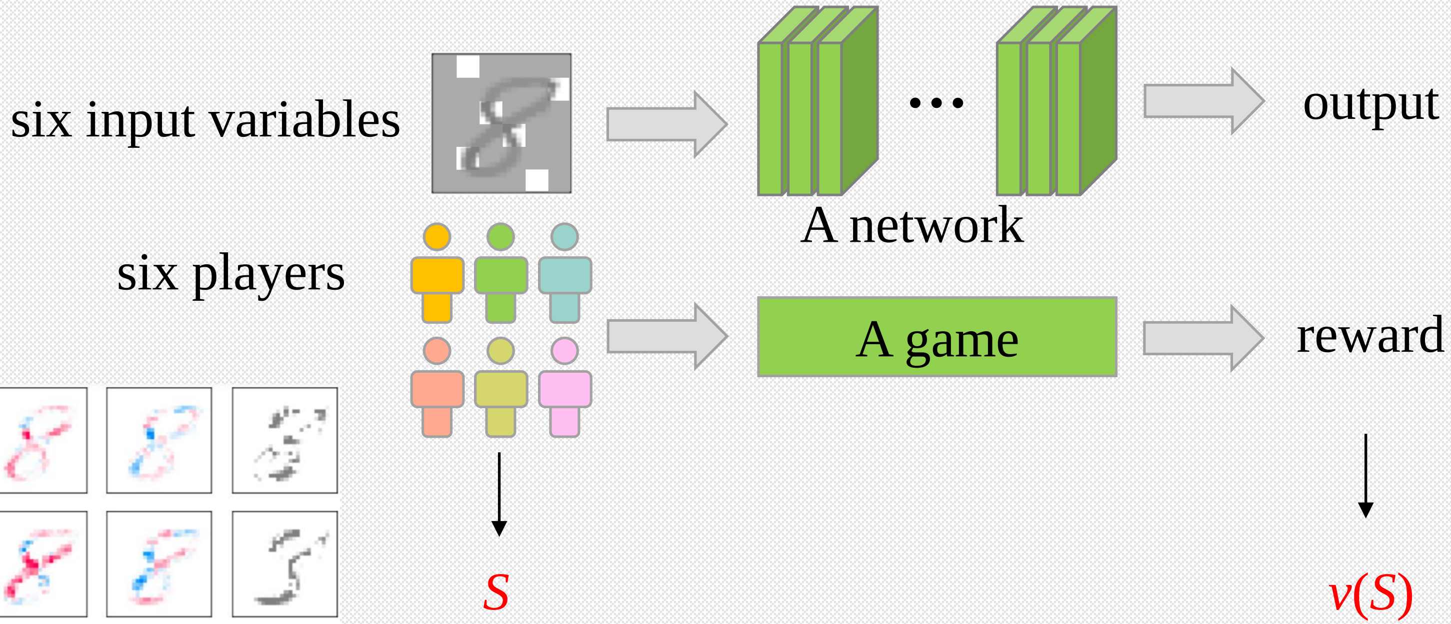


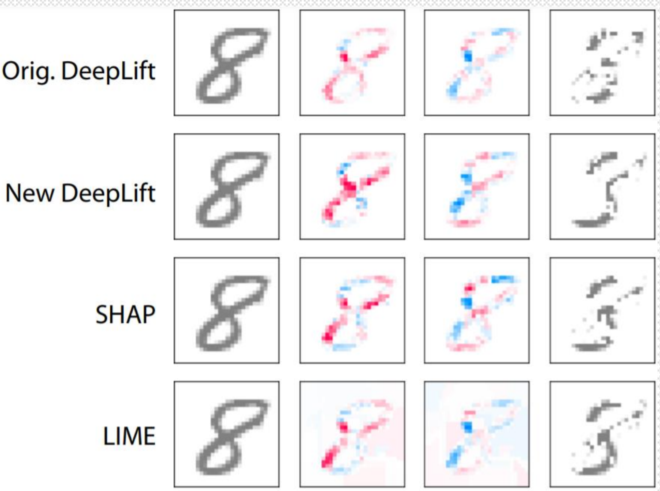
Zhang et al. "Interpreting Multivariate Shapley Interactions in DNNs" in AAAI 2021

Preliminaries: Shapley values

☐ Game

- Input variables → players
- Scalar network output/loss → total rewards of players in the game





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

Preliminaries: Shapley values

- Given a game, how to fairly allocate contribution of each player?

The **Shapley value** is considered as a method that fairly allocates the reward to players.

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

$$v(N) = v(\emptyset) + \sum_{i \in N} \phi(i|N)$$

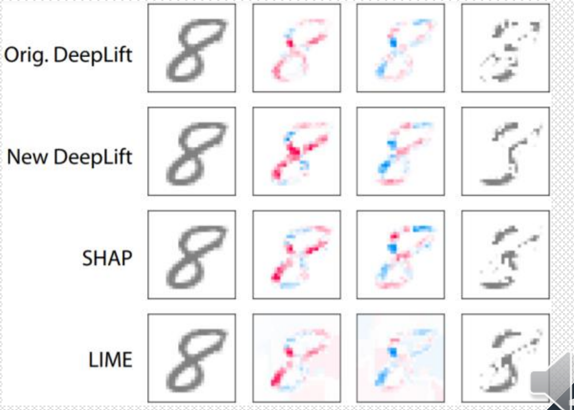
Lloyd S Shapley. “A value for n-person games”. In: Contributions to the Theory of Games 2.28 (1953), pp. 307–317.
Scott M. Lundberg, and Su-In Lee, “A unified approach to interpreting model predictions” in NeurIPS 2017

Preliminaries: Shapley values

Question: Given a game, how to fairly allocate contribution of each player?

Several **desirable axioms** ensure the fairness of allocation:

- **Linearity axiom**
  If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $\phi_u(i|N) = \phi_v(i|N) + \phi_w(i|N)$
- **Dummy axiom**
  If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, then $\phi(i|N) = v(\{i\}) - v(\emptyset)$
- **Symmetry axiom**
  If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\phi(i|N) = \phi(j|N)$
- **Efficiency axiom**
  $\sum_{i \in N} \phi(i|N) = v(N) - v(\emptyset)$



Lloyd S Shapley. "A value for n-person games". Contributions to the Theory of Games 2.28 (1953), pp. 307–317.

Preliminaries: Shapley values

- **Question**: Given a game, how to fairly allocate contribution of each player?

Several **desirable axioms** ensure the fairness of allocation:

- **Linearity axiom**

  If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $\phi_u(i|N) = \phi_v(i|N) + \phi_w(i|N)$

  If two independent games v and w can be merged into one game, then the Shapley value of the player i in game v and game w also can be merged.

- **Dummy axiom**

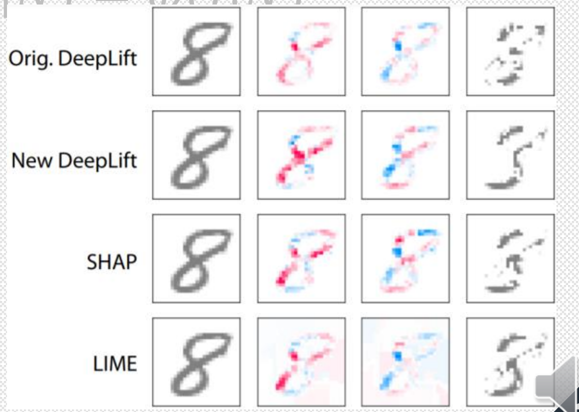
  If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, then $\phi(i|N) = v(\{i\}) - v(\emptyset)$

- **Symmetry axiom**

  If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\phi(i|N) = \phi(j|N)$

- **Efficiency axiom**

  $\sum_{i \in N} \phi(i|N) = v(N) - v(\emptyset)$



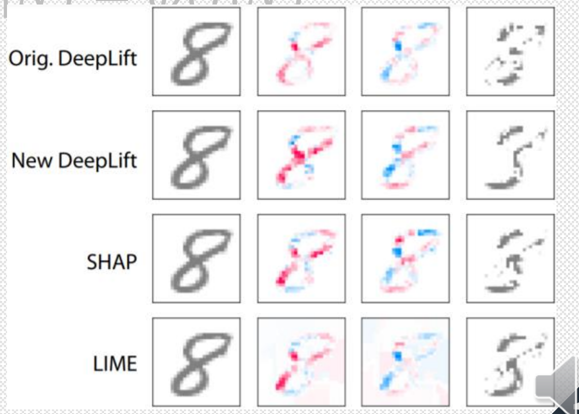
Lloyd S Shapley. "A value for n-person games". Contributions to the Theory of Games 2.28 (1953), pp. 307–317.

Preliminaries: Shapley values

Question: Given a game, how to fairly allocate contribution of each player?

Several **desirable axioms** ensure the fairness of allocation:

- **Linearity axiom**
  If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $\phi_u(i|N) = \phi_v(i|N) + \phi_w(i|N)$
- **Dummy axiom**
  If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, then $\phi(i|N) = v(\{i\}) - v(\emptyset)$
  A dummy player i satisfies that the player i has no interaction with other players.
- **Symmetry axiom**
  If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\phi(i|N) = \phi(j|N)$
- **Efficiency axiom**
  $\sum_{i \in N} \phi(i|N) = v(N) - v(\emptyset)$



Lloyd S Shapley. "A value for n-person games". Contributions to the Theory of Games 2.28 (1953), pp. 307–317.

Preliminaries: Shapley values

- **Question**: Given a game, how to fairly allocate contribution of each player?

Several **desirable axioms** ensure the fairness of allocation:

- **Linearity axiom**

  If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $\phi_u(i|N) = \phi_v(i|N) + \phi_w(i|N)$

- **Dummy axiom**

  If $\forall S \subseteq N \setminus \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, then $\phi(i|N) = v(\{i\}) - v(\emptyset)$

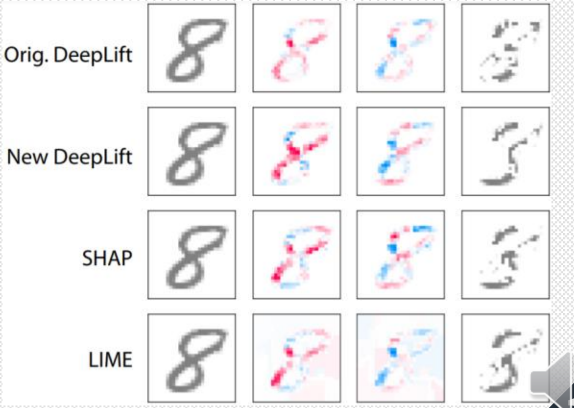
- **Symmetry axiom**

  If $\forall S \subseteq N \setminus \{i\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\phi(i|N) = \phi(j|N)$

  If two players i, j have same collaborations with other players, then they have the same Shapley value.

- **Efficiency axiom**

  $\sum_{i \in N} \phi(i|N) = v(N) - v(\emptyset)$



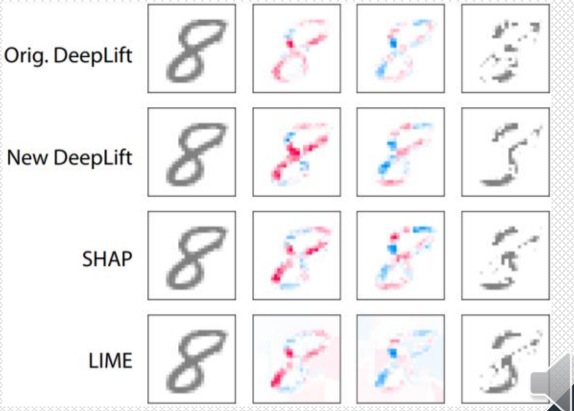
Lloyd S Shapley. "A value for n-person games". Contributions to the Theory of Games 2.28 (1953), pp. 307–317.

Preliminaries: Shapley values

- **Question**: Given a game, how to fairly allocate contribution of each player?

Several **desirable axioms** ensure the fairness of allocation:

- **Linearity axiom**
  If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $\phi_u(i|N) = \phi_v(i|N) + \phi_w(i|N)$
- **Dummy axiom**
  If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, then $\phi(i|N) = v(\{i\}) - v(\emptyset)$
- **Symmetry axiom**
  If $\forall S \subseteq N \backslash \{i\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\phi(i|N) = \phi(j|N)$
- **Efficiency axiom**
  $\sum_{i \in N} \phi(i|N) = v(N) - v(\emptyset)$
  The overall reward can be allocated to all players in the game.



Lloyd S Shapley. "A value for n-person games". Contributions to the Theory of Games 2.28 (1953), pp. 307–317.

Preliminaries: Shapley values

Remaining issues

- How to determine reasonable baseline values?
- How to determine the reasonable partition of players?

Lloyd S Shapley. "A value for n-person games". Contributions to the Theory of Games 2.28 (1953), pp. 307–317.

How to define interactions in game theory?

How to determine baseline values for the Shapley value?
What is the relationship between interactions and visual concepts?
What is the relationship between interactions and the aesthetic appreciation?
What is the relationship between interactions and the generalization?
What is the relationship between interactions and adversarial transferability?
What is the relationship between interactions and adversarial robustness?

Game-theoretic interactions

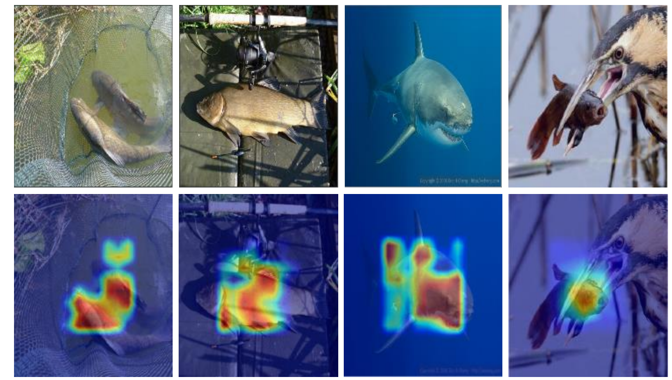
it ' s a remarkably solid and subtly satirical tour de force .

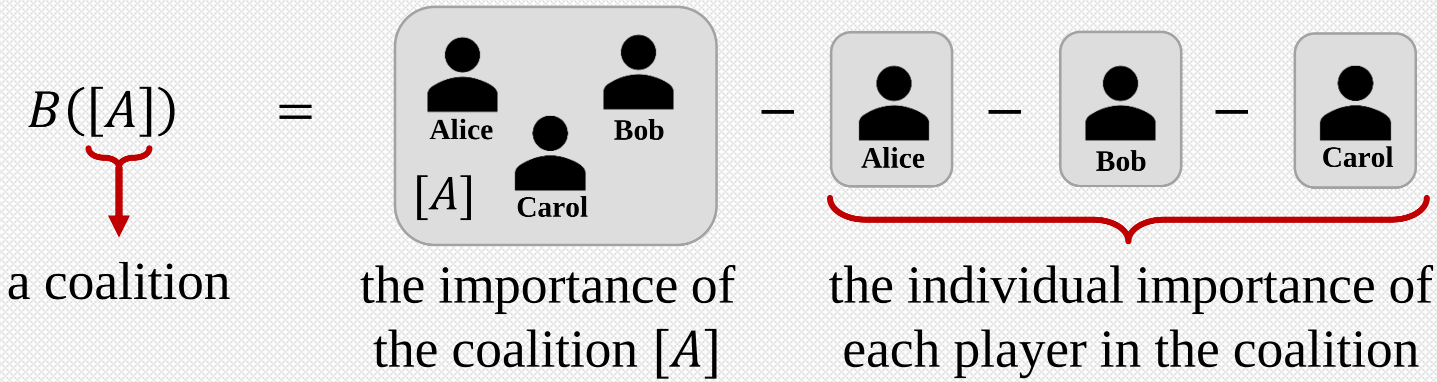
this is a good script , good dialogue , funny even for adults

dull, lifeless, and amateurishly assembled .

a warm but realistic meditation on friendship , family and affection .

no telegraphing is too obvious or simplistic for this movie .



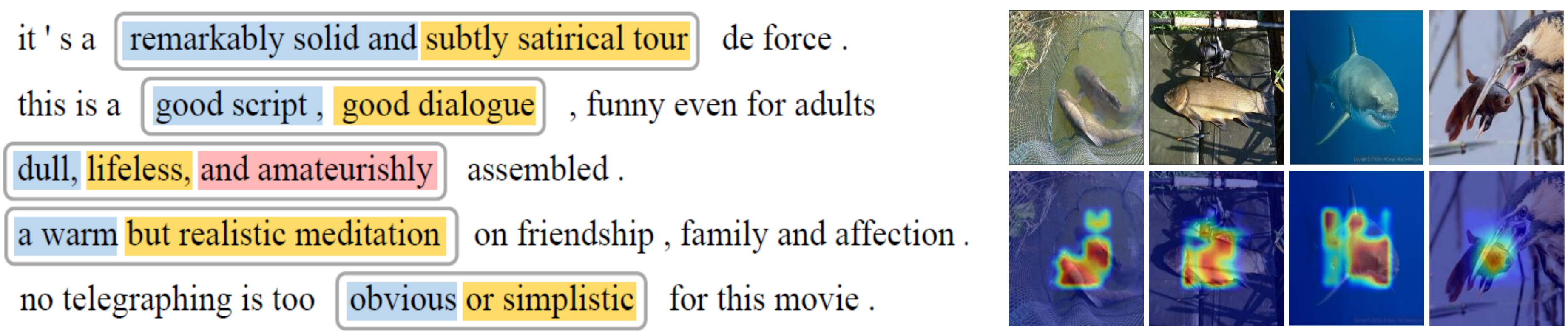


$B([A]) > 0$: Players in $[A]$ mainly have **cooperative** relationship.

$B([A]) < 0$: Players in $[A]$ mainly have **adversarial** relationship.

Zhang et al, "Interpreting Multivariate Shapley Interactions in DNNs" in AAAI 2021

Game-theoretic interactions



- **Input words of a sentence (or the pixels of an image) usually cooperate with each other, rather than work individually to make inferences.**
- **The cooperative input words (or pixels) have strong interactions.**
- **Shapley interactions between two players (i,j):** the change of the importance of *i* when *j* is present, w.r.t. the importance when *j* is absent.

$$I(i,j) = \phi_{w/j}(i|N) - \phi_{w/oj}(i|N)$$

Zhang et al, "Interpreting Multivariate Shapley Interactions in DNNs" in AAAI 2021

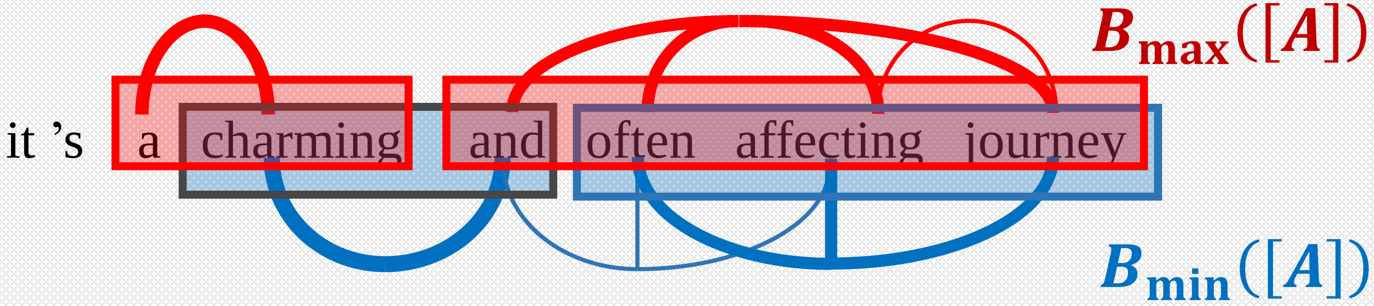
Multivariate Shapley interactions: properties

Properties of multivariate Shapley interactions $B([A])$:

- **Linearity property :**
  If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $\forall A \subsetneq N, B_u([A]) = B_v([A]) + B_w([A])$.
- **Dummy property :** the dummy player has **no** interaction with other players. If $\forall S \subseteq N\backslash\{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, then $\forall A \subsetneq N\backslash\{i\}, B([A \cup \{i\}]) = B([A])$.
- **Symmetry property :** symmetric players have **same** interaction with other players.
  If $\forall S \subseteq N\backslash\{i, j\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\forall A \subsetneq N, B([A \cup \{i\}]) = B([A \cup \{j\}])$

Zhang et al, "Interpreting Multivariate Shapley Interactions in DNNs" in AAAI 2021

Multivariate Shapley interactions

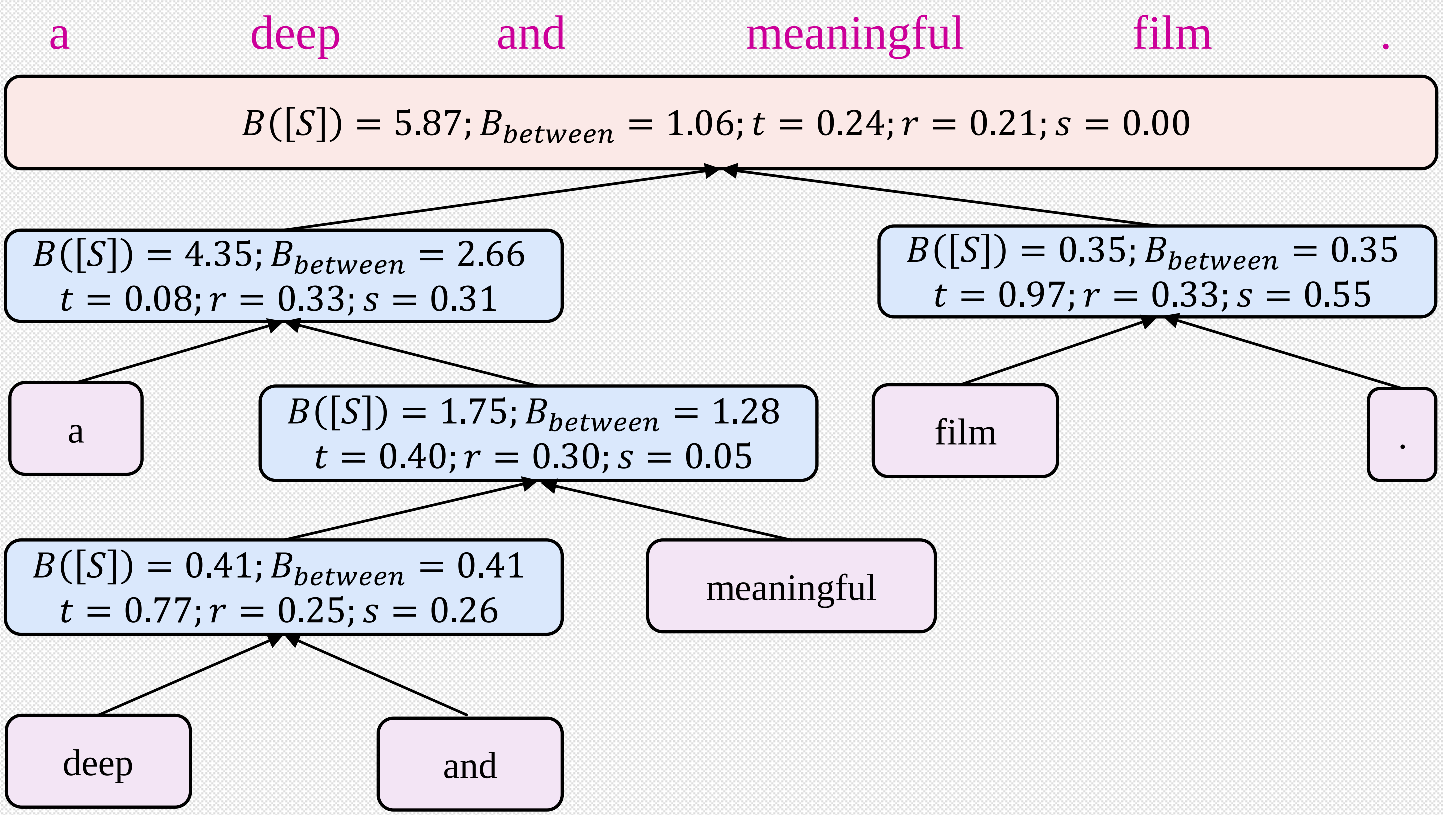


- $B_{max}([A])$ reflects **positive interaction** inside $[A]$.
- $B_{min}([A])$ reflects **negative interaction** inside $[A]$.
- $T([A]) = B_{max}([A]) - B_{min}([A])$
- $T([A])$ can measure both **positive** and **negative** interactions.
- We design an **effective** method to estimate the optimal partition and approximate $T([A])$.

Zhang et al, "Interpreting Multivariate Shapley Interactions in DNNs" in AAAI 2021

Explain the rationale of incorrect prediction

- **Multivariate interactions** can be used to extract tree structures that encoded interactions among words inside different DNNs.



Zhang et al, “Building Interpretable Interaction Trees for Deep NLP Models” in AAAI 2021

Explain the rationale of incorrect prediction

- **Multivariate interactions** show extract prototype features to help us understand the **incorrect predictions** of DNNs

maximum (prototypes towards incorrect predictions):

if steven soderbergh ' s ' solaris ' is a failure it is a glorious failure . predict: negative

minimum (prototypes towards correct predictions):

if steven soderbergh ' s ' solaris ' is a failure it is a glorious failure . label: positive

maximum (prototypes towards incorrect predictions):

the longer the movie goes , the worse it gets , but it ' s actually pretty good in the first few minutes. predict: positive

minimum (prototypes towards correct predictions):

the longer the movie goes , the worse it gets , but it ' s actually pretty good in the first few minutes. label: negative

maximum (prototypes towards incorrect predictions):

on the heels of the ring comes a similarly morose and humorless horror movie that , although flawed , is to be commended for its straight - ahead approach to creepiness . predict: negative

minimum (prototypes towards correct predictions):

on the heels of the ring comes a similarly morose and humorless horror movie that , although flawed , is to be commended for its straight - ahead approach to creepiness . label: positive

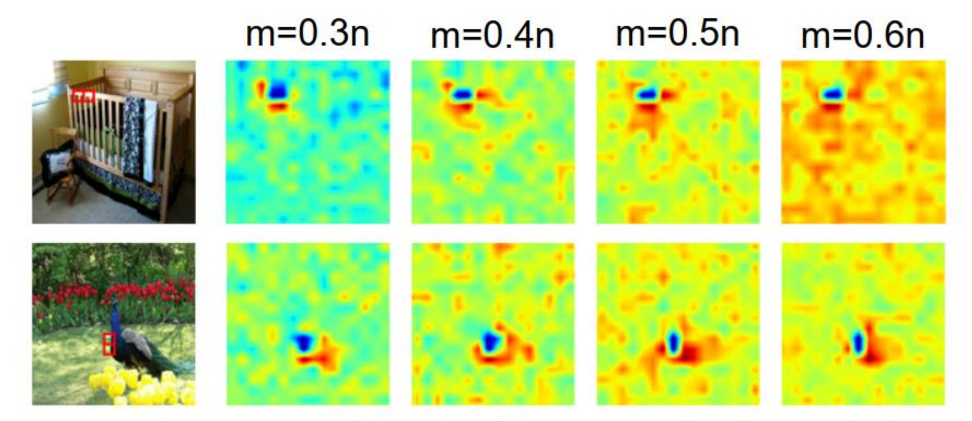
Zhang et al, "Interpreting Multivariate Shapley Interactions in DNNs" in AAAI 2021

Multi-order interactions to represent the complexity of representations

- We further define interactions of different orders as follows.

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

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



Low order m: simple contextual collaborations with a few pixels → represents simple concepts;

High order m: complex contextual collaborations with massive pixels → represents complex concepts.

Ren et al. Game-theoretic Understanding of Adversarially Learned Features. in arXiv:2103.07364.

Multi-order interactions: properties

Properties of multi-order interactions

- **Marginal contribution property :** $\forall i,j \in N, i \neq j, \phi^{(m+1)}(i|N) - \phi^{(m)}(i|N) = \mathop{\mathbb{E}}_{j\in N\setminus\{i\}}[I^{(m)}(i,j)]$
- **Accumulation property :** $\phi^{(m)}(i|N) = \mathop{\mathbb{E}}_{j\in N\setminus\{i\}}\left[\sum_{k=0}^{m-1} I^{(k)}(i,j)\right] + \phi^{(0)}(i|N)$
- **Efficiency property :** $v(N) - v(\emptyset) = \sum_{i\in N}\phi^{(0)}(i|N) + \sum_{i\in N}\sum_{j\in N\setminus\{i\}}\left[\sum_{k=0}^{n-2}\frac{n-1-k}{n(n-1)} I^{(k)}(i,j)\right]$
- **Linearity property :** If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $I_u^{(m)}(i,j) = I_v^{(m)}(i,j) + I_w^{(m)}(i,j)$
- **Independency property :** If $\forall S \subseteq N\setminus\{i\}, v(S\cup\{i\}) = v(S) + v(\{i\})$, then $\forall j \in N, I^{(m)}(i,j) = 0$
- **Symmetry property :** If $\forall S \subseteq N, v(S\cup\{i\}) = v(S\cup\{j\})$, then $\forall k \in N\setminus\{i,j\}, I^{(m)}(i,k) = I^{(m)}(j,k)$
- **Summability property :** $\phi^{(n-1)}(i|N) - \phi^{(0)}(i|N) = \mathop{\mathbb{E}}_{j\in N\setminus\{i\}}\left[\sum_{m=0}^{n-2} I^{(m)}(i,j)\right] = I(N\setminus\{i\}, i) = \sum_{j\in N\setminus\{i\}} I(i,j)$

Ren et al. Game-theoretic Understanding of Adversarially Learned Features. in arXiv:2103.07364.

How to define interactions in game theory?

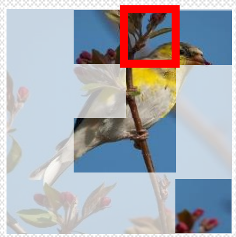
How to determine baseline values for the Shapley value?

What is the relationship between interactions and visual concepts?
What is the relationship between interactions and the aesthetic appreciation?
What is the relationship between interactions and the generalization?
What is the relationship between interactions and adversarial transferability?
What is the relationship between interactions and adversarial robustness?

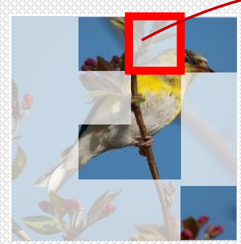
Problem with baseline values

The marginal effects of the additional variable (red square)

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} w_S [\, v(S \cup \{i\}) \; - \; v(S) \,]$$



−



Baseline values: the value representing **the absence of the variables** (providing no signal to the model inference).

Previous settings of baseline values

- Zero
- Mean
- Blurring
- Depending on neighboring contexts S[1]:

$$v(S) = E_{p(x'_{\bar{S}} | x_S)} \left[f\left(x_S \sqcup x'_{\bar{S}} \right) \right]$$

$\neq$

Remove all information of variables w/o generating new edges/dots.

[1] Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In International Conference on Learning Representations, 2021.

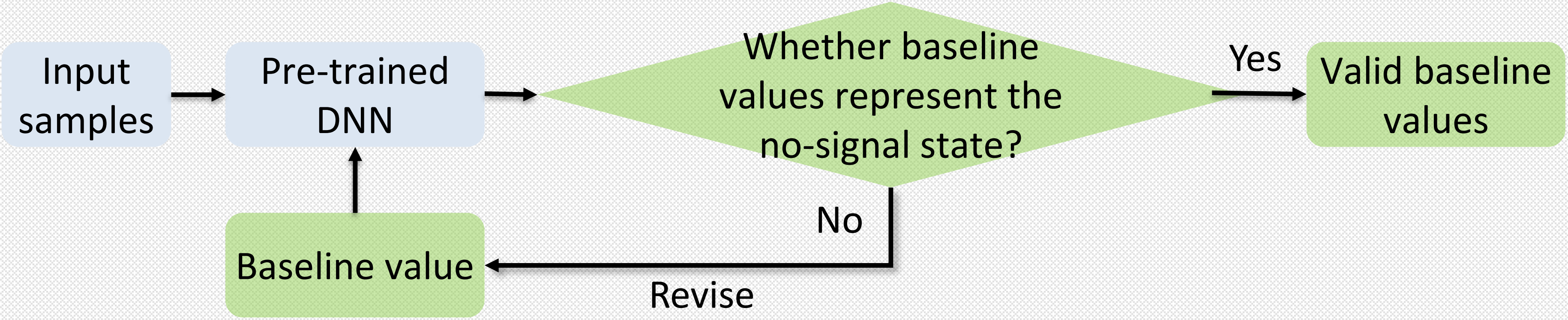
Objective of learning baseline values

Input: A trained model and input samples

Output: Baseline values that satisfy the following two requirements:

(1) retain the four axioms of Shapley values

(2) push the baseline value towards representing no-signal state as much as possible.



Multi-variate interaction

- The multi-variate interaction should ensure that

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

Network output
the benefit from all variables

a constant bias

the marginal benefit from the interaction of all variables in S

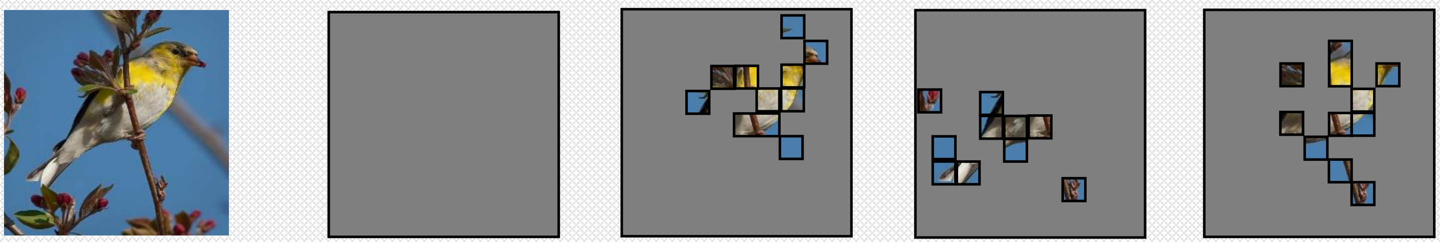
- Solution:

$$I(S) = \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L)$$

A new multi-variate interaction

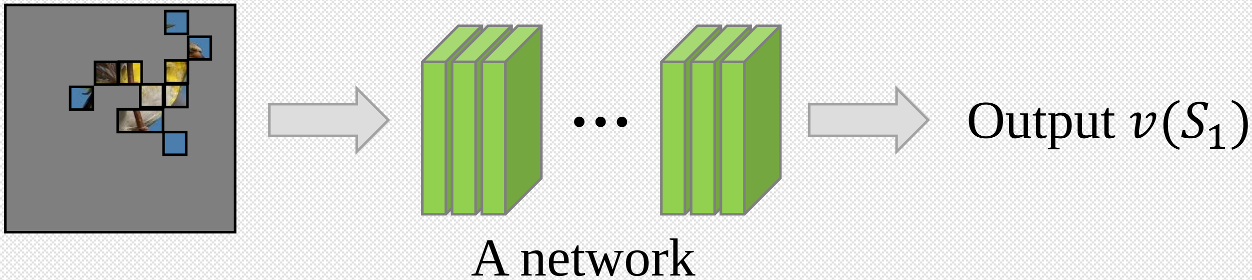
- **Transforming a DNN into an AND-OR representation.**
- **Decompose the overall utility of a DNN into utilities of different multi-variate interactions**

$$v(N) = v(\emptyset) + I(S_1) + I(S_2) + I(S_3) + \cdots$$



Network output | **Constant bias** | **Elementary interaction component** | **Elementary interaction component** | **Elementary interaction component**

For $I(S_1) = \sum_{L \subseteq S_1} (-1)^{|S_1|-|L|} v(L)$



Using interaction patterns to represent the no-signal state

- **Salient patterns I(S)** with significant influences, |I(S)| is large
- **Noisy patterns I(S):** with little influences, |I(S)| is small

$$v(N) - v(\emptyset) = \sum_{S \subseteq N} I(S)$$

Learning baseline values that activate the least salient patterns → most likely to represent the no-signal state.

Learning baseline values → to minimize the number of salient patterns

Conclusion 1

The optimal baseline values

↓ aim to

represent no-signal state

Conclusion 2

How to represent signal state?

↓ $v(N) - v(\emptyset) = \sum_{S \subseteq N} I(S)$

Using the number of salient patterns

Therefore, we can learn the baseline values that minimize the number of salient patterns

Activation rate of different interaction patterns

Let $\delta_i = 1$ denote the presence of the variable i, and let δ_i =0 represent the absence of i. Let us consider a set of m variables. Let $P(\delta_i = 1) = \frac{1}{2}$

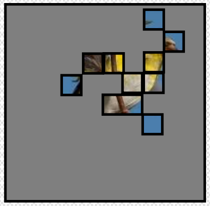
We can rewrite $I(S)$ as

$$I(S) = \delta_1 \delta_2 \cdots \delta_m \cdot w_S \qquad m = |S|$$

Then $P(I(S) \neq 0) = P(\delta_1 = 1) P(\delta_2 = 1) \cdots P(\delta_m = 1) = 0.5^m$

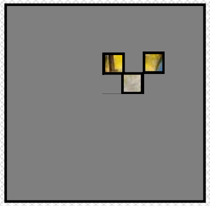
We define m = $|S|$ as the order of the interaction $I(S)$.

- For high-order interactions, where m = $|S|$ is large:



$$\mathrm{P}(I(S) \neq 0) = \frac{1}{2} * \frac{1}{2} * \cdots = \frac{1}{2^{11}}$$

- For low-order interactions, where m = $|S|$ is small:



$$\mathrm{P}(I(S) \neq 0) = \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{2^3}$$

Relationship between low-order interactions and high-order interactions

$$v(N) = v(\emptyset) + \underbrace{\sum_{S \in \Omega_{low}} I(S)}_{\text{Low-order interactions}} + \underbrace{\sum_{S \in \Omega_{high}} I(S)}_{\text{High-order interactions}}$$

- $\Omega_{low} = \{S \,|\, |S| \leq threshold\}, \ \Omega_{high} = \{S \,|\, |S| > threshold\}$

Learning baseline values → to reduce low-order interaction patterns

$$v(N) = v(\emptyset) + \sum_{S \in \Omega_{low}} I(S) + \sum_{S \in \Omega_{high}} I(S)$$

High-order interactions
→ Low activation rate
→ Sparse activations

Reduce signals represented by low-order interactions

||

Strengthen signal represented by high-order interactions

||

Make most signals sparsely activated

How to reduce low-order interactions

We prove that **low-order Shapley values only contain low-order interactions.**

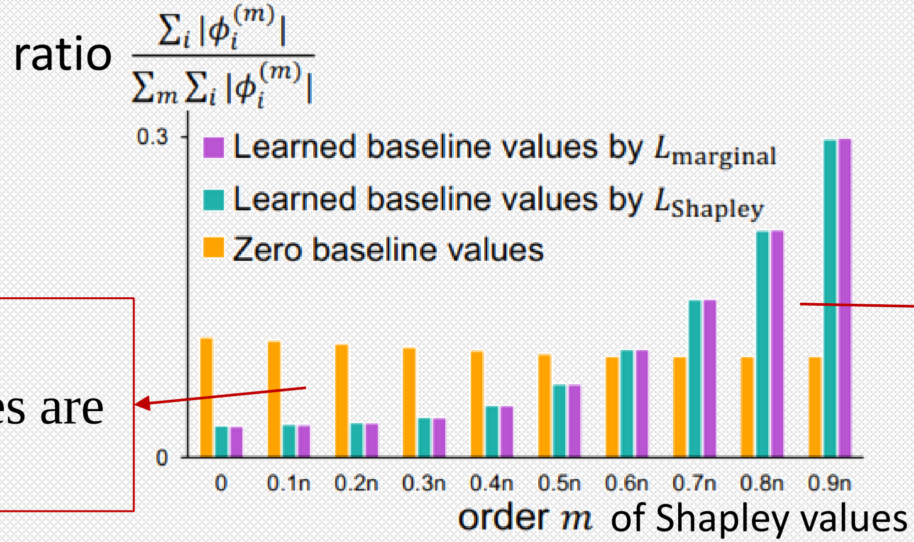
The m-order Shapley value $\phi^{(m)}(i) = E_{S\subseteq N\setminus\{i\},|S|=m}[v(S\cup\{i\}) - v(S)]$

The approximate-yet-efficient solution: penalizing low-order Shapley values

$$L_{Shapley} = \sum_{m\sim Unif(0;\lambda)} \sum_{x\in X} \sum_{i\in N} |\phi^{(m)}(i)|$$

Verification

Objective: we aim to verify whether or not we can successfully reduce the ratio of low-order Shapley values and boost the influence of high-order Shapley values



Low-order Shapley values are reduced

High-order Shapley values are strengthened.

Connections between multi-variate interaction and other metrics

- Connecting the interaction I(S) to the Shapley value:

The Shapley value $\phi(i) = \sum_{S: i \in S} \frac{1}{|S|} I(S)$

- Connecting the interaction I(S) to the Shapley interaction index:

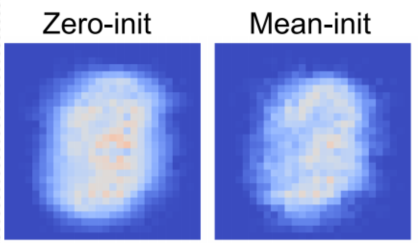
$$I_{shap}(S) = \sum_{T \subseteq N \setminus S} p(T) \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L \cup T) = \sum_{T \subseteq N \setminus S} p(T) I(S|env(T))$$

where $I(S|env(T))$ denotes the specific interaction $I(S)$ when variables in T are always presents.

Experiments: learned baseline values and Shapley values

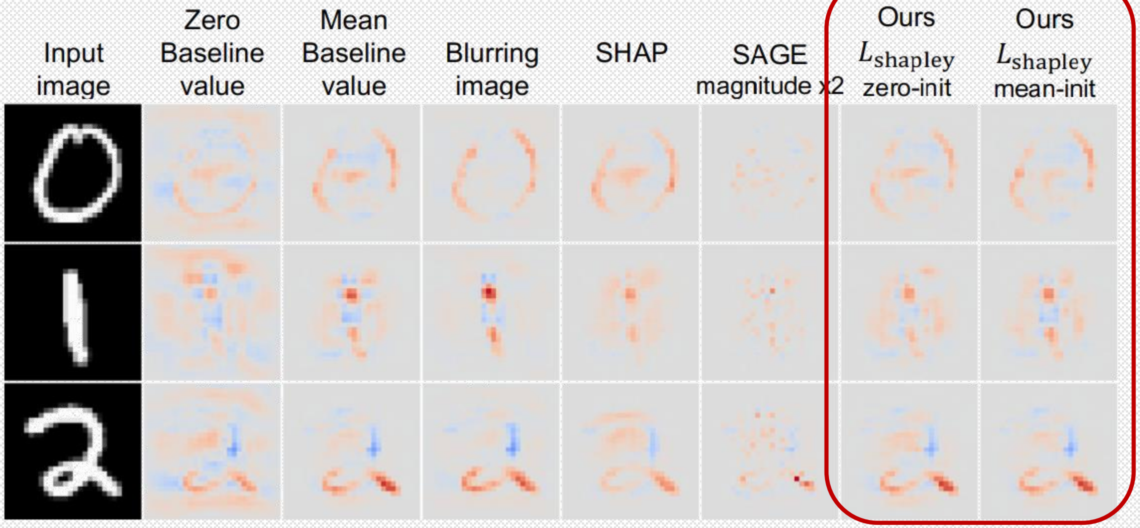
The baseline values learned by our method generated less noisy Shapley values than other methods

- On MNIST dataset
  - ➢ Learned baseline value: (shared by all MNIST images)



Focus on foreground

Shapley values based on different baseline values

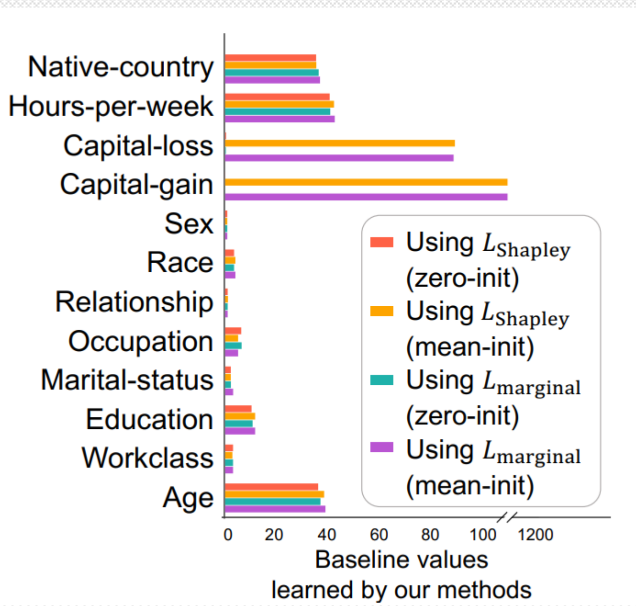


Less noise

Experiments: learned baseline values and Shapley values

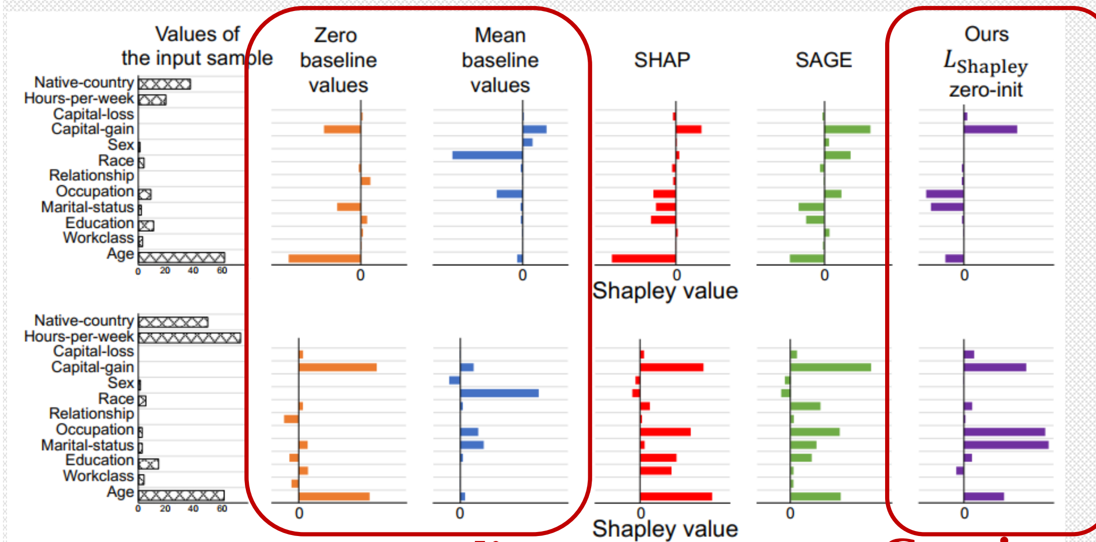
The learned baseline values generate Shapley values, which are consistent with SHAP and SAGE.

- On the UCI Census Income dataset
  - ➢ Learned baseline value:



Shapley values on other baseline values

Shapley value: Ours



Contradict with other methods

Consistent with SHAP and SAGE

Experiments: verification of the learned baseline values

Verify the correctness of the learned baseline values

- On images, there are **no ground-truth baseline values** for verification.
- We generated functions, whose ground truth of baseline values could be easily determined.

Functions ($\forall i \in N, x_i \in \{0,1\}$)
$-0.185x_1(x_2+x_3)^{2.432} - x_4x_5x_6x_7$
$-sigmoid(-4x_1 - 4x_2 - 4x_3 + 2.00) - 0.011x_4(x_5+x_6+x_7+x_8+x_9)^{2.341}$
$-x_1x_2x_3 + sigmoid(-5x_4x_5x_6x_7 + 2.50) - x_8x_9$
$-sigmoid(+4x_1 - 4x_2 + 4x_3 - 6.00) - x_4x_5x_6x_7 - x_8x_9x_{10}$

The ground truth of baseline values
$b_i^* = 0$ for $i \in \{1,2,3,4,5,6,7\}$
$b_i^* = 1$ for $i \in \{1,2,3\}$, $b_i^* = 0$ for $i \in \{4,5,6,7,8,9\}$
$b_i^* = 1$ for $i \in \{4,5,6,7\}$, $b_i^* = 0$ for $i \in \{1,2,3,8,9\}$
$b_i^* = 1$ for $i = 2$, $b_i^* = 0$ for $i \in \{1,3,4,5,6,7,8,9,10\}$

Experiments: verification of the learned baseline values

Verify the correctness of the learned baseline values ($\boldsymbol{b_i^*} \in \{\mathbf{0}, \mathbf{1}\}$)

- Metric: accuracy of the learned baseline values

$$\frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}(b_i^* = 1 \;\&\; b_i > 0.5) + \mathbf{1}(b_i^* = 0 \;\&\; b_i < 0.5)]$$

Table 3: Accuracy of learned baseline values.

	L_{Shapley}			L_{marginal}		
	initialize with 0	initialize with 0.5	initialize with 1	initialize with 0	initialize with 0.5	initialize with 1
Synthetic functions	98.06%	98.70%	98.70%	98.06%	98.14%	98.14%
Functions in [47]	88.52%	91.80%	90.16%	86.89%	91.80%	90.16%

In most cases, the accuracy was above 90%, showing that **our method could effectively learn correct baseline values.**

Can we unify all attribution methods using game-theoretic interactions?

Huiqi Deng
Sun Yat-sen University

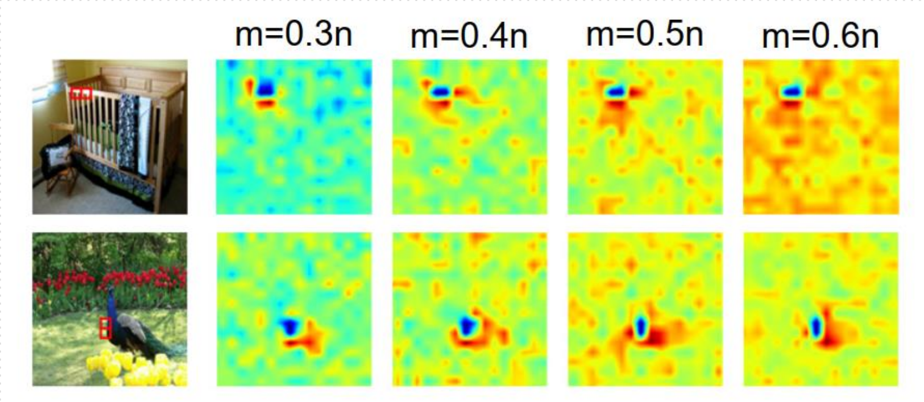
How to define interactions in game theory?
How to determine baseline values for the Shapley value?

What is the relationship between interactions and visual concepts?

What is the relationship between interactions and the aesthetic appreciation?
What is the relationship between interactions and the generalization?
What is the relationship between interactions and adversarial transferability?
What is the relationship between interactions and adversarial robustness?

Explaining textures, shapes, and beyond

- Multi-order interaction: measures the average interaction between pixels (i,j) under all contexts consisting of m pixels.



M-order interaction $$I^{(m)}(i,j) \stackrel{\text{def}}{=} \mathbb{E}_{S\subseteq N\setminus\{i,j\},|S|=m}[\Delta v(S,i,j)]$$

- ➢ Low-order interactions mainly reflect simple and common concepts.
- ➢ Middle-order interactions mainly represent middle complex concepts.
- ➢ High-order interactions mainly represent the memory of specific large-scale concepts.

Cheng et al, "A Game-Theoretic Taxonomy of Visual Concepts in DNNs" in arXiv:2106.10938, 2021.

Multi-order interactions: properties

Properties of multi-order interactions

- **Marginal contribution property :** $\forall i,j \in N, i \neq j, \phi^{(m+1)}(i|N) - \phi^{(m)}(i|N) = \mathbb{E}_{j\in N\setminus\{i\}}[I^{(m)}(i,j)]$
- **Accumulation property :** $\phi^{(m)}(i|N) = \mathbb{E}_{j\in N\setminus\{i\}}\left[\sum_{k=0}^{m-1} I^{(k)}(i,j)\right] + \phi^{(0)}(i|N)$
- **Efficiency property :** $v(N) - v(\emptyset) = \sum_{i\in N}\phi^{(0)}(i|N) + \sum_{i\in N}\sum_{j\in N\setminus\{i\}}\left[\sum_{k=0}^{n-2}\frac{n-1-k}{n(n-1)} I^{(k)}(i,j)\right]$
- **Linearity property :** If $\forall S \subseteq N, u(S) = v(S) + w(S)$, then $I_u^{(m)}(i,j) = I_v^{(m)}(i,j) + I_w^{(m)}(i,j)$
- **Independency property :** If $\forall S \subseteq N\setminus\{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, then $\forall j \in N, I^{(m)}(i,j) = 0$
- **Symmetry property :** If $\forall S \subseteq N, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\forall k \in N\setminus\{i,j\}, I^{(m)}(i,k) = I^{(m)}(j,k)$
- **Summability property :** $\phi^{(n-1)}(i|N) - \phi^{(0)}(i|N) = \mathbb{E}_{j\in N\setminus\{i\}}\left[\sum_{m=0}^{n-2} I^{(m)}(i,j)\right] = I(N\setminus\{i\}, i) = \sum_{j\in N\setminus\{i\}} I(i,j)$

Ren et al. Game-theoretic Understanding of Adversarially Learned Features. in arXiv:2103.07364.

What is the relationship between interactions and visual concepts?

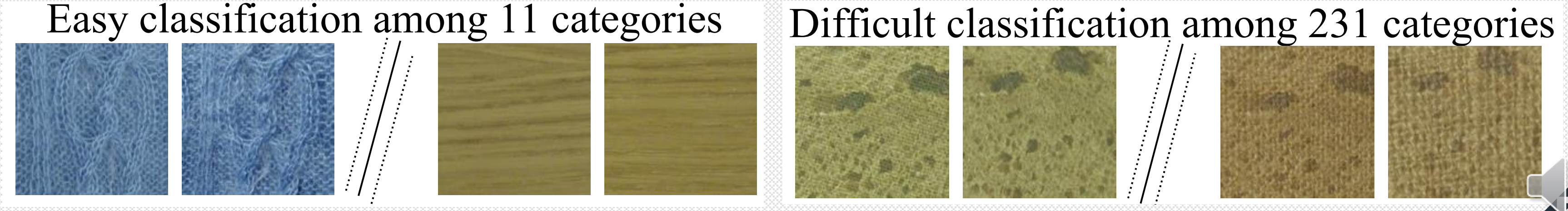
- **Understanding the encoding of textures**
- Understanding the difference between textures & shapes
- Understanding large-scale visual concepts
- Understanding outliers

Understanding the encoding of textures

- How does a DNN encodes textures?
  - **Low-order** interactions usually represent **common and widely-shared local** textures.
  - **Middle-order** interactions usually represent more **complex** textures.

- Hypothesis:

Compared to classify a few textures using low-order (simple) interactions, the classification of massive **fine-grained textures** usually forced a DNN to encode **fewer middle-order** interactions, which **subtly distinguish** fine-grained textures.

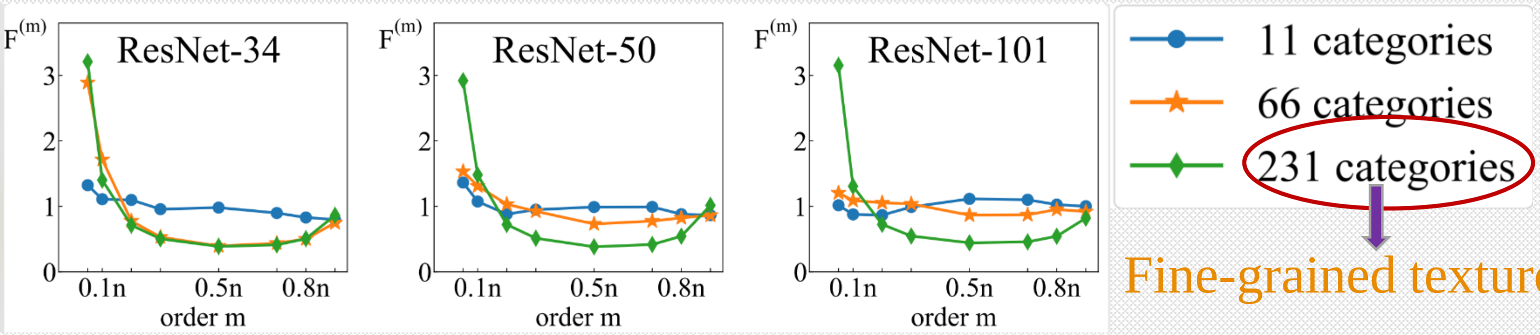


Understanding the encoding of textures

- In order to verify hypothesis that **fine-grained texture classification** made the DNN encode **fewer but more complex middle-order** interactions.
  - The **metric** $\boldsymbol{F^{(m)}}$ → the relative strength of the m-th order

$$F^{(m)} = I^{(m)}_{\text{strength}} / \mathbb{E}_{m'}[I^{(m')}_{\text{strength}}], \qquad I^{(m)}_{\text{strength}} = \mathbb{E}_{x\in\Omega}\left[\mathbb{E}_{i,j}[|I^{(m)}(i,j|x)|]\right]$$

- Verification:



Conclusion: The stricter encoding of fine-grained textures usually leads to fewer middle-order interactions.

Cheng et al, "A Game-Theoretic Taxonomy of Visual Concepts in DNNs" in arXiv:2106.10938, 2021.

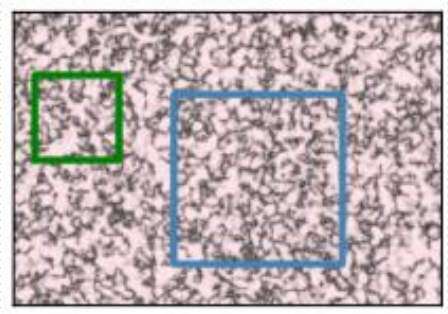
What is the relationship between interactions and visual concepts?

- Understanding the encoding of textures
- **Understanding difference between textures & shapes**
- Understanding large-scale visual concepts
- Understanding outliers

Difference between textures & shapes

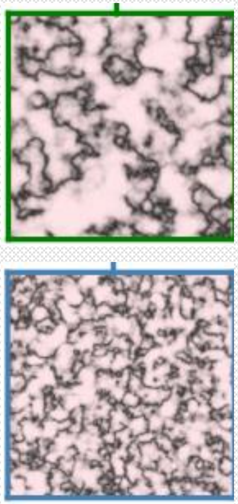
- **Encoding textures is more flexible than encoding shapes.**

➢ A **large-scale texture**



Can be modeled either as the ensemble of massive local textures.

Or as the ensemble of a few middle-complex textures.



➢ A **large-scale shape** is usually encoded as the ensemble of middle-complex shapes.

Cheng et al, "A Game-Theoretic Taxonomy of Visual Concepts in DNNs" in arXiv:2106.10938, 2021.

Difference between textures & shapes

- **Hypothesis:**
  If DNNs learned under **different noisy** conditions have **similar distributions** of the interaction orders, we consider the encoding of concepts is **not flexible**; otherwise, it is flexible.

- **Metric to verify hypothesis:**
  $\Delta F^{(m)} = |F^{(m,noise)} - F^{(m)}|$ measures the difference of multi-order interaction strength between the DNN learned with noise and the DNN learned without noise.

  ➡ A **large** $\Delta F^{(m)}$ indicates the encoding of concepts is **flexible**.

Cheng et al, "A Game-Theoretic Taxonomy of Visual Concepts in DNNs" in arXiv:2106.10938, 2021.

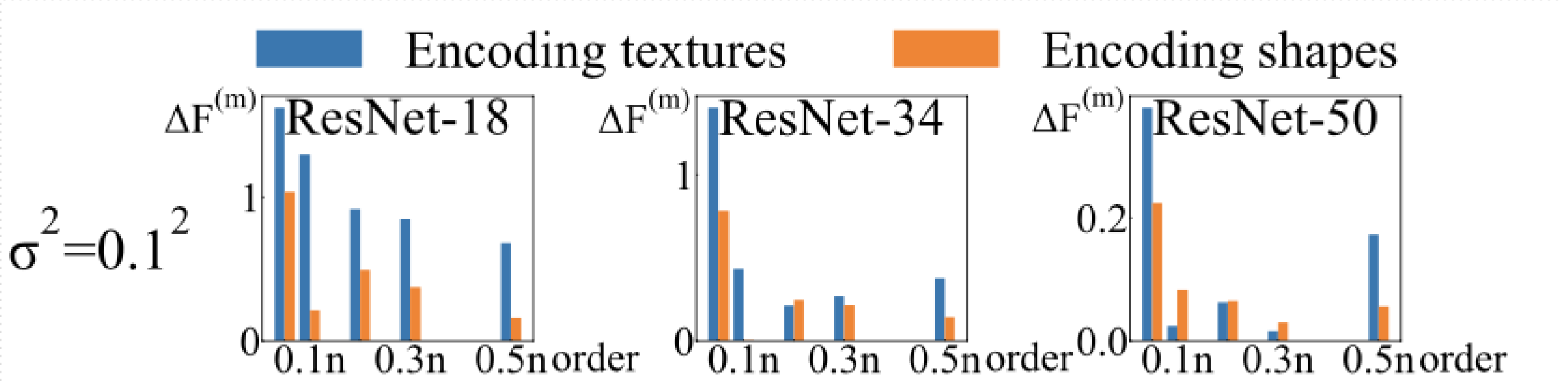
Difference between textures & shapes

- **Verification:**

  Compared with encoding shapes, encoding textures usually had **large $\Delta F^{(m)}$** values.

Conclusion: Compared with encoding shapes, a DNN encodes textures with more flexibility.

Noise level



Cheng et al, "A Game-Theoretic Taxonomy of Visual Concepts in DNNs" in arXiv:2106.10938, 2021.

What is the relationship between interactions and visual concepts?

- Understanding the encoding of textures
- Understanding the difference between textures & shapes
- **Understanding large-scale visual concepts**
- Understanding outliers

Understanding large-scale visual concepts

- **Concepts encoded as high-order interactions usually satisfy two requirements:**

  1. Frequently appear in images, such as *sky or ocean*;
  2. The interaction between the background and the foreground is used for inference, such as *the interaction between the ocean and the red-breasted merganser*.

red-breasted merganser

(a) (b) (c) (d)

Either only the foreground or only the background is not discriminative enough for inference.

Cheng et al, "A Game-Theoretic Taxonomy of Visual Concepts in DNNs" in arXiv:2106.10938, 2021.

Understanding large-scale visual concepts

- **Hypothesis:**

If a DNN **memorizes large-scale** concepts for inference, then this DNN is supposed to encode **more high-order interactions**.

- In order to verify this hypothesis, we construct two datasets.
  - One dataset of classifying entire bird heads and partial bird heads forces the DNN to **hard memorize** the **entire large-scale concepts** for inference.
  - The other dataset for the estimation of whether or not an image contains bird heads.

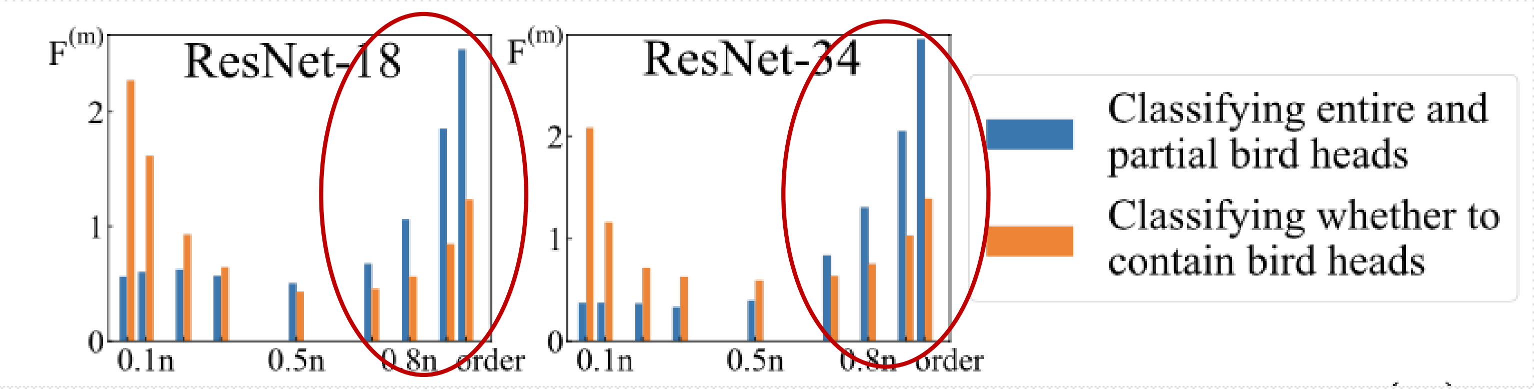
Understanding large-scale visual concepts

- **Metrics for verification:** Multi-order interaction strength $\boldsymbol{F}^{(m)}$.

- **Verification:**

The classification of entire and partial bird heads encoded more high-order interactions.

Conclusion: The DNN memorized large-scale concepts for inference usually encode more high-order interactions



Cheng et al, "A Game-Theoretic Taxonomy of Visual Concepts in DNNs" in arXiv:2106.10938, 2021.

What is the relationship between interactions and visual concepts?

- Understanding the encoding of textures
- Understanding the difference between textures & shapes
- Understanding large-scale visual concepts
- **Understanding outliers**

Understanding outliers

- Hypothesis:

The classification of outliers mainly depends on high-order interactions.

- In order to verify this hypothesis, we construct **synthetic outliers**.

  ✓ We add **negligible noises** to 50/100/200/300 randomly chosen training samples from Tiny ImageNet dataset, and assigned these noisy images with **random labels** to generate outliers.

Understanding outliers

- Two **metrics** to verify the above hypothesis:
  - $\boldsymbol{I_{avg}^{(m)}}$: measures the **average** m-order interaction.

$$I_{\text{avg}}^{(m)} = \mathbb{E}_{x\in\Omega}[\mathbb{E}_{i,j\in N}[I^{(m)}(i,j|x)]],$$

A **large** $I_{\text{avg}}^{(m)}$ value indicates that the m-order interaction made a **significant contribution** to the classification.

  - $\boldsymbol{P^{(m)}}$: measures the ratio of m-order interactions having **positive** effects among all m-order interactions.

$$P^{(m)} = \frac{\mathbb{E}_{x\in\Omega}\mathbb{E}_{i,j\in N}[\max(I^{(m)}(i,j|x),0)]}{\mathbb{E}_{x\in\Omega}[\mathbb{E}_{i,j}[|I^{(m)}(i,j|x)|]]}.$$

A **large** $P^{(m)}$ value indicates **more** m-order interactions contribute to the classification **positively**, *i.e.* being **more useful**.

Understanding outliers

- In order to verify the hypothesis:
  - ✓ We compare the difference of metrics $\boldsymbol{I_{avg}^{(m)}}$ and $\boldsymbol{P^{(m)}}$ between **outliers and normal samples**, i.e.

$$\Delta \boldsymbol{I_{avg}^{(m)}} = I_{\text{avg}}^{(m,outlier)} - I_{\text{avg}}^{(m,normal)},$$
$$\Delta \boldsymbol{P^{(m)}} = P^{(m,outlier)} - P^{(m,normal)}$$

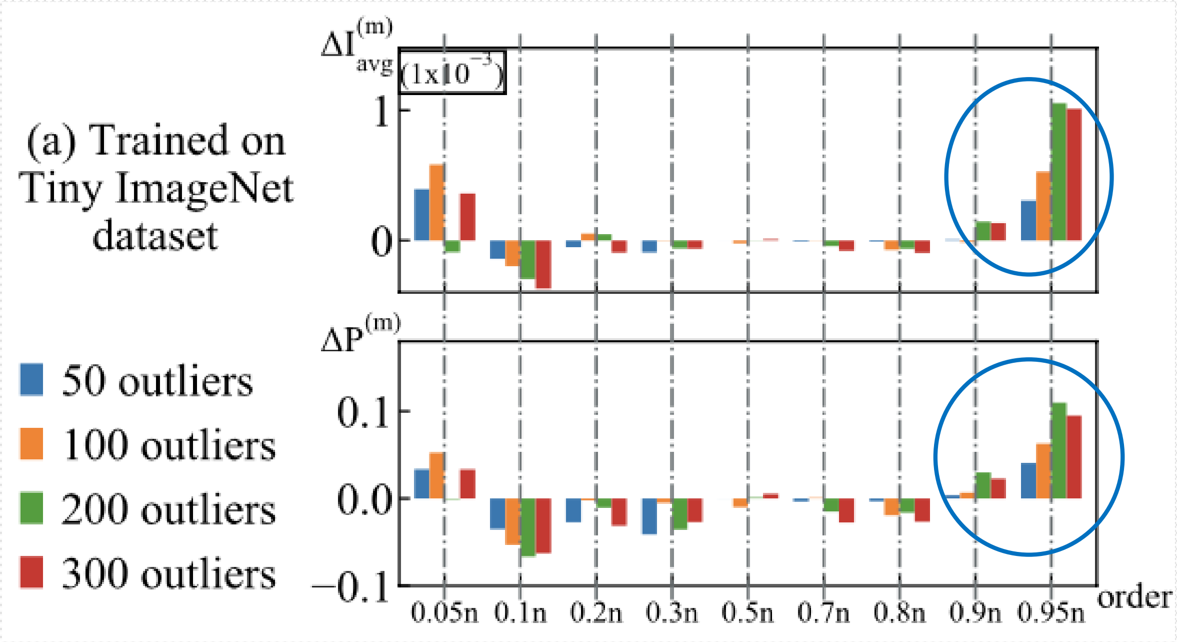
  - ✓ If $\Delta \boldsymbol{I_{avg}^{(m)}} > \mathbf{0}\ \boldsymbol{and}\ \Delta \boldsymbol{P^{(m)}} > \mathbf{0}$ for high order m, then the classification of outliers mainly depends on high-order interactions.

Understanding outliers

- Verification:

  For DNNs trained using datasets contained 50/100/200/300 outliers, $\Delta I_{avg}^{(m)} > 0$ **and** $\Delta P^{(m)} > 0$, when the order m > 0.8n.

Conclusion: Compared to normal samples, the classification of outliers mainly depends on high-order interactions.



Cheng et al, "A Game-Theoretic Taxonomy of Visual Concepts in DNNs" in arXiv:2106.10938, 2021.

Can we learn meaningful features based on interactions?

Wen Shen
Tongji University

How to define interactions in game theory?
How to determine baseline values for the Shapley value?
What is the relationship between interactions and visual concepts?
What is the relationship between interactions and the aesthetic appreciation?

What is the relationship between interactions and the generalization?

What is the relationship between interactions and adversarial transferability?
What is the relationship between interactions and adversarial robustness?

The Link between Interactions and the Network's Generalization Ability

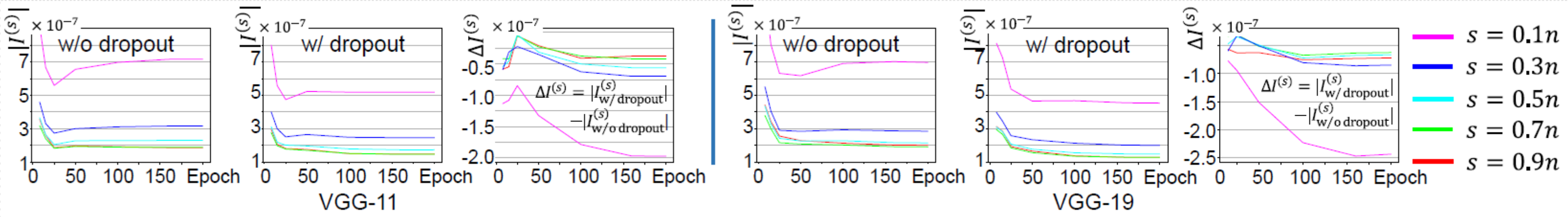
- Theoretically prove that Dropout can decrease the strength of interactions modeled by DNNs

- There is a negative correlation between the strength of interactions and the generalization ability of the network

- The generalization ability of the network can be enhanced by directly controlling the strength of interactions

Overfitting → Strong Interactions

Dropout can **decrease** the **strength of interactions** modeled by DNNs



The relationship between interactions and the generalization ability:

over-fitting → more interactions

Dataset	Model	Ordinary	Over-fitted
MNIST	RN-44	2.17×10^{-3}	$\mathbf{3.64\times10^{-3}}$
Tiny-ImageNet	RN-34	2.57×10^{-3}	$\mathbf{2.89\times10^{-3}}$
CelebA	RN-34	6.46×10^{-3}	$\mathbf{1.17\times10^{-2}}$

Zhang et al. "Interpreting and Boosting Dropout from a Game-Theoretic View" in ICLR, 2021

Suppressing Interactions → Boosting the generalization power

Enhance the generalization ability of the network by directly suppressing the interactions modeled by the network:

$$\text{Loss} = \text{Loss}_{\text{classification}} + \lambda \text{Loss}_{\text{interaction}}$$

$$\text{Loss}_{\text{interaction}} = \mathbb{E}_{i,j\in N, i\neq j}[|I(i,j)|]$$

$$= \mathbb{E}_{i,j\in N, i\neq j}\left[\left|\sum_{S\subseteq N\backslash\{i,j\}} P_{\text{Shapley}}(S|N\backslash\{i,j\})[\Delta f(S,i,j)]\right|\right]$$

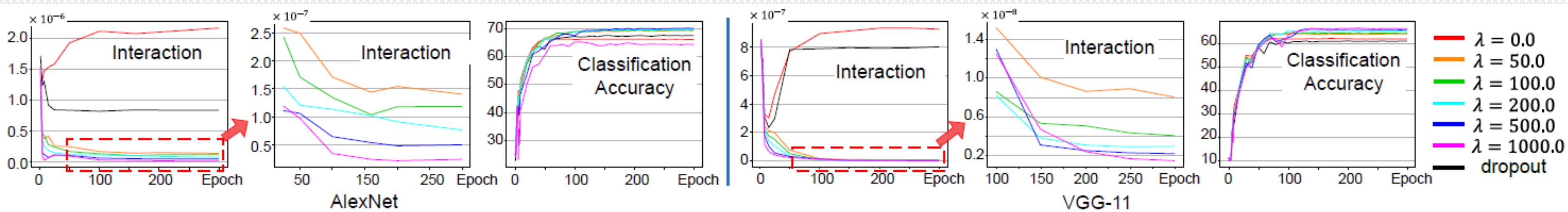
Based on the interactions, we improve the utility of dropout

- **Explicitly control the DNN between over-fitting and under-fitting.**
- **Solve the issue that dropout is not compatible with batch normalization**

Two advantages

Zhang et al. "Interpreting and Boosting Dropout from a Game-Theoretic View" in ICLR, 2021

Suppressing Interactions → Boosting the generalization power



CIFAR-10 dataset

λ	AlexNet[2]	VGG-11[2]	VGG-13[2]	VGG-16[2]
0.0	66.2	61.9	60.8	62.0
50.0	69.2	63.9	64.0	63.8
100.0	69.6	64.3	65.4	64.5
200.0	69.6	65.3	65.9	64.7
500.0	**70.0**	65.9	**66.2**	**64.9**
1000.0	64.3	**66.3**	66.0	64.5
Dropout	67.5	60.9	60.9	63.0

Tiny ImageNet

λ	RN-18[2]	RN-34[2]	λ	VGG-16	VGG-19
0.0	48.8	45.6	0.0	33.4	37.6
0.001	50.0	48.4	50.0	38.4	38.2
0.003	49.6	49.0	100.0	38.0	38.6
0.01	**52.2**	**49.6**	200.0	38.2	39.0
0.03	50.4	48.8	500.0	**42.8**	41.8
			1000.0	40.8	**45.2**
Dropout	47.4	46.0	Dropout	36.8	32.6

Gender estimation

λ	VGG-13	VGG-16	λ	RN-18
0.0	94.6	93.7	0.0	92.7
5.0	94.8	93.8	0.001	93.0
10.0	94.7	**94.6**	0.003	**93.1**
20.0	**94.9**	94.1	0.01	93.0
50.0	94.7	94.08	0.03	92.9
100.0	94.7	94.3		
Dropout	94.6	92.4	Dropout	92.1

Over-fitting ↑ / Under-fitting ↓

Zhang et al. "Interpreting and Boosting Dropout from a Game-Theoretic View" in ICLR, 2021

How to define interactions in game theory?
How to determine baseline values for the Shapley value?
What is the relationship between interactions and visual concepts?
What is the relationship between interactions and the aesthetic appreciation?
What is the relationship between interactions and the generalization?

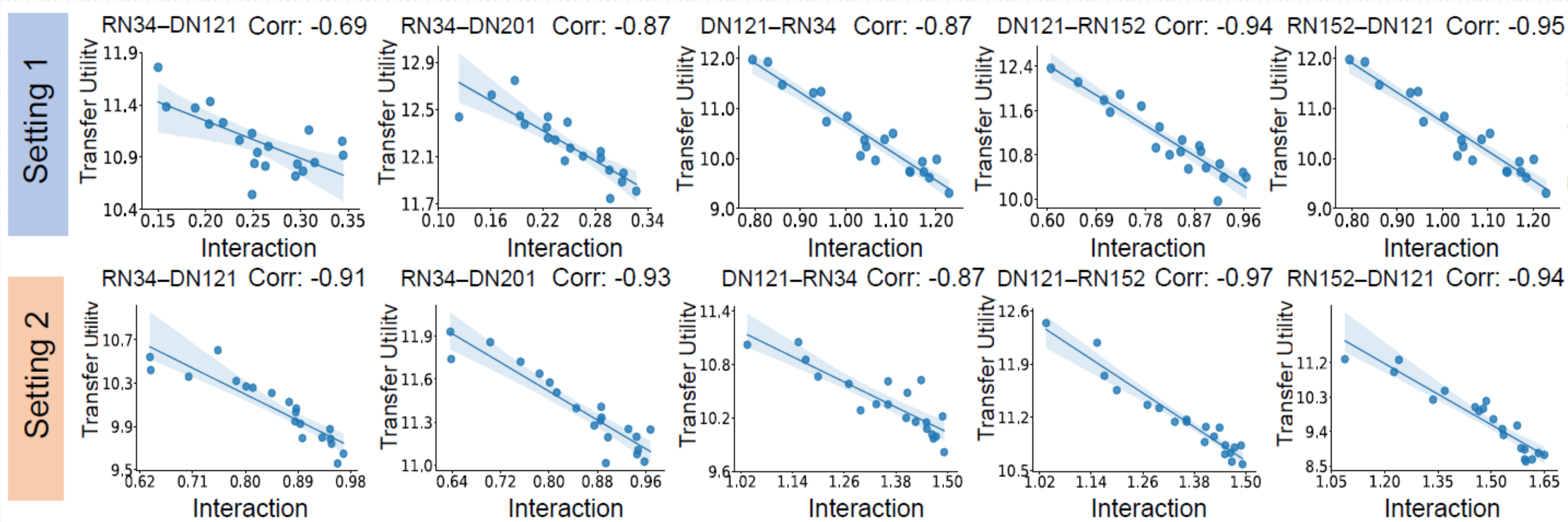
What is the relationship between interactions and adversarial transferability?

What is the relationship between interactions and adversarial robustness?

The negative correlation between the interaction and the adversarial transferability

- Theoretical foundations： Multi-step attacks vs. Single-step attacks
  - Interaction： Multi-step attacks > Single-step attacks
  - Overfitting： Multi-step attacks > Single-step attacks[1]

- Empirical verification：



[1] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2730–2739, 2019.

去芜存菁 Common essence: the reduction of interactions is the common mechanism of previous transferability-boosting methods

- Existing transferability-boosting methods can be approximately explained as the reduction of interactions.
  - Theoretically prove the attack based on momentum (MI Attack)[2]
  - Theoretically prove the attack based on smooth of gradients (VR Attack)[3]
  - Theoretically prove the attack based on skip connections (SGM Attack)[4]
  - Empirically verify the attack based on Translation-invariant (TI Attack)[5]
  - Empirically verify the attack based on Input diversity (DI Attack)[6]

Proposition 1 The adversarial perturbation generated by the multi-step attack is given as $\delta^m_{multi} = \alpha \sum_{t=0}^{m-1} \nabla_x l(h(x + \delta^t_{multi}), y)$, where δ^t_{multi} denotes the perturbation after the t-th step of updating, and m is referred to as the total number of steps. The adversarial perturbation generated by the single-step attack is given as $\delta_{single} = \alpha m \nabla_x l(h(x), y)$. Then, the expectation of interactions between perturbation units in δ^m_{multi}, $\mathbb{E}_{a,b}[I_{ab}(\delta^m_{multi})]$, is larger than $\mathbb{E}_{a,b}[I_{ab}(\delta_{single})]$, i.e. $\mathbb{E}_{a,b}[I_{ab}(\delta^m_{multi})] \geq \mathbb{E}_{a,b}[I_{ab}(\delta_{single})]$.

Proposition 2 The adversarial perturbation generated by the multi-step attack is given as $\delta^m_{multi} = \alpha \sum_{t=0}^{m-1} \nabla_x l(h(x + \delta^t_{multi}), y)$. The adversarial perturbation generated by the VR Attack is computed as $\delta^m_{vr} = \alpha \sum_{t=0}^{m-1} \nabla_x \hat{l}(h(x + \delta^t_{vr}), y)$, where $\hat{l}(h(x +$

Proposition 3 The adversarial perturbation generated by the multi-step attack is given as $\delta^m_{multi} = \alpha \sum_{t=0}^{m-1} \nabla_x l(h(x + \delta^t_{multi}), y)$. The adversarial perturbation generated by the multi-step attack incorporating the momentum is computed as $\delta^m_{mi} = \alpha \sum_{t=0}^{m-1} g^t_{mi}$. Perturbation units of δ^m_{mi} tend to exhibit smaller interactions than δ^m_{multi}, i.e. $\mathbb{E}_x \mathbb{E}_{a,b}[I_{ab}(\delta^m_{mi})] < \mathbb{E}_x \mathbb{E}_{a,b}[I_{ab}(\delta^m_{multi})]$.

[2] Yinpeng Dong, Fangzhou Liao, and et al. Boosting adversarial attacks with momentum. In CVPR, 2018.
[3] Lei Wu, Zhanxing Zhu, and Cheng Tai. Understanding and enhancing the transferability of adversarial examples. arXiv preprint arXiv:1802.09707, 2018.
[4] Dongxian Wu, Yisen Wang, and et al. Skip connections matter: On the transferability of adversarial examples generated with resnets. In ICLR, 2020.
[5] Yinpeng Dong, Tianyu Pang, and et al. Evading defenses to transferable adversarial examples by translation-invariant attacks. In CVPR, 2019.
[6] Cihang Xie, Zhishuai Zhang, and et al. Improving transferability of adversarial examples with input diversity. In CVPR, 2019.

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. In arXiv:2010.04055, 2020

Application： Penalizing interactions to improve adversarial transferability

- With the additional interaction-reduction loss, the PGD attack improves more than 10% adversarial transferability.

- Combining existing methods with the interaction-reduction loss, the adversarial transferability is improved from 54.6%-98.8% to 70.2%-99.1%

Source	Method	VGG-16	RN152	DN-201	SE-154	IncV3	IncV4	IncResV2
RN-34	MI	80.1±0.5	73.0±2.3	77.7±0.5	48.9±0.8	46.2±1.2	39.9±0.5	34.8±2.5
	VR	88.8±0.2	86.4±1.6	87.9±2.4	62.1±1.5	58.4±3.0	56.3±2.3	49.7±0.9
	SGM	91.8±0.6	89.0±0.9	90.0±0.4	68.0±1.4	63.9±0.3	58.2±1.1	54.6±1.2
	SGM+IR	94.7±0.6	91.7±0.6	93.4±0.8	72.7±0.4	68.9±0.9	64.1±1.3	61.3±1.0
	HybridIR	**96.5±0.1**	**94.9±0.3**	**95.6±0.6**	**79.7±1.0**	**77.1±0.8**	**73.8±0.1**	**70.2±0.5**
RN-152	MI	70.3±0.6	–	74.8±1.4	51.7±0.8	47.1±0.9	40.5±1.6	36.8±2.7
	VR	83.9±3.4	–	91.1±0.9	70.0±3.7	63.1±0.9	58.8±0.1	56.2±1.3
	SGM	88.2±0.5	–	90.2±0.3	72.7±1.4	63.2±0.7	59.1±1.5	58.1±1.2
	SGM+IR	92.0±1.0	–	92.5±0.4	79.3±0.1	69.6±0.8	66.2±1.0	63.6±0.9
	HybridIR	**95.3±0.4**	–	**96.9±0.2**	**84.7±0.7**	**80.0±1.2**	**77.5±0.8**	**75.6±0.6**
DN-121	MI	83.0±4.9	72.0±0.7	91.5±0.2	58.4±2.6	54.6±1.6	49.2±2.4	43.9±1.5
	VR	91.5±0.5	88.7±0.5	98.8±0.2	75.1±1.3	74.3±1.7	75.6±3.0	69.8±1.3
	SGM	88.7±0.9	88.1±1.0	98.0±0.4	78.0±0.9	64.7±2.5	65.4±2.3	59.7±1.7
	SGM+IR	91.7±0.2	90.4±0.4	94.3±0.1	87.0±0.4	78.8±1.3	79.5±0.2	75.8±2.7
	HybridIR	**96.9±0.4**	**96.8±0.4**	**99.1±0.4**	**90.9±0.5**	**88.4±0.8**	**87.8±0.8**	**87.1±0.4**
DN-201	MI	77.3±0.8	74.8±1.4	–	64.6±1.0	56.5±2.5	51.1±2.1	47.8±1.9
	VR	87.3±1.1	90.4±1.2	–	78.0±1.5	75.8±2.1	75.8±1.3	71.3±1.2
	SGM	87.3±0.3	92.4±1.0	–	82.9±0.2	72.3±0.3	71.3±0.6	68.8±0.5
	SGM+IR	89.5±0.9	91.8±0.7	–	87.3±1.2	82.5±0.8	80.3±0.3	81.5±0.5
	HybridIR	**94.4±0.1**	**96.9±0.5**	–	**91.7±0.2**	**89.6±0.6**	**88.3±0.3**	**87.3±0.7**

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. In arXiv:2010.04055, 2020

How to define interactions in game theory?
How to determine baseline values for the Shapley value?
What is the relationship between interactions and visual concepts?
What is the relationship between interactions and the aesthetic appreciation?
What is the relationship between interactions and the generalization?
What is the relationship between interactions and adversarial transferability?

What is the relationship between interactions and adversarial robustness?

Previous explanations of adversarial robustness

- Previous explanations lack an essential and unified explanation.

What is the essence of adversarial attacks and defense?

- Explaining **adversarial examples**
  - Linearity of feature representations
  - Non-robust but discriminative features
- Explaining **adversarial training**
  - Learning general shapes of objects
  - Enumeration of all possible adversarial examples

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

- Explaining **adversarial robustness**
  - Proving the theoretical bound

[1] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. ICLR, 2016.
[2] Lei Wu, Zhanxing Zhu, and Cheng Tai. Understanding and enhancing the transferability of adversarial examples. arXiv preprint arXiv:1802.09707, 2018.
[3] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In 28th USENIX Security Symposium USENIX Security, pp. 321–338, 2019.

Contributions of this paper

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

- The adversarial **training** boosts the robustness of DNNs by learning more discriminative low-order interactions.

- We propose a **unified explanation** for several adversarial defense methods.

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. in ICLR 2021

Contributions of this paper

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

- The adversarial **training** boosts the robustness of DNNs by learning more discriminative low-order interactions.

- We propose a **unified explanation** for several adversarial defense methods.

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. in ICLR 2021

Adversarial attacks mainly affect high-order interactions

Given an normal sample x, let $\tilde{x} = x + \delta$ denotes its adversarial example.

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

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

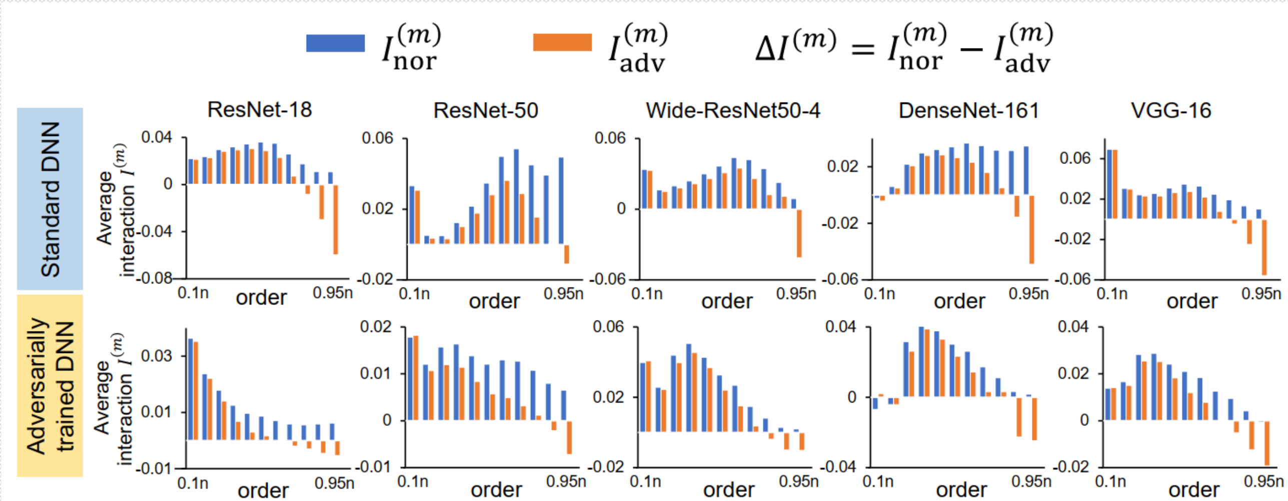
$$\Delta J_{ij}^{(m)} = \frac{n-1-m}{n(n-1)} \Delta I_{ij}^{(m)}$$

Small and can be ignored

$$\Delta I_{ij}^{(m)} = I_{ij}^{(m)}(x) - I_{ij}^{(m)}(\tilde{x})$$

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. in ICLR 2021

Adversarial attacks mainly affect high-order interactions



(a) Untargeted attack

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.

Contributions of this paper

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

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. in ICLR 2021

Adversarial training boosts the robustness of high-order interactions

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

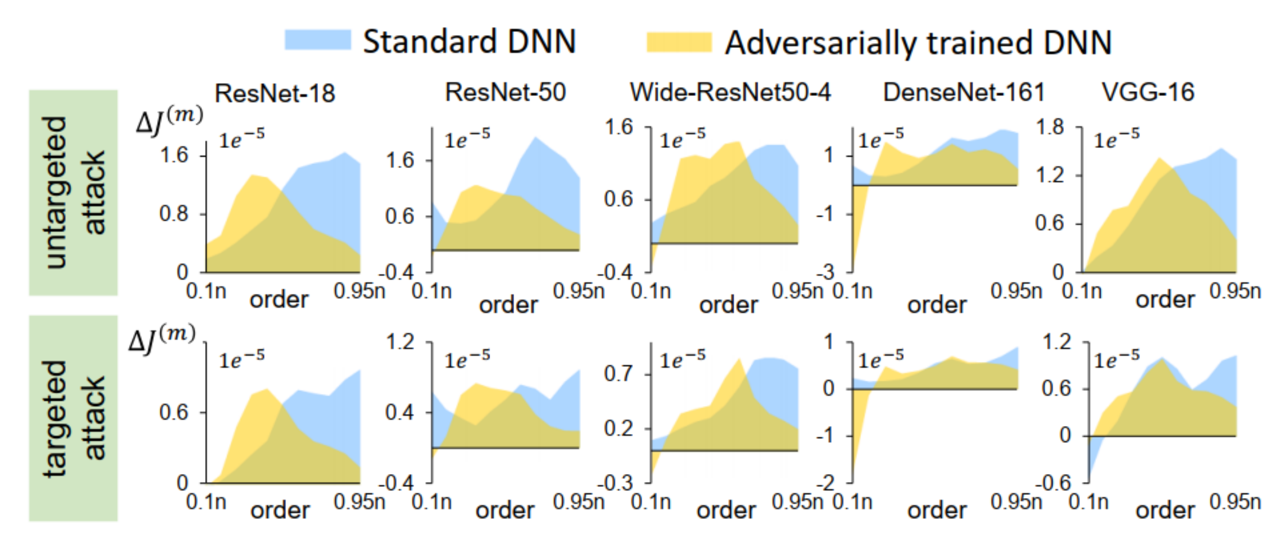


Figure: Distribution of compositional attacking utilities caused by interactions of different orders in standard DNNs and adversarially trained DNNs.

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

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. in ICLR 2021

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

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

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

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

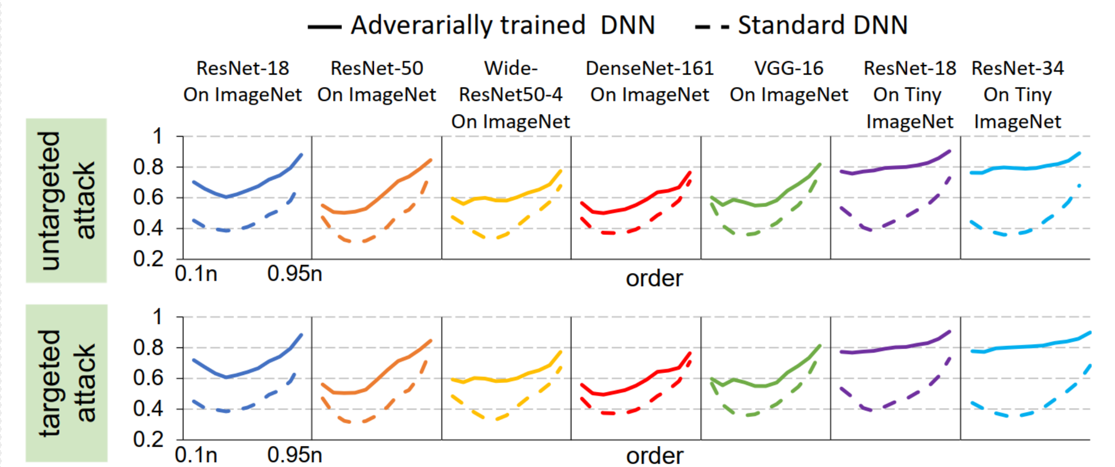


Figure: The interaction disentanglement.

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. in ICLR 2021

Contributions of this paper

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

- The adversarial **training** boosts the robustness of DNNs by learning more discriminative low-order interactions.

- We propose a **unified explanation** for several adversarial defense methods.

Wang et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. in ICLR 2021

去芜存菁 The unified explanation for previous adversarial defenses

- Attribution-based method for detecting adversarial examples: ML-LOO [1]
- Rank-based method for detecting adversarial examples [2]

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

- Cutout method [3]
- High recoverability of adversarial examples in adversarially trained DNNs

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

[1] 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.
[2] Malhar Jere, Maghav Kumar, and Farinaz Koushanfar. A singular value perspective on model robustness. arXiv preprint arXiv:2012.03516, 2020.
[3] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

Thank you