Learning Label Shift Correction for Test-Agnostic Long-Tailed Recognition

Tong Wei[1] Zhen Mao[1] Zi-Hao Zhou[1] Yuanyu Wan[2] Min-Ling Zhang[1]

[1]Southeast University [2]Zhejiang University

Background

- Real-world data often exhibits a **long-tail class distribution**.
- Moreover, test label distribution may change across different tasks, a.k.a. test-agnostic long-tail learning (Figure 1).

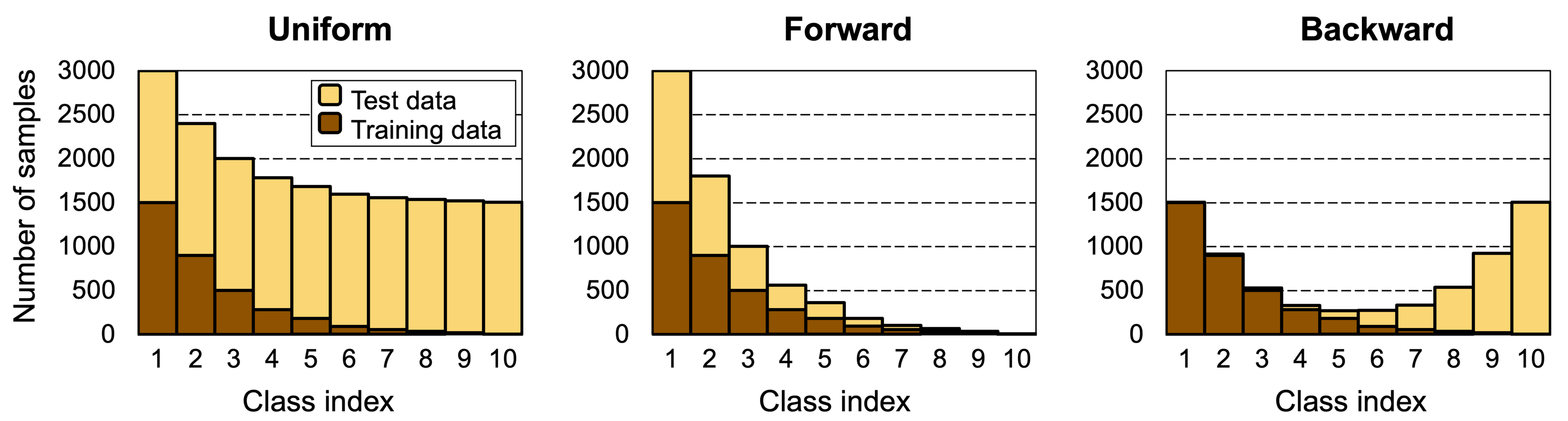

Figure 1. Test-agnostic long-tail learning.

Adjusting Model Predictions Helps Reduce Generalization Error

Denote $\mathbb{P}_{\mathcal{D}_S}(Y=y)$ and $\mathbb{P}_{\mathcal{D}_T}(Y=y)$ the label distributions of train and test domain, respectively. We construct the "post-adjusted" model outputs for each sample x:

$$\widetilde{f}_y(x) = f_y(x) + \log\left(\frac{\mathbb{P}_{\mathcal{D}_T}(Y=y)}{\mathbb{P}_{\mathcal{D}_S}(Y=y)}\right), \quad y \in [K]. \tag{1}$$

If we can accurately estimate $\mathbb{P}_{\mathcal{D}_T}(Y=y)$, we can seamlessly adapt pre-trained models to the specific test dataset using Eq. eq. (1).

Starting with an estimated test label distribution $\widehat{\mathbb{P}}_{\mathcal{D}_T}(Y)$, our adapted model $\widetilde{f}$ induces a hypothesis:

$$\begin{aligned} h_{\widetilde{f}}(x) &= \arg\max_{y\in[K]} f_y(x) + \log\left(\frac{\widehat{\mathbb{P}}_{\mathcal{D}_T}(Y=y)}{\mathbb{P}_{\mathcal{D}_S}(Y=y)}\right) \\ &= \arg\max_{y\in[K]} \widehat{\mathbb{P}}(y \mid x)\frac{\widehat{\mathbb{P}}_{\mathcal{D}_T}(Y=y)}{\mathbb{P}_{\mathcal{D}_S}(Y=y)}. \end{aligned} \tag{2}$$

Theorem (Error gap between $h_{\widetilde{f}}$ and Bayes-optimal classifier)

Given an estimated label distribution of test data $\widehat{\mathbb{P}}_{\mathcal{D}_T}(Y)$, a pre-trained scoring function f, and a hypothesis $h_{\widetilde{f}}$ induced by $\widetilde{f}$, we can bound the error gap by:

$$\begin{aligned} \epsilon_T(h_{\widetilde{f}}) - \epsilon_T(h^*) \le & \left\|\widehat{\mathbb{P}}(Y \mid X) - \mathbb{P}_{\mathcal{D}_S}(Y \mid X)\right\|_{L^1,w} \\ & + \mathrm{BPE}(h_f)\left\|\widehat{\mathbb{P}}_{\mathcal{D}_T}(Y) - \mathbb{P}_{\mathcal{D}_T}(Y)\right\|_{L^1}, \end{aligned}$$

where $w = \left(\frac{\widehat{\mathbb{P}}_{\mathcal{D}_T}(Y=1)}{\mathbb{P}_{\mathcal{D}_S}(Y=1)}, \frac{\widehat{\mathbb{P}}_{\mathcal{D}_T}(Y=2)}{\mathbb{P}_{\mathcal{D}_S}(Y=2)}, \cdots, \frac{\widehat{\mathbb{P}}_{\mathcal{D}_T}(Y=K)}{\mathbb{P}_{\mathcal{D}_S}(Y=K)}\right)$, *and* $\mathrm{BPE}(h_f) = \max_{y\in[K]} \mathbb{P}_{\mathcal{D}_S}(h_f(X) \neq y \mid Y=y)$ *denotes balanced posterior error.*

Learning Label Shift Correction

We introduce a simple estimation method that employs a shallow neural network within the framework of generalized blackbox shift estimation:

- **STEP 1:** Train a neural estimator by simulating various label distributions using the training dataset. The neural network takes the predicted logits from any pre-trained model as input and learns to approximate the true label distribution of these constructed subsets of training data.
- **STEP 2:** During testing time, the neural estimator provides an estimation of the test label distribution, which is used to adjust the pre-trained model's outputs.

Algorithm 1 Meta algorithm for label shift correction

Input: Training data: (X_S, Y_S), unlabeled test data: X_T, pre-trained model f

{`Sample from training data by varying class priors for` Q `times`}

Initialize $\widetilde{S} = \emptyset$
for $q = 1$ to Q **do**
  $(\widetilde{X}, \widetilde{Y}) \leftarrow$ SampleByClassPrior(X_S, Y_S, π^q)
  Compute class-wise average logits by $\widetilde{Z} = f(\widetilde{X})$ and $\widetilde{z}_{\text{avg}} = \frac{1}{|\widetilde{X}|}\sum_{i=1}^{|\widetilde{X}|} \widetilde{Z}_i$
  $\widetilde{S} = \widetilde{S} \cup (\widetilde{z}_{\text{avg}}, \pi^q)$
end for
Train *neural estimator* g_θ on $\widetilde{S}$ by minimizing $\mathcal{L}(\widetilde{S}, g_\theta) = \frac{1}{|\widetilde{S}|}\sum_{(\widetilde{z},\pi^q)\in\widetilde{S}} \ell(\pi^q, g_\theta(\widetilde{z}))$
Obtain predicted logits for test data using the pre-trained model by $Z_T \leftarrow f(X_T)$
Apply adaptive logits clipping on Z_T with the value of k set by Eq. (3) and obtain $\widehat{Z}_T$
Estimate test label distribution by $\widetilde{\pi} \leftarrow g_\theta(\widehat{z}_T)$, where $\widehat{z}_T$ is the class-wise average of $\widehat{Z}_T$

Output: Adjusted predictions $\widehat{Y}_T = \arg\max(Z_T + \log\widetilde{\pi})$

Overconfidence of Base Models on Tail Classes

In practice, f can be achieved by many long-tail learning methods. Intriguingly, we discover that these methods tend to produce overconfident logits for tail classes while inhibiting head classes. The bias towards tail classes can lead to undesirable label distribution predictions by **neural estimator**.

To rectify the bias, we introduce **logit clipping**, which truncates the small predicted logits for each sample to zero. The parameter k controls how many of the smallest logits are clipped to zero. Specifically, we determine k based on a comparison between head and tail classes:

$$\begin{aligned} k &= \arg\max_{k\in\mathcal{K}} \mathbb{I}(\pi_0^h > \lambda\pi_0^t)\widehat{Z}^h + \mathbb{I}(\pi_0^h < \lambda\pi_0^t)\widehat{Z}^t, \\ \text{s.t.} \quad & \widehat{Z} = logitClip(Z, k). \end{aligned} \tag{3}$$

Theorem (Bayes error when using pseudo-label for estimation)

Given a hypothesis h_f, let $C_{h_f(X)|Y} \in \mathbb{R}^{K\times K}$ denote the conditional confusion matrix, i.e., $C_{h_f(X)|Y}(i,j) = \mathbb{P}(h_f(X) = i \mid Y = j)$. Suppose $C_{h_f(X)|Y}$ is invertible and the test label distribution π is sampled uniformly at random, the error of Bayes function g^* holds following inequality:

$$\frac{K-1}{K(M+K+1)} \le \epsilon_L(g^*) \le \frac{K-1}{K(M+K+1)|\det(C_{h_f(X)|Y})|\sigma^2_{\min}}. \tag{4}$$

Empirical Results

- Our method sets new state-of-the-art on commonly used long-tail learning datasets.
- Our method can be seamlessly integrated with many existing models.
- Our method can tackle both offline and online settings.

Table 1. Test accuracy (%) on CIFAR100-LT (ResNet32), ImageNet-LT (ResNeXt50), and Places-LT (ResNet152). Prior: test class distribution. ∗: Prior estimated from test data.

		CIFAR100-LT-100					ImageNet-LT					Places-LT				
		Forward		Uni.	Backward		Forward		Uni.	Backward		Forward		Uni.	Backward	
Method	Prior	50	5	1	5	50	50	5	1	5	50	50	5	1	5	50
Softmax	✗	63.3	52.5	41.4	30.5	17.5	66.1	56.6	48.0	38.6	27.6	45.6	38.0	31.4	25.4	19.4
MiSLAS	✗	58.8	53.0	46.8	40.1	32.1	61.6	56.3	51.4	46.1	39.5	40.9	39.6	38.3	36.7	34.4
LADE	✗	56.0	51.0	45.6	40.0	34.0	63.4	57.4	52.3	46.8	40.7	42.8	40.8	39.2	37.6	35.7
RIDE	✗	63.0	53.6	48.0	38.1	29.2	67.6	61.7	56.3	51.0	44.0	43.1	42.0	40.3	38.7	36.9
PaCo	✗	62.0	57.6	52.2	47.0	40.7	66.6	62.7	58.9	54.1	48.7	-	-	-	-	-
SADE	✗	58.4	53.1	49.4	42.6	35.0	65.5	62.0	58.8	54.7	49.8	-	-	-	-	-
BalPoE	✗	65.1	54.8	**52.0**	44.6	36.1	67.6	63.3	59.8	55.7	50.8	-	-	-	-	-
BBSE	*	63.9	48.3	20.5	30.1	24.1	63.5	54.9	48.2	42.5	36.3	43.0	36.2	30.9	26.2	20.5
RLLS	*	67.2	53.8	41.7	29.3	16.4	65.2	55.0	45.3	35.2	23.6	43.4	35.1	27.9	20.9	13.5
MLLS	*	65.6	54.4	46.0	38.8	33.9	60.9	52.1	46.3	41.7	39.0	41.8	35.1	30.5	26.6	22.9
LADE	✓	62.6	52.7	45.6	41.1	41.6	65.8	57.5	52.3	48.8	49.2	46.3	41.2	39.4	39.9	43.0
SADE	*	65.9	54.8	49.8	44.7	42.4	69.4	63.0	58.8	55.5	53.1	46.4	42.6	40.9	41.1	41.6
LSC (ours)	*	**68.1**	**58.4**	51.9	**46.0**	**48.3**	**72.3**	**65.6**	**60.5**	**58.2**	**57.3**	**47.7**	**43.7**	**41.4**	**41.5**	**44.4**

Table 2. Test accuracy (%) by combining LSC with existing methods on CIFAR100-LT.

	Forward		Uniform	Backward	
Methods	50	5	1	5	50
RIDE	64.1	55.9	48.6	40.8	31.5
RIDE + LSC	**66.2**	**56.2**	48.6	**41.3**	**33.2**
PaCo	62.0	57.6	52.2	47.0	40.7
PaCo + LSC	**63.3**	**57.7**	52.2	**47.6**	**42.0**
NCL	66.4	59.8	54.3	48.0	41.4
NCL + LSC	**71.5**	**61.1**	54.3	**49.8**	**47.9**
SHIKE	67.8	60.1	53.8	46.6	38.4
SHIKE + LSC	**70.3**	**60.5**	53.8	**48.8**	**43.2**

Table 3. Test accuracy (%) on ImageNet-LT in the online setting with varying batch size.

	Forward		Uniform	Backward	
Methods (setting)	50	5	1	5	50
No adaptation	70.9	65.6	60.5	55.1	48.4
Offline model (ours)	72.3	65.6	60.5	58.2	57.3
SADE ($B=64$)	68.7	63.2	58.8	55.2	51.9
Ours ($B=64$)	**71.8**	**65.7**	**60.8**	**56.5**	**52.8**
SADE ($B=8$)	69.7	63.1	58.8	55.5	**53.0**
Ours ($B=8$)	**71.8**	**65.7**	**60.8**	**56.5**	52.8
SADE ($B=1$)	69.7	63.1	58.5	55.2	**52.9**
Ours ($B=1$)	**71.9**	**65.7**	**60.7**	**56.1**	52.3

Take-Home Messages

诚聘英才
东南大学计软智学院
欢迎您！

Paper&Code WeChat

weit@seu.edu.cn