ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Pure Noise to the Rescue of Insufficient Data:

Improving Imbalanced Classification by Training on Random Noise Images

Shiran Zada Itay Benou Michal Irani

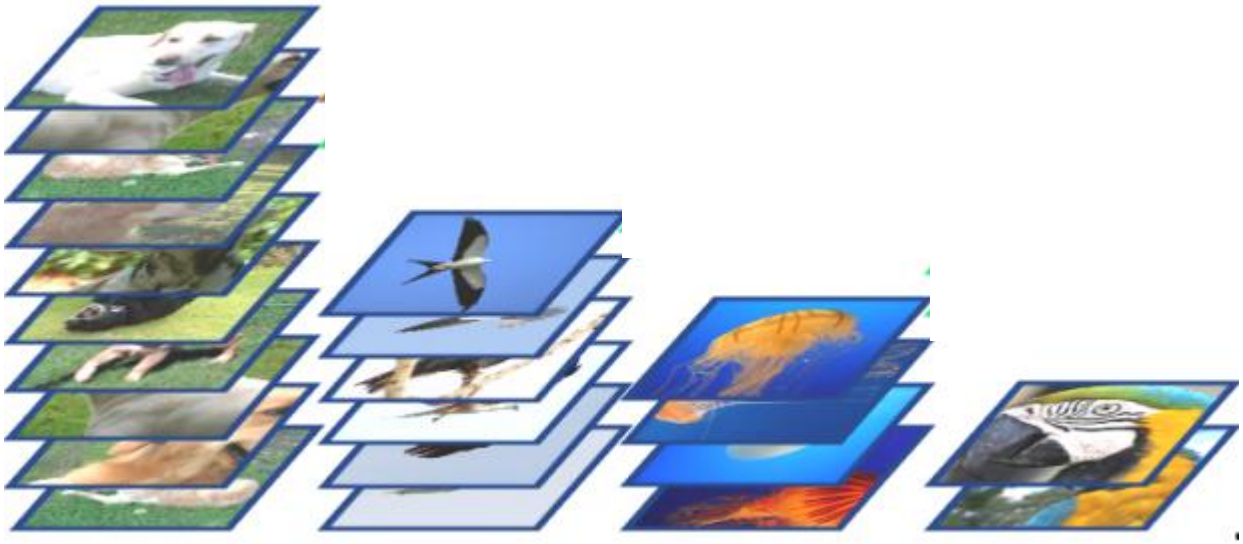
Dept. of Computer Science and Applied Math, The Weizmann Institute of Science, Israel

ICML 2022

Imbalanced Data

Solutions

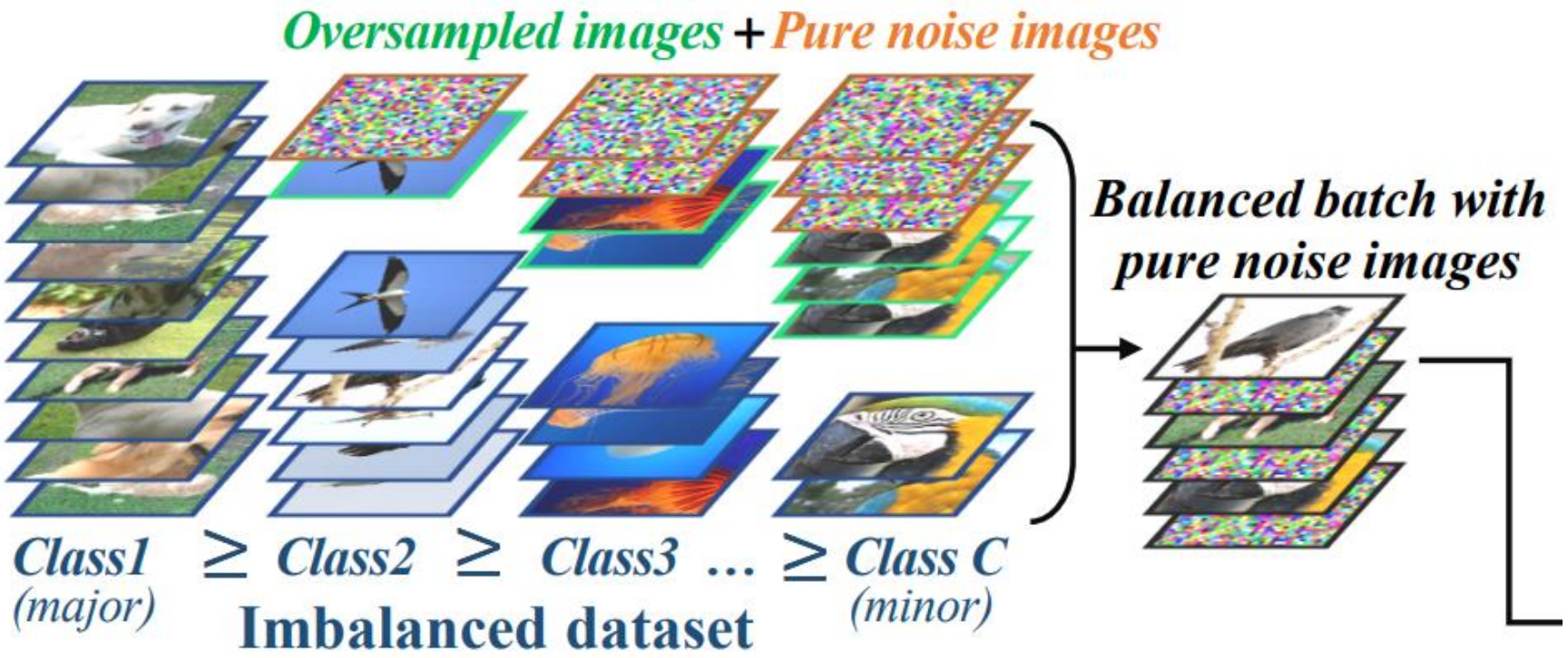
1. Over-sampling
2. Re-weighting
3. Decoupled training



Main idea

ParNeC

- Re-balances an imbalanced dataset with **pure-noise** images, in addition to oversampled natural images.

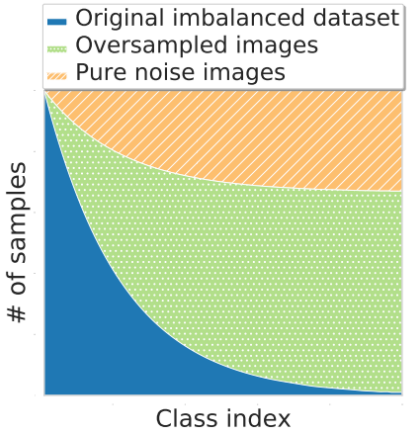


Method

1. How many noise images should be generated?

$$\mathbb{P}(\text{replace } x \text{ with } x_{noise} | c_i) = (1 - \rho_i) \cdot \delta$$

Class i Imbalanced ratio



2. How to generate?

a) Compute mean and var. of each color channel $l \in \{1,2,3\}$

$$\mu_{\mathcal{D},l} = \mathbb{E}(\mathcal{X}[l, :]) \quad , \quad \sigma_{\mathcal{D},l} = \sqrt{Var(\mathcal{X}[l, :])}$$

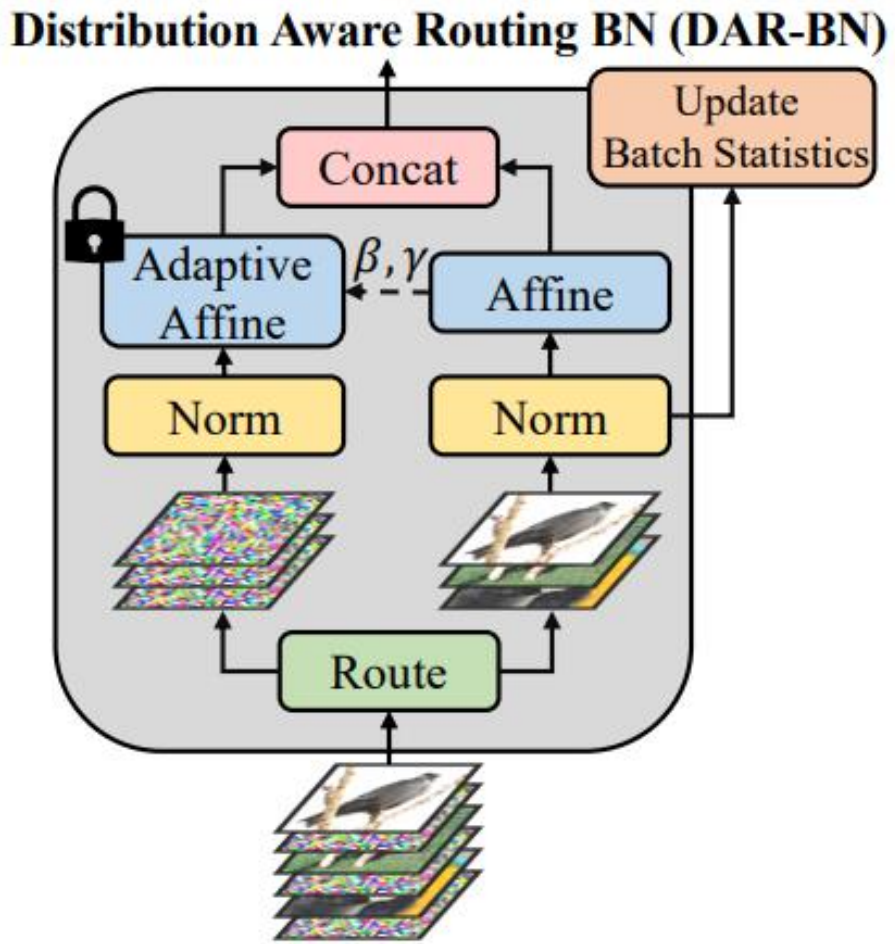
b) Generate Gaussian noise

$$\hat{x}_{noise} \sim \mathcal{N}(\mu_{\mathcal{D}}, \sigma_{\mathcal{D}})$$

$$x_{noise} = \min(\max(\hat{x}_{noise}, 0), 1)$$

Method

Since the noise images behave differently with the real images. A new batch normalization method is proposed to handle this problem.



Experiments

PartNeC

✓ Datasets

Dataset	# of classes	Imbalance-ratio (IR)	Largest class size	Smallest class size	# of samples
CIFAR-10-LT [5]	10	{50 , 100}	5,000	{100 , 50}	{13,996 , 12,406}
CIFAR-100-LT [5]	100	{50 , 100}	500	{10 , 5}	{12,608 , 10,847}
ImageNet-LT [36]	1,000	256	1,200	5	115,846
Places-LT [36]	365	996	4,980	5	62,500
CelebA-5 [27]	5	10.7	2423	227	6651

✓ Baselines

- **ERM** : Directly train the network on the imbalanced dataset.
- **Oversampling** : Oversampling the minority classes.
- **ERM + AutoAugment**: Employ auto-augmentation in training.
- **Other SOTA**.

✓ Implementation details

1. The proposed method is applied only on the last 15~30 epochs
2. δ is fixed as 1/3 , the experiments are repeated 4 times.

Results

Methods	CIFAR-10-LT		CIFAR-100-LT		ImageNet-LT	Places-LT	CelebA-5
	IR=100	IR=50	IR=100	IR=50			
Empirical Risk Minimization (ERM)	79.6±0.2	84.9±0.4	47.0±0.5	52.4±0.4	51.1	29.9	78.6 ±0.1
Oversampling	75.1±0.4	82.2±0.4	42.5±0.3	48.0±0.2	49.0	38.1	76.4 ±0.2
LADM-DRW [5]$^{\odot}$	77.1	81.1	42.1	46.7	-	-	-
M2m [27]§	79.1±0.2	-	43.5±0.2	-	43.7	-	75.9±1.1
Balanced Meta-Softmax (BALMS) [40]§	-	-	-	-	41.8	38.7	-
LADE [22]§	-	-	45.4	50.5	53.0	38.8	-
MisLAS [54]§	82.1	85.7	47.0	52.3	52.7	40.4	-
OPeN (ours)	**84.6**±0.2	**87.9**±0.2	**51.5**±0.4	**56.3**±0.4	**55.1**	**40.5**	**79.7** ±0.2
ERM + AutoAugment	81.4±0.3	86.4±0.2	49.9±0.4	55.7±0.4	52.2	29.2	79.3 ±0.5
BALMS [40] + AutoAugment§	84.9	-	50.8	-	-	-	-
OPeN (ours) + AutoAugment	**86.1**±0.1	**89.2**±0.2	**54.2**±0.5	**59.8**±0.5	**56.1**	39.6	**80.9**±0.4

Table 2. **Results & Comparison on imbalanced benchmark datasets.** *Mean accuracy over all classes per dataset. OPeN outperforms*

Results

ParNeC

Norm Layer	CIFAR-10-LT	CIFAR-100-LT
Standard BN [25]	81.45±0.70	49.18±0.54
Auxiliary BN [48]	83.38±0.16	50.13±0.06
DAR-BN (ours)	**84.64**±0.16	**51.50**±0.44

Table 3. **Ablation study: Comparing different Batch-Norm layers.** *Mean accuracy on CIFAR-10/100-LT with IR=100. Each type of BN is plugged into OPeN (with same training parameters). DAR-BN outperforms the other normalization layers.*

Explanation

- From a training point of view, oversampling with pure noise images increases the magnitude of minority gradient components.

- Although the pure noise does not contain discriminative information, it implies the class prior to the model.

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Discovering and Explaining the Representation Bottleneck Of DNNs

Huiqi Deng,* Qihan Ren,* Hao Zhang, Quanshi Zhang†
Shanghai Jiao Tong University
{denghq7,renqihan,1603023-zh,zqs1022}@sjtu.edu.cn

ICLR 2022 (rating: 8/10/8/8)

Preliminaries

- v DNN
- $N = \{1, \dots, n\}$ A set of n variables, e.g., an input image with n pixels
- $v(N)$ the network output of all input variables

Example: Given an image x, $v(x_S)$ can be implemented as any scalar output of the $\left(e.g., \log \frac{P(\hat{y}=y^{\text{truth}}|x_S)}{1-P(\hat{y}=y^{\text{truth}}|x_S)} \text{ of the true category}\right)$

Definition: m-order interaction

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

where

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

It quantifies the marginal effects (the importance) of the variable j that are changed by the presence or absence of the variable i, i.e., the collaboration between i, j in a context S.

Representation bottleneck

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

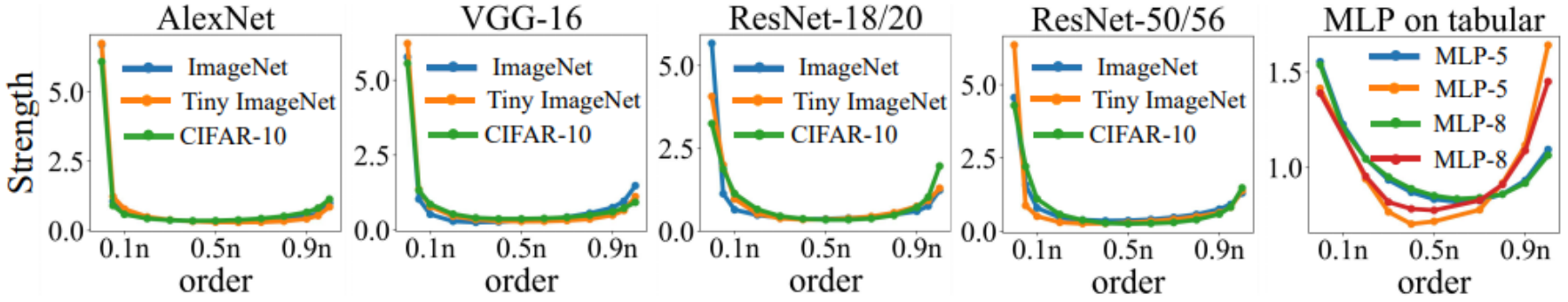
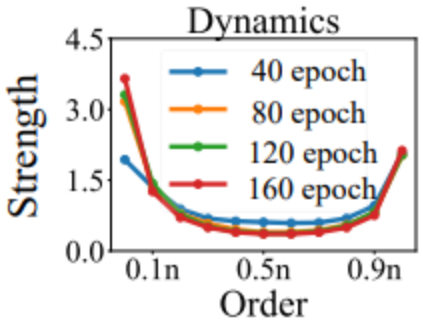
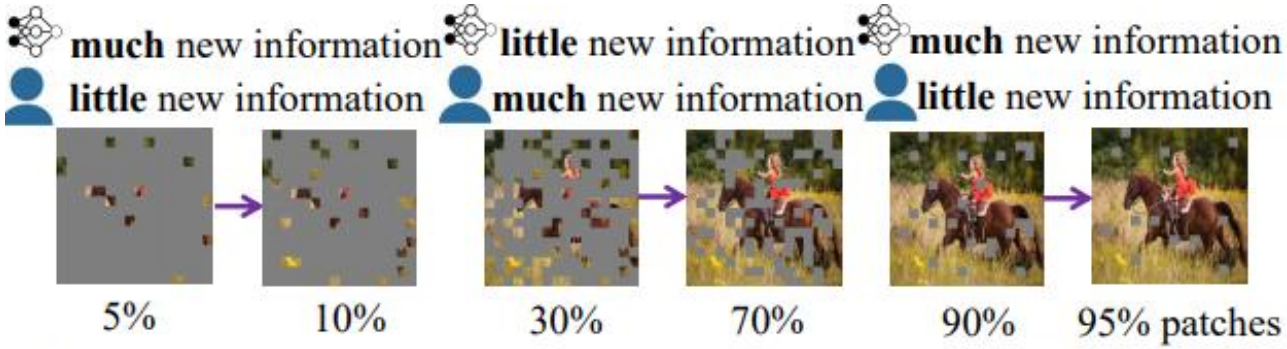


Figure 2: The distributions of interaction strength $J^{(m)}$ of different DNNs trained on various image datasets and tabular datasets.[2]



(a) Dynamics



ıeck(c) Whether humans/DNNs extract new information from patches.

Theoretical Analysis

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, $\mathbb{E}_{i,j}[\Delta W^{(m)}(i,j)] = \mathbf{0}$ and the variance of each dimension of $\Delta W^{(m)}(i,j)$ is $(\eta \frac{\partial L}{\partial v(N)} \frac{n-m-1}{n(n\ \ 1)})^2 \sigma^2 / \binom{n-2}{m}$. Therefore, $\mathbb{E}_{i,j}[\|\Delta W^{(m)}(i,j)\|_2^2] = K(\eta \frac{\partial L}{\partial v(N)} \frac{n-m-1}{n(n-1)})^2 \sigma^2 / \binom{n-2}{m}$, where K is the dimension of the network parameter W.

定理一证明了该权重改变量在各阶交互上的分量 $\Delta W^{(m)}(i,j)$ 的标准差正比于 $F^{(m)} = \frac{n-m-1}{n(n-1)} \sqrt{\binom{n-2}{m}}$。可以看出，权重改变量在中阶交互上的分量的标准差显著更小。即，中阶交互的学习强度更低，神经网络难以建模中阶交互，相对来说更易于建模低阶和高阶交互。

Control interactions of specific order

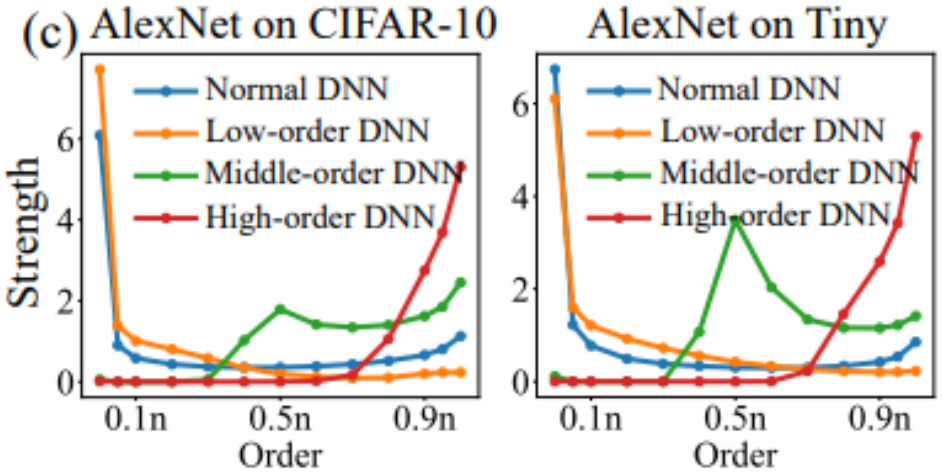
$$L^{+}(r_1, r_2) = -\frac{1}{|\Omega|}\sum\nolimits_{x\in\Omega}\sum\nolimits_{c=1}^{C} P(y^{*}=c|x)\log P(\hat{y}=c|\Delta u_c(r_1, r_2|x)),$$

$$L^{-}(r_1, r_2) = \frac{1}{|\Omega|}\sum\nolimits_{x\in\Omega}\sum\nolimits_{c=1}^{C} P(\hat{y}=c|\Delta u_c(r_1, r_2|x))\log P(\hat{y}=c|\Delta u_c(r_1, r_2|x)),$$

where $\Delta u_c(r_1, r_2|x) = v_c(S_2|x) - r_2/r_1 \cdot v_c(S_1|x)$

Such that $\emptyset \subseteq S_1 \subsetneq S_2 \subseteq N, |S_1| = r_1 n, |S_2| = r_2 n$, and $0 \leq r_1 < r_2 \leq 1$.

- $\mathcal{L}^{-}(r_1, r_2)$ usually could successfully **penalize** interactions of the $[r_1 n, r_2 n]$-th orders
- $\mathcal{L}^{+}(r_1, r_2)$ could **encourage** interactions of the $[r_1 n, r_2 n]$-th orders



Empirical Studies

Table 1: (left) Classification accuracies of four types of DNNs, including the normally trained DNNs, and the other three types of DNNs mainly encoding low-order, middle-order, and high-order

	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

The similar performance indicated that it was not necessary for a DNN to encode low-order interactions and high-order interactions to make inferences. Middle-order interactions could also provide discriminative information.

Empirical Studies

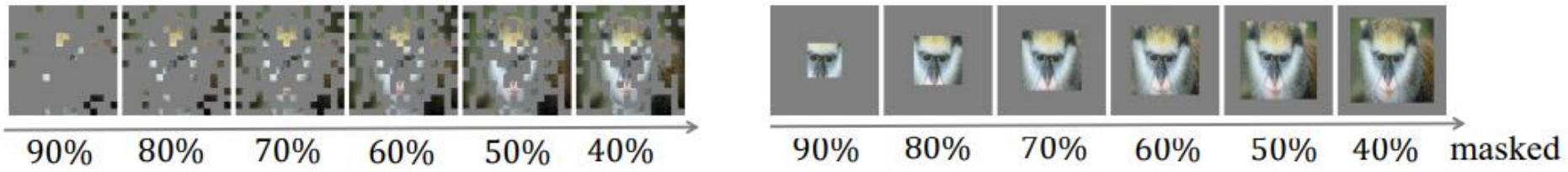


Figure 5: Tested images by random masking (left) and centrally-surrounding masking (right).

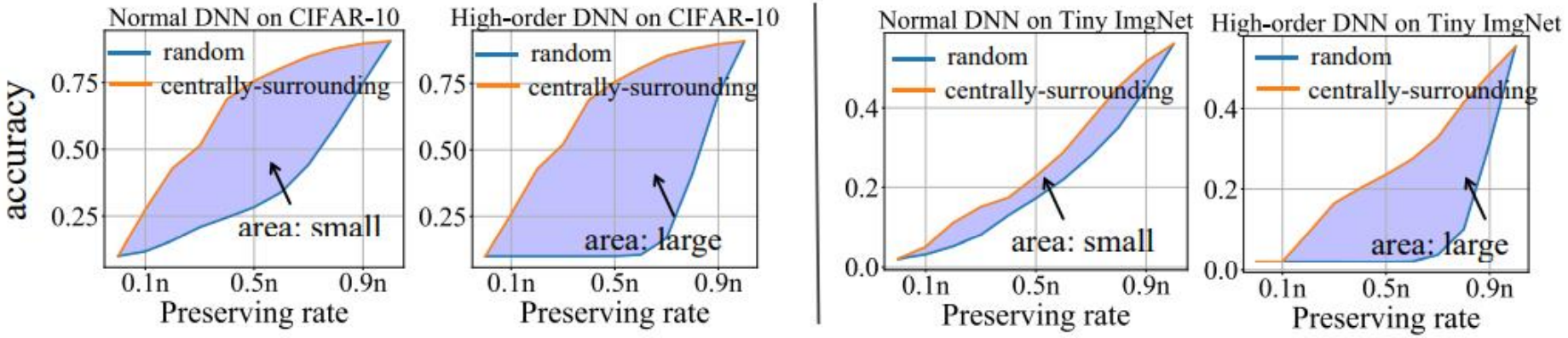


Figure 6: Classification accuracies using VGG-16 on images with different numbers of patches being masked. The Appendix C provides more results.

- Normally trained DNN usually encoded local patterns
- High-order DNN encoded more structural information

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

THANKS