Predictive Consistency Learning for Long-Tailed Recognition

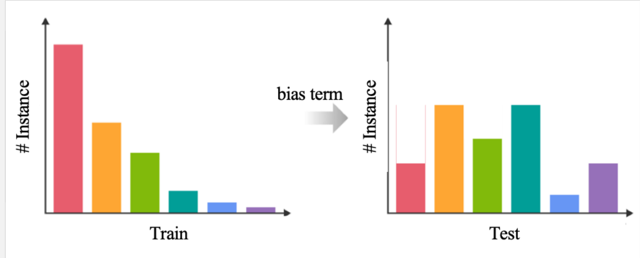
Nan Kang, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, Xilin Chen

nan.kang@vipl.ict.ac.cn   Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS).

BMVC 2023

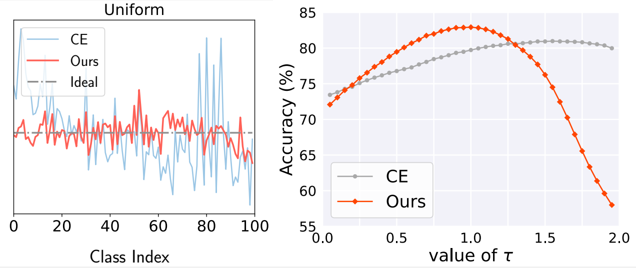
Motivation

- Previous post-hoc correction methods focus on the design of correction bias, but little attention has been paid to the estimation of $\hat{p}(y|x)$, which limits the effectiveness of correction.



- The devil is the one-hot CE loss that uniformly push all predictions to 1.0, which violates the calibration of $\hat{p}(y|x)$.

- Exploring how to optimize the estimation of $\hat{p}(y|x)$ so as to enhance the effectiveness of post-hoc methods.



① distribution mismatch   ② τ value shift

Methodology: Predictive Consistency Learning (PCL)

1. Introducing soft labels to adaptively assign flatter targets for hard/tail samples, which is calculated from the aggregation of historical predictions.

$$\mathcal{L}(x,\mathcal{T}) = -\sum_{j=1}^{K} \mathcal{T}_j(x)\cdot\log\left[\hat{p}_s(y=j|x)\right], \qquad \mathcal{T}_j^e(x) = (1-\alpha_{e,i})\cdot\delta_{i,j} + \alpha_{e,i}\cdot\bar{\mathcal{H}}_j^e(x),$$

$$\bar{\mathcal{H}}_j^e(x) = (1-\beta)\cdot\mathcal{H}_j^{e-1}(x) + \beta\cdot\bar{\mathcal{H}}_j^{e-1}(x),$$

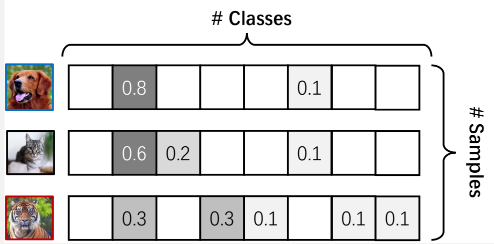
$$\mathcal{H}_j^{e-1}(x) = \hat{p}_s(y=j|x;\theta^{e-1}),$$

2. Class-aware weight adjustment to progressively interpolate the weights between ground-truth and soft labels.

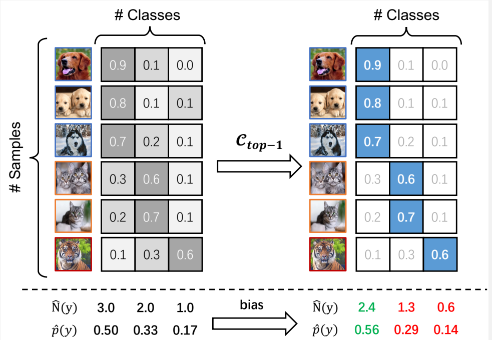
$$\alpha_{e,i} = \alpha\cdot\left(\frac{e}{E}\right)^{\lambda\cdot(1-q_i)}$$

3. Label compression by confidence to reduce the space complexity from $\mathcal{O}(N\cdot K)$ to $\mathcal{O}(N\cdot k)$, where $k \ll K$.

$$S(x;\mathcal{H},\gamma) = \min\{c\in[1,K] : \sum_{j=1}^{c}\mathcal{H}_{o(j)}(x) \geq \gamma\}$$



4. Eliminate the accociated class bias from compression with no cost by introducing the *effective class distribution*.



$$\hat{N}_{\mathrm{ecn}}(c) = \sum_x \hat{p}_s(y=c|x;\theta^E)$$

$$\hat{p}_{\mathrm{ecd}}(y) = \hat{N}_{\mathrm{ecn}}(y)/\sum_c \hat{N}_{\mathrm{ecn}}(c)$$

$$\arg\max_c f_\theta(x)[c] - \tau\cdot\log p_s(c)$$

$$\Downarrow$$

$$\arg\max_c f_\theta(x)[c] - \tau\cdot\log \hat{p}_{ecd}(c)$$

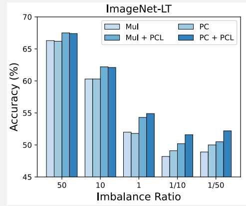
Experiments

Table 1: Top-1 accuracy (%) on CIFAR-10-LT and CIFAR-100-LT.

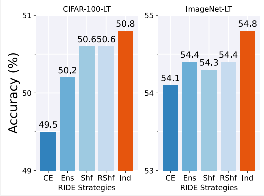
Dataset	CIFAR-10-LT			CIFAR-100-LT		
Imbalance Ratio	100	50	10	100	50	10
Softmax	70.4	74.8	86.4	38.4	43.9	55.8
LDAM-DRW [2]	77.1	81.1	88.4	42.1	46.7	58.8
MiSLAS [39]	82.1	85.7	90.0	47.0	52.3	**63.2**
TSC [25]	79.7	82.9	88.7	43.8	47.4	59.0
MetaSAug [24]	80.7	84.3	89.7	48.0	52.3	61.3
PC-Softmax [12]	79.4 ±0.5	82.8 ±0.3	88.4 ±0.4	45.5 ±0.7	50.3 ±0.4	60.0 ±0.3
PCL	**83.8** ±0.42	**86.1** ±0.21	**90.1** ±0.10	**49.4** ±0.37	**54.0** ±0.20	62.9 ±0.15
BALMS [28]	81.5 ±0.0	-	91.3 ±0.1	50.8 ±0.0	-	63.0 ±0.1
PaCo [5]	-	-	-	52.0	56.0	64.2
CC-SAM [41]	83.9	86.2	-	50.8	53.9	-
DCRNets [13]	85.0	-	-	51.4	-	-
PCL + AA	**85.5** ±0.34	**87.5** ±0.21	**91.3** ±0.22	**52.1** ±0.16	**57.0** ±0.24	**65.0** ±0.20

Table 2: Top-1 accuracy (%) on ImageNet-LT and Places-LT.

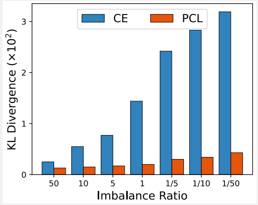
Dataset	ImageNet-LT				Places-LT			
Method	Many	Med.	Few	All	Many	Med.	Few	All
Softmax	64.0	33.8	5.8	41.6	**45.9**	22.4	0.4	27.2
cRT [15]	58.8	44.0	26.1	47.3	42.0	37.6	24.9	36.7
OLTM [10]	-	-	-	52.4	-	-	-	-
TSC [25]	63.5	49.7	30.4	52.4	-	-	-	-
MiSLAS [39]	61.7	51.3	35.8	52.7	39.6	43.3	36.1	40.4
PC-Softmax [12]	64.1	48.4	32.4	52.2	43.1	39.7	33.9	39.8
PCL	**66.2**	**53.0**	**36.1**	**55.8**	43.5	**42.6**	**38.0**	**42.0**
CC-SAM [41]	61.4	49.5	37.1	52.4	41.2	42.1	36.4	40.6
PaCo [5]	65.0	55.7	38.2	57.0	36.1	**47.9**	35.3	41.2
PaCo + DLSA [33]	64.6	54.9	41.8	56.9	**44.4**	44.6	32.3	42.1
PaCo + DCRNets [13]	-	-	-	58.0	-	-	-	41.7
PCL + SAM + RA	**67.3**	**58.8**	**43.5**	**60.0**	43.5	44.0	**39.9**	**43.0**



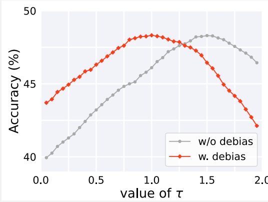
Test-time shift accuracy



PCL + Enemble



Better matched distribution



Effective debias