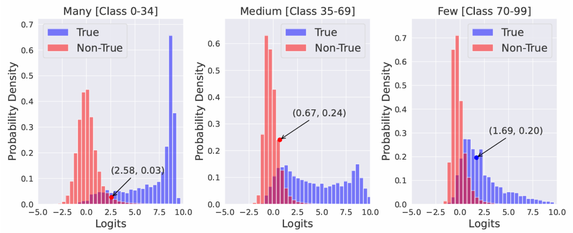
RETHINKING CLASSIFIER RE-TRAINING IN LONG-TAILED RECOGNITION: LABEL OVER-SMOOTH CAN BALANCE

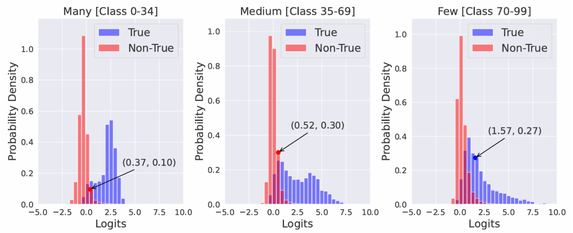
Siyu Sun*, Han Lu*, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, Junchi Yan

ICLR International Conference On Learning Representations

The example of how the classifier works



(a) **Vanilla Cross-Entropy (CE).** "Many" class logit 2.58 is larger than the ground truth "Few" class logit 1.69, leading to the misclassification.



(b) **LOS (ours).** "Few" class logit 1.57 outperforms the other classes.

The blue columns represent the logits of the true samples of current class, while the red columns correspond to the logits of the non-true samples.

Logits Magnitude and Regularized Standard Deviation

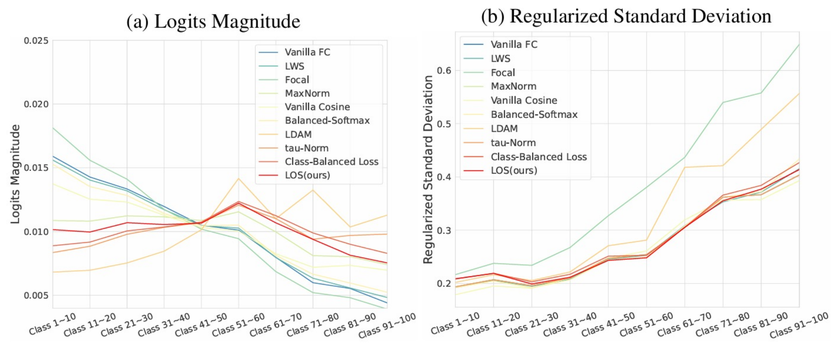


Figure 3: **Overview of our proposed metrics. Left: Overview of Logits Magnitude** for various methods on CIFAR100-LT with an imbalanced ratio of 100. Classes are grouped into segments of 10, and mean values are computed for comparative analysis. **The methods in the legend are sorted in ascending order of performance from top to bottom.** The difference in means between samples from true and non-true classes is evaluated for each class in the test set. Magnitude regularization using the 1-norm is employed to enhance comparability. **Right: Overview of Regularized Standard Deviation**. The true distribution of logits $\mathbf{z}$ for each class is not available considering the lack of the samples, so the displayed results represent computations on the test set.

Rethink

Methods	Formulation	Accuracy All	Many	Medium	Few
Vanilla FC	$g_i = \boldsymbol{W}_i^\top f(\mathbf{x})$	48.44	**78.74**	47.37	14.33
Vanilla Cosine	$g_i = \boldsymbol{W}_i^\top / \|\|\boldsymbol{W}_i^\top\|\| \cdot f(\mathbf{x}) / \|\|f(\mathbf{x})\|\| \cdot 1/\tau$	52.20	77.21	50.52	24.98
Learnable Weight Scaling (Kang et al., 2019)	$g_i = \mathbf{c}_i \cdot \boldsymbol{W}_i^\top f(x)$, where $\mathbf{c}_i$ is learnable	48.99	78.51	47.51	16.26
LDAM Loss (Cao et al., 2019)	$g_i = \boldsymbol{W}_i^\top f(\mathbf{x}) - \mathbf{1}\{i = y\} \cdot C/n_i^\gamma$	52.94	72.37	**54.31**	28.67
Balanced-Softmax Loss (Ren et al., 2020)	$g_i = \boldsymbol{W}_i^\top f(\mathbf{x}) + \log(n_i)$	52.23	77.63	50.74	24.33
Re-Sampling (Shen et al., 2016)	$r_s(i) \propto 1/n_i$, $r_s(i)$ is the sample ratio of class i	48.44	78.74	47.37	14.33
Focal Loss (Lin et al., 2017)	$r(y) = (1 - \mathbf{p}_y)^\gamma$, $\mathbf{p}_y$ is the predicted probability of y	50.94	78.00	49.97	20.50
Class-Balanced Loss (Cui et al., 2019)	$r(y) = (1-\beta)/(1-\beta_y^n)$ applied with BCE loss	**53.31**	70.88	50.25	**36.41**
MaxNorm (Alshammari et al., 2022)	$\|\|\boldsymbol{W}_i\|\|_2^2 \leq \delta^2$	51.64	77.60	49.37	23.99
τ-normalized (Kang et al., 2019)	$h_i = (\boldsymbol{W}_i / \|\|\boldsymbol{W}_i\|\|^\tau)^\top f(\mathbf{x})$	53.01	74.74	47.31	34.30
Post-hoc Logit Adjustment (Menon et al., 2021)	$h_i = \boldsymbol{W}_i^\top f(\mathbf{x}) - \tau \log(n_i)$	52.45	75.57	47.71	31.00

How does imbalance occur

The expected value of probability $\mathbf{s}_i'$ can be expressed as:

$$\mathbb{E}[\mathbf{s}_i'] = \mathbb{E}\left[\frac{1}{\sum_j e^{\mathbf{z}_j' - \mathbf{z}_i'}}\right] = \mathbb{E}\left[\frac{1}{\sum_j e^{(\Delta_j - \Delta_i)} e^{(\mathbf{z}_j - \mathbf{z}_i)}}\right]. \tag{8}$$

In the balanced datasets, expected with $\mathbf{R}_1 \approx \mathbf{R}_2 \approx \cdots \approx \mathbf{R}_K$, $(\mathbf{R}_j \mathbf{L}_j - \mathbf{R}_i \mathbf{L}_i)$ is close to zero.

$$\mathbb{E}[\mathbf{s}_i'] \approx \frac{1}{\mathbb{E}[\exp(\Delta_j - \Delta_i)]} \mathbb{E}\left[\frac{1}{\sum_j e^{\mathbf{z}_j - \mathbf{z}_i}}\right] = \frac{\mathbb{E}[\mathbf{s}_i]}{\mathbb{E}[\exp(\Delta_j - \Delta_i)]} \tag{9}$$

where $\mathbb{E}[\exp(\Delta_j - \Delta_i)]$ can be considered as a constant $\mathcal{E}$. See the Appendix C for a detailed analysis of the approximation. For balanced datasets, the impact of disturbances is consistent across different classes i.

$$\mathbb{E}[\mathbf{s}_i'] \,/\, \mathbb{E}[\mathbf{s}_i] \approx 1/\mathcal{E}. \tag{10}$$

However, for imbalanced datasets, Eq. 9 no longer holds. There is a significant disparity in the values of $\exp(\Delta_j - \Delta_i)$ for different class j. If $\mathbf{R}_j < \mathbf{R}_k$, $\mathbb{E}[\exp(\Delta_j - \Delta_i)] < \mathbb{E}[\exp(\Delta_k - \Delta_i)]$. For instance, when ϵ follows a normal distribution we have

$$\mathbb{E}[\exp(\Delta_j - \Delta_i)] = \exp((\sigma(\Delta_j - \Delta_i))^2/2), \tag{11}$$

Abstract of This Paper

In the field of long-tailed recognition, the Decoupled Training paradigm has shown exceptional promise by dividing training into two stages: representation learning and classifier re-training. While previous work has tried to improve both stages simultaneously, this complicates isolating the effect of classifier re-training. Recent studies reveal that simple regularization can produce strong feature representations, highlighting the need to reassess classifier re-training methods. In this study, we revisit classifier re-training methods based on a unified feature representation and re-evaluate their performances. We propose two new metrics, Logits Magnitude and Regularized Standard Deviation, to compare the differences and similarities between various methods. Using these two newly proposed metrics, we demonstrate that when the Logits Magnitude across classes is nearly balanced, further reducing its overall value can effectively decrease errors and disturbances during training, leading to better model performance. Based on our analysis using these metrics, we observe that adjusting the logits could improve model performance, leading us to develop a simple label over-smoothing approach to adjust the logits without requiring prior knowledge of class distribution. This method softens the original one-hot labels by assigning a probability slightly higher 1/K to the true class and slightly lower than 1/K to the other classes, where K is the number of classes. Our method achieves state-of-the-art performance on various imbalanced datasets, including CIFAR100-LT, ImageNet-LT, and iNaturalist2018.

Label Over-Smooth

Suppose the number of the categories is K. Given the input image $\mathbf{x}$ and label y, feature extractor backbone $f(\cdot)$ and classifier $\theta = (\boldsymbol{W}, \boldsymbol{b})$, the formulation of our "Label Over-Smooth" is as follows:

$$\mathcal{L}(\theta; f(\mathbf{x}), y) = \sum_{i=1}^{K} -\tilde{\mathbf{y}}_i \cdot \log\left(\frac{e^{\mathbf{z}_i}}{\sum_j e^{\mathbf{z}_j}}\right), \qquad \tilde{\mathbf{y}}_i = \begin{cases} 1 - \delta + \frac{\delta}{K}, & i = y \\ \frac{\delta}{K}, & i \neq y \end{cases} \tag{12}$$

where $\delta \in [0, 1)$ is a constant value to control the overall non-true class probability. Interestingly the above formulation is fundamentally consistent with the conventional Label Smoothing

Method	Reference	CIFAR100-LT IR=100	IR=50	IR=10	ImageNet-LT Many	Medium	Few	All	iNaturalist2018 Many	Medium	Few	All
CE	—	38.3	43.8	55.7	**65.9**	37.5	7.7	44.4	**72.2**	63.0	57.2	61.7
Focal (Lin et al., 2017)	ICCV17	38.4	44.3	55.7	36.4	29.9	16.0	30.5	—	—	—	61.1
LDAM-DRW (Cao et al., 2019)	NIPS19	42.0	46.6	58.7	—	—	—	48.8	—	—	—	—
cRT (Kang et al., 2019)	ICLR19	—	—	—	61.8	46.2	27.3	49.6	69.0	66.0	63.2	65.2
τ-norm (Kang et al., 2019)	ICLR19	47.7	52.5	63.8	59.1	46.9	30.7	49.4	65.6	65.3	65.5	65.6
CB-CE (Cui et al., 2019)	CVPR19	39.6	45.3	58.0	39.6	32.7	16.8	33.2	53.4	54.8	53.2	54.0
De-confound (Tang et al., 2020)	NIPS20	44.1	50.3	59.6	62.7	48.8	31.6	51.8	—	—	—	—
BALMS (Ren et al., 2020)	NIPS20	50.8	—	63.0	61.1	48.5	31.8	50.9	65.5	67.5	67.5	67.2
BBN (Zhou et al., 2020)	CVPR20	42.6	47.0	59.1	—	—	—	—	49.4	70.8	65.3	66.3
LogitAjust (Menon et al., 2021)	ICLR21	42.0	47.0	57.7	—	—	—	51.1	—	—	—	69.4
DisAlign (Zhang et al., 2021a)	CVPR21	—	—	—	61.3	52.2	31.4	52.9	69.0	**71.1**	70.2	70.6
LTWB (Alshammari et al., 2022)	CVPR22	53.4	57.7	68.7	62.5	50.4	41.5	53.9	71.2	70.4	69.7	70.2
AREA (Chen et al., 2023)	ICCV23	48.8	51.8	60.1	—	—	—	49.5	—	—	—	68.4
RBL (Peifeng et al., 2023)	ICML23	53.1	57.2	—	64.8	49.6	34.2	53.3	—	—	—	—
EWB-FDR (Hasegawa & Sato, 2024)	ICLR24	53.0	—	—	63.4	50.0	35.1	53.2	—	—	—	—
LOS (ours)		**54.9**	**58.8**	**69.7**	63.2	**50.7**	**42.3**	**54.4**	69.2	70.7	**71.3**	**70.8**

Plug-in with other methods

Method	Reference	CIFAR100-LT IR=100	IR=50	IR=10	ImageNet-LT Many	Medium	Few	All
Vallina CE + LOS (ours)		**54.9**	**58.8**	**69.7**	63.2	**50.7**	**42.3**	**54.4**
RIDE (Wang et al., 2020b)	ICLR20	49.1	—	—	67.9	52.3	36.0	56.1
SSD (Li et al., 2021)	ICCV21	46.0	50.5	62.3	66.8	53.1	35.4	56.0
PaCo (Cui et al., 2021)	ICCV21	52.0	56.0	64.2	63.2	51.6	39.2	54.4
BCL (Zhu et al., 2022)	CVPR22	51.0	54.9	64.4	65.3	53.5	36.3	55.6
NCL (Li et al., 2022a)	CVPR22	53.3	56.8	—	—	—	—	57.4
GML (Suh & Seo, 2023)	ICML23	53.0	57.6	65.7	68.7	55.7	38.6	58.3
ProCo (Du et al., 2024)	TPAMI24	52.8	57.1	65.5	—	—	—	58.0
RIDE+LOS(ours)		50.2(+1.1)	—	—	66.9(-1.0)	53.2(+0.9)	37.9(+1.9)	56.5(+0.4)
SSD+LOS(ours)		47.3(+1.3)	51.4(+0.9)	63.1(+0.8)	66.2(-0.6)	53.8(+0.7)	36.7(+1.3)	56.3(+0.3)
PaCo+LOS(ours)		52.9(+0.9)	56.8(+0.8)	65.3(+1.1)	62.7(-0.5)	52.5(+0.9)	41.2(+2.0)	55.0(+0.6)
BCL+LOS(ours)		52.2(+1.2)	56.0(+1.1)	65.1(+0.7)	64.7(-0.6)	54.1(+0.8)	28.1(+1.8)	56.1(+0.5)
NCL+LOS(ours)		54.1(+0.8)	57.5(+0.7)	—	—	—	—	57.7(+0.3)
GML+LOS(ours)		54.0(+1.0)	58.4(+0.8)	66.6(+0.9)	68.0(-0.7)	55.7(+0.6)	40.7(+2.1)	58.7(+0.4)
ProCo+LOS(ours)		53.9(+1.1)	58.0(+0.9)	66.7(+1.2)	—	—	—	58.5(+0.5)

Star us on GitHub!
Code Available at https://github.com/Thinklab-SJTU/LOS

Work done by 上海交通大学 SHANGHAI JIAO TONG UNIVERSITY