Supplementary: "Prior2Posterior: Model Prior Correction for Long-Tailed Learning"

S Divakar Bhat* Amit More* Mudit Soni Surbhi Agrawal
Honda R&D Co., Ltd.
Tokyo, Japan

1. Additional training details

We train Stage 1 models using cross-entropy loss for CIFAR10-LT and CIFAR100-LT datasets for 20,000 iterations. In Stage 2, for both classifier and feature tuning cases, we train the models for 4000 iterations. For ImageNet-LT dataset, we train the Stage 1 and both the Stage 2 methods for 200 and 20 epochs, respectively. For iNaturalist18 dataset, we train the Stage 1 and both Stage 2 models for 200 and 10 epochs, respectively. For all models the batch size of 64 is used. Rest of the experimental settings are borrowed from [9].

2. Dissimilarity of data distributions: $P(X) \neq P^t(X)$

In the paper we clearly maintain the distributions $P(X)$ and $P^t(X)$ as distinct in nature. For the sake of completeness we provide a mathematical justification using the approach of moment matching. In particular, we show that the first moment of train and test data distribution is not the same.

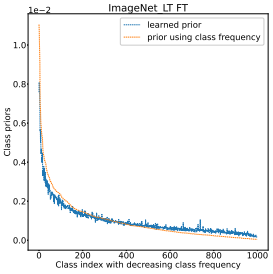
$$\mu_x = \int_x \int_y x P(x,y) dx dy \tag{1}$$

$$= \int_y \left(\int_x x P(x|y) dx \right) P(y) dy \tag{2}$$

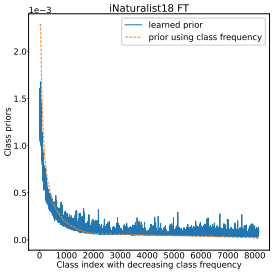
$$= \int_y \mu_{x|y} P(y) dy \tag{3}$$

$$\tag{4}$$

where, $\mu_{x|y}$ represents the mean of conditional distribution $P(x|y)$ and μ_x represents the mean of data distribution $P(x)$.



(a) Feature tuning in Stage 2 on ImageNet-LT



(b) Feature tuning in Stage 2 on iNaturalist18

Figure 1. We show model biases for different cases. Bias estimated using class frequencies and proposed method are shown for ImageNet-LT and iNaturalist18 datasets for logit-adjusted feature tuning in Stage 2.

Similarly we can show that for the test time distributions,

$$\mu_x^t = \int_y \mu_{x|y}^t P^t(y) dy \tag{5}$$

$$\tag{6}$$

Given that the class conditional data distributions $P(x|y)$ and $P^t(x|y)$ are the same, we have,

$$\mu_{x|y} = \mu_{x|y}^t \tag{7}$$

However, since the class prior distributions are not the same by definition, i.e. $P(y) \neq P^t(y)$, we have

$$\mu_x \neq \mu_x^t \tag{8}$$

And hence,

$$P(X) \neq P^t(X) \tag{9}$$

*Equal contribution

3. The effective prior for ImageNet-LT and iNaturalist18 datasets

In Figure 1 we show the model bias estimated using class frequencies and the effective prior calculated using proposed approach on ImageNet-LT and iNaturalist18 datasets. From the figure it is clearly observed that model bias is quite different from empirical bias estimated using class frequencies. In particular, the effective prior is higher for low frequency classes than the class frequency based prior for both datasets.

ImageNet-LT				
Method	Many	Medium	Few	All
Bal-Soft [9]	64.10	48.20	33.40	52.30
RSG [14]	63.20	48.20	32.20	51.80
LADE [3]	65.10	48.90	33.40	53.00
DisAlign [16]	62.70	52.10	31.40	53.40
ResLT [2]	63.00	53.30	35.50	52.90
WB+MaxNorm [1]	62.50	50.40	**41.50**	53.90
MARC [15]	60.40	50.30	36.60	52.30
RBL [6]	64.80	49.60	34.20	53.30
CCL [12]]	60.70	52.90	39.00	54.00
NC-DRW-cRT [4]	65.60	51.20	35.40	54.20
CE	**68.11**	42.56	14.85	48.63
CE + P2P	63.36	49.99	36.03	53.24
CL	63.85	49.95	34.75	53.23
CL + P2P	62.12	51.18	37.79	53.57
FT	65.83	51.32	28.22	53.82
FT + P2P	62.44	**53.34**	36.06	**54.67**

Table 1. The table show many, medium and few shot accuracies on ImageNet-LT dataset. Best and Second best results are shown in **bold faces** and underlined.

4. Multishot accuracies

In Table 1 and Table 2 we show multi-shot accuracies for ImageNet-LT and iNaturalist18 datasets and compare it with some of the recently published methods. We note from the table that proposed approach achieves highest overall accuracy while shot-wise accuracies are not affected much. We also show in Figure 2 the performance on iNaturalist18 for models trained with plain CE and with logit-adjustment (CL and FT). It can be noted that, P2P outperforms baseline class frequency based adjustment in all the cases.

5. Additional results on test time shifted imbalance

In Table. 3 we compare model performance for test-time shifted distributions with additional baselines and a few more distribution shifts. We note the superior performance of proposed algorithm.

iNaturalist18				
Method	Many	Medium	Few	All
DisAlign [16]	69.00	71.10	70.20	70.60
LDAM+DRW+SAM [8]	64.10	70.50	71.20	70.10
WB+MaxNorm [1]	71.20	70.40	69.70	70.20
ResLT [2]	68.50	69.90	70.40	70.20
SWA+SRepr [5]	70.70	70.83	70.76	70.79
CC-SAM [17]	65.40	70.90	72.20	70.9
CE	**76.33**	68.15	60.66	66.03
CE + P2P	67.02	71.05	**72.36**	71.15
CL	70.35	70.98	71.06	70.81
CL + P2P	68.09	71.15	72.19	71.43
FT	71.81	71.46	70.16	71.12
FT + P2P	66.63	**71.73**	72.32	**71.78**

Table 2. The table show many, medium and few shot accuracies on iNaturalist18 dataset. Best and Second best results are shown in **bold faces** and underlined.

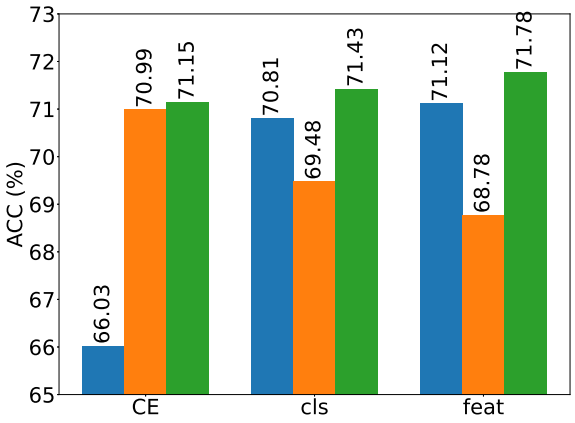


Figure 2. The performance on iNaturalist18 for Stage 1 baseline (CE) and Stage 2 (CL and FT) are shown. The effect of post-hoc using class frequency and proposed approach can be observed.

6. Discussion on Distribution Matching

Recent works like [7] have proposed to tackle this distribution misalignment problem from an optimisation perspective, employing the concept of optimal transport. Although the work provides interesting mathematical insights into relation between the distribution alignment problem and optimal transport, the method assumes that the marginal distribution is consistent with a uniform distribution. Unlike this we impose no such constraint in our proposed approach rendering further flexibility and simplicity in its implementation.

Similarly [13] also propose optimal transport based distribution matching framework for imbalanced partial label learning. They propose to refine the pseudo-labels in order to align with the true class prior by reducing the optimal

	Forward					Uniform	Backward				
Imbalance ratio	50	25	10	5	2	1	2	5	10	25	50
CE	66.3	63.9	60.4	57.1	52.3	48.63	44.2	38.9	35.0	30.5	27.9
De-Confound [11]	64.1	62.5	60.1	57.8	54.6	52.0	49.3	45.8	43.4	40.4	38.4
Bal-Soft [9]	62.5	60.9	58.8	57.0	54.4	52.3	49.6	46.5	44.1	41.4	39.7
PC Causal Norm [3]	66.7	64.3	60.9	58.1	54.6	52.0	49.8	47.9	47.0	46.7	46.7
PC-Balanced Softmax [3]	65.5	63.1	59.9	57.3	54.3	52.1	50.2	48.8	48.3	48.5	49.0
PC-Softmax [3]	66.6	63.9	60.6	58.1	55.0	52.8	51.0	49.3	48.8	48.5	49.0
LADE [3]	67.4	64.8	61.3	58.6	55.2	53.0	51.2	49.8	49.2	49.3	50.0
Our FT+P2P	**67.6**	**64.9**	**61.5**	**58.7**	**56.4**	**54.67**	**52.3**	**51.0**	**50.5**	**50.8**	**51.1**

Table 3. Top 1 Accuracy on test time shifted ImageNet-LT dataset.

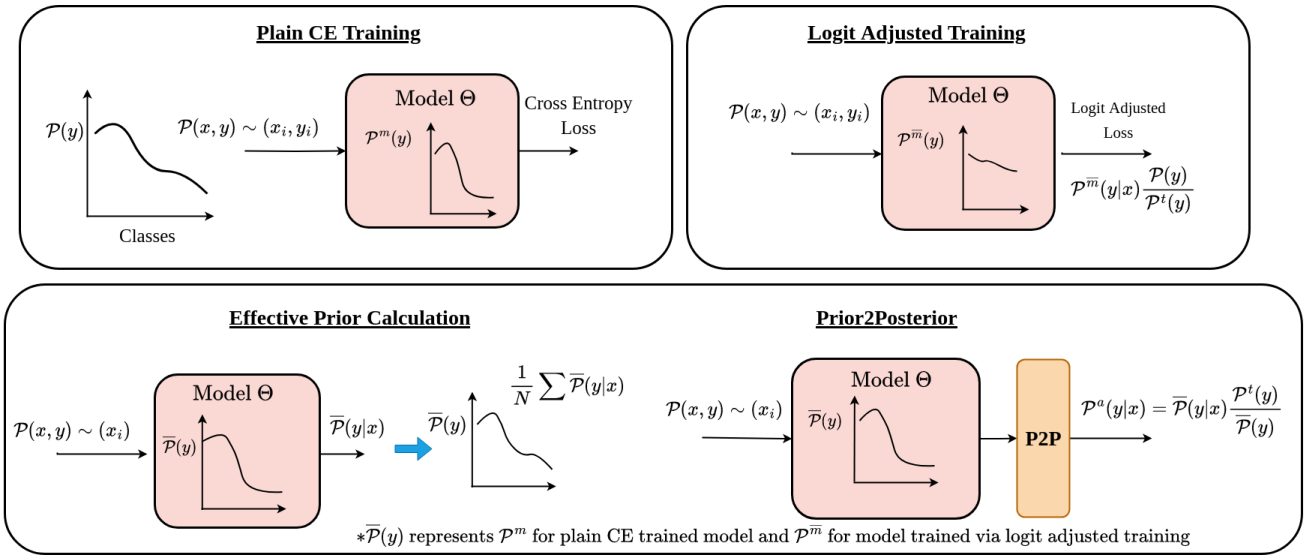


Figure 3. Proposed approach is summarized in the figure. The top row illustrates plain CE loss and logit-adjusted loss showing that model accumulates some bias due to imbalanced training data. We show effective prior calculation for trained model in the bottom row. Once the prior is calculated, *a posteriori* probabilities can be corrected using proposed approach as shown in the bottom row.

transport objective function.

[10] present a novel variant of the optimal transport called, Relative Entropic Optimal Transport to learn matching with a specified prior. The manually specified smoothing guidance matrix $\mathcal{Q}$ can be seen as a generic representation for the effective prior.

7. Flow of the proposed approach

We summarise the proposed approach in a block diagram as shown in Figure3. The block diagram illustrates the different stages involved in the process starting from bias accumulation in traditional training to bias removal using the proposed method. Both Logit adjusted training and Prior2Posterior is depicted along with an illustration showing the Effective Prior computation.

References

[1] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *IEEE CVPR*, pages 6897–6907, 2022. 2

[2] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE TPAMI*, 45(3):3695–3706, 2022. 2

[3] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *IEEE CVPR*, pages 6626–6636, 2021. 2, 3

[4] Xuantong Liu, Jianfeng Zhang, Tianyang Hu, He Cao, Yuan Yao, and Lujia Pan. Inducing neural collapse in deep long-tailed learning. In *AISTATS*, pages 11534–11544. PMLR, 2023. 2

[5] Giung Nam, Sunguk Jang, and Juho Lee. Decoupled training

for long-tailed classification with stochastic representations. *arXiv preprint arXiv:2304.09426*, 2023. 2

[6] Gao Peifeng, Qianqian Xu, Peisong Wen, Zhiyong Yang, Huiyang Shao, and Qingming Huang. Feature directions matter: Long-tailed learning via rotated balanced representation. 2023. 2

[7] Hanyu Peng, Mingming Sun, and Ping Li. Optimal transport for long-tailed recognition with learnable cost matrix. In *International conference on learning representations*, 2021. 2

[8] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, et al. Escaping saddle points for effective generalization on class-imbalanced data. *Advances in NeurIPS*, 35:22791–22805, 2022. 2

[9] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in NeurIPS*, 33:4175–4186, 2020. 1, 2, 3

[10] Liangliang Shi, Haoyu Zhen, Gu Zhang, and Junchi Yan. Relative entropic optimal transport: a (prior-aware) matching perspective to (unbalanced) classification. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[11] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in NeurIPS*, 33:1513–1524, 2020. 3

[12] Anthony Meng Huat Tiong, Junnan Li, Guosheng Lin, Boyang Li, Caiming Xiong, and Steven CH Hoi. Improving tail-class representation with centroid contrastive learning. *Pattern Recognition Letters*, 168:123–130, 2023. 2

[13] Haobo Wang, Mingxuan Xia, Yixuan Li, Yuren Mao, Lei Feng, Gang Chen, and Junbo Zhao. Solar: Sinkhorn label refinery for imbalanced partial-label learning. *Advances in neural information processing systems*, 35:8104–8117, 2022. 2

[14] Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu. Rsg: A simple but effective module for learning imbalanced datasets. In *IEEE CVPR*, pages 3784–3793, 2021. 2

[15] Yidong Wang, Bowen Zhang, Wenxin Hou, Zhen Wu, Jindong Wang, and Takahiro Shinozaki. Margin calibration for long-tailed visual recognition. In *ACML*, pages 1101–1116. PMLR, 2023. 2

[16] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *IEEE CVPR*, pages 2361–2370, 2021. 2

[17] Zhipeng Zhou, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Wei Gong. Class-conditional sharpness-aware minimization for deep long-tailed recognition. In *IEEE CVPR*, pages 3499–3509, 2023. 2