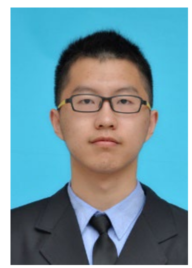
Discovering and Explaining the Representation Bottleneck of DNNs





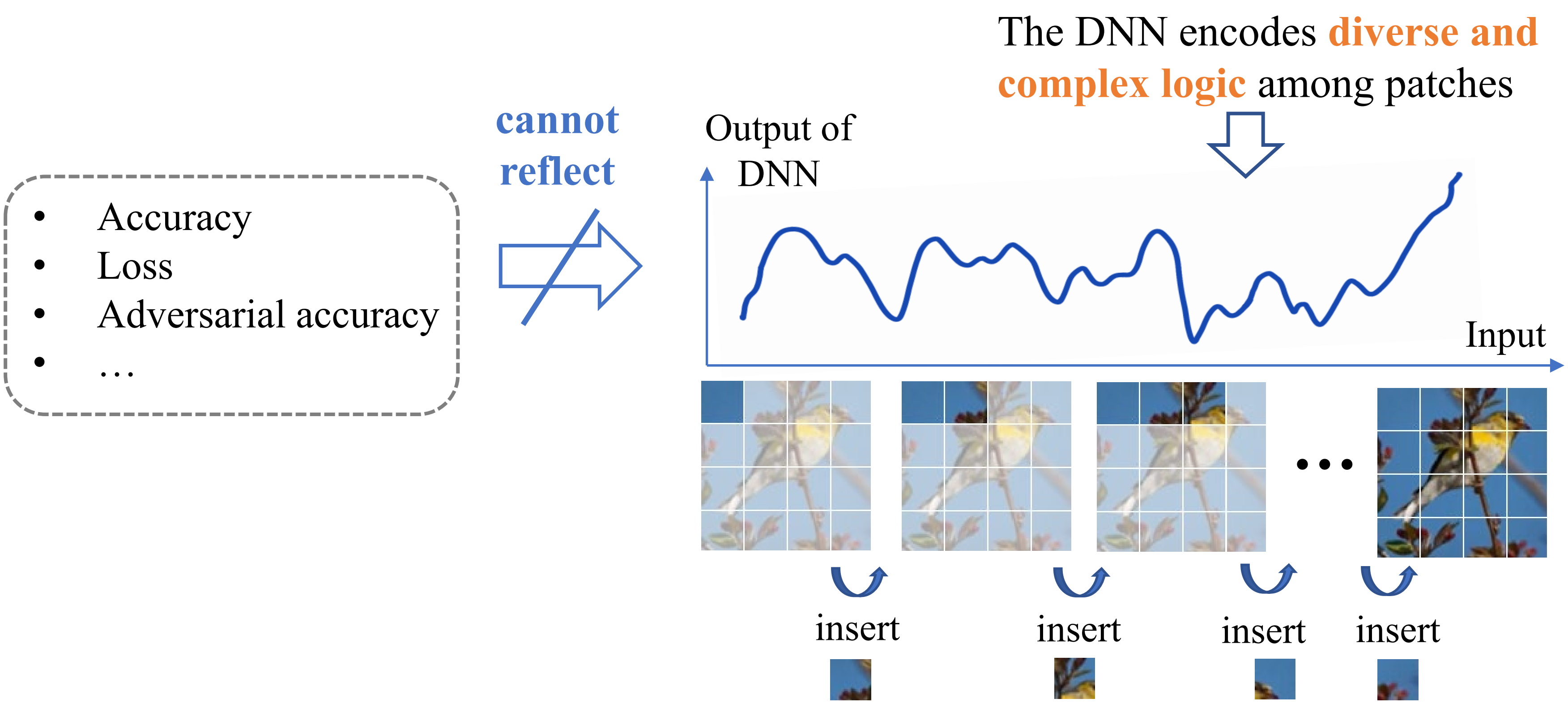

Huiqi Deng*, Qihan Ren*, Hao Zhang, Quanshi Zhang†

Shanghai Jiao Tong University

* Equal contribution.
† Correspondence.

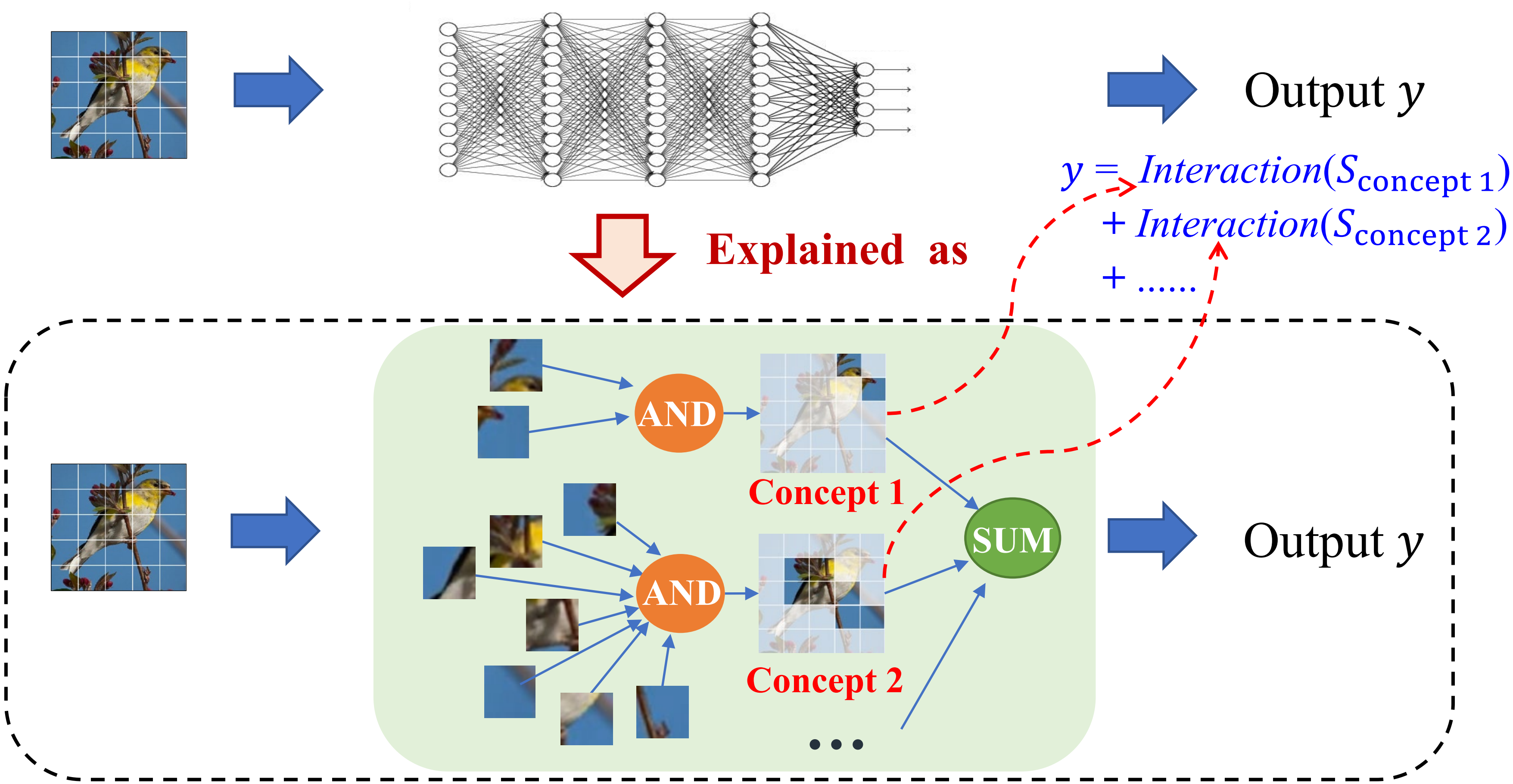
Previous studies used a single scalar metric to analyze an entire complex DNN

- A **single scalar** metric **cannot reflect** the **diversity of reasons** that contribute to the performance of an **entire complex DNN.**



Instead, we aim to explore the diverse reasons for the performance of an entire complex DNN

- In theory, we prove that the **performance**, such as accuracy/loss/… can be **explained** and **decomposed** into massive **multi-order interaction concepts**, which reflects the **diverse reasons** for the performance.

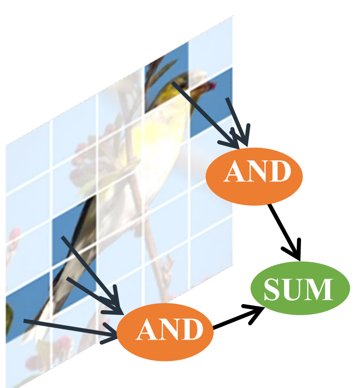


Are there common tendencies of DNNs in encoding concepts?

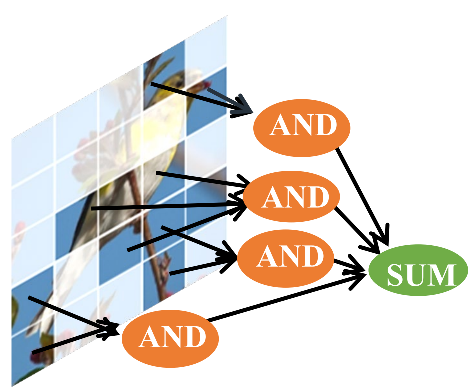
The representation bottleneck phenomenon:

- A DNN is **more likely** to encode both **too simple and too complex** interaction concepts.
- A DNN is **less likely** to encode **moderately complex** interaction concepts.

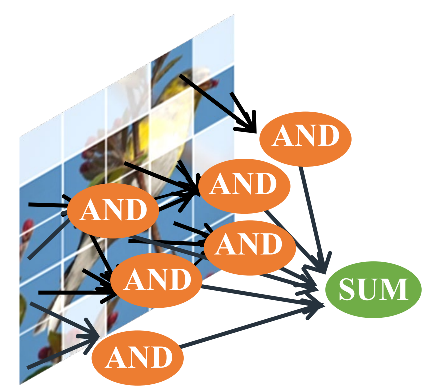
simple concepts
(composed of
a few patches)



moderately complex concepts
(composed of
a middle number of patches)



complex concepts
(composed of
massive patches)



Definition of interaction concepts

- **Background:** Input variables do not contribute to the network output independently, but **interact with each other** to form **interaction concepts** for inference.

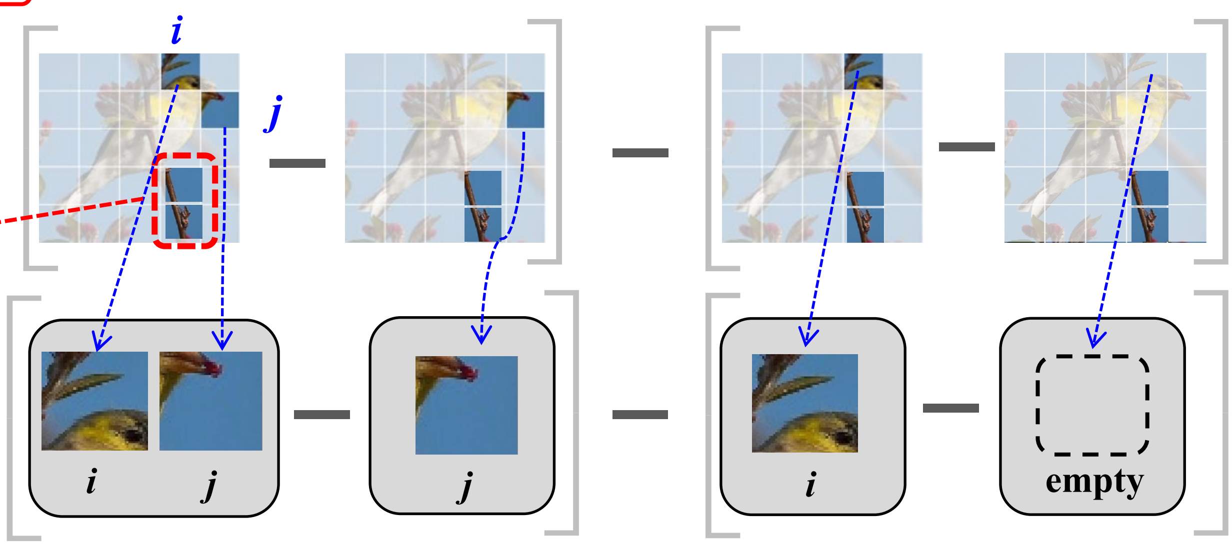
- The **interaction** between patch i and patch j is defined as

[the importance of i when j is present] − [the importance of i when j is absent]

$$I(i,j) = \mathbb{E}_{\boxed{S\subseteq N\setminus\{i,j\}}}\left[\overbrace{[v(S\cup\{i,j\}) - v(S\cup\{j\})]} - \overbrace{[v(S\cup\{i\}) - \boxed{v(S\cup\emptyset)}]}\right]$$

Sampling various contexts S

Context S



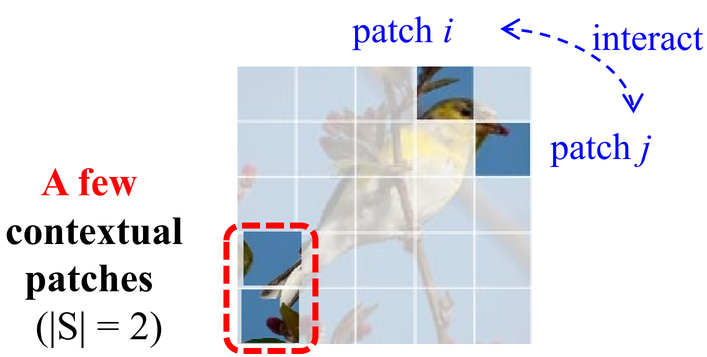
The network output when only **variables in $S\cup\emptyset$ are present, and other variables are masked**.

Complexity of interaction concepts

- The interaction between i, j can be further **decomposed** into the sum of **multi-order interactions**. Here, **the order *m*** (the number of contextual variables S) reflects the **complexity** of interaction concepts.

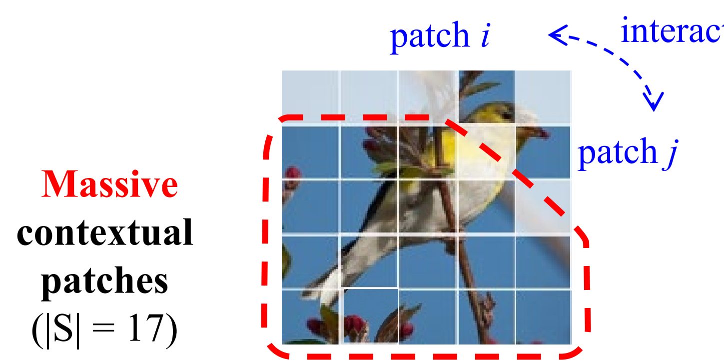
$$I(i,j) = \frac{1}{n-1} \sum_{m=0}^{n-2} I^{(m)}(i,j)$$

$$I^{(m)}(i,j) = \mathbb{E}_{S \subseteq N \setminus \{i,j\}, \boxed{|S|=m}} \left[[v(S \cup \{i,j\}) - v(S \cup \{j\})] - [v(S \cup \{i\}) - v(S)] \right]$$



Simple interaction concept

$= Interaction(i, j|\text{context with } \mathbf{2} \text{ patches})$



Complex interaction concept

$= Interaction(i, j|\text{context with } \mathbf{17} \text{ patches})$

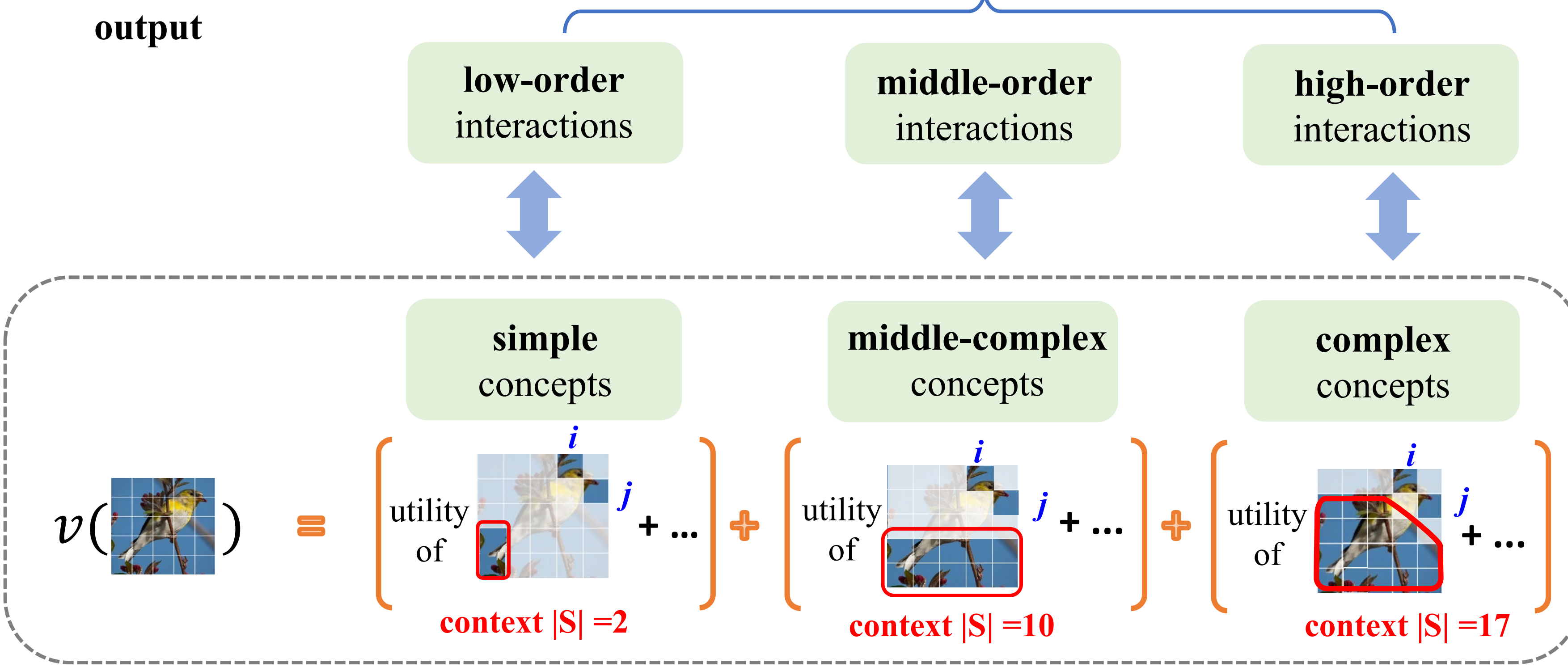
The network output can be decomposed into utilities of interaction concepts of different complexities

- We theoretically prove the following efficiency axiom:

$$\boxed{v(N)} = v(\emptyset) + \sum_{i\in N}\mu_i + \boxed{\sum_{i,j\in N, i\neq j}\sum_{m=0}^{n-2} w^{(m)} I^{(m)}(i,j)}$$

output

overall utilities of multi-order interactions



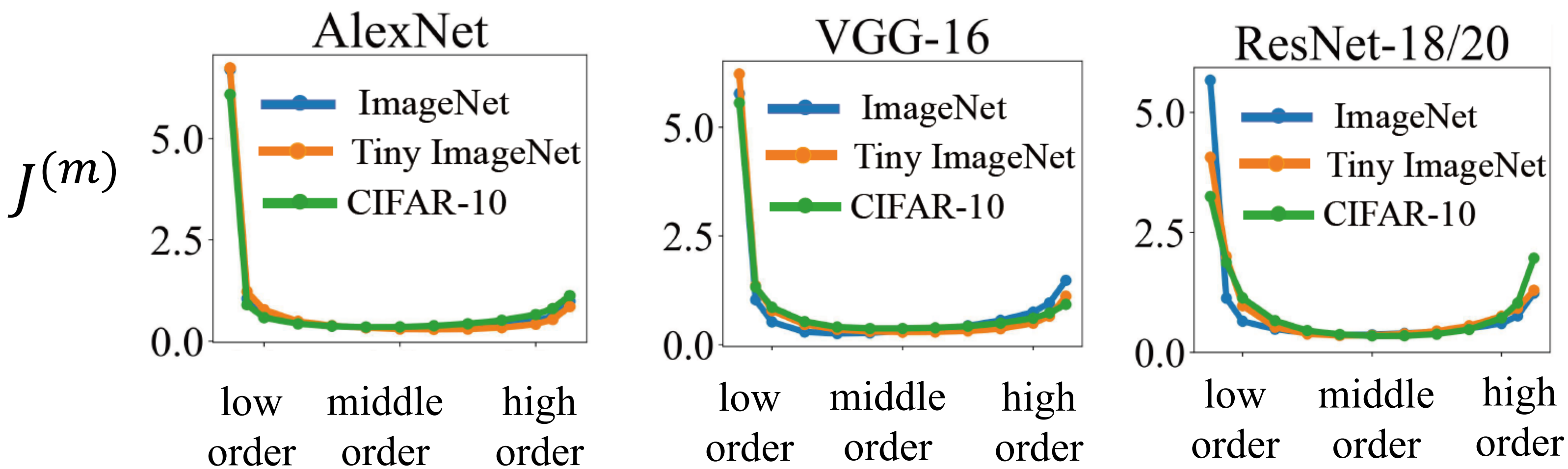
Discovering the representation bottleneck of DNNs

Representation bottleneck: a DNN usually encodes **strong low-order and high-order interactions**, but encodes **weak middle-order interactions**.

- We define the **relative strength** of the m-th order interaction.

$$J^{(m)} = \frac{\mathbb{E}_{x\in\Omega}[\mathbb{E}_{i,j}[|I^{(m)}(i,j|x)|]]}{\mathbb{E}_{m'}[\mathbb{E}_{x\in\Omega}[\mathbb{E}_{i,j}[|I^{(m')}(i,j|x)|]]]}$$

where the strength of interaction $I^{(m)}(i,j)$ is defined as $|I^{(m)}(i,j)|$



Theoretically explaining the representation bottleneck

- The learning effects of the entire DNN can be decomposed into the sum of the learning effects of multi-order interactions.

$$\underbrace{\boxed{\Delta W = -\eta \frac{\partial L}{\partial v(N)} \frac{\partial v(N)}{\partial W}}}_{\text{the learning effects of the entire DNN}} = \Delta W_U + \sum_{m=0}^{n-2} \sum_{i,j \in N, i \neq j} \underbrace{\boxed{R^{(m)} \frac{\partial I^{(m)}(i,j)}{\partial W}}}_{\text{the learning effects of } I^{(m)}(i,j)},$$

Theorem 1. (*Proof in Appendix B*) Assume $\mathbb{E}_{i,j,S}[\frac{\partial \Delta v(i,j,S)}{\partial W}] = \mathbf{0}$. Let σ^2 denote the variance of each dimension of $\frac{\partial \Delta v(i,j,S)}{\partial W}$. Then, we have $\mathbb{E}_{i,j}[\Delta W^{(m)}(i,j)] = \mathbf{0}$ and $\mathrm{Var}_{i,j}[\Delta W^{(m)}(i,j)] = \mathbf{1} \cdot (\eta \frac{\partial L}{\partial v(N)} \frac{n-m-1}{n(n-1)})^2 \sigma^2 / \binom{n-2}{m}$. Besides, $\mathbb{E}_{i,j}[\|\Delta W^{(m)}(i,j)\|_2^2] = K(\eta \frac{\partial L}{\partial v(N)} \frac{n-m-1}{n(n-1)})^2 \sigma^2 / \binom{n-2}{m}$, where K is the dimension of the network parameter W.

Theorem 1 indicates that:

The strength of learning $I^{(m)}(i,j)$ is proportional to $\frac{n-m-1}{n(n-1)} / \sqrt{\binom{n-2}{m}}$

- The strength of learning **extremely low-order** or **extremely high-order** interactions is **much higher**
- The strength of learning **middle-order** interactions is **much lower**

Theoretically explaining the representation bottleneck

- The learning effects of the entire DNN can be decomposed into the sum of the learning effects of multi-order interactions.

$$\underbrace{\Delta W = -\eta \frac{\partial L}{\partial v(N)} \frac{\partial v(N)}{\partial W}}_{\text{the learning effects of the entire DNN}} = \Delta W_U + \sum_{m=0}^{n-2} \sum_{i,j \in N, i \neq j} \underbrace{R^{(m)} \frac{\partial I^{(m)}(i,j)}{\partial W}}_{\text{the learning effects of } I^{(m)}(i,j)},$$

Theorem 1. (*Proof in Appendix B*) Assume $\mathbb{E}_{i,j,S}[\frac{\partial \Delta v(i,j,S)}{\partial W}] = \mathbf{0}$. Let σ^2 denote the variance of each dimension of $\frac{\partial \Delta v(i,j,S)}{\partial W}$. Then, we have $\mathbb{E}_{i,j}[\Delta W^{(m)}(i,j)] = \mathbf{0}$ and $\mathrm{Var}_{i,j}[\Delta W^{(m)}(i,j)] = \mathbf{1} \cdot (\eta \frac{\partial L}{\partial v(N)} \frac{n-m-1}{n(n-1)})^2 \sigma^2 / \binom{n-2}{m}$. Besides, $\mathbb{E}_{i,j}[\|\Delta W^{(m)}(i,j)\|_2^2] = K(\eta \frac{\partial L}{\partial v(N)} \frac{n-m-1}{n(n-1)})^2 \sigma^2 / \binom{n-2}{m}$, where K is the dimension of the network parameter W.

Theorem 1 indicates that:

The strength of learning $I^{(m)}(i,j)$ is proportional to $\boxed{\frac{n-m-1}{n(n-1)} / \sqrt{\binom{n-2}{m}}}$

The strength of learning **extremely low-order** or **extremely high-order** interactions is **much higher**

The strength of learning **middle-order** interactions is **much lower**

Theoretically simulated distribution of interaction concepts *vs.* the true distribution of interaction concepts

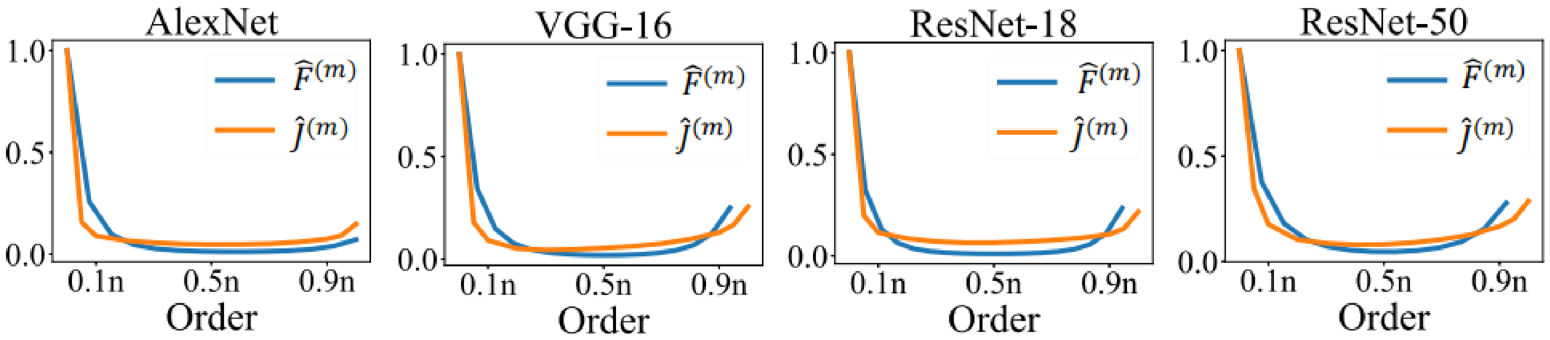
Theoretically simulated distribution
of interaction concepts of different orders

$$F^{(m)} = \frac{n-m-1}{n(n-1)} / \sqrt{\binom{n-2}{m}}$$

Simulate

True Distribution
of interaction concepts of different orders

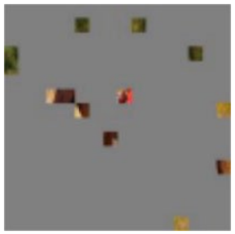
$$J^{(m)} = \frac{\mathbb{E}_{x\in\Omega}[\mathbb{E}_{i,j}[|I^{(m)}(i,j|x)|]]}{\mathbb{E}_{m'}[\mathbb{E}_{x\in\Omega}[\mathbb{E}_{i,j}[|I^{(m')}(i,j|x)|]]]}$$



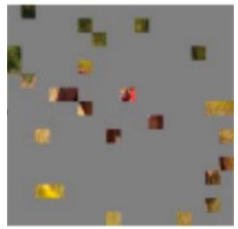
The two distributions are **well matched**.

Human cognition vs. concepts encoded by a DNN

Given **a few** patches:



additional patches



DNNs: extract **strong interactions**

Humans: extract **little information**

Given **middle number of** patches:

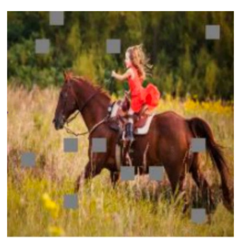
additional patches

DNNs: extract **weak interactions**

Humans: extract **much information**

Given **massive** patches:

additional patches



DNNs: extract **strong interactions**

Humans: extract **little information**

Breaking the representation bottleneck

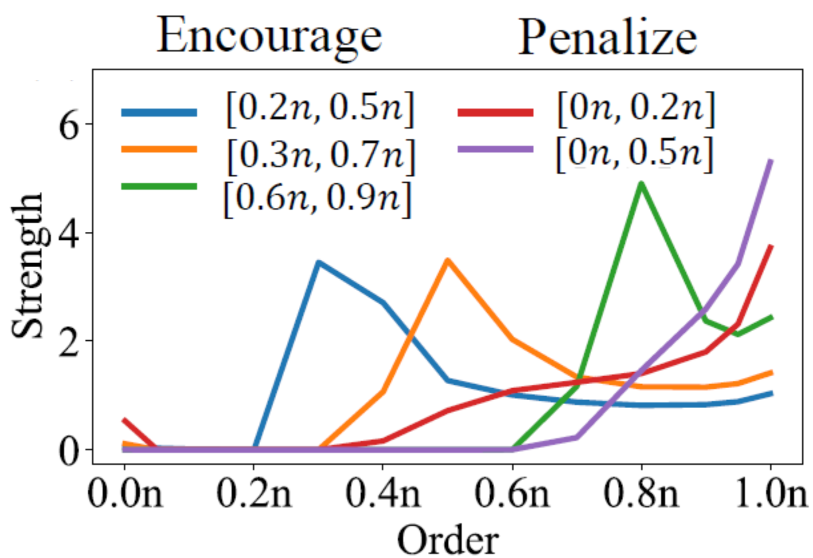
- We propose losses to encourage/penalize interactions of specific orders

$$\text{Loss} = \text{Loss}_{\text{classification}} + \lambda_1 \boxed{L^+(r_1, r_2)} + \lambda_2 \boxed{L^-(r_1, r_2)}$$

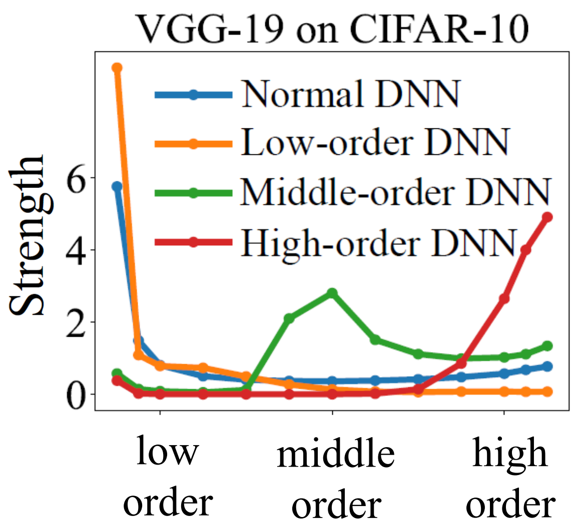
encourage interactions of orders in the range $[r_1 n, r_2 n]$

penalize interactions of orders in the range $[r_1 n, r_2 n]$

Experimental verification:



DNNs encoding interactions of different orders achieve similar accuracies

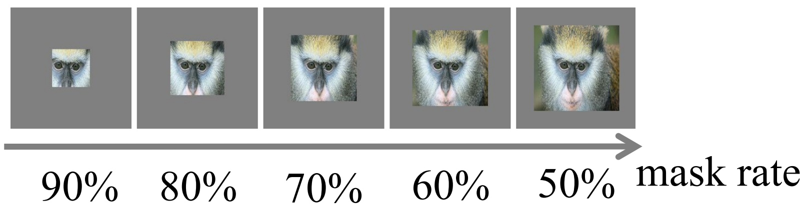


	CIFAR-10			Tiny-ImageNet		
Model	AlexNet	VGG16	VGG19	AlexNet	VGG16	VGG19
Normal training	88.52	90.50	90.61	56.00	56.16	52.56
Low interaction	86.97	89.99	89.74	58.68	55.60	55.04
Mid interaction	86.65	90.29	90.03	53.88	55.84	53.36
High interaction	88.68	90.84	90.79	56.12	55.36	53.28

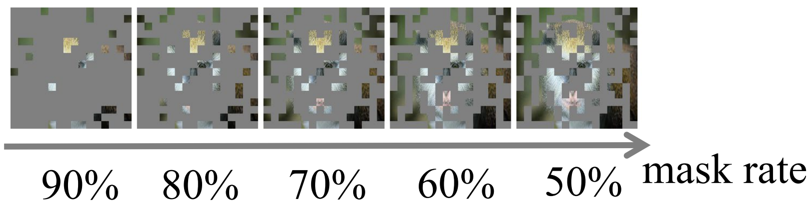
High-order interactions are more vulnerable to adversarial attacks

	Normal DNN	High-order DNN
Model	Normal training	Penalize low-order & boost high-order
MLP-5 on census	38.22	**7.31**
MLP-8 on census	39.33	**2.02**
MLP-5 on commer	27.01	**22.00**
MLP-8 on commer	25.92	**20.58**

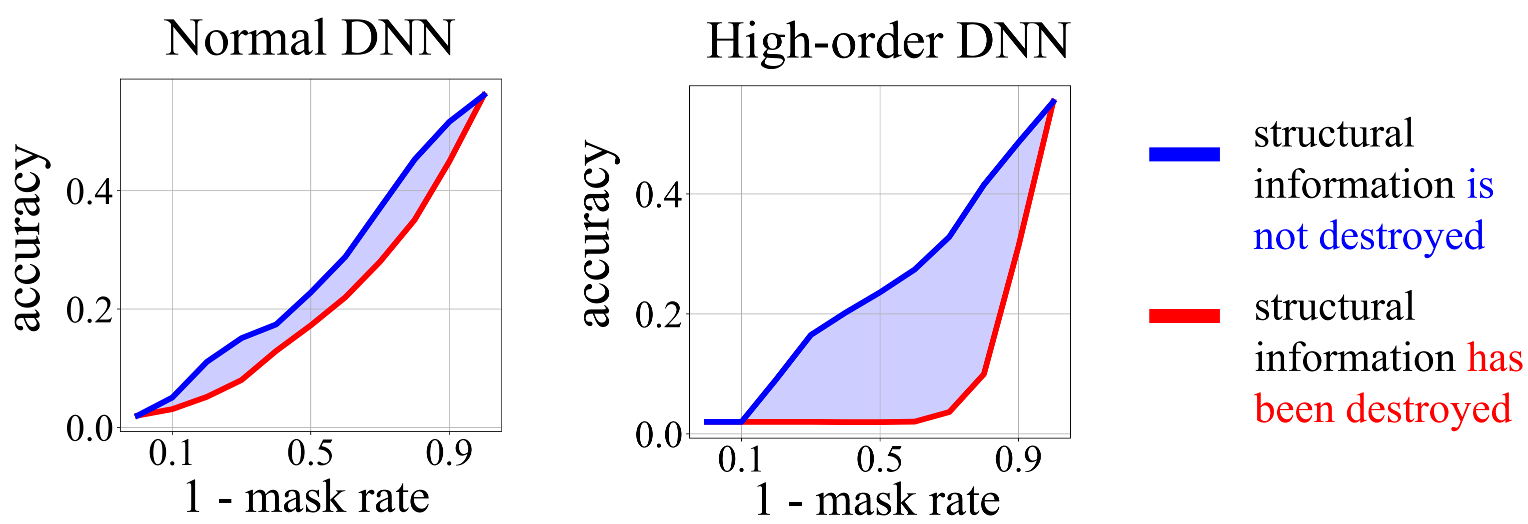
High-order interactions encode more structural information



structural information **is not destroyed**



structural information **has been destroyed**



Conclusions

- We discover the **representation bottleneck phenomenon** of DNNs
- We **theoretically explain** the representation bottleneck phenomenon
- We propose losses to force DNNs to **encode interactions of specific orders**
- We **investigate the representation capacities** of DNNs mainly encoding low-order, middle-order, and high-order interactions