南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Breaking Long-Tailed Learning Bottlenecks: A Controllable Paradigm with Hypernetwork-Generated Diverse Experts

Zhe Zhao[1,3], **Haibin Wen**[5], **Zikang Wang**[1], **Pengkun Wang**[1,2]*, **Fanfu Wang**[6], **Song Lai**[3], **Qingfu Zhang**[3], **Yang Wang**[1,2]*

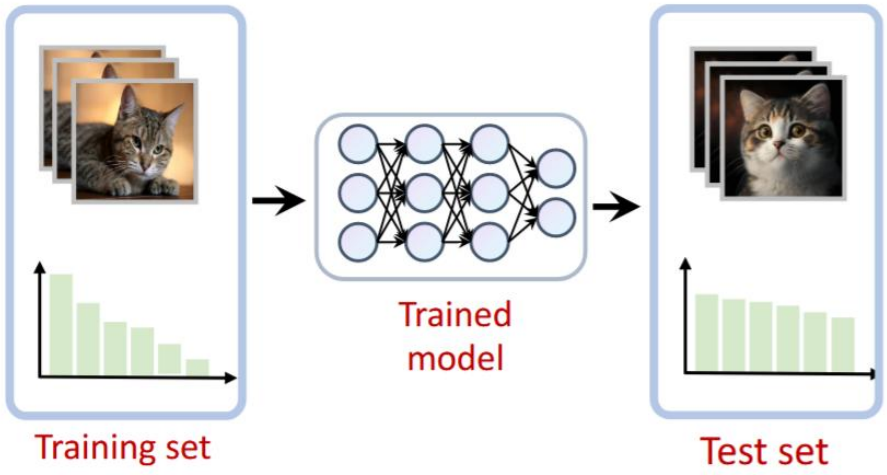
[1]University of Science and Technology of China (USTC), Hefei, China
[2]Suzhou Institute for Advanced Research, USTC, Suzhou, China
[3]City University of Hong Kong, Hong Kong, China
[4]Harbin Institute of Technology, Harbin, China
[5]MorongAI, Suzhou, China
[6]Lanzhou University, Lanzhou, China

NeurIPS2024

Background

Long-tailed methods: **re-sampling** and modifying the **loss function** focusing on improving the performance of the tail classes.

Assume that the distributions of the training and test data remain invariant. Cannot well handle the **distribution shift** between training and testing in real-world scenarios.



(a) Traditional long-tailed learning method

Background

Some more recent works propose using multiple expert models to obtain stronger distribution adaptability. That is, the models learned by various losses are skilled in handling class distributions with different skewness.

The forward expert: directly simulates the original long-tailed training class distribution:

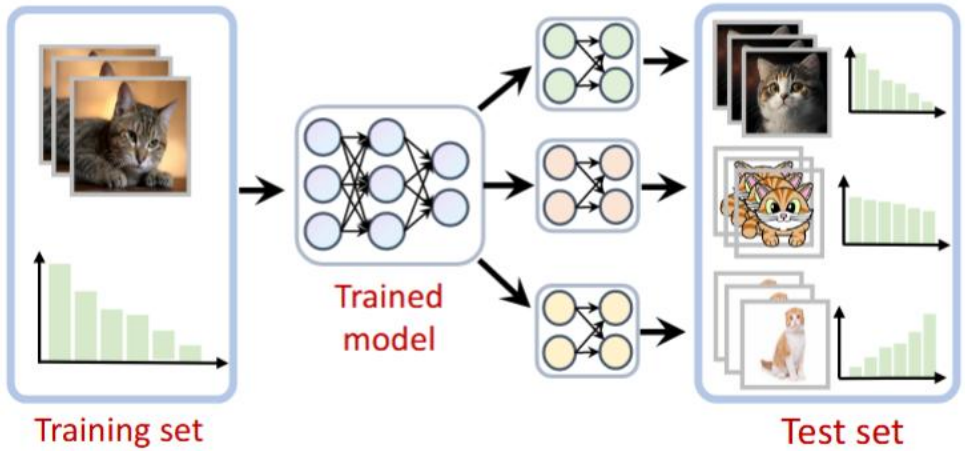
$$\mathcal{L}_{\text{ce}} = \frac{1}{n_s} \sum_{x_i \in \mathcal{D}_s} -y_i \log \sigma(v_1(x_i)),$$

The uniform expert: simulate the uniform class distribution

$$\mathcal{L}_{\text{bal}} = \frac{1}{n_s} \sum_{x_i \in \mathcal{D}_s} -y_i \log \sigma(v_2(x_i) + \log \pi).$$

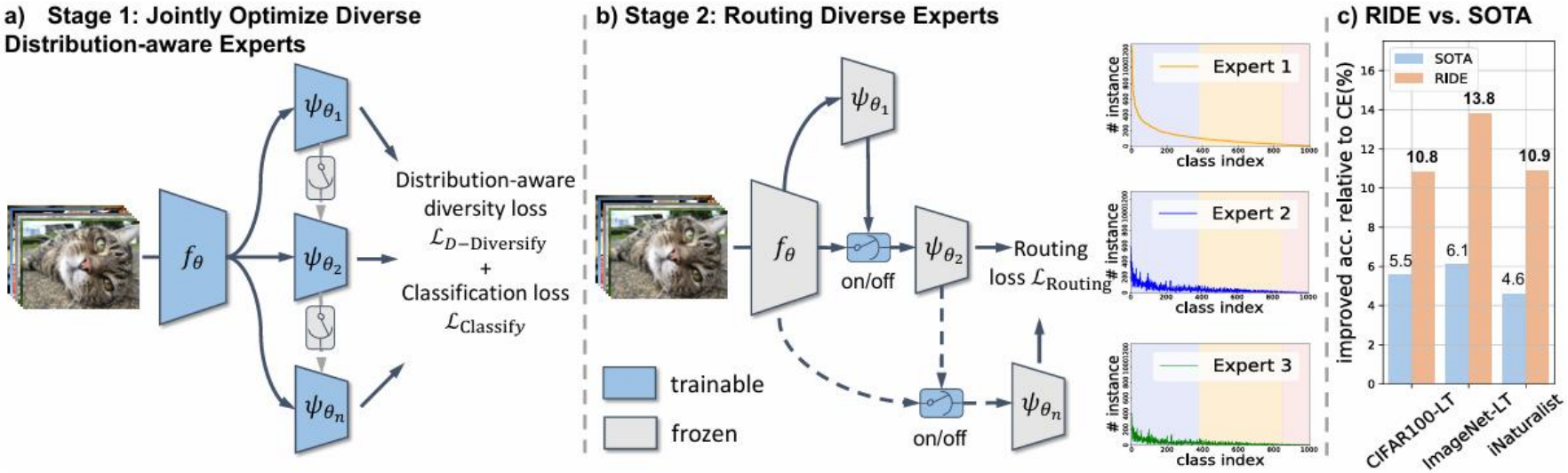
The backward expert (inverse softmax loss): simulate the inversely long-tailed class distribution:

$$\mathcal{L}_{\text{inv}} = \frac{1}{n_s} \sum_{x_i \in \mathcal{D}_s} -y_i \log \sigma(v_3(x_i) + \log \pi - \lambda \log \bar{\pi}),$$



(b) Multi-expert long-tailed learning method

Background



RIDE: train a router that dynamically assigns ambiguous samples to additional experts on an as-needed basis. The distribution of instances seen by each expert shows that head instances need fewer experts and the imbalance between classes gets reduced for later experts.

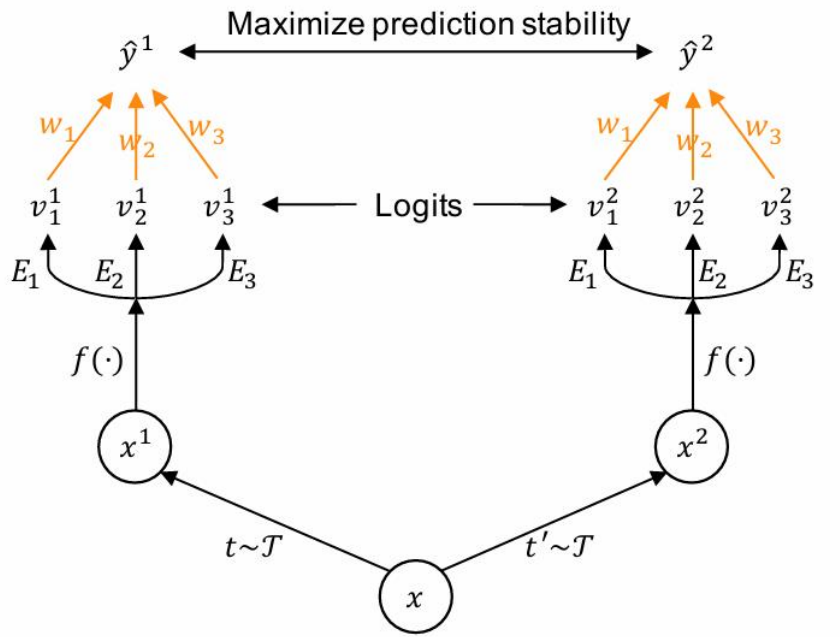
Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. ICLR 2021

Background

SADE (Test-time self-supervised):

Empirical observation: stronger experts have higher prediction similarity between different views of samples from their favorable classes.

alleviate the problem of distribution mismatch between training and testing to some extent.



Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. NeuraIPS 2022

Introduction

Goal: Maximize the overall performance.

New requirements: users may have **different preferences** and needs for the relative trade-off between head and tail classes.

Examples:

lung cancer, we may also be willing to moderately increase the false positive rate in exchange for higher coverage of the tail classes, to ensure that no patients are missed.

nature reserves, we want the model to accurately detect common species (i.e., head classes) to understand their population sizes. But when looking for rare species (i.e., tail classes), we care more about covering all species, even at the cost of some false detections.

Contribution: a new interpretable and controllable long-tailed learning method that can acquire the ability to overcome test distribution shift from a single distribution dataset and satisfy user preferences in any shifted distribution scenario.

Method

$$R_{test}(f) = R_m(f) + \sum_{i=1}^{K}(\pi_i^{test} - \pi_i^m) \cdot \mathbb{E}_{\boldsymbol{x}\sim P_{test}(\boldsymbol{x}|y=i)}[\ell(f(\boldsymbol{x};\boldsymbol{\theta}), i)]$$

$$R_{test}(\hat{f}) \leq \frac{1}{N}\sum_{m=1}^{N} R_m(f_m) + 2M \cdot \left(\frac{1}{N}\sum_{m=1}^{N}\delta(\mathcal{E}_m, \mathcal{E}_{test}) + \frac{N-1}{N}\Delta(\mathcal{E}_1, \ldots, \mathcal{E}_N)\right)$$

where $M = \max_{i,\boldsymbol{x}} \ell(\hat{f}(\boldsymbol{x}), i)$.

$$\delta(\mathcal{E}_i, \mathcal{E}_j) = \frac{1}{2}\sum_{k=1}^{K}|\pi_k^i - \pi_k^j|$$

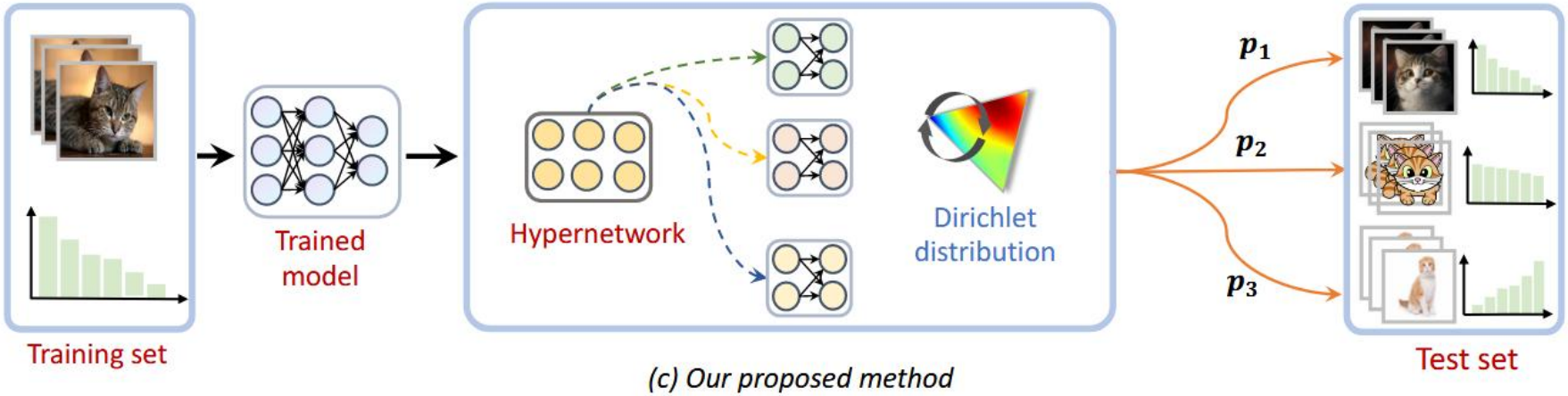
1. the average empirical risk of all experts
2. the average total variation distance between the training environments and the test environment.
3. the weighted average of ETVD among the training environments.

Environment's Total Variation Distance (ETVD)

$$\Delta(\mathcal{E}_1, \ldots, \mathcal{E}_M) = \max_{i,j\in\{1,\ldots,M\}} \delta(\mathcal{E}_i, \mathcal{E}_j)$$

the experts method capture the distributional characteristics of different environments, which can reduce the distribution discrepancy between the training environments and the test environment

Method



(c) Our proposed method

Training phrase:

$$f_i(\mathbf{x}) = g_{w_i}(\phi_\theta(\mathbf{x})), \quad i = 1, \ldots, T$$

To generate diverse experts, we introduce a hypernetwork

$$\boldsymbol{\alpha}_i \sim \mathrm{Dir}(\boldsymbol{\beta}_i),\ \boldsymbol{\beta}_i \in \mathbb{R}_+^T \qquad \hat{\boldsymbol{r}}_i = \frac{\boldsymbol{r} \odot \boldsymbol{\alpha}_i}{\boldsymbol{r}^\top \boldsymbol{\alpha}_i}$$

$$\mathbf{w}_i = h_\psi(\hat{\boldsymbol{r}}_i), \quad i = 1, \ldots, T$$

- $\beta_1 > \beta_2 = \beta_3$ simulate the long-tailed distribution.
- $\beta_1 = \beta_2 = \beta_3$ simulate the uniform distribution.
- $\beta_1 < \beta_2 = \beta_3$ simulate the inverse long-tailed distribution.

$\min_{\boldsymbol{\Theta}} \left(\sum_{i=1}^{T} \mathcal{L}_i(\boldsymbol{\Theta}, \mathcal{D}) + \lambda \cdot \log \sum_{i=1}^{T} \exp\left(\frac{1}{\lambda} \mathcal{L}_i(\boldsymbol{\Theta}, \mathcal{D}) \right) \right)$ promotes diversity among experts.

Smooth Tchebycheff scalarization for multi-objective optimization ICLR. 2024.

Method

Testing phrase:

user-input preference vector: $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \alpha_3^*)^\top$

$$\hat{\boldsymbol{r}} = \frac{\boldsymbol{r} \odot \boldsymbol{\alpha}^*}{\boldsymbol{r}^\top \boldsymbol{\alpha}^*}$$

the value space of the preference vector

the model's performance on three distributions

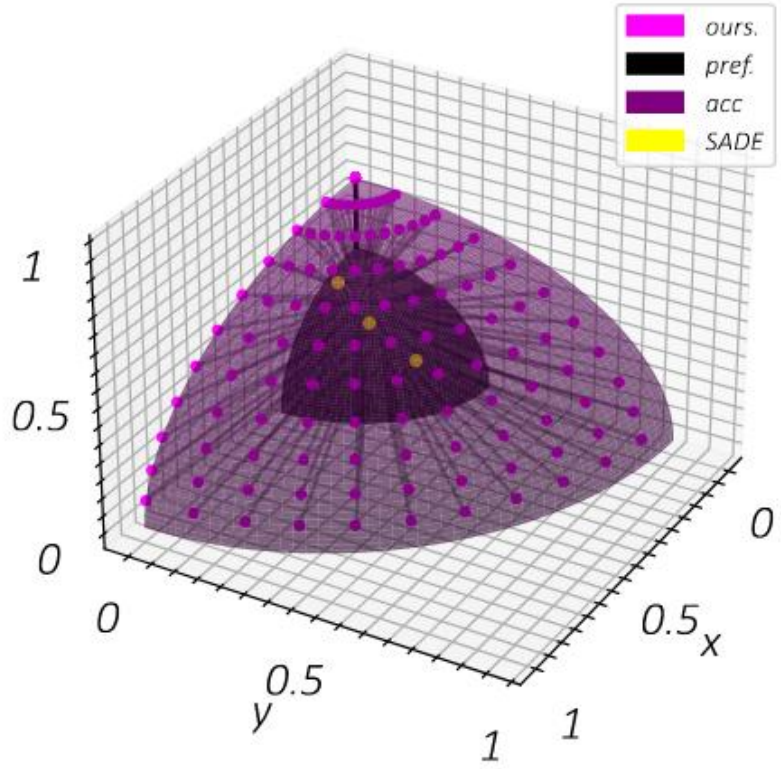


Figure 2: Mapping from preference to model properties.

Experiment

Table 1: Top-1 accuracy on CIFAR100-LT, Places-LT, iNaturalist 2018, and ImageNet-LT, where the test class distribution is uniform.

Method	CIFAR100-LT			Places-LT	iNaturalist 2018	ImageNet-LT
	IR=10	IR=50	IR=100			
Softmax	59.1	45.6	41.4	31.4	64.7	48.0
Causal [28]	59.4	48.8	45.0	32.2	64.4	50.3
Balanced Softmax [15]	61.0	50.9	46.1	39.4	70.6	52.3
MiSLAS [41]	62.5	51.5	46.8	38.3	70.7	51.4
LADE [12]	61.6	50.1	45.6	39.2	69.3	52.3
RIDE [32]	61.8	51.7	48.0	40.3	71.8	56.3
SADE [38]	63.6	53.8	48.8	40.9	72.7	58.8
LSC [33]	65.0	56.5	51.8	41.3	73.9	60.2
BalPoE [1]	64.8	56.3	52.0	40.8	75.0	59.3
PRL(ours)	**65.6**	**57.3**	**52.8**	**41.6**	**75.1**	**60.8**

Experiment

Table 2: Top-1 accuracy on CIFAR100-LT (IR100) with various unknown test class distributions.

Method	Prior	Forward-LT					Uni.	Backward-LT				
		50	25	10	5	2	1	2	5	10	25	50
Softmax	✗	63.3	62.0	56.2	52.5	46.4	41.4	36.5	30.5	25.8	21.7	17.5
BS	✗	57.8	55.5	54.2	52.0	48.7	46.1	43.6	40.8	38.4	36.3	33.7
MiSLAS	✗	58.8	57.2	55.2	53.0	49.6	46.8	43.6	40.1	37.7	33.9	32.1
LADE	✗	56.0	55.5	52.8	51.0	48.0	45.6	43.2	40.0	38.3	35.5	34.0
LADE	✓	62.6	60.2	55.6	52.7	48.2	45.6	43.8	41.1	41.5	40.7	41.6
RIDE	✗	63.0	59.9	57.0	53.6	49.4	48.0	42.5	38.1	35.4	31.6	29.2
SADE	✗	65.2	62.5	58.8	55.4	51.2	48.8	43.0	43.9	42.4	42.2	42.0
LSC	✗	67.8	64.2	60.2	58.1	53.2	51.6	44.7	45.7	44.2	44.7	48.0
BalPoE	✗	69.0	65.2	61.2	59.0	54.2	51.7	45.7	46.6	45.2	45.2	45.8
PRL (ours)	✗	**69.5**	**65.7**	**61.7**	**59.5**	**54.7**	**52.2**	**46.2**	**47.1**	**45.7**	**45.7**	**48.5**

Experiment

Table 4: Control of trade-off preference for long-tailed classes with different preferences, **bold text**, underlined text, and dashed underline respectively indicate the highest performance of the head, middle, and tail classes in this line.

Dist.		R=(1.0, 2.7)			R=(0.5, 2.5)			R=(1.9, 1.1)		
		Many	Middle	Few	Many	Middle	Few	Many	Middle	Few
Forward	50	**61.4**	50.4	36.5	61.0	52.6	31.5	61.1	48.9	40.3
	25	**61.6**	48.3	28.4	60.6	49.6	31.5	59.7	49.4	33.1
Uni	1	61.6	51.4	33.2	61.6	51.5	33.2	61.6	51.4	33.2
Backward	25	**63.8**	49.4	31.1	60.2	48.2	32.1	63.2	48.2	32.2
	50	**66.6**	47.1	30.6	66.1	48.9	30.9	64.6	47.8	31.7

Experiment

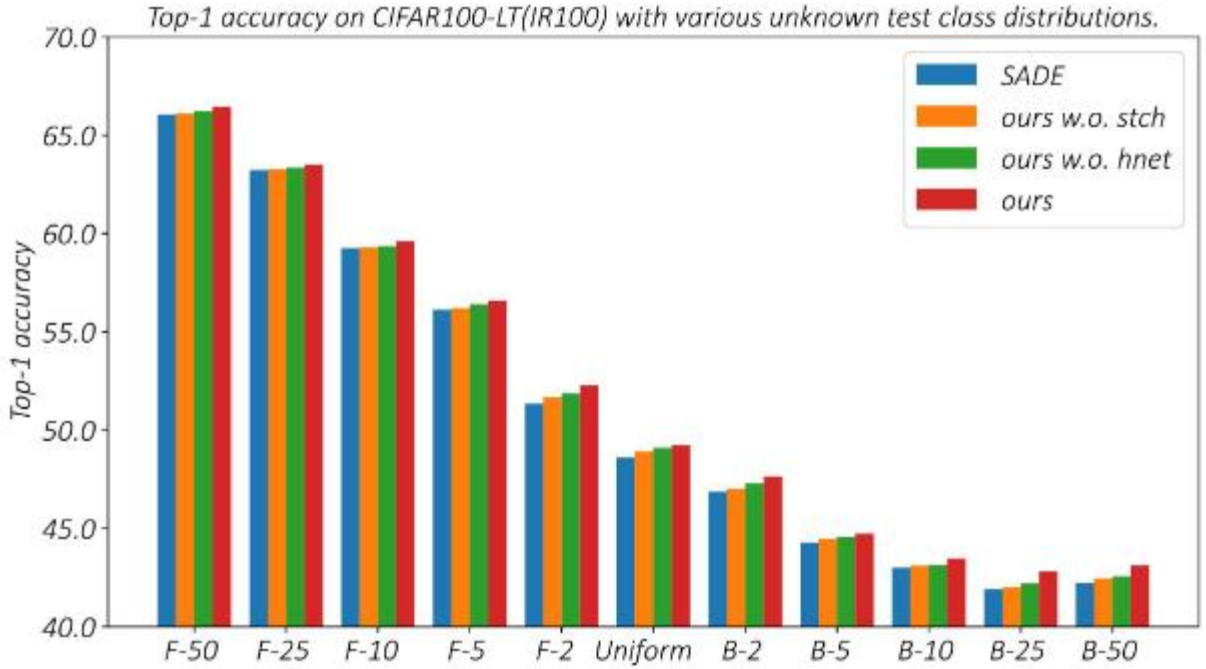


Figure 5: Ablation analysis, including the ablation of the hypernetwork and Chebyshev polynomials.

Thank you