Long-Tailed Recognition via Information-Preservable Two-Stage Learning

Fudong Lin and Xu Yuan

Department of Computer & Information Sciences
University of Delaware

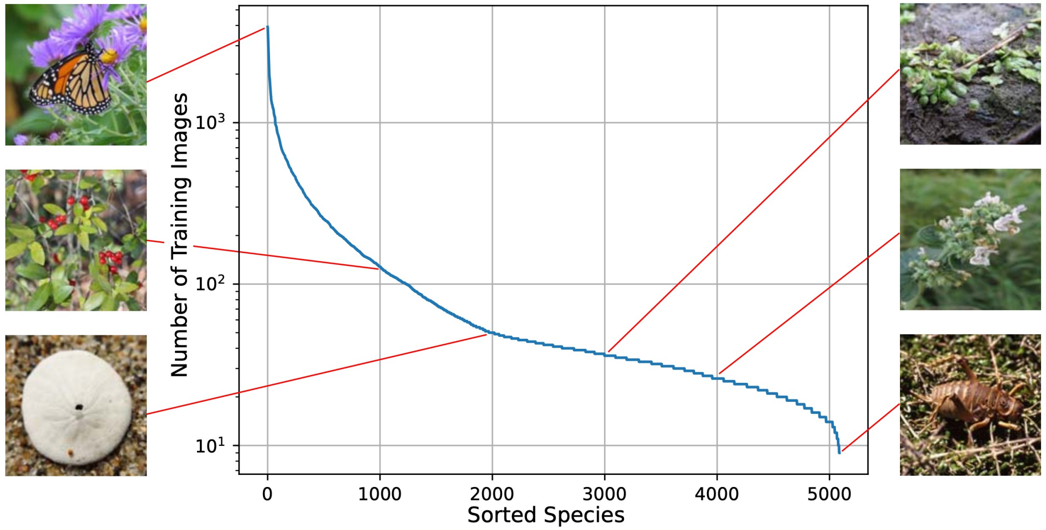
NEURAL INFORMATION PROCESSING SYSTEMS 2025

UNIVERSITY OF DELAWARE

Real-World Data Distribution

iNaturalist Dataset

Long-Tailed Data Distribution



Long-Tailed Data Distribution and Its Challenge

- Real-world data follows an imbalanced (or long-tailed) distribution.
- DNNs trained on highly skewed data suffer from majority-biased decision boundary.

Stage 1: BNS for Representation Learning

Our Key Idea and Objective

- Learn effective and well-separated feature space by capturing instance-level and class-level semantics
- Maximize mutual information between two feature spaces for the original and the augmented data, respectively

$$\operatorname*{argmin}_{\theta} MI(\mathbf{Q}(\mathbb{X}_Q;\theta), \mathbf{V}(\mathbb{X}_V;\theta)).$$

Balanced Negative Sampling

- Sample (m+1) positive pairs and n(m+1) negative pairs: $\mathcal{L}_{\text{BNS}} = -\frac{1}{m+1}\left[\sum_{q_* \in \{q_i\} \,\cup\, \mathbf{Q}^+_{i,m}} \log\sigma(\frac{\mathbf{q}_*^\top \mathbf{v}^+_{j,i}}{\tau}) + \sum_{q_* \in \{q_i\} \,\cup\, \mathbf{Q}^+_{i,m}} \sum_{j=1}^{n} \log\sigma(-\frac{\mathbf{q}_*^\top \mathbf{v}^-_{j}}{\tau})\right].$
- Capture instance-level and class-level semantics: $\mathcal{L}_{\text{BNS}} = -\frac{1}{m+1}\boxed{\left\{\underbrace{\log\sigma(\frac{\mathbf{q}_i^\top \mathbf{v}^+_{j,i}}{\tau}) + \sum_{j=1}^{n}\log\sigma(-\frac{\mathbf{q}_i^\top \mathbf{v}^-_{j}}{\tau})}_{\text{instance-level}} + \underbrace{\sum_{q_k \in \mathbf{Q}^+_{i,m}}\left[\log\sigma(\frac{\mathbf{q}_k^\top \mathbf{v}^+_{j,i}}{\tau}) + \sum_{j=1}^{n}\log\sigma(-\frac{\mathbf{q}_k^\top \mathbf{v}^-_{j}}{\tau})\right]}_{\text{class-level}}\right\}}$

Intra-Class Distance Mutual Information Theorem

Theorem 4.1. *(Intra-Class Distance Mutual Information Theorem) Let $\mathbb{X}^c_Q$ and $\mathbb{X}^c_V$ be two sets of images with the same label c, obtained by different data augmentation techniques. Given a feature extractor $f_{\boldsymbol{\theta}}(\cdot)$, we define $\boldsymbol{Q}^c$ and $\boldsymbol{V}^c$ as the representation spaces for $\mathbb{X}^c_Q$ and $\mathbb{X}^c_V$, respectively. Then, any pair of $\boldsymbol{q}^c_i \in \boldsymbol{Q}^c$ and $\boldsymbol{v}^c_j \in \boldsymbol{V}^c$ is a positive pair. Let $MI(\cdot)$ and $D(\cdot)$ respectively denote the mutual information and a distance metric, we have:*

$$\max MI(\boldsymbol{Q}^c, \boldsymbol{V}^c) \propto \min D(\boldsymbol{Q}^c, \boldsymbol{V}^c),$$

where $D(\boldsymbol{Q}^c, \boldsymbol{V}^c)$ can be considered as the intra-class distance because they have the same label.

Stage 2: IP-DPP for Imbalance Classification

Information-Preservable Determinantal Point Process (IP-DPP)

- Rectify majority-biased decision boundary by sampling balanced and mathematically informative instances:

$$\mathcal{P}_{\mathbf{S}}^{k}(\mathbb{Y}) = \frac{\det(\mathbf{S}_{\mathbb{Y}})}{\sum_{|\mathbb{Y}'|}\det(\mathbf{S}_{\mathbb{Y}'})}. \quad \longleftarrow \quad \mathbf{S}_{i,j} = \begin{cases} \frac{p(i)p(j)}{N}, & i \neq j \\ 1 - \sum_{k\neq j}\frac{p(k)p(j)}{N}, & i = j \end{cases}.$$

Mathematical Properties for a Valid DPP

Lemma 4.4. *(Positive Semi-definiteness) Let $p(i) = p(y_i|\boldsymbol{x}_i)$ represent the probability of y_i given $\boldsymbol{x}_i$. $\boldsymbol{S} \in \mathbb{R}^{N\times N}$ is a symmetric stochastic matrix, where each row (or column) sums to 1. Then, we have:*

$$\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v} \geq 0, \quad \forall \boldsymbol{v} \in \mathbb{R}^N.$$

In other words, $\boldsymbol{S}$ is positive semi-definite.

Lemma 4.5. *(Bounds on Eigenvalues) Let $\{\lambda_i\}_{i=1}^N$ be the eigenvalues of the symmetric stochastic matrix $\boldsymbol{S} \in \mathbb{R}^{N\times N}$, we have:*

$$0 \leq \lambda_i \leq 1, \quad \forall \lambda_i.$$

Theorem 4.6. *(Bounded Determinant Probability Measurement) Let $\mathbb{X}$ be a ground set with N items and $\boldsymbol{S} \in \mathbb{R}^{N\times N}$ denote a symmetric stochastic matrix, indexed by elements in $\mathbb{X}$. Here, $\boldsymbol{S}$ is positive semi-definite and satisfies $\boldsymbol{0} \preceq \boldsymbol{S} \preceq \boldsymbol{I}$, where $\boldsymbol{I}$ is the $N \times N$ identity matrix. Let $\boldsymbol{S}_{\mathbb{Y}}$ denote the principal submatrix of $\boldsymbol{S}$ corresponding to $\mathbb{Y}$, for any subset $\mathbb{Y} \subseteq \mathbb{X}$, the following holds:*

$$0 \leq \frac{\det(\boldsymbol{S}_{\mathbb{Y}})}{\det(\boldsymbol{S}+\boldsymbol{I})} \leq 1.$$

Information-Preservable Property

Remark 4.7 (Information-Preserving Sampling Principle). Let $I(\boldsymbol{x}) = -\log[p(y|\boldsymbol{x})]$ denote the information content of item $\boldsymbol{x}$ relevant to its correct classification. $\mathcal{P}_{\boldsymbol{S}}(\mathbb{Y} \cup \{\boldsymbol{x}\})$ denotes the probability that item $\boldsymbol{x}$ is sampled by our DPP approach, which prioritizes sampling elements with higher information content, as expressed by:

$$\mathcal{P}_{\boldsymbol{S}}\left(\mathbb{Y} \cup \{\boldsymbol{x}\}\right) \propto I(\boldsymbol{x}).$$

Comparison to Baseline Methods

Experimental Settings

- **Dataset:** CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018
- **Baseline Method:** Focal Loss, LDAM Loss, τ-norm, RIDE, KCL, TSC, SBCL, OTmix, and DisA
- **Metric:** Many-shot, Medium-shot, Few-shot, and Overall Accuracies

Performance Results on Small-Scale Datasets

Methods	CIFAR-10-LT				CIFAR-100-LT			
	Many-shot	Medium-shot	Few-shot	Overall	Many-shot	Medium-shot	Few-shot	Overall
Focal Loss	86.3	60.6	46.3	69.2	71.1	43.9	10.5	43.5
LDAM Loss	85.8	64.8	51.9	71.5	71.4	44.5	11.7	44.1
τ-norm	85.2	64.4	51.7	70.9	60.7	54.4	14.8	44.7
RIDE	86.2	63.6	56.1	73.4	73.1	47.6	16.4	47.2
KCL	83.7	63.8	53.6	71.7	72.3	46.1	14.8	45.8
TSC	81.5	71.9	56.3	71.9	71.3	43.9	10.5	43.5
SBCL	81.6	72.4	57.6	72.6	72.7	48.5	20.0	48.5
OTmix	**87.9**	67.8	47.3	73.8	**73.1**	48.0	19.1	48.1
DisA	86.1	68.3	50.3	73.6	72.4	49.3	21.9	49.2
Ours	82.0	**76.3**	**67.2**	**76.4**	62.4	**59.7**	**31.9**	**52.4**

Performance Results on Large-Scale Datasets

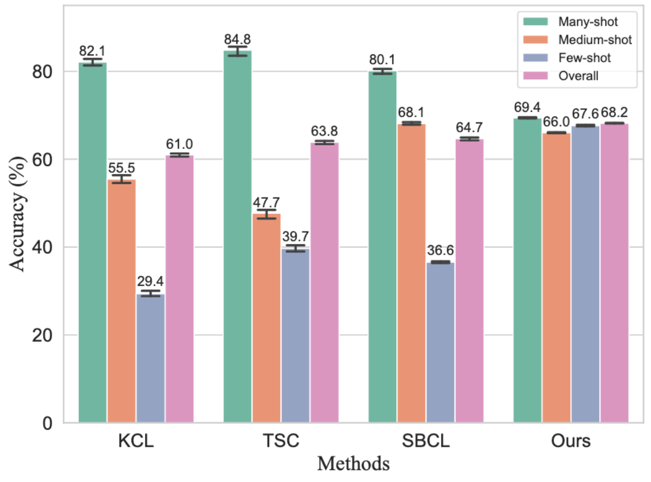
Methods	ImageNet-LT				iNaturalist 2018			
	Many-shot	Medium-shot	Few-shot	Overall	Many-shot	Medium-shot	Few-shot	Overall
Focal Loss	51.4	41.2	16.0	41.7	61.2	62.7	64.4	63.2
LDAM Loss	55.0	46.4	16.7	45.7	65.1	66.8	61.7	64.6
τ-norm	56.6	44.2	27.4	46.7	71.3	65.8	69.1	67.7
RIDE	56.7	46.4	25.7	47.6	67.5	68.6	69.3	68.8
KCL	55.0	42.6	25.4	45.0	61.2	62.7	64.4	63.2
TSC	57.1	45.2	29.3	47.6	66.4	65.7	64.0	65.1
SBCL	55.8	45.7	27.1	47.1	**73.4**	70.2	69.8	70.4
OTmix	50.9	46.0	25.7	45.1	70.1	70.9	68.6	69.9
DisA	**61.0**	47.0	25.3	49.4	70.7	70.8	68.4	69.8
Ours	59.7	**50.8**	**32.4**	**51.7**	72.7	**72.9**	**75.7**	**74.0**

Our approach outperformed baseline methods on both small- and large-scale datasets.

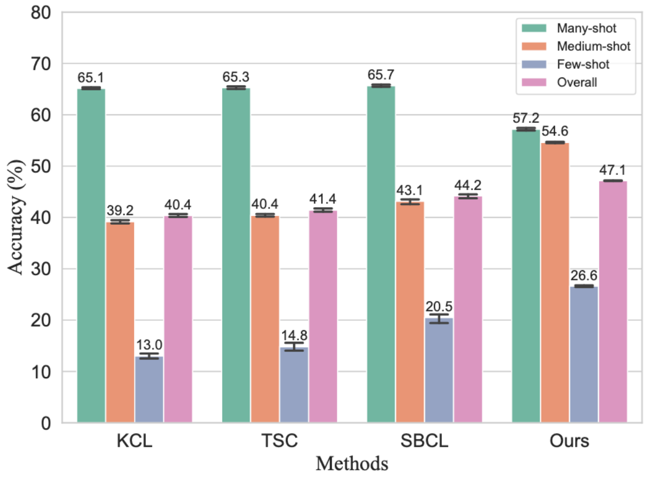
Evaluation on Representation Learning

- **Dataset:** CIFAR-10-LT and CIFAR-100-LT
- **Baseline Method:** KCL, TSC, and SBCL
- **Metric:** Many-shot, Medium-shot, Few-shot, and Overall Accuracies

Linear Probing Accuracy

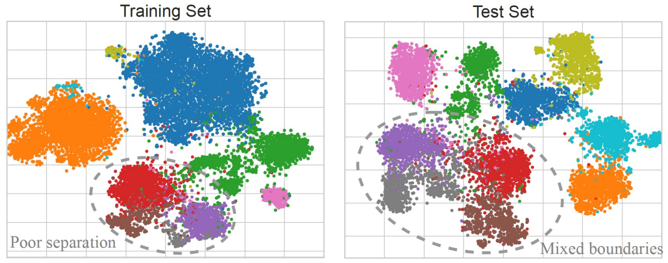


(a) CIFAR-10-LT

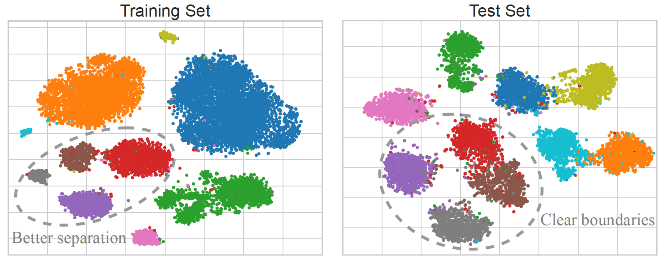


(b) CIFAR-100-LT

t-SNE Visualization on CIFAR-10-LT



(a) SBCL (b) SBCL



(c) Ours (d) Ours

Our approach learned effective and well-separated feature space.

Conclusion

- This work has proposed an information-preservable two-stage learning approach for addressing the long-tailed data distribution in real scenarios.
- Our key contributions include Balanced Negative Sampling (BNS) and Information-Preservable Determinantal Point Process (IP-DPP).
- BNS learned effective and well-separated feature space by capturing instance-level and class-level semantics. Meanwhile, IP-DPP rectified majority-biased decision boundary by sampling mathematically informative instances.
- Our approach demonstrates theoretical soundness and empirical effectiveness in addressing imbalanced classification.

Paper **Code**

Thank you!
Q & A