IJCAI 2020 Tutorial on Trustworthiness of Interpretable Machine Learning

Efforts in Pushing XAI Towards Science

Quanshi Zhang
Shanghai Jiao Tong University

Zhou et al. "Interpreting Deep Visual Representations via Network Dissection" in IEEE Trans. on PAMI 2018
Kim et al. "Sanity Checks for Saliency Maps" in NIPS 2018
Lapuschkin et al. “unmasking clever hans predictors and assessing what machines really learn“ in Nat Commun 10 1096, 2019
Lundberg et al., "A unified approach to interpreting model predictions" in NeurIPS 2017
Fong et al. "Net2Vec: Quantifying and Explaining how Concepts are encoded by filters in deep neural networks" in CVPR 2018
Dhamdhere et al., "The Shapley Taylor Interaction Index" in arXiv:1902.05622, 2019
Lloyd S Shapley, "A value for n-person games" in contributions to the Theory of Games 2.28 (1953), pp. 307–317.
Pitas et al. "Pac-bayesian margin bounds for convolutional neural networks" in arXiv:1801.00171, 2018
Zhang et al. "Examining CNN Representations with respect to Dataset Bias" in AAAI 2018
Zhang et al. "Interpreting Multivariate Shapley Interactions in DNNs" in AAAI 2021
Zhang et al. "Extracting an Explanatory Graph to Interpret a CNN" in IEEE Trans. on PAMI 2020
Zhang et al. "Interpretable CNNs for Object Classification" in IEEE Trans. on PAMI, 2020
Ma et al. "Quantifying Layerwise Information Discarding of Neural Networks" in arXiv:1906.04109, 2019
Guan et al. "Towards A Deep and Unified Understanding of Deep Neural Models in NLP" in ICML 2019
Cheng et al. "Explaining Knowledge Distillation by Quantifying the Knowledge" in CVPR 2020
Liang et al. "Knowledge Consistency between Neural Networks and Beyond" in ICLR 2020
Ren et al. "Interpreting and Disentangling Feature Components of Various Complexity from DNNs" in arXiv:2006.15920, 2020
Ren et al. "Towards Theoretical Analysis of Transformation Complexity of ReLU DNNs" in arXiv
Zhang et al. "Interpreting and Boosting Dropout from a Game-Theoretic View" in arXiv:2009.11729, 2020
Zhang et al. "A Unified Approach to Interpreting and Boosting Adversarial Transferability" in arXiv:2010.04055, 2020

Outline

- XAI studies and vision of XAI science
- Explanation based on strict and fine-grained concepts
- Quantification of the representation power of a DNN
- Proof of mathematic essence of existing DL methods

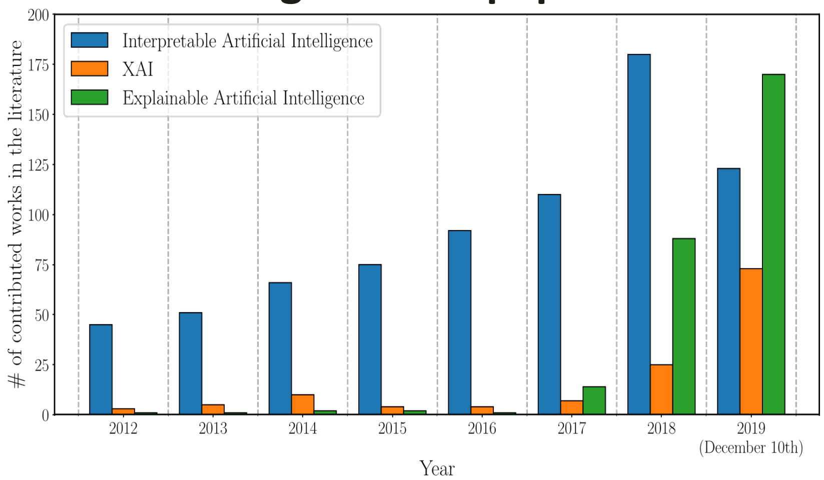
Outline

- **XAI studies and vision of XAI science**
- Explanation based on strict and fine-grained concepts
- Quantification of the representation power of a DNN
- Proof of mathematic essence of existing DL methods

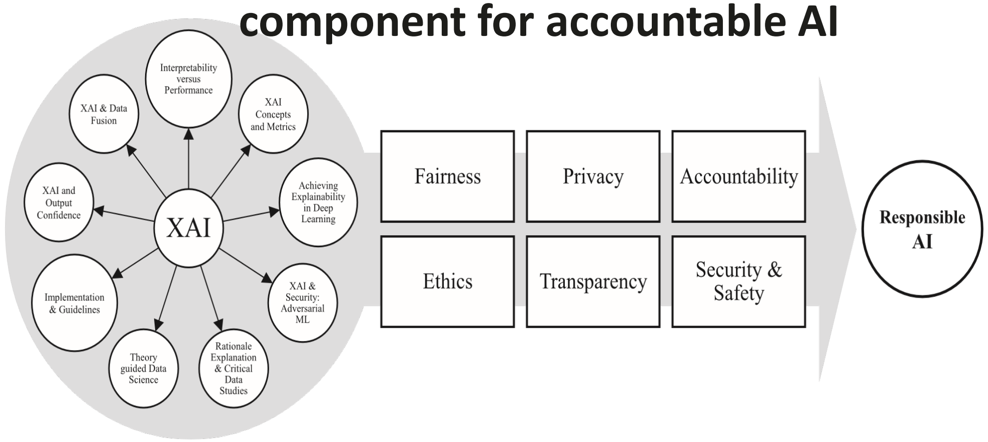
Why XAI is important ?

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

The growth of papers in XAI



Interpretability is a necessary component for accountable AI



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

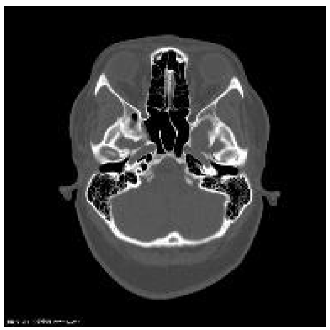
Mathematical explanation

- 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

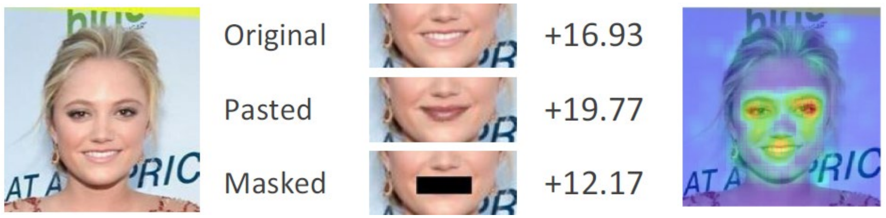
Semantic explanation

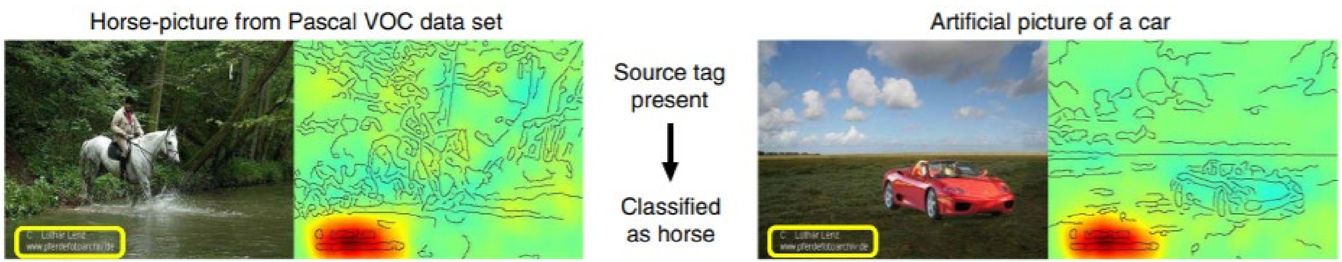
- Which semantic concepts are modeled and used for prediction
- How to quantify and improve the trustworthiness of a DNN
- End-to-end learn interpretable features
- Communicative learning at the semantic level
- How to evaluate the explanation

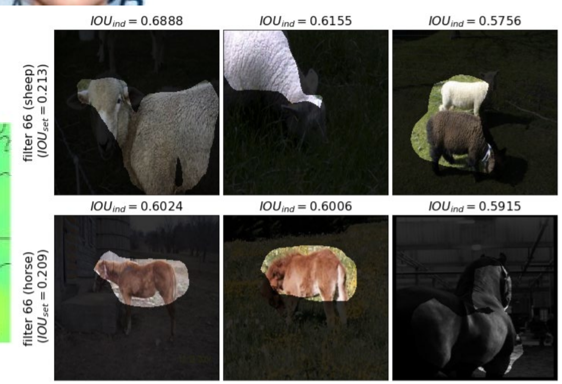


Make a surgery. Score=0.9
It is because
1) From Organ A. Score=0.2
2) From Organ B. Score=0.1
…

Score of lipstick





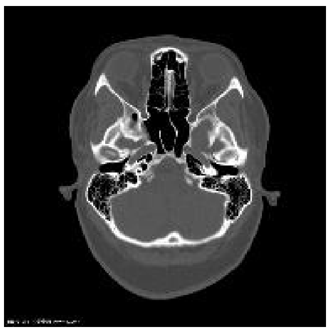


Lapuschkin et al. “unmasking clever hans predictors and assessing what machines really learn“ in Nat Commun 10 1096, 2019
Fong et al. "Net2Vec: Quantifying and Explaining how Concepts are encoded by filters in deep neural networks" in CVPR 2018
Zhang et al. “Examining CNN Representations with respect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

XAI Topics

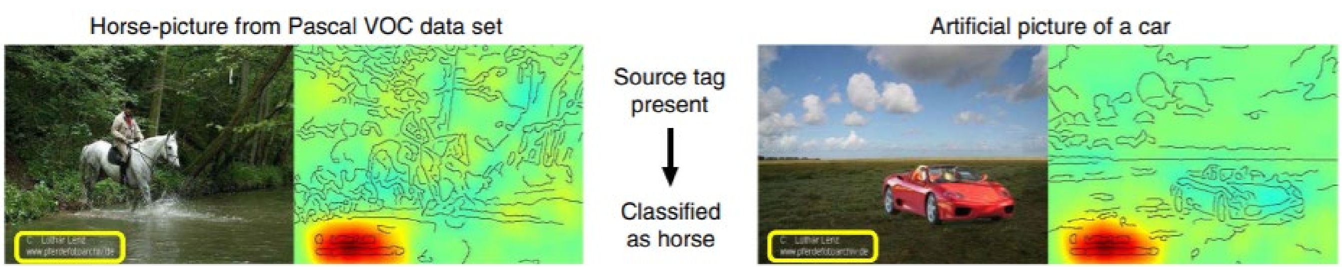
Semantic explanation

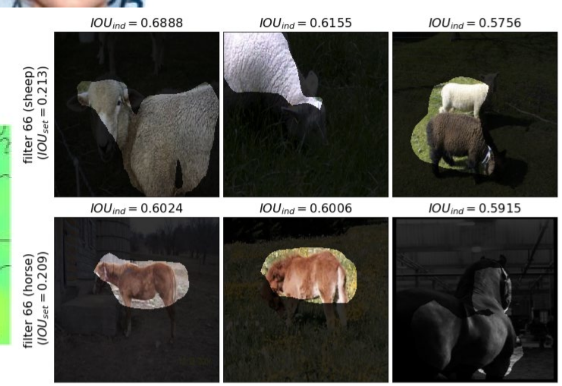
Which semantic concepts are modeled and used for prediction	How to quantify and improve the trustworthiness of a DNN	End-to-end learn interpretable features	Communicative learning at the semantic level	How to evaluate the explanation



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

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

Mathematical explanation

- 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?

- How to evaluate the generalization power of a DNN?
- Why does a specific DNN architecture perform better than 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 of semantic explanations

Many semantic explanations are still heuristic technologies, rather than science

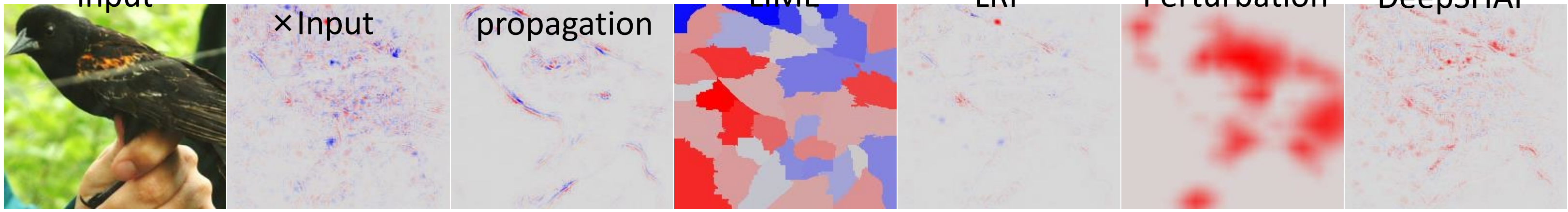
Only self-consistency, no mutuality between XAI methods

Very few theoretic foundations

Difficult to improve DNNs

Lack of convincing enough evaluation metrics

Explanation results conflict with each other.



Many existing attribution-based explanations seem like edge detection

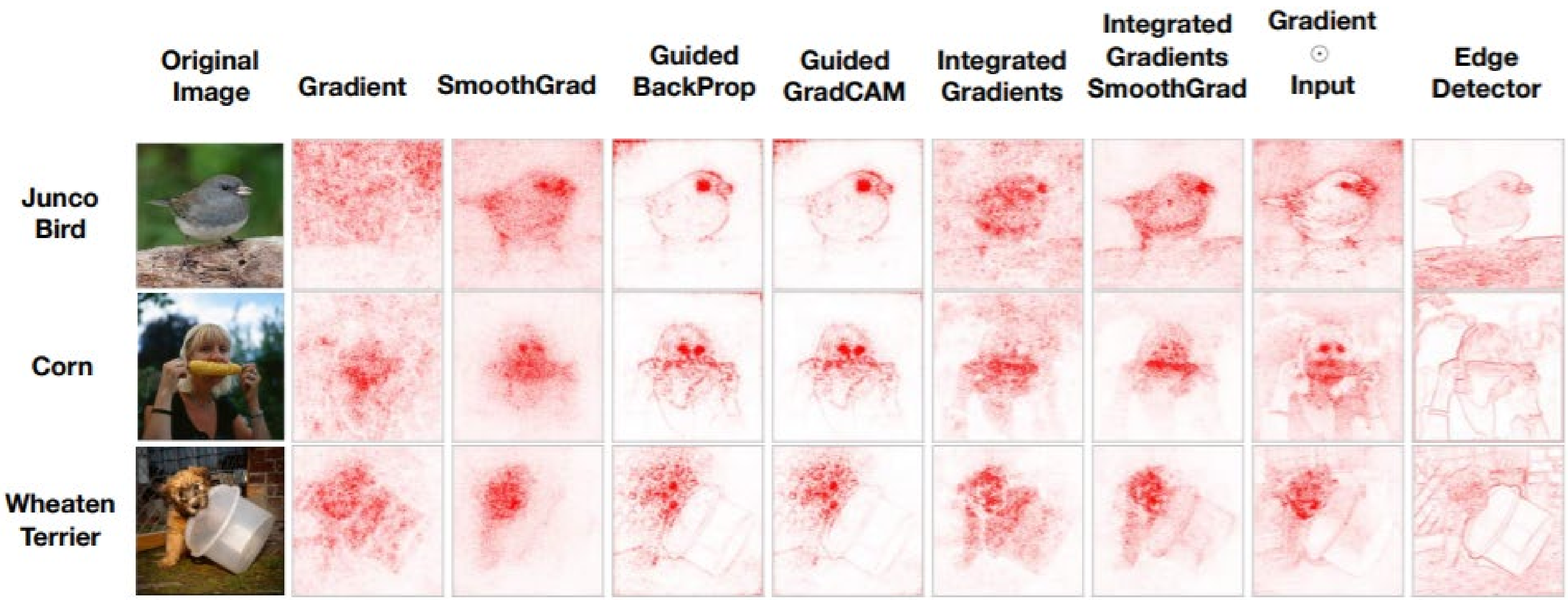


Figure 1: **Saliency maps for some common methods compared to an edge detector.** Saliency

Kim et al. "Sanity Checks for Saliency Maps" in NIPS 2018

Problems of 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

Outline

- XAI studies and vision of XAI science
- **Explanation based on strict and fine-grained concepts**
- Quantification of the representation power of a DNN
- Proof of mathematic essence of existing DL methods

Strict and fine-grained explanations

- XAI studies and vision of XAI science
- **Explanation based on strict and fine-grained concepts**
  - Strictness
    - Shapley values
    - Game-theoretic interactions
  - Fine-grained
    - Explanatory graph
    - Interpretable filters
- Quantification of the representation power of a DNN
- Proof of mathematic essence of existing DL methods

Strict and fine-grained explanations

- XAI studies and vision of XAI science
- **Explanation based on strict and fine-grained concepts**
  - Strictness
    - **Shapley values**
    - Game-theoretic interactions
  - Fine-grained
    - Explanatory graph
    - Interpretable filters
- Quantification of the representation power of a DNN
- Proof of mathematic essence of existing DL methods

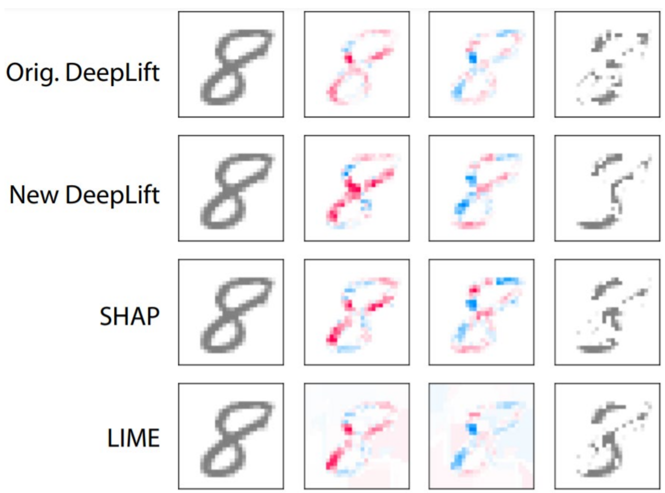
Strict attributions: Shapley values

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

- 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.

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



$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i',$$

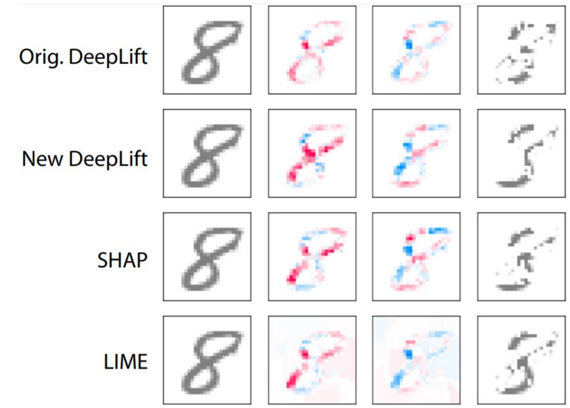
Lloyd S Shapley. "A value for n-person games". In: Contributions to the Theory of Games 2.28 (1953), pp. 307–317.
Scott M. Lundberg, and Su-In Lee, "A unified approach to interpreting model predictions" in NeurIPS 2017

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



☐ Remaining issues

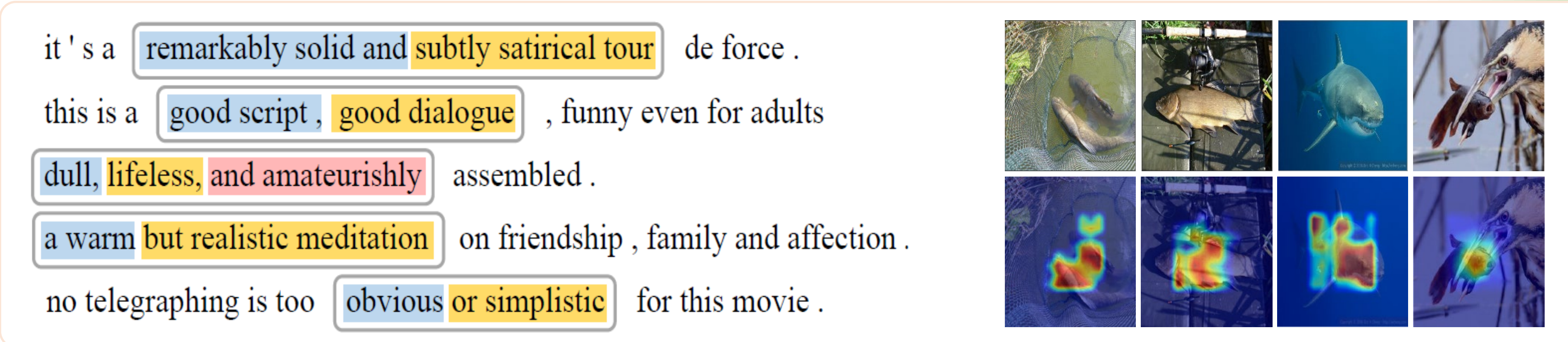
- How to determine reasonable reference values?
- How to determine the reasonable partition of players?

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

Strict and fine-grained explanations

- XAI studies and vision of XAI science
- **Explanation based on strict and fine-grained concepts**
  - Strictness
    - Shapley values
    - **Game-theoretic interactions**
  - Fine-grained
    - Explanatory graph
    - Interpretable filters
- Quantification of the representation power of a DNN
- Proof of mathematic essence of existing DL methods

Game-theoretic interactions



- **The input words of a sentence (or the input pixels of an image) into a DNN 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 (a,b):** the change of the importance (Shapley value) of a when b is present, w.r.t. the importance (Shapley value) when b is absent.
- Each word/pixel can be considered as player.

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

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

Multi-order interactions

The interaction of the m-th order: the interaction two players considering collaborations with m contextual players

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

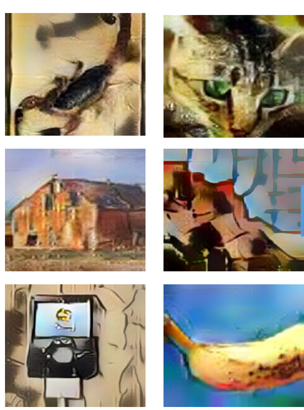
- Marginal contribution property
  $\forall i,j \in N, i \neq j,\ \phi^{(m+1)}(i|N) - \phi^{(m)}(i|N) = \mathbb{E}_{j\in N\setminus\{i\}}[I^{(m)}(i,j)]$
- Accumulation property
  $\phi^{(m)}(i|N) = \mathbb{E}_{j\in N\setminus\{i\}}[\sum_{k=0}^{m-1} I^{(k)}(i,j)] + \phi^{(0)}(i|N)$
- Efficiency property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

Connections between interactions & visual concepts

$I^{(m)}(i,j)$

- **Small m： Low-order interactions**
  - Simple features, such as edges, colors
- **Middle m： Middle-order interactions**
  - Complex features, such as complex structure
- **Large m： High-order interactions**
  - Global textures, outliers, noises

Style-transferred images



Multivariate interactions

The Link between Interactions and the Network's Semantic Representation ——explain the abnormal behavior of the network

- 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

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

Shapley Taylor Interaction Index

- Dhamdhere and Sundararajan defined a new type of interactions between multiple variables.

1. **Linearity axiom:** $\mathcal{I}^k(\cdot)$ is a linear function; i.e. for two functions $F_1, F_2 \in \mathcal{G}^N$, $\mathcal{I}^k_S(F_1 + F_2) = \mathcal{I}^k_S(F_1) + \mathcal{I}^k_S(F_2)$ and $\mathcal{I}^k_S(c \cdot F_1) = c \cdot \mathcal{I}^k_S(F_1)$.
2. **Dummy axiom:** If i is a dummy feature for F, i.e. $F(S) = F(S \setminus i) + F(i)$ for any $S \subseteq N$ with $i \in S$, then
   (i) $\mathcal{I}^k_i(F) = F(i)$
   (ii) for every $S \subseteq N$ with $i \in S$, we have $\mathcal{I}^k_S(F) = 0$
3. **Symmetry axiom:** for all functions $F \in \mathcal{G}^N$, for all permutations π on N, :

$$\mathcal{I}^k_S(F) = \mathcal{I}^k_{\pi S}(\pi F)$$

where $\pi S := \{\pi(i) | i \in S\}$ and the function πv is defined by $(\pi F)(\pi S) = F(S)$, i.e. it arises from relabeling of features $1, \ldots, n$ with the labels $\pi(1), \ldots, \pi(n)$.

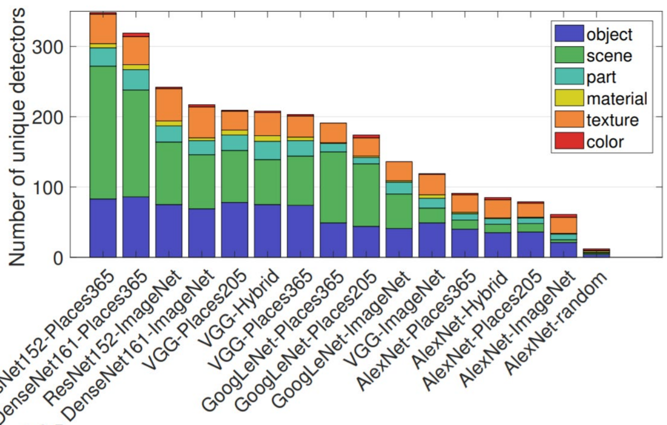
Dhamdhere et al., "The Shapley Taylor Interaction Index" in arXiv:1902.05622, 2019

Strict and fine-grained explanations

- XAI studies and vision of XAI science
- **Explanation based on strict and fine-grained concepts**
  - Strictness
    - Shapley values
    - Game-theoretic interactions
  - Fine-grained
    - **Explanatory graph**
    - Interpretable filters
- Quantification of the representation power of a DNN
- Proof of mathematic essence of existing DL methods

Semantics in intermediate layers

Distribution of various semantics encoded in convolutional layers

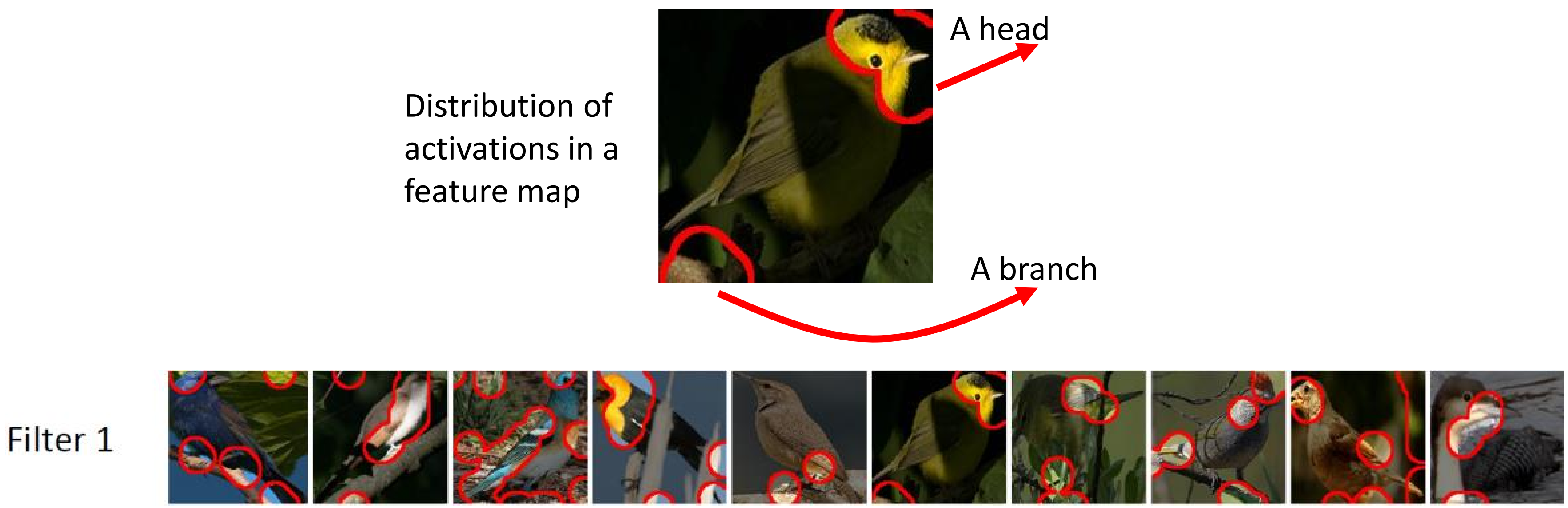


Visualization of semantic meanings of convolutional filters

Zhou et al. “Interpreting Deep Visual Representations via Network Dissection” in IEEE Trans. on PAMI 2018

Representing CNNs as an explanatory graph

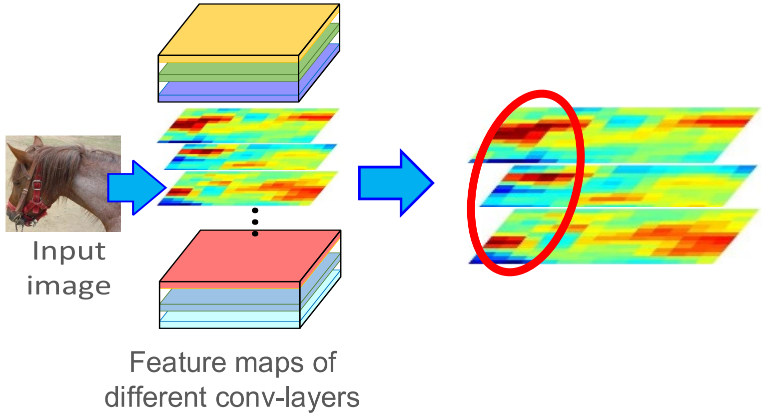
- Given a CNN that is pre-trained for object classification
  - How many types of patterns (visual concepts) are memorized by a convolutional filter of the CNN?



Quanshi Zhang et al. "Extracting an Explanatory Graph to Interpret a CNN" in IEEE Trans. on PAMI 2020

Representing CNNs as an explanatory graph (2)

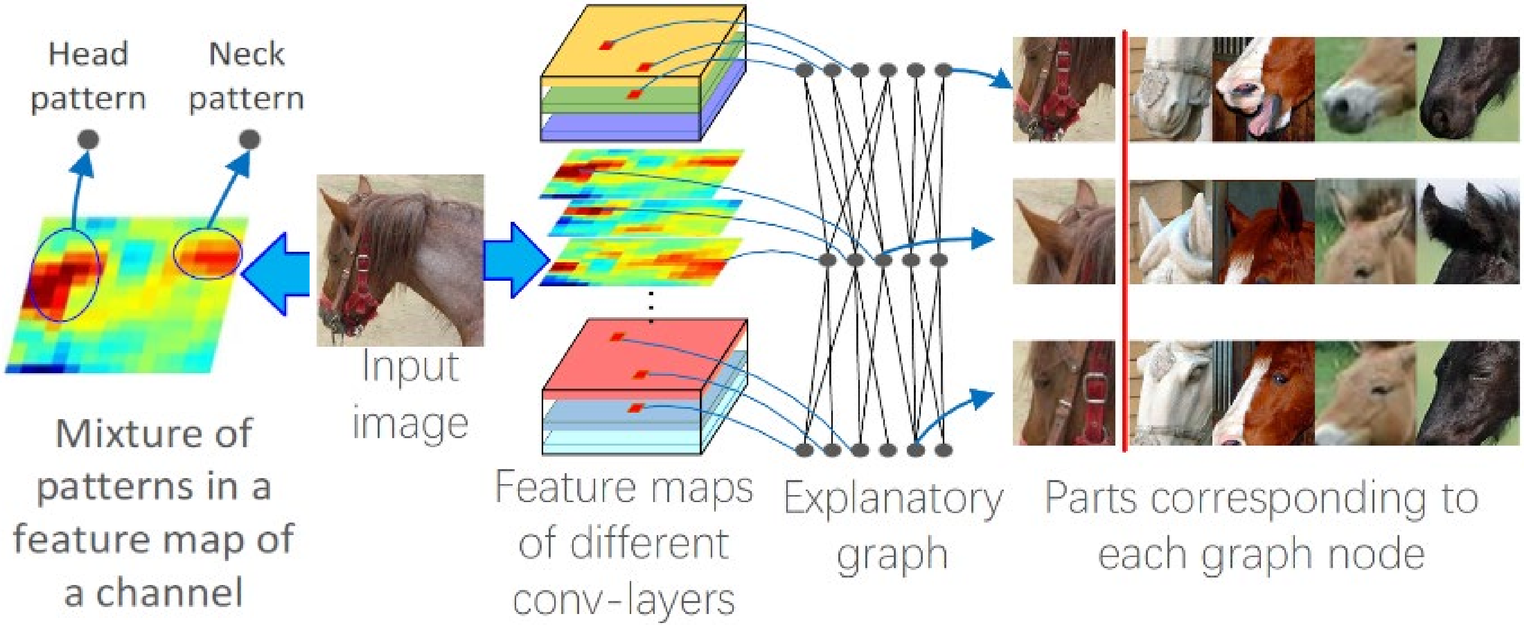
- How many types of patterns (visual concepts) are memorized by a convolutional filter of the CNN?
- Which concepts are co-activated to describe a part?
- What is the spatial relationship between two patterns?



These filters are co-activated in certain area to represent the head of a horse.

Quanshi Zhang et al. “Extracting an Explanatory Graph to Interpret a CNN” in IEEE Trans. on PAMI 2020

Explanatory graph for a CNN

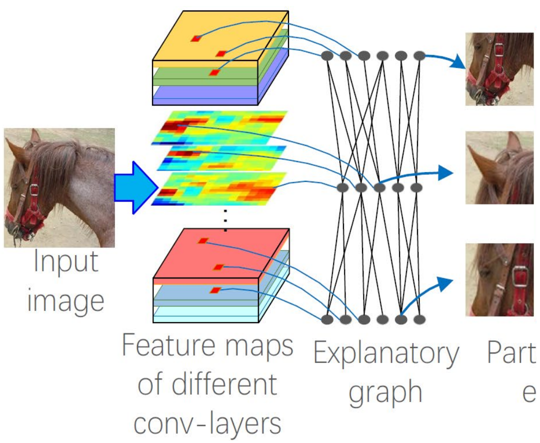


- The graph has multiple layers → multiple conv-layers of the CNN
- Each node → a pattern of an object part
- A filter may encode multiple patterns (nodes)→ disentangle a mixture of patterns from the feature map of a filter
- Each edge → co-activation relationships and spatial relationships between two patterns

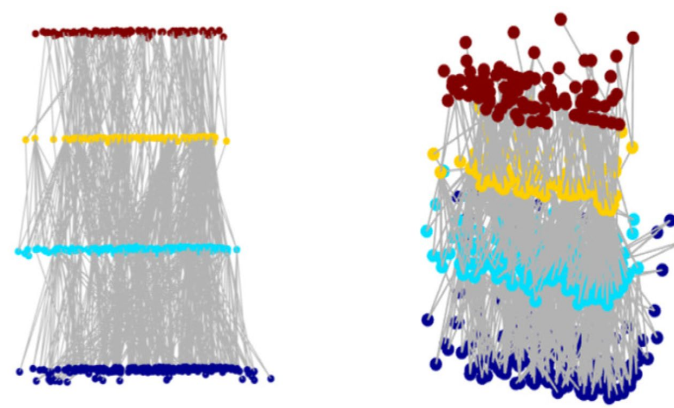
Quanshi Zhang et al. “Extracting an Explanatory Graph to Interpret a CNN” in IEEE Trans. on PAMI 2020

Task

- Input: a pre-trained CNN
  - Trained for classification, segmentation, or ...
  - AlexNet, VGG-16, ResNet-50, ResNet-152, and etc.
  - **Without any annotations of parts or textures**
- Output: an explanatory graph



Explanatory graph for four conv-layers of a VGG-16 network



For clarity, we only show 10% of the patterns

Quanshi Zhang et al. “Extracting an Explanatory Graph to Interpret a CNN” in IEEE Trans. on PAMI 2020

Disentangling object parts from raw filters

Nodes in the explanatory graph

Raw filters in the CNN

Quanshi Zhang et al. “Extracting an Explanatory Graph to Interpret a CNN” in IEEE Trans. on PAMI 2020

Strict and fine-grained explanations

- XAI studies and vision of XAI science
- **Explanation based on strict and fine-grained concepts**
  - Strictness
    - Shapley values
    - Game-theoretic interactions
  - Fine-grained
    - Explanatory graph
    - **Interpretable filters**
- Quantification of the representation power of a DNN
- Proof of mathematic essence of existing DL methods

Background

Feature maps of Filter 1

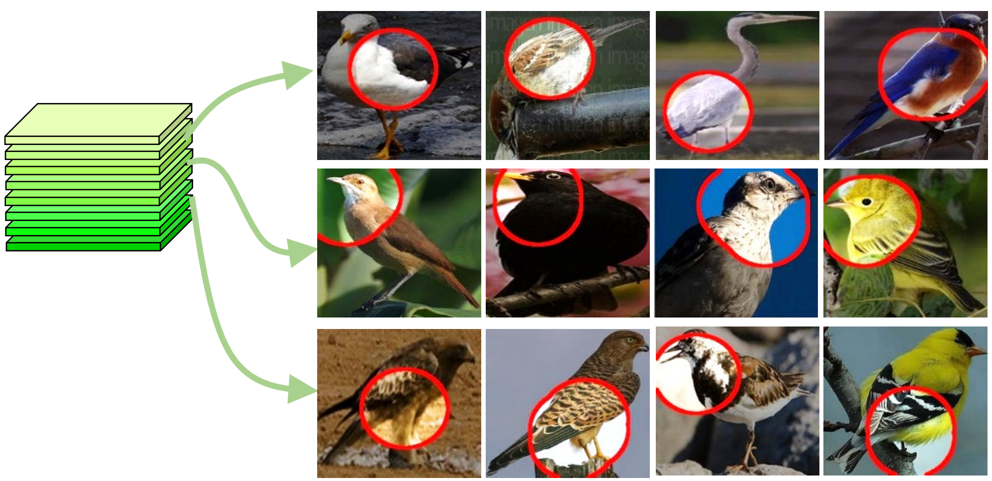
Feature maps of Filter 2

Feature maps of Filter 3

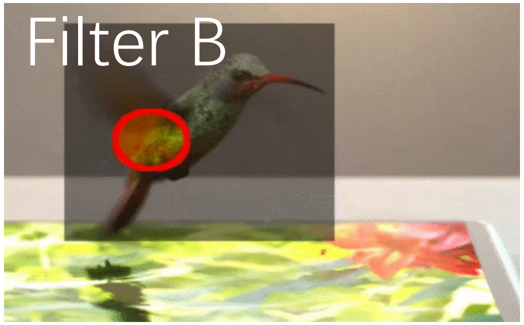
Quanshi Zhang et al. “Interpretable CNNs for Object Classification” in IEEE Trans. on PAMI, 2020

Objective

Without additional part annotations, learn a CNN, where each filter represents a specific part through different objects.



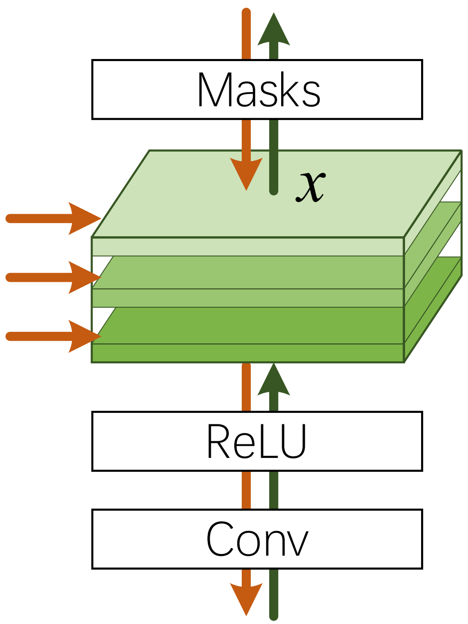
Neural activations of 3 interpretable filters



Quanshi Zhang et al. "Interpretable CNNs for Object Classification" in IEEE Trans. on PAMI, 2020

Force filters to represent object parts without part annotations

We add a loss to each channel to construct an interpretable layer



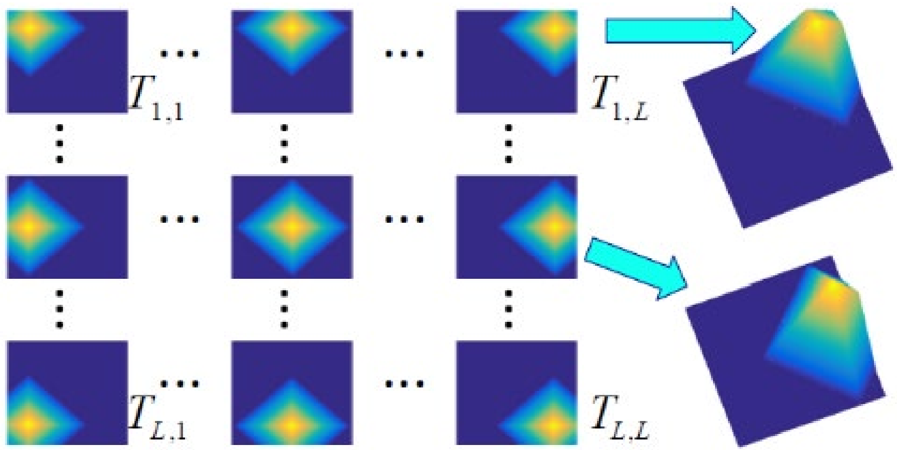
$$Loss = \underbrace{Loss(\hat{y}, y^*)}_{\text{task loss}} + \sum_f \underbrace{Loss_f(x)}_{\text{filter loss}}$$

The filter loss boosts the mutual information between feature maps $\boldsymbol{X}$ and a set of pre-defined part locations $\boldsymbol{T}$.

$$\mathbf{Loss}_f = -MI(\mathbf{X}; \mathbf{T}) \quad \text{for filter } f$$

Quanshi Zhang et al. “Interpretable CNNs for Object Classification” in IEEE Trans. on PAMI, 2020

Filter loss



$$Loss = \underbrace{Loss(\hat{y}, y^*)}_{\text{task loss}} + \sum_f \underbrace{Loss_f(x)}_{\text{filter loss}}$$

$$-Loss_f(x) = \mathrm{MI}(\mathrm{X}, \mathrm{T}) = \underbrace{-H(\mathbf{T})}_{\text{A constant}} + \underbrace{H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X})}_{\text{Entropy of Inter-category activations}} + \sum_x p(\mathbf{T}^+, x) \underbrace{H(\mathbf{T}^+ = \{T_\mu\}|X = x)}_{\text{Entropy of the spatial distribution of activations}}$$

Quanshi Zhang et al. “Interpretable CNNs for Object Classification” in IEEE Trans. on PAMI, 2020

Activation regions of interpretable filters

Filter 1

Filter 2

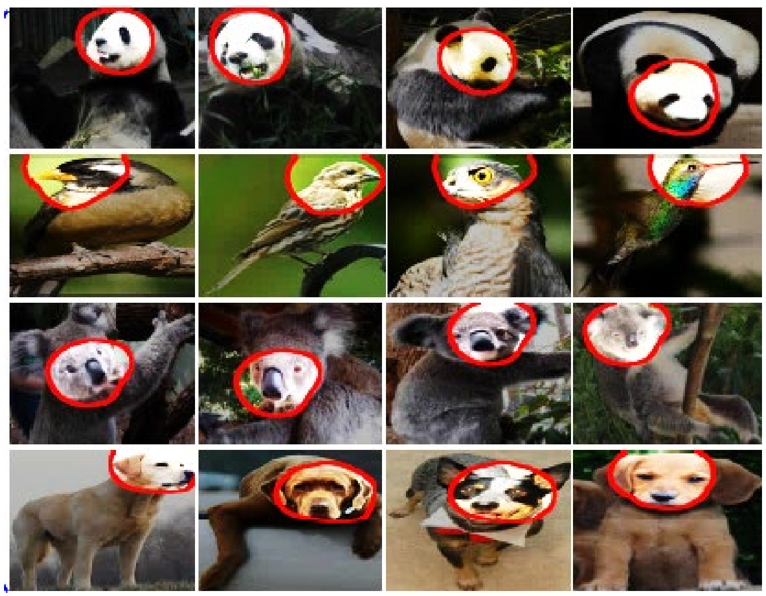
Filters 3 & 4

Filter

Filter

Filter

Filter

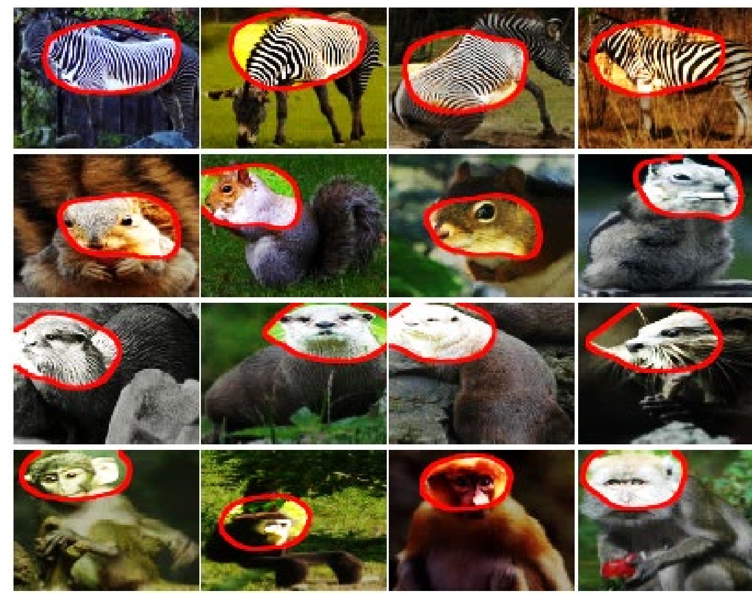


Filter

Filter

Filter

Filter



Quanshi Zhang et al. “Interpretable CNNs for Object Classification” in IEEE Trans. on PAMI, 2020

Outline

- ☐ XAI studies and vision of XAI science
- ☐ Explanation based on strict and fine-grained concepts
- ☐ **Quantification of the representation power of a DNN**
- ☐ Proof of mathematic essence of existing DL methods

Outline

- XAI studies and vision of XAI science
- Explanation based on strict and fine-grained concepts
- **Quantification of the representation power of a DNN**
  - **Metric of layerwise and pixel-wise information discarding**
  - Metric of knowledge consistency
  - Metric of feature/transformation complexity
- Proof of mathematic essence of existing DL methods

Layerwise and pixel-wise information discarding in DNNs

- As a generic metric, the information encoded in intermediate layers of DNNs
  - Show information-processing behaviors in classic deep models
  - Explain existing deep learning techniques
    - Network compression
    - Knowledge distillation
    - Modification of neural network architecture

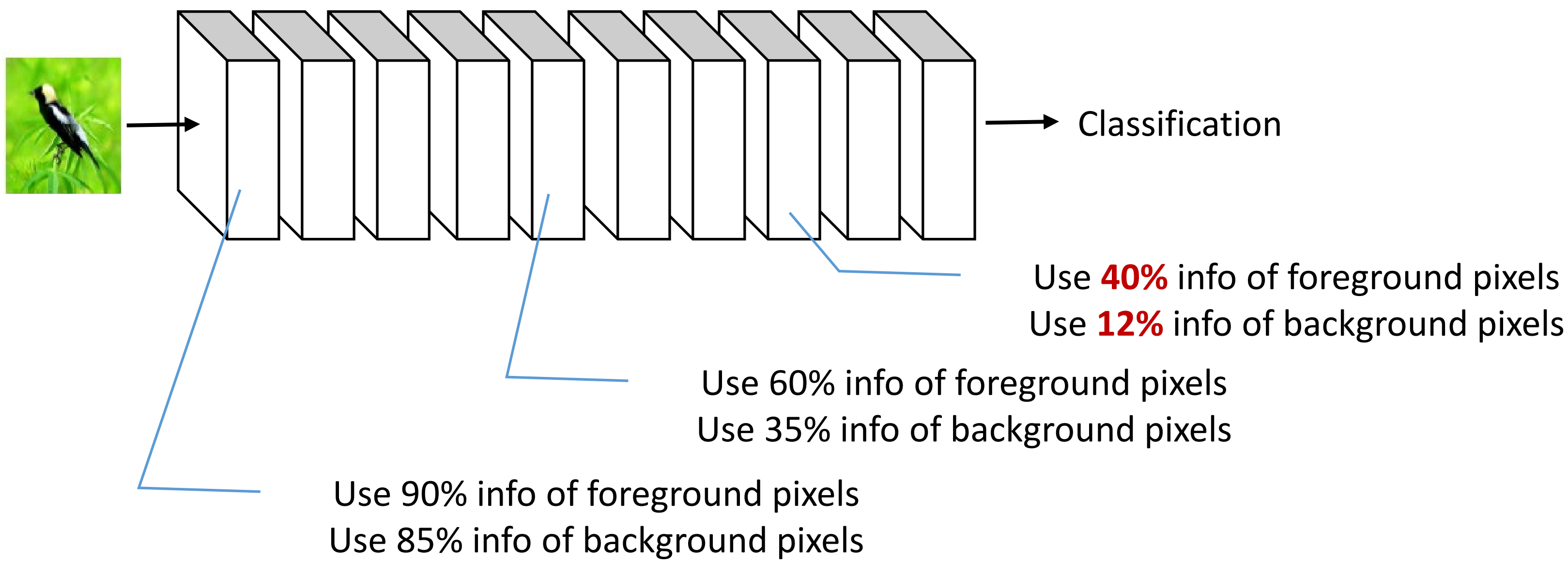
Ma et al. “Quantifying Layerwise Information Discarding of Neural Networks” in arXiv:1906.04109, 2019
Guan et al. “Towards A Deep and Unified Understanding of Deep Neural Models in NLP” in ICML 2019

Understanding DNNs as layerwise discarding of input information

- ◻ A DNN → layerwise discarding of input information
  - Discard less foreground information
  - Discard more background information

  } Enable reliable predictions

- ◻ Measure two types of information discarding
  - How much information of the input **is used to** compute the feature
  - How much information of the input **can be recovered from** the feature

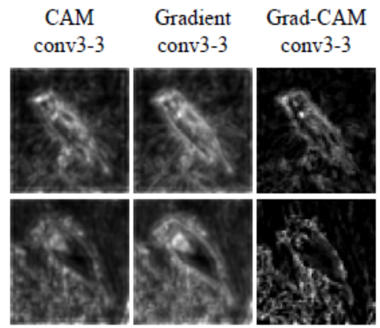


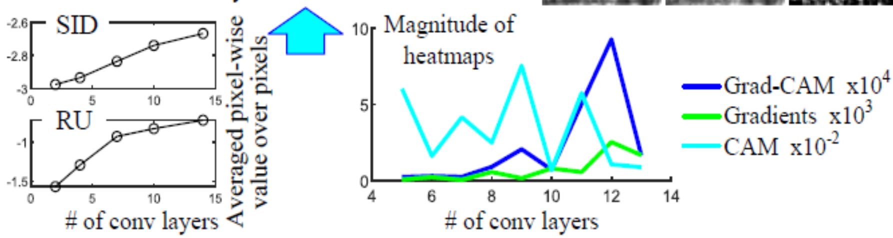
Ma et al. "Quantifying Layerwise Information Discarding of Neural Networks" in arXiv:1906.04109, 2019

Generality & coherency → enable comprehensive comparisons

	Coherency		Generality
	Layers	Nets	
Gradient-based	No	No	No
Perturbation-based	No	No	No
CAM-based	No	No	No
ours	Yes	Yes	Yes

Comparisons of different methods in terms of generality and coherency. Our method provides coherent results across layers and networks.





Signal magnitudes of layerwise heatmaps

Not enable fair layerwise comparisons

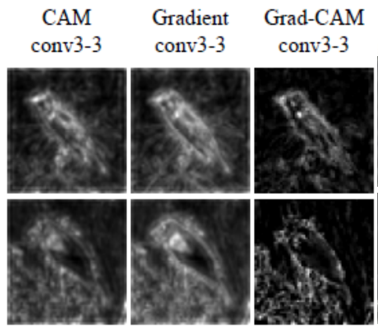
- Coherency: How to enable fair comparisons between layerwise attentions?
  - Previous methods of computing the pixel-wise attention / saliency / attribution / importance
    - Grad-CAM
    - Gradients-based
    - CAM
    - etc.

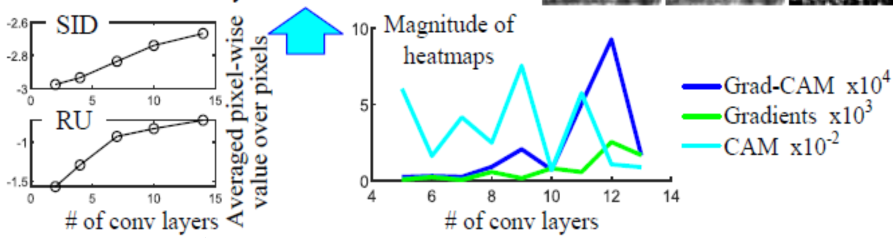
Ma et al. “Quantifying Layerwise Information Discarding of Neural Networks” in arXiv:1906.04109, 2019

Generality & coherency → enable comprehensive comparisons

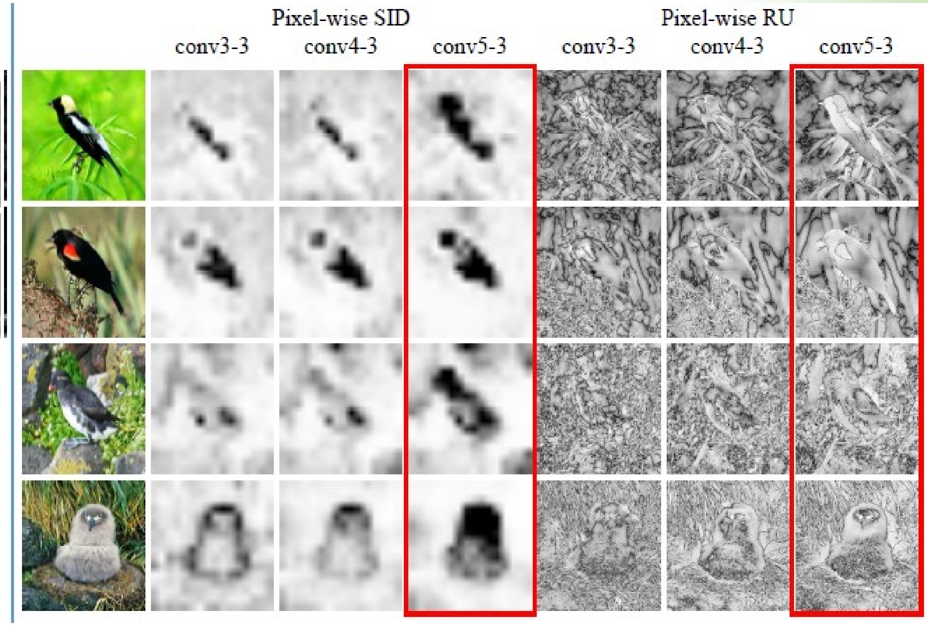
	Coherency		Generality
	Layers	Nets	
Gradient-based	No	No	No
Perturbation-based	No	No	No
CAM-based	No	No	No
ours	Yes	Yes	Yes

Comparisons of different methods in terms of generality and coherency. Our method provides coherent results across layers and networks.





gnal magnitudes of layerwise heatmaps



Enable fair layerwise comparisons

Not enable fair layerwise comparisons

Pixel-wise entropies of input information

Pixel-wise entropies of input reconstruction

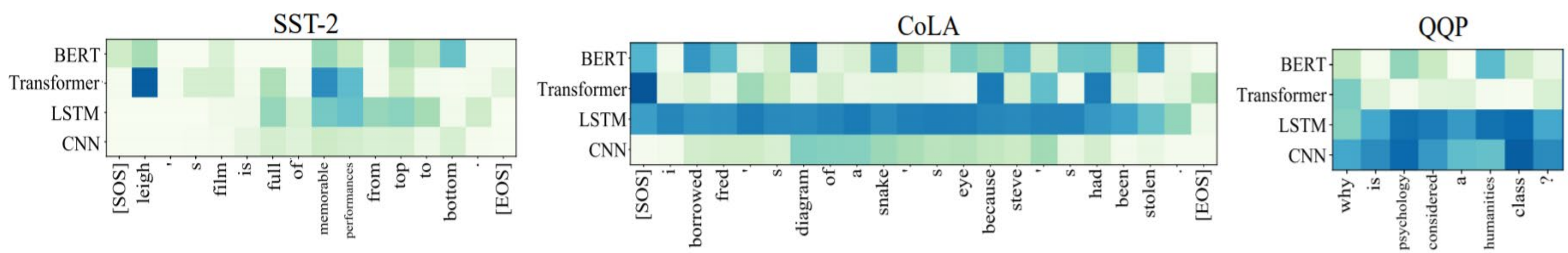
Comparing the discarding of the foreground / background information

1. Enable fair layerwise comparisons within a specific DNN
2. Enable fair comparisons between specific layers of different DNNs
3. Enable fair comparisons between different DNNs learned using the same input but for different tasks

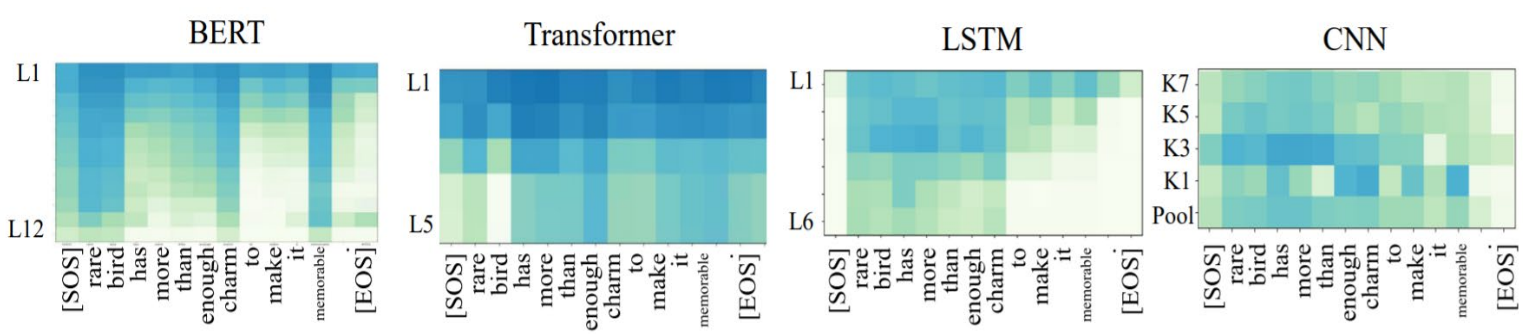
Ma et al. "Quantifying Layerwise Information Discarding of Neural Networks" in arXiv:1906.04109, 2019

Analysis on Deep Neural Models in NLP

Visualization of word importance. CNN and LSTM usually use sub-sequences of consecutive words for prediction, while BERT and Transformer select important word individually.



Layerwise information discarding. There is no specific information-discarding layer in the CNN. LSTM cannot distinguish important words. BERT and Transformer usually discard meaningless words in the first third of layers.



Guan et al. “Towards A Deep and Unified Understanding of Deep Neural Models in NLP” in ICML 2019

Metric to quantify knowledge points

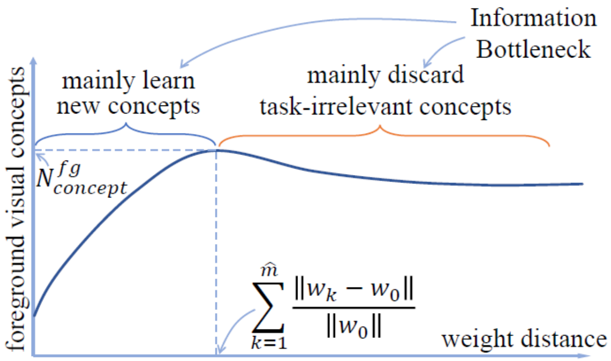
We propose a metric to **quantify knowledge points** in intermediate layer.

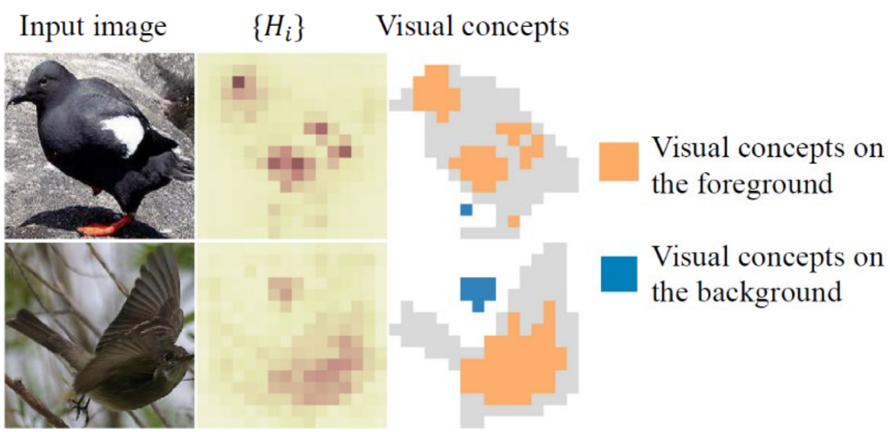
- Image regions that discarding much less information than most other regions.

Based on this metric, we **verify there hypothesis** for knowledge distillation

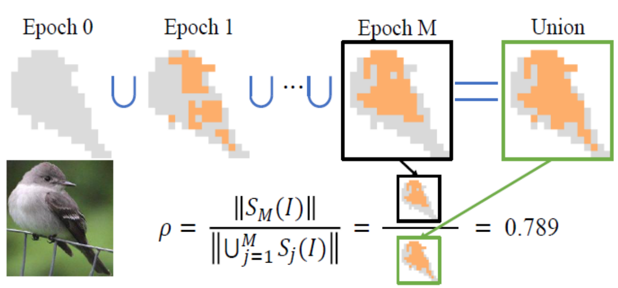
Compared with learning from scratch, knowledge distillation

- 1. learn more knowledge
- 2. learn diverse knowledge
- 3. less detour in learning





Quantifying detours in learning



Cheng et al. “Explaining Knowledge Distillation by Quantifying the Knowledge” in CVPR 2020

Outline

- ☐ XAI studies and vision of XAI science
- ☐ Explanation based on strict and fine-grained concepts
- ☐ **Quantification of the representation power of a DNN**
  - Metric of layerwise and pixel-wise information discarding
  - **Metric of knowledge consistency**
  - Metric of feature/transformation complexity
- ☐ Proof of mathematic essence of existing DL methods

Knowledge consistency

- As a generic metric, knowledge consistency can
  - **Quantify and evaluate the reliability** of intermediate-layer features of DNNs.
    - Without any additional testing samples or annotations.
  - Further **boost the performance of DNNs without additional annotations.**
  - Explain the success of existing deep-learning techniques
    - Knowledge distillation
    - Network compression
    - Network adversarial attack

Liang et al., "Knowledge Consistency between Neural Networks and Beyond" in ICLR, 2020

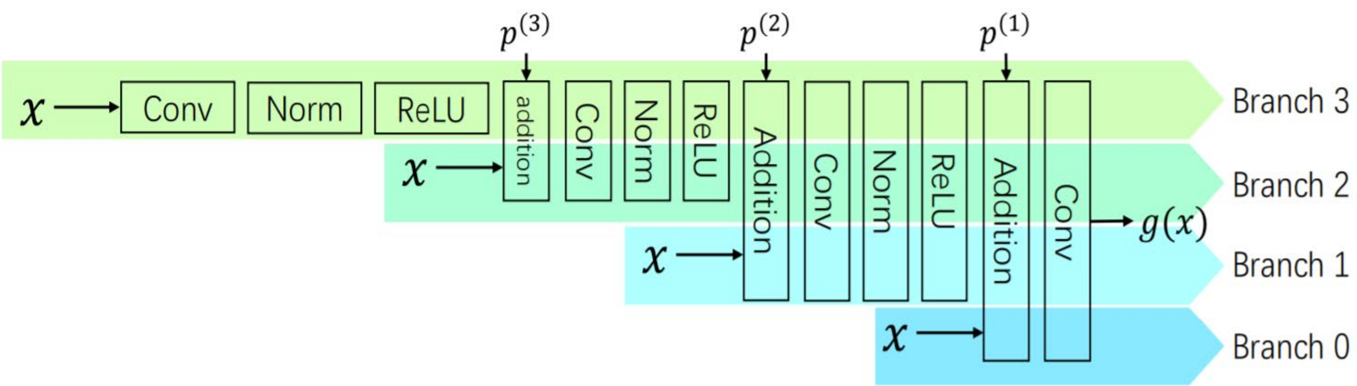
Knowledge consistency

x_A: an intermediate-layer feature of DNN A.
x_B: an intermediate-layer feature of DNN B.

If x_B can be reconstructed by x_A via

- a linear transformation → x_A and x_B are 0-order consistent
- one non-linear operation → x_A and x_B are 1-order consistent
- n non-linear operations → x_A and x_B are n-order consistent

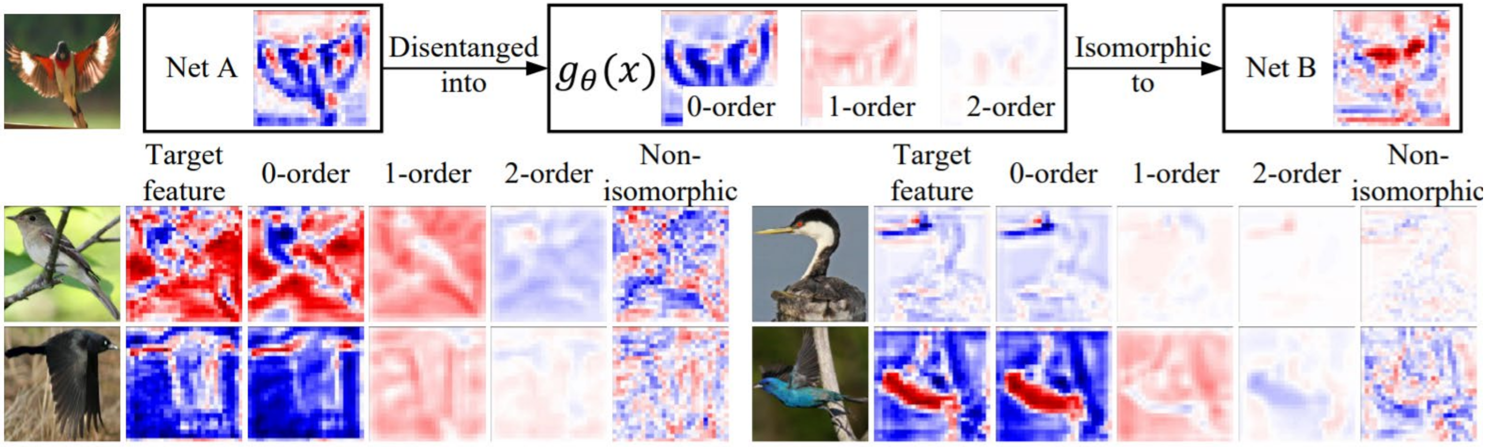
$$x^* = g_\theta(x) + x^\Delta, \qquad g_\theta(x) = x^{(0)} + x^{(1)} + \cdots + x^{(K)}$$



Liang et al., "Knowledge Consistency between Neural Networks and Beyond" in ICLR, 2020

Knowledge consistency with different orders

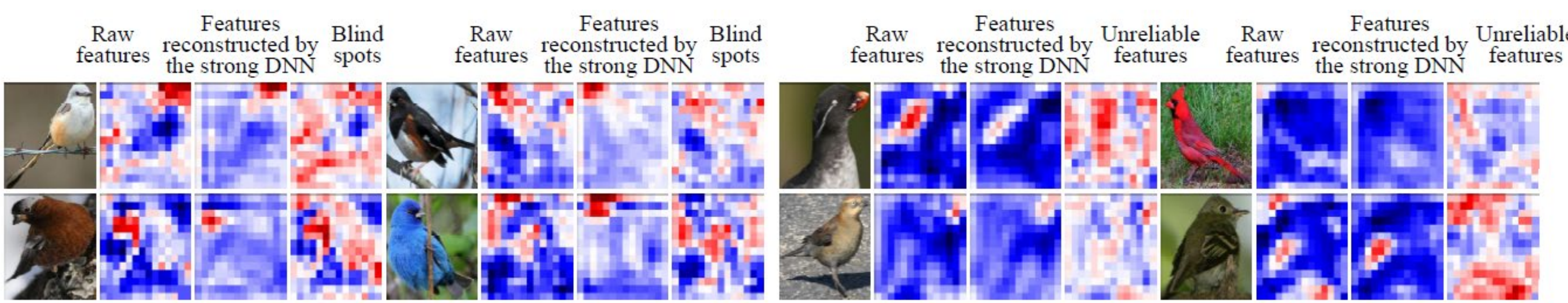
- If we trained **multiple DNNs for the same task**
  - Consistent feature components → reliable knowledge
  - Consistent feature components → boost the performance
- The following figure shows 0/1/2-order consistent feature components.
  - A low-order consistent feature components → **reliable features**
  - Inconsistent feature components → **noises**



Liang et al., "Knowledge Consistency between Neural Networks and Beyond" in ICLR, 2020

Detect blind spots and unreliable features

- Given a weak DNN and a well-trained DNN for a same task, we can disentangle and visualize the **blind spots** and **unreliable features** of the weak DNN using knowledge consistency.
- **Blind spots** of the weak DNN are defined as feature components that are encoded by the well-trained DNN, but are not encoded by the week DNN.
- **Unreliable features** of the DNN are defined as feature components that are encoded by the weak DNN, but are not encoded by the well-trained DNN.



(a) Blind spots of the weak DNN

(a) Unreliable/noisy features of the weak DNN

Liang et al., "Knowledge Consistency between Neural Networks and Beyond" in ICLR, 2020

Stable of learning DNNs (overfitting risk)

- Disentangling and quantifying inconsistent feature components can be used to measure the instability of learning DNNs.
  - **Overfitting risk is low:** DNNs can converge to the same knowledge representation from different initialization states.
- Given a relatively small training set, the learning of shallow DNNs was usually more stable than the learning of deep DNNs.

conv4 @ AlexNet	conv5 @ AlexNet	conv4-3 @ VGG-16	conv5-3 @ VGG-16	last conv @ ResNet-34
0.086	0.116	0.124	0.196	0.776
Learning DNNs using different training data				
conv4 @ AlexNet	conv5 @ AlexNet	conv4-3 @ VGG-16	conv5-3 @ VGG-16	last conv @ ResNet-34
0.089	0.155	0.121	0.198	0.275

Table 1: Instability of learning DNNs from different initializations and instability of learning DNNs using different training data. Without a huge dataset for training, networks with more layers usually suffered more from the over-fitting problem.

Liang et al., "Knowledge Consistency between Neural Networks and Beyond" in ICLR, 2020

Remove redundant features from pre-trained DNNs

- **Input:** Pre-trained DNNs——for various categories
  - Fine-grained classification for both 200 bird categories and 120 dog categories
- **Task:** Finetune DNNs——for several specific categories
  - Fine-grained classification for either 200 bird categories or 120 dog categories
- **Objective:** Detect and **remove redundant features** from pre-trained DNNs during the finetune process, in order to **improve the stability of intermediate-layer features.**

	VGG-16 conv4-3			VGG-16 conv5-2		
	VOC-animal	Mix-CUB	Mix-Dogs	VOC-animal	Mix-CUB	Mix-Dogs
Features from the network A	51.55	44.44	15.15	51.55	44.44	15.15
Features from the network B	50.80	45.93	15.19	50.80	45.93	15.19
$x^{(0)} + x^{(1)} + x^{(2)}$	**59.38**	**47.50**	**16.53**	**60.18**	**46.65**	**16.70**
	ResNet-18			ResNet-34		
	VOC-animal	Mix-CUB	Mix-Dogs	VOC-animal	Mix-CUB	Mix-Dogs
Features from the network A	37.65	31.93	14.20	39.42	30.91	12.96
Features from the network B	37.22	32.02	14.28	35.95	27.74	12.46
$x^{(0)} + x^{(1)} + x^{(2)}$	**53.52**	**38.02**	**16.17**	**49.98**	**33.98**	**14.21**

Liang et al., “Knowledge Consistency between Neural Networks and Beyond” in ICLR, 2020

Outline

- XAI studies and vision of XAI science
- Explanation based on strict and fine-grained concepts
- **Quantification of the representation power of a DNN**
  - Metric of layerwise and pixel-wise information discarding
  - Metric of knowledge consistency
  - **Metric of feature/transformation complexity**
- Proof of mathematic essence of existing DL methods

Metric of feature/transformation complexity

- Definitions of feature complexity and transformation complexity
- Relationship between complexity and other metrics
  - Reliability
  - Generalization power
  - Feature disentanglement

Definition of feature complexity

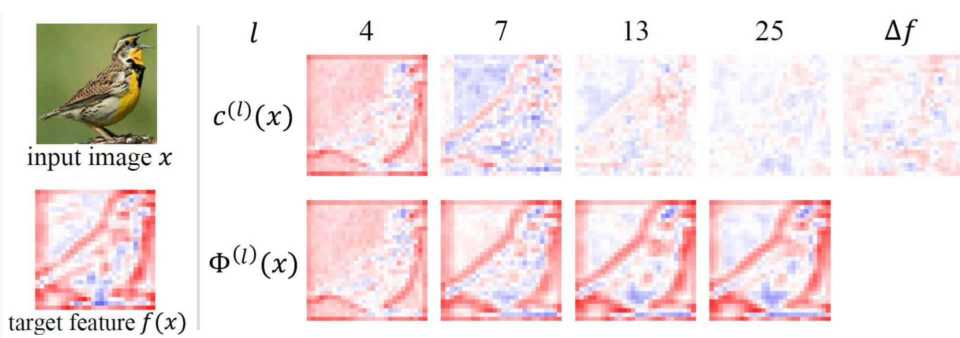
- Given an intermediate-layer feature of a DNN, the feature complexity is defined as the minimum number of non-linear layers that are required to compute the feature using another benchmark DNN with a fixed width.

$$l = \operatorname{argmin}_{l',\Phi}\left\{\Phi^{(l')}(x) = c\right\}$$

Theoretical maximum complexity $\neq$ Real feature complexity

- Disentangle intermediate-layer features into components of different complexity orders.

$$f(x) = c^{(1)}(x) + c^{(2)}(x) + \cdots + c^{(L)}(x) + \Delta f$$



Simple component:
Global shape
Complex component:
Details and noises

Ren et al. Interpreting and Disentangling Feature Components of Various Complexity from DNNs in arXiv:2006.15920, 2020

Relationship between complexity and reliability

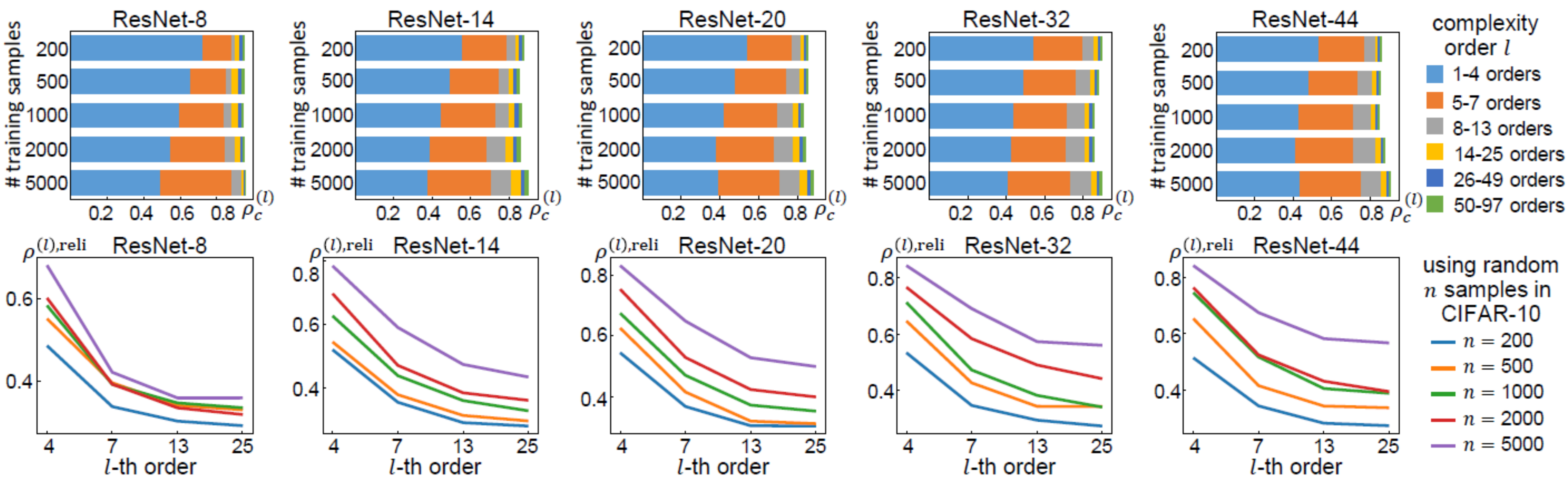
- Increasing training data → boosting reliability, not the complexity

# of training samples ↑

✔ **Boost feature commonness**

✘ **Not significantly increase complexity**

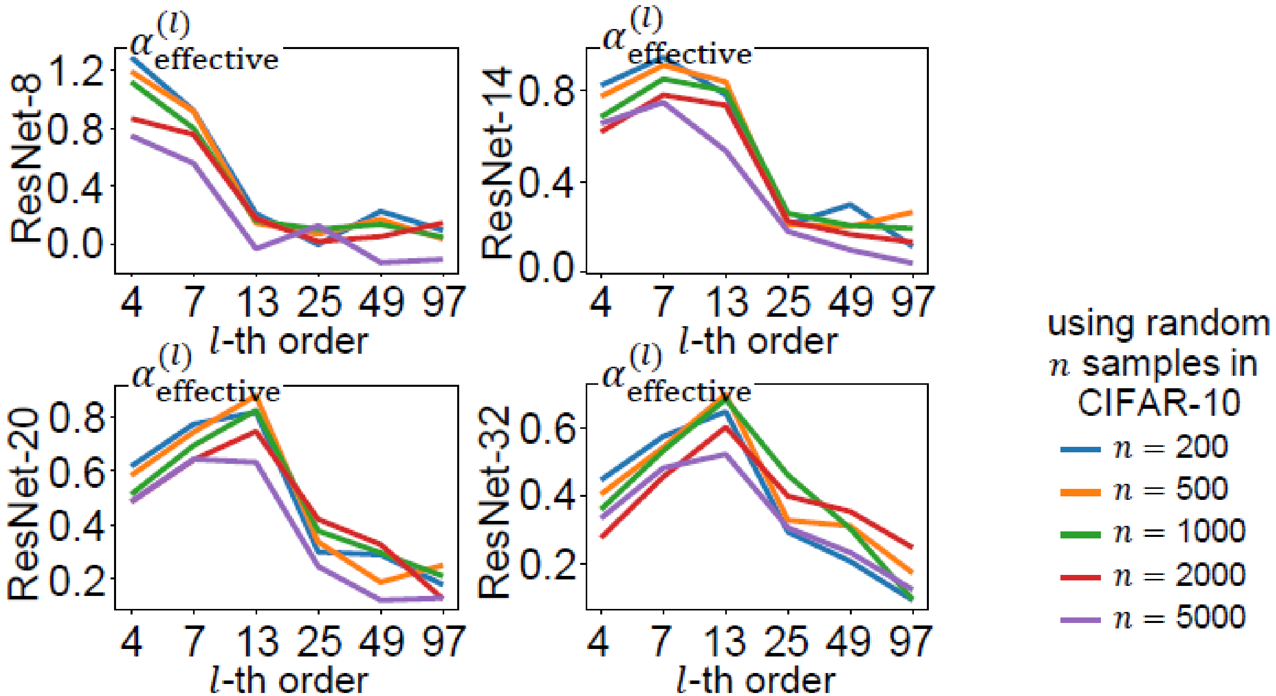
Improve performance ↑



Ren et al. Interpreting and Disentangling Feature Components of Various Complexity from DNNs in arXiv:2006.15920, 2020

Relationship between complexity and effectiveness

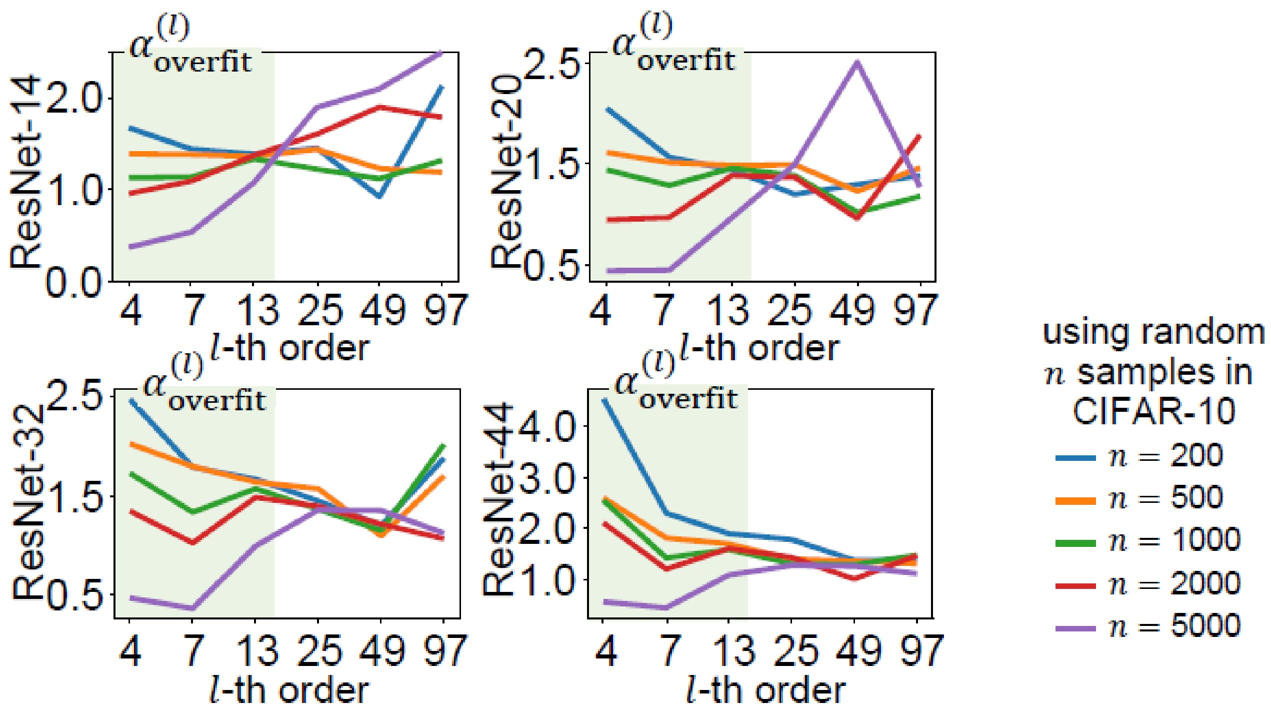
- Feature components of the complexity of **about the half depth** is the most effective (most influence to classification).
- Complex features are not always effective.



Ren et al. Interpreting and Disentangling Feature Components of Various Complexity from DNNs in arXiv:2006.15920, 2020

Relationship between complexity and overfitting

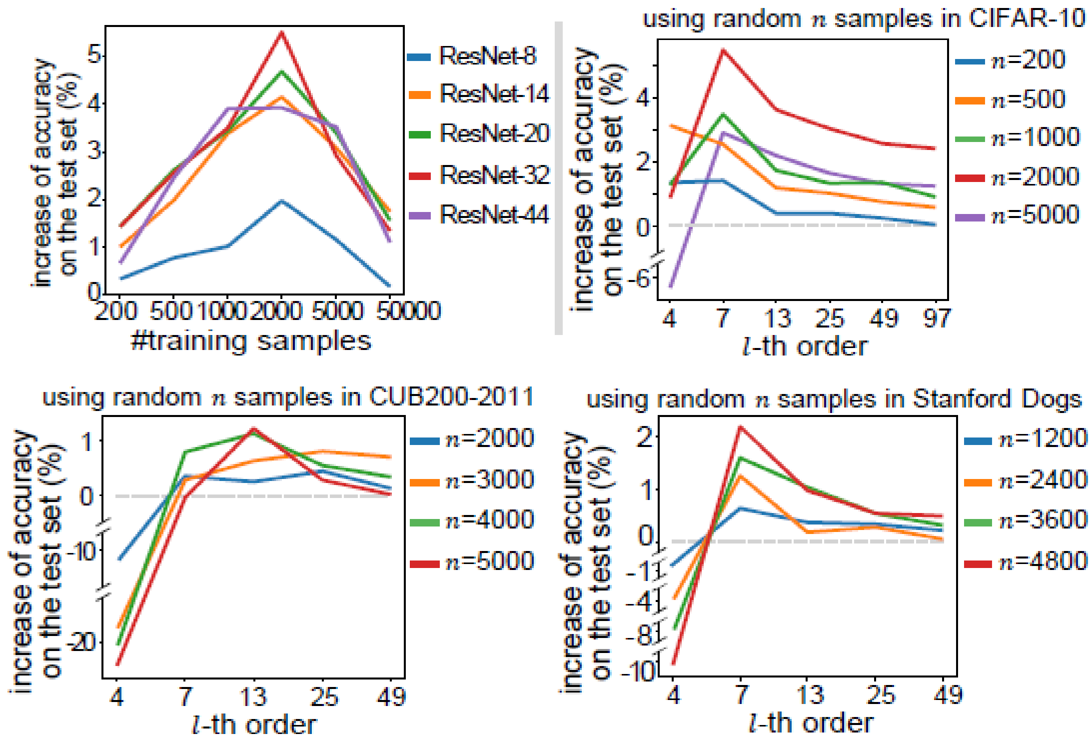
- For low-complexity feature components, the significance of overfitting can be reduced by adding more training samples.
- For high-complexity feature components, their overfitting level is insensitive to the sample number.



Ren et al. Interpreting and Disentangling Feature Components of Various Complexity from DNNs in arXiv:2006.15920, 2020

Using reliable feature components to boost performance

- Using the disentangled most effective components to boost the classification accuracy by 5%.



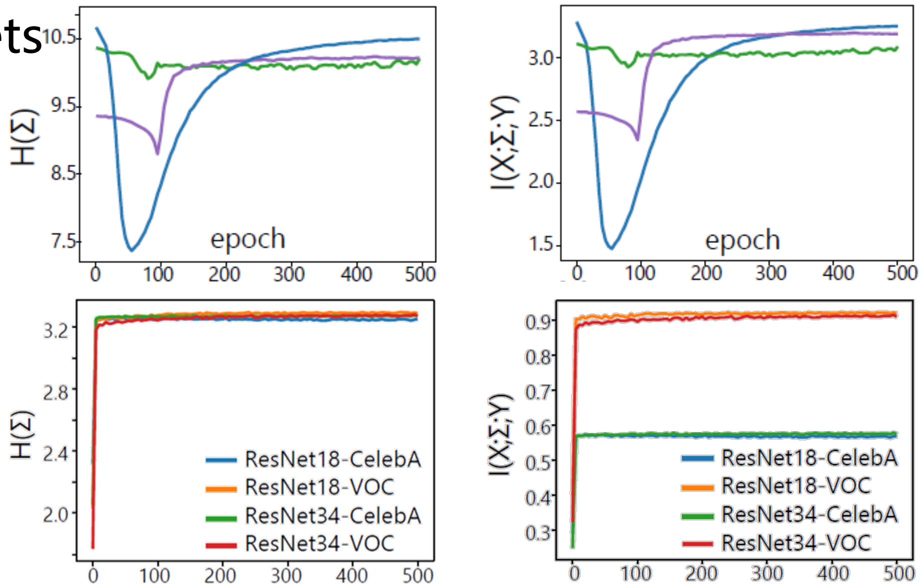
Ren et al. Interpreting and Disentangling Feature Components of Various Complexity from DNNs in arXiv:2006.15920, 2020

Definition of the transformation complexity

- Transformation complexity: the entropy of ReLU gating states.
  - $H(\Sigma)$ the entropy of gating states in all layers
  - $I(X;\Sigma)$ the mutual information of gating states and the input
  - $I(X;\Sigma;Y)$ the mutual information of gating states, the input, and the output

Cascaded nets

ResNets



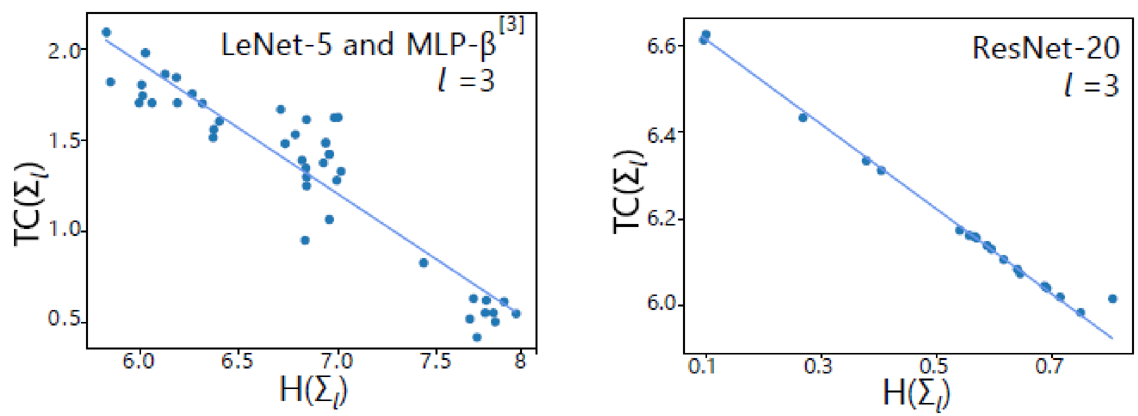
Ren et al. "Towards Theoretical Analysis of Transformation Complexity of ReLU DNNs" in arXiv

Proof of the relationship between feature disentanglement and the complexity

- Theoretically prove the negative relationship between feature disentanglement and transformation complexity.

Entanglement: $TC(\Sigma_l) = KL(p(\sigma_l) \| \prod_d p(\sigma_l^d))$

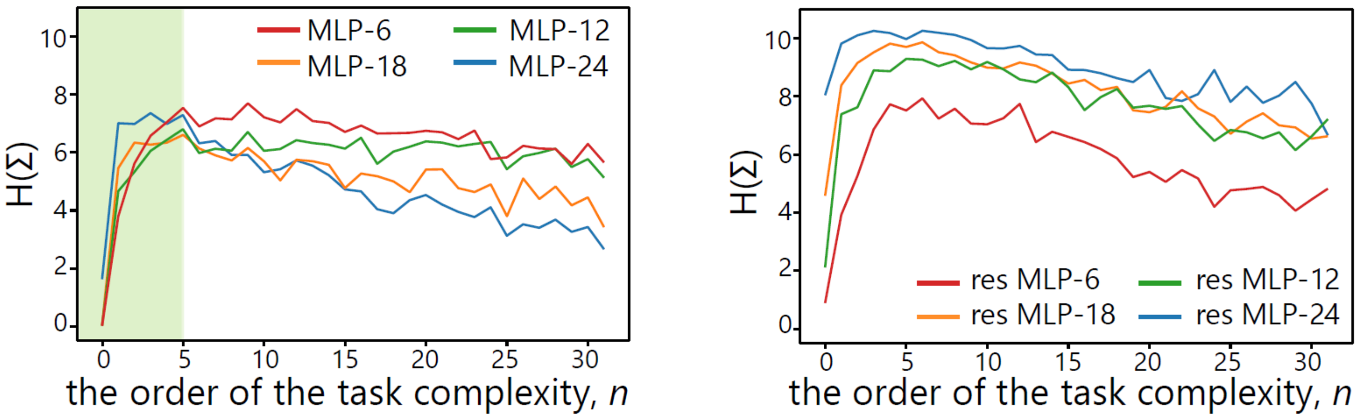
$$H(\Sigma_l) + TC(\Sigma_l) = C_l, \quad C_l = -\mathbb{E}_{\sigma_l}\left[\log \prod_d p\left(\sigma_l^d\right)\right]$$



Ren et al. “Towards Theoretical Analysis of Transformation Complexity of ReLU DNNs” in arXiv

Exploring the maximum complexity of a DNN

- The transformation complexity of a DNN is limited due to the optimization power of a DNN.
  - The transformation complexity does not monotonously increase along with the complexity of the task.
  - The transformation complexity begins to be saturated and decrease when the task is too complex.



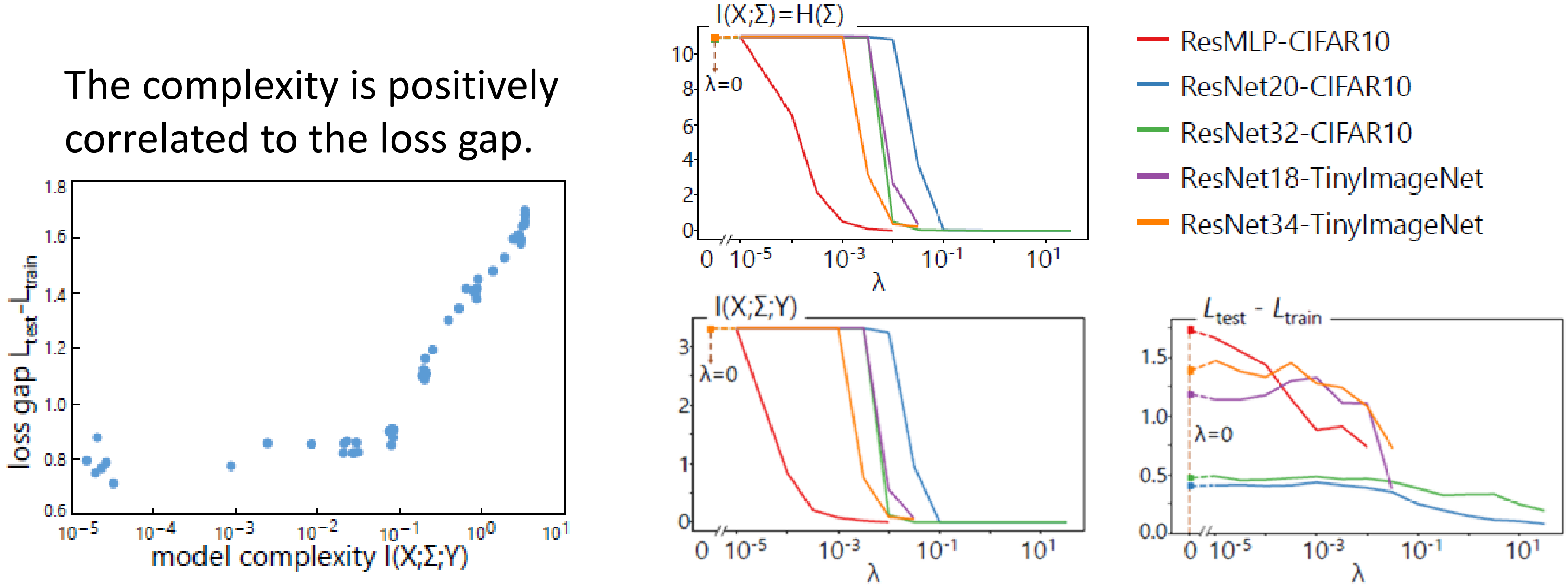
Ren et al. "Towards Theoretical Analysis of Transformation Complexity of ReLU DNNs" in arXiv

A loss to penalize the transformation complexity

- A loss to penalize the transformation complexity
  - It also reduce the gap between training loss and the testing loss.

$$L_{\text{complexity}} = \sum_{l=1}^{L} H(\Sigma_l) = \sum_{l=1}^{L} \{-\mathbb{E}_{\sigma_l}[\log p(\sigma_l)]\}$$

The complexity is positively correlated to the loss gap.



Ren et al. “Towards Theoretical Analysis of Transformation Complexity of ReLU DNNs” in arXiv

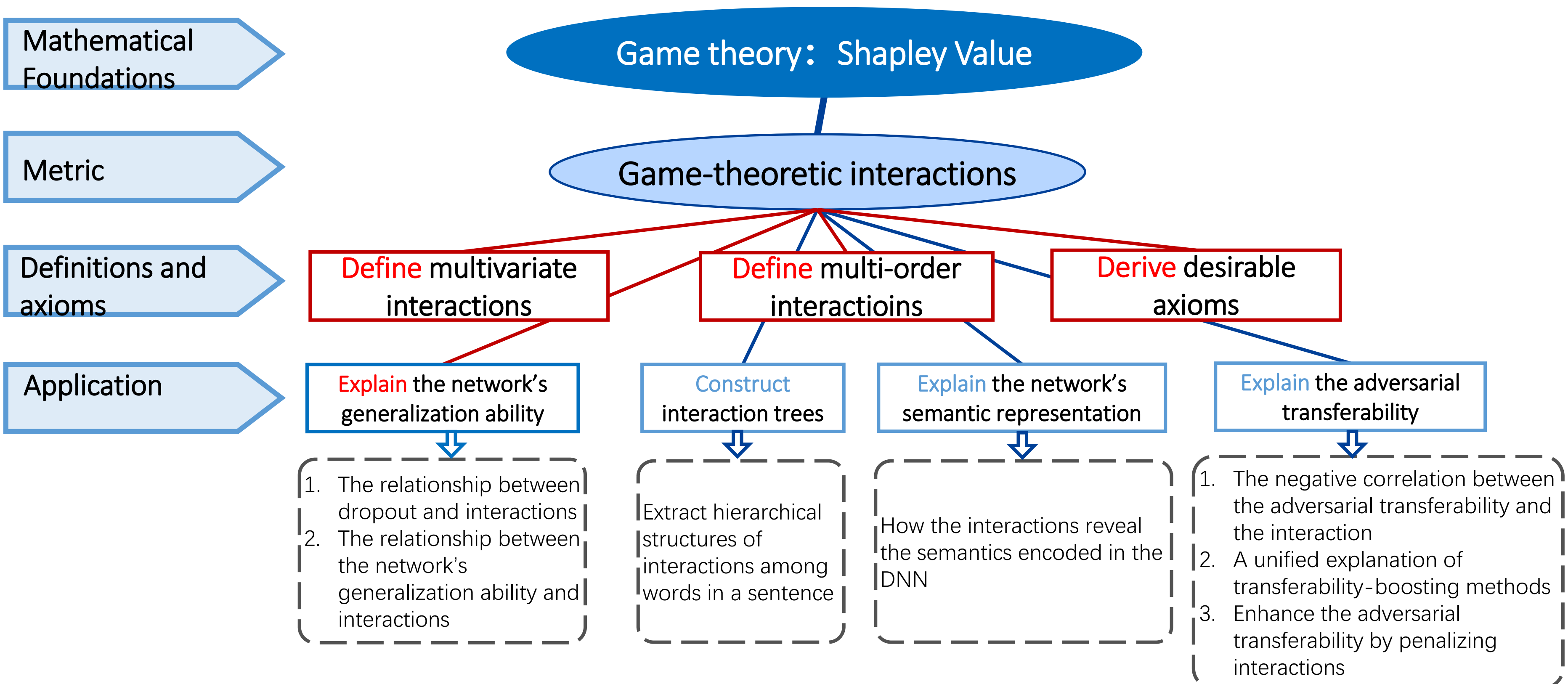
Outline

- ☐ XAI studies and vision of XAI science
- ☐ Explanation based on strict and fine-grained concepts
- ☐ Quantification of the representation power of a DNN
- ☐ **Proof of mathematic essence of existing DL methods**

Outline

- ☐ XAI studies and vision of XAI science
- ☐ Explanation based on strict and fine-grained concepts
- ☐ Quantification of the representation power of a DNN
- ☐ **Proof of mathematic essence of existing DL methods**
  - **Essence of methods of boosting adversarial transferability**
  - **Essence of the dropout operation**

Game-theoretic Interactions



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

Game-theoretic Interactions

- The input words of a sentence (or the input pixels of an image) into a DNN usually cooperate with each other, rather than work individually to make inferences.
- The cooperative input words (or pixels) have strong interactions.

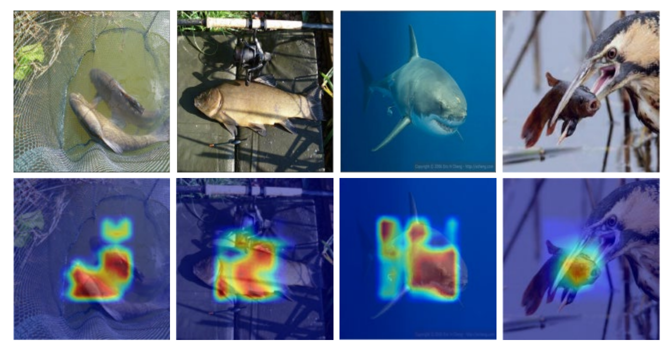
it ' s a remarkably solid and subtly satirical tour de force .

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 .



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

The Link between Interactions and the Network's Generalization Ability

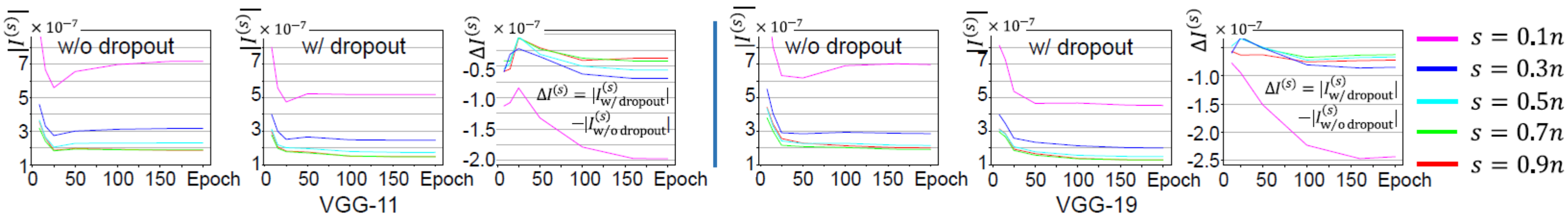
- Theoretically prove that Dropout can decrease the strength of interactions modeled by DNNs
- There is a negative correlation between the strength of interactions and the generalization ability of the network
- The generalization ability of the network can be enhanced by directly controlling the strength of interactions

Zhang et al. "Interpreting and Boosting Dropout from a Game-Theoretic View" in arXiv:2009.11729, 2020

The Link between Interactions and the Network's Generalization Ability

——relationships among dropout，interactions，and the generalization ability

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}	$\mathbf{3.64 \times 10^{-3}}$
Tiny-ImageNet	RN-34	2.57×10^{-3}	$\mathbf{2.89 \times 10^{-3}}$
CelebA	RN-34	6.46×10^{-3}	$\mathbf{1.17 \times 10^{-2}}$

Zhang et al. "Interpreting and Boosting Dropout from a Game-Theoretic View" in arXiv:2009.11729, 2020

The Link between Interactions and the Network's Generalization Ability

——directly suppress the interactions

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

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

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

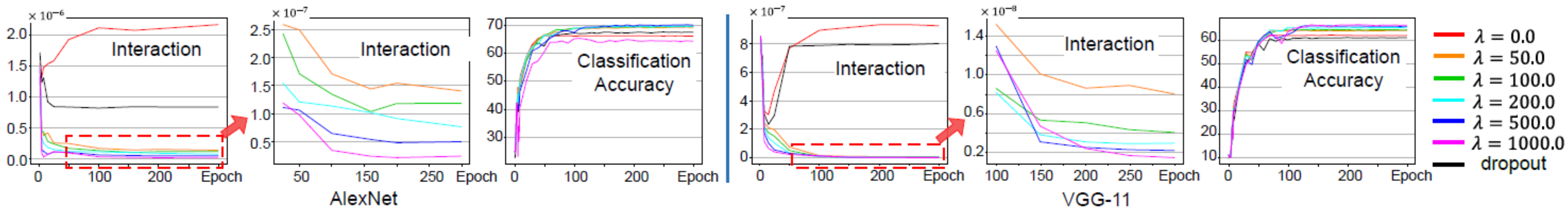
Based on the interactions, we improve the utility of dropout

- Control the utility of dropout by penalizing the strength of interactions, to explicitly control the DNN between over-fitting and under-fitting.
- Solve the issue that dropout is not compatible with batch normalization

Zhang et al. “Interpreting and Boosting Dropout from a Game-Theoretic View” in arXiv:2009.11729, 2020

The Link between Interactions and the Network's Generalization Ability

——directly suppress the interactions



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

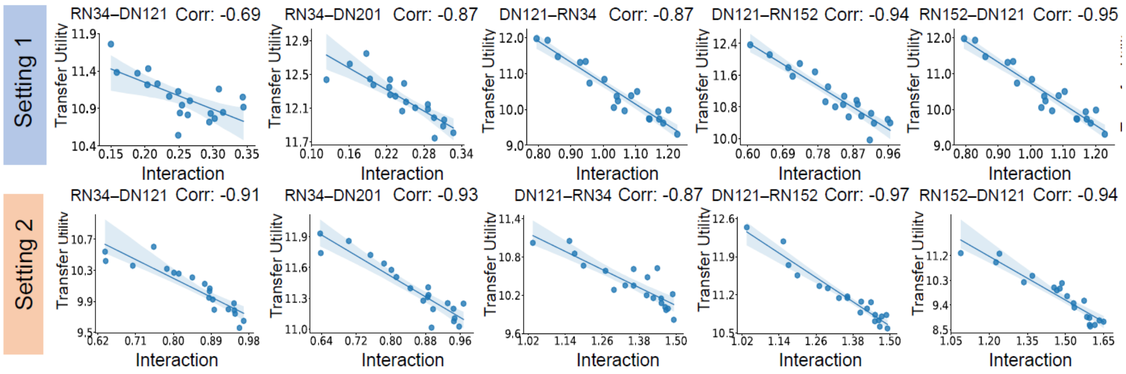
Under-fitting

Zhang et al. "Interpreting and Boosting Dropout from a Game-Theoretic View" in arXiv:2009.11729, 2020

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.
[2] Want et al. A Unified Approach to Interpreting and Boosting Adversarial Transferability. In arXiv:2010.04055, 2020

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

Essence: the reduction of interactions is the common mechanism of previous transferability-boosting methods

- Many previous transferability-boosting methods (mainly based on intuitions) 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 $\delta_{multi}^{m}=\alpha\sum_{t=0}^{m-1}\nabla_x l(h(x+\delta_{multi}^{t}),y)$, where δ_{multi}^{t} 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_{multi}^{m}, $\mathbb{E}_{a,b}[I_{ab}(\delta_{multi}^{m})]$, is larger than $\mathbb{E}_{a,b}[I_{ab}(\delta_{single})]$, i.e. $\mathbb{E}_{a,b}[I_{ab}(\delta_{multi}^{m})]\geq\mathbb{E}_{a,b}[I_{ab}(\delta_{single})]$.

Proposition 2

The adversarial perturbation generated by the multi-step attack is given as $\delta_{multi}^{m}=\alpha\sum_{t=0}^{m-1}\nabla_x l(h(x+\delta_{multi}^{t}),y)$. The adversarial perturbation generated by the VR Attack is computed as $\delta_{vr}^{m}=\alpha\sum_{t=0}^{m-1}\nabla_x\hat{l}(h(x+\delta_{vr}^{t}),y)$, where $\hat{l}(h(x+\delta_{vr}^{t}),y)=\mathbb{E}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}[l(h(x+\delta_{vr}^{t}+\xi),y)]$. Perturbation units of δ_{vr}^{m} tend to exhibit smaller interactions than δ_{multi}^{m}, i.e. $\mathbb{E}_x\mathbb{E}_{a,b}[I_{ab}(\delta_{vr}^{m})]<\mathbb{E}_x\mathbb{E}_{a,b}[I_{ab}(\delta_{multi}^{m})]$.

Proposition 3

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

[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

Future of pushing XAI towards science

Thank you

Although XAI is still far from science

Regional explanation with strict meanings

XAI metrics for representation power of DNNs

Well-proved theoretic foundation