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

Open-Sampling Exploring Out-of-Distribution Data for Re-balancing Long-tailed Datasets

Hongxin Wei [1] Lue Tao [2] Renchunzi Xie [1] Lei Feng [3] Bo An [1]

ICML 2022

Introduction

In the literature, a popular direction in long-tailed learning is to re-balance the data distribution by data re-sampling.

Over-sampling repeats samples from under-presented classes, but it usually causes over-fitting to the minority classes.

To alleviate the over-fitting issue:

- synthesized novel samples to augment the minority classes(error-prone due to noise).
- [1] introduced unlabeled-in-distribution data + semi-supervised (require in-distribution data, but the cost is expensive).

Motivation -> out-of-distribution (OOD) data for long-tailed imbalanced learning

[1]Rethinking the value of labels for improving class-imbalanced learning. NeurIPS, 2020.

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

Notation

label space: $\mathcal{Y} = \{1, \ldots, K\}$

training data: $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{N} \in \mathcal{X} \times \mathcal{Y}$

source distribution: $P_{\mathbf{s}}(X, Y)$

target distribution: $P_{\boldsymbol{t}}(X, Y)$

same class conditional probability: $P_{\mathbf{s}}(X|Y) = P_{\boldsymbol{t}}(X|Y)$

class priors are different: $P_{\mathrm{s}}(Y) \neq P_{\mathrm{t}}(Y)$

an unlabeled auxiliary dataset: $\mathcal{D}_{\mathrm{out}}^{(\boldsymbol{x})} = \{\tilde{x}_i\}_{i=1}^{M} \in \mathcal{X}, M \gg N, \mathcal{D}_{\mathrm{out}} = \{(\tilde{x}_i, y_i)\}_{i=1}^{M}$

Motivation

$$y^* = \arg\max_{y\in\mathcal{Y}} P(y|\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} P(\boldsymbol{x}|y)P(y), \qquad (1)$$

$$\arg\max_{y\in\mathcal{Y}} P_{\text{mix}}(\boldsymbol{x}|y)P_{\text{mix}}(y) = \arg\max_{y\in\mathcal{Y}} P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(y).$$

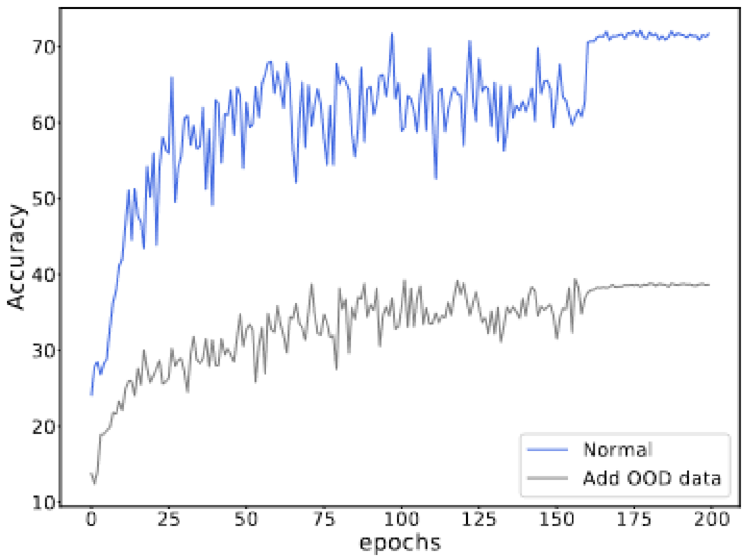
$$\begin{aligned}
&P_{\text{mix}}(\boldsymbol{x}|y)P_{\text{mix}}(y)\\
=&\frac{N}{M+N}P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(y)+\frac{M}{M+N}P_{\text{out}}(\boldsymbol{x}|y)P_{\text{out}}(y)\\
=&\frac{N}{M+N}P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(y)+\frac{M}{M+N}P_{\text{out}}(\boldsymbol{x})P_{\text{out}}(y)\\
=&\frac{N}{M+N}P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(\boldsymbol{x},y)+\frac{1}{K}\cdot\frac{M}{M+N}P_{\text{out}}(\boldsymbol{x}),
\end{aligned}$$

$$\begin{aligned}
&\arg\max_{y\in\mathcal{Y}} P_{\text{mix}}(\boldsymbol{x}|y)P_{\text{mix}}(y)\\
=&\arg\max_{y\in\mathcal{Y}}\left\{\frac{N}{M+N}P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(\boldsymbol{x},y)+\frac{M/K}{M+N}P_{\text{out}}(\boldsymbol{x})\right\}\\
=&\arg\max_{y\in\mathcal{Y}}\frac{N}{M+N}P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(\boldsymbol{x},y).\\
=&\arg\max_{y\in\mathcal{Y}}P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(\boldsymbol{x},y).
\end{aligned}$$

motivates us to exploit the potential value of OOD instances to repair class imbalance.

Method

$P_{mix}(Y)$ still remains largely imbalanced.



the trade-off between re-balancing the class priors and keeping the non-toxicity of the added noisy labels.

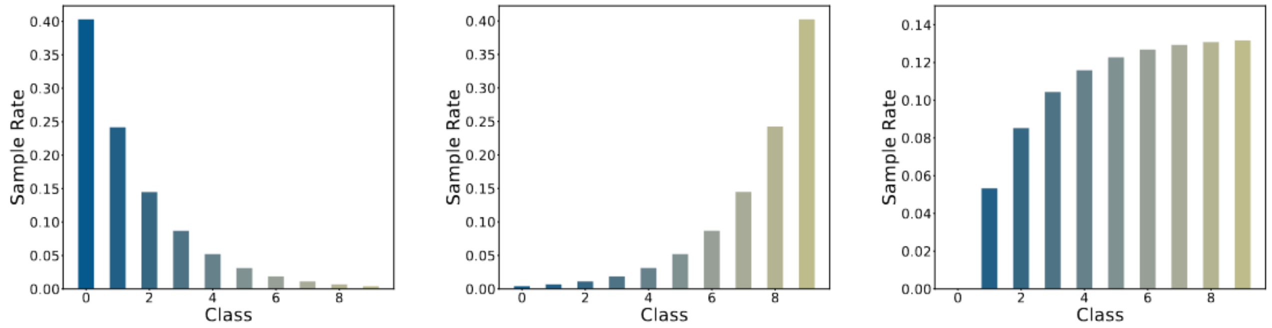
Complementary Distribution: a label distribution for the auxiliary dataset to re-balance the class priors of the original dataset.

Minimum Complementary Distribution(MCD): the smallest number of auxiliary instances.

We use instances from 300K Random Images as OOD data and label as a minority class—“9”.
Simply adding OOD data into training may downgrade the generalization performance.

Method

Proposition 2.3 (Complementary Sampling Rate). $\Gamma_j = (\alpha - \beta_j)/(K \cdot \alpha - 1)$, *where* $\beta_j = \frac{n_j}{\sum_{i=1}^{K} n_i}$. *Then, (i)* $\sum_{i=1}^{K} \Gamma_i = 1$; *(ii)* $\Gamma = \Gamma^m$ *if* $\alpha = \max_j(\beta_j)$; *(iii)* $\Gamma_j \to 1/K$ *as* $\alpha \to \infty$.



(a) Original (b) CB (c) MCD

Method

Issue: consume too much capacity of the network on fitting the open-set noisy labels, making it hard to converge. Especially, M >> N

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\widetilde{\boldsymbol{x}} \sim P_{\text{out}}(X)} \left[\omega_{\widetilde{y}} \cdot \ell \left(f(\widetilde{\boldsymbol{x}}; \boldsymbol{\theta}), \widetilde{y} \right) \right], \qquad \tilde{y} \sim \Gamma,\ \omega_{\tilde{y}} = \Gamma_{\tilde{y}} \cdot K$$

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \mathbb{E}_{((\boldsymbol{x}, y) \sim P_s(X, Y))} \left[\ell \left(f(\boldsymbol{x}; \boldsymbol{\theta}), y \right) \right] \\ & + \eta \cdot \mathbb{E}_{(\widetilde{\boldsymbol{x}}) \sim P_{\text{out}}(X)} \left[\omega_{\widetilde{y}} \cdot \ell \left(f(\widetilde{\boldsymbol{x}}; \boldsymbol{\theta}), \widetilde{y} \right) \right], \end{aligned}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{imb}} + \eta \cdot \mathcal{L}_{\text{reg}}.$$

Algorithm 1 Open-sampling

Require: Training dataset $\mathcal{D}_{\text{train}}$. Open-set auxiliary dataset $\mathcal{D}_{\text{out}}^{(x)}$;

1: **for** each iteration **do**
2: Sample a mini-batch of original training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=0}^{n}$ from $\mathcal{D}_{\text{train}}$;
3: Sample a mini-batch of open-set instances $\{\widetilde{\boldsymbol{x}}_i\}_{i=0}^{m}$ from $\mathcal{D}_{\text{out}}^{(x)}$;
4: Generate random noisy label $\widetilde{y}_i \sim \Gamma$ for each open-set instance $\widetilde{\boldsymbol{x}}_i$;
5: Perform gradient descent on f with $\mathcal{L}_{\text{total}}$ from Equation (2);
6: **end for**

Experiments

Table 1. Test accuracy (%) of ResNet-32 on long-tailed CIFAR-10 and CIFAR-100 with various imbalance ratios. "†" indicates the reported results from (Kim et al., 2020). The bold indicates the improved results by integrating our regularization.

Dataset	Long-tailed CIFAR-10			Long-tailed CIFAR-100		
Imbalance Ratio	100	50	10	100	50	10
Standard	71.61 ± 0.21	77.30 ± 0.13	86.74 ± 0.41	37.59 ± 0.19	43.20 ± 0.30	56.44 ± 0.12
SMOTE †	71.50 ± 0.57	-	85.70 ± 0.25	34.00 ± 0.33	-	53.80 ± 0.93
CB-RW	72.57 ± 1.30	78.19 ± 1.79	87.18 ± 0.95	38.11 ± 0.78	43.26 ± 0.87	56.40 ± 0.40
CB-Focal	70.91 ± 0.39	77.71 ± 0.57	86.89 ± 0.21	37.84 ± 0.80	42.96 ± 0.77	56.09 ± 0.15
Ours	**77.62 ± 0.28**	**81.76 ± 0.51**	**89.38 ± 0.46**	**40.26 ± 0.65**	**44.77 ± 0.25**	**58.09 ± 0.29**
LDAM-RW	74.21 ± 0.61	78.86 ± 0.65	86.44 ± 0.78	29.02 ± 0.34	36.41 ± 0.84	54.23 ± 0.72
+ Ours	**75.19 ± 0.34**	**79.76 ± 0.44**	**87.28 ± 0.61**	**35.85 ± 0.62**	**42.18 ± 0.82**	**55.48 ± 0.59**
LDAM-DRW	78.08 ± 0.38	81.88 ± 0.44	87.49 ± 0.18	42.84 ± 0.25	47.13 ± 0.28	57.18 ± 0.47
+ Ours	**79.82 ± 0.31**	**82.22 ± 0.45**	**87.83 ± 0.38**	**44.07 ± 0.75**	**47.5 ± 0.24**	**57.43 ± 0.31**
Balanced Softmax	78.03 ± 0.28	81.63 ± 0.39	88.10 ± 0.32	42.11 ± 0.70	46.79 ± 0.24	58.06 ± 0.40
+ Ours	**79.05 ± 0.20**	**82.76 ± 0.52**	**88.89 ± 0.21**	**42.86 ± 0.27**	**47.28 ± 0.58**	**58.80 ± 0.72**
SSP	74.58 ± 0.16	79.20 ± 0.43	88.50 ± 0.24	43.00 ± 0.51	47.04 ± 0.60	59.08 ± 0.46
+ Ours	**79.38 ± 0.65**	**82.18 ± 0.33**	**88.80 ± 0.43**	**43.57 ± 0.29**	**48.66 ± 0.57**	**59.78 ± 0.91**

Experiments

Table 5. OOD detection performance comparison on long-tailed CIFAR-10. All values are percentages and are averaged over the ten test datasets described in Appendix E. "↑" indicates larger values are better, and "↓" indicates smaller values are better. Bold numbers are superior results. Detailed results for each OOD test dataset can be found in Appendix F.

Method	Test Accuracy ↑	FPR95 ↓	AUROC ↑	AUPR ↑
MSP	71.83	56.1	75.2	32.71
OE	66.74	32.38	84.15	36.86
Ours	**77.62**	**20.68**	92.40	58.38
Ours ($\alpha = 5$)	75.16	22.13	**94.26**	**75.91**

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}\left[-\log f_y(x)\right]+\lambda\mathbb{E}_{x\sim\mathcal{D}_{\text{out}}}\left[H(P(Y);f(x))\right]$$

Experiments

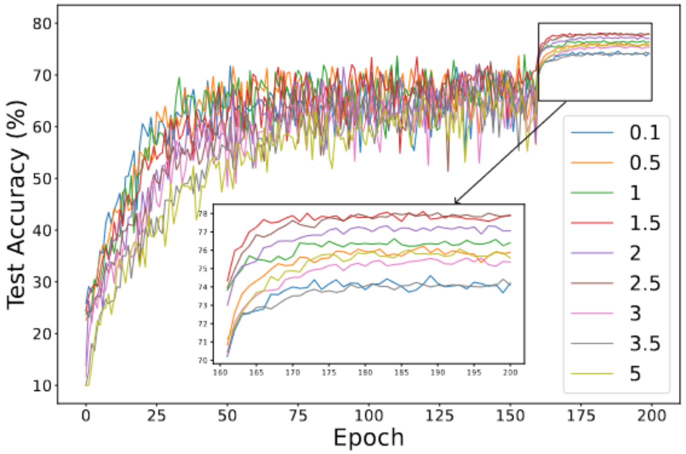
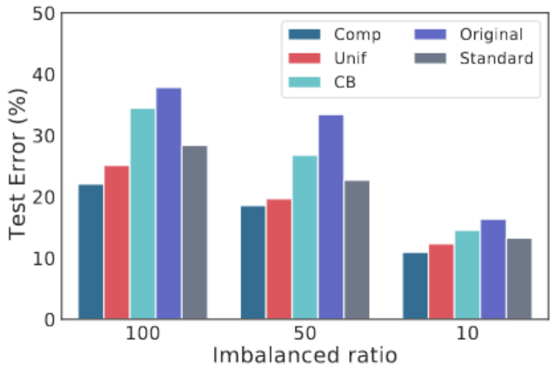


Figure 3. Results of sensitivity analysis on long-tailed CIFAR-10 with various values for η.

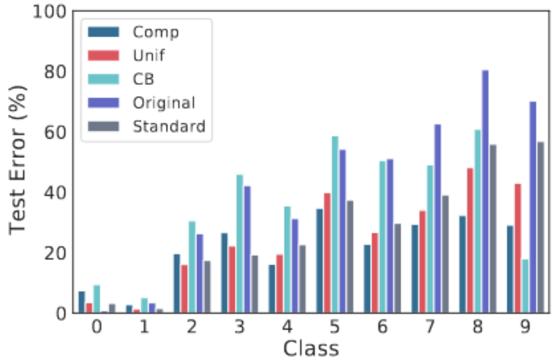
Table 6. Results of ablation study on long-tailed CIFAR-10 for the class-dependent weighting factor w_j.

Imbalance Factor	100	50	10
Standard	71.61	77.30	86.74
Ours w/o w_j	76.57	81.18	88.57
Ours	**77.62**	**81.76**	**89.38**

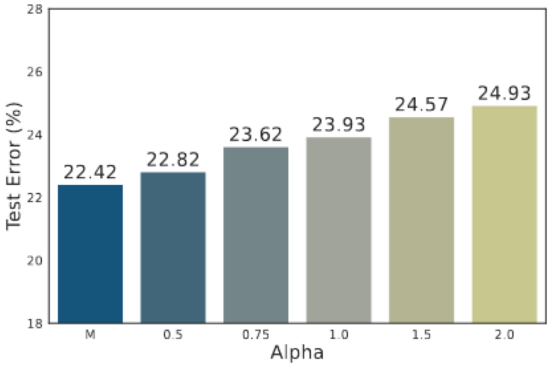
Experiments



(a) Label distribution.

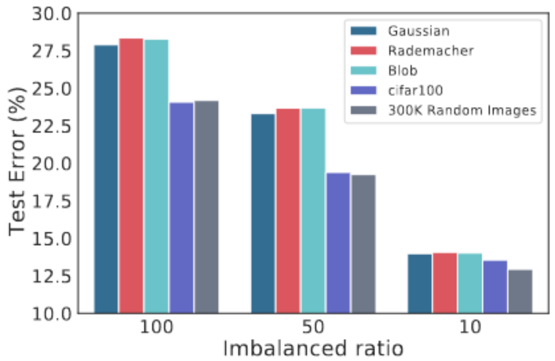


(b) Label distribution.

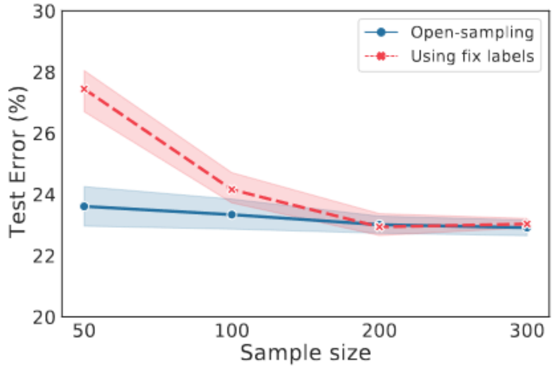


(c) Alpha.

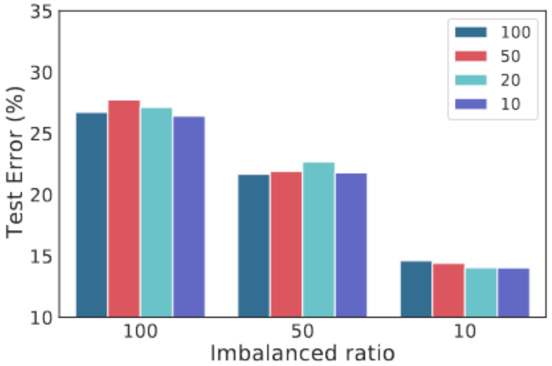
M $\alpha = (\max_j \beta_j + \min_j \beta_j)$.



(d) Auxiliary dataset.

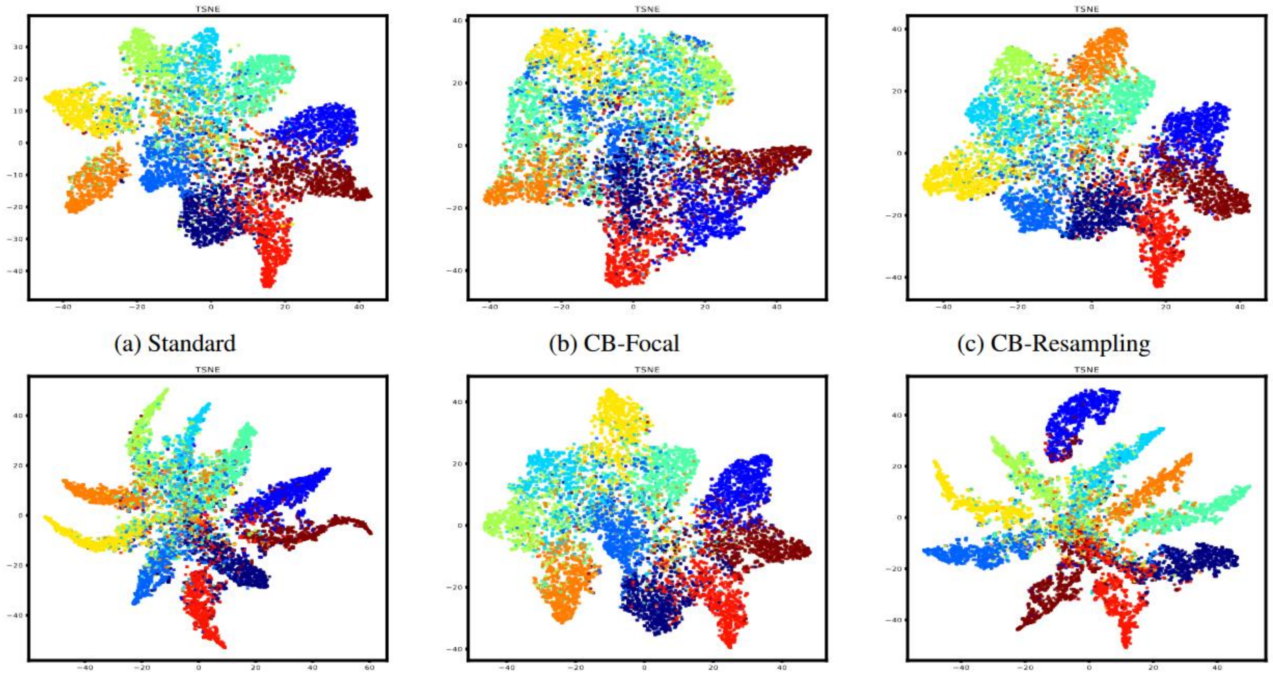


(e) Sample size.



(f) Number of Classes.

Experiments

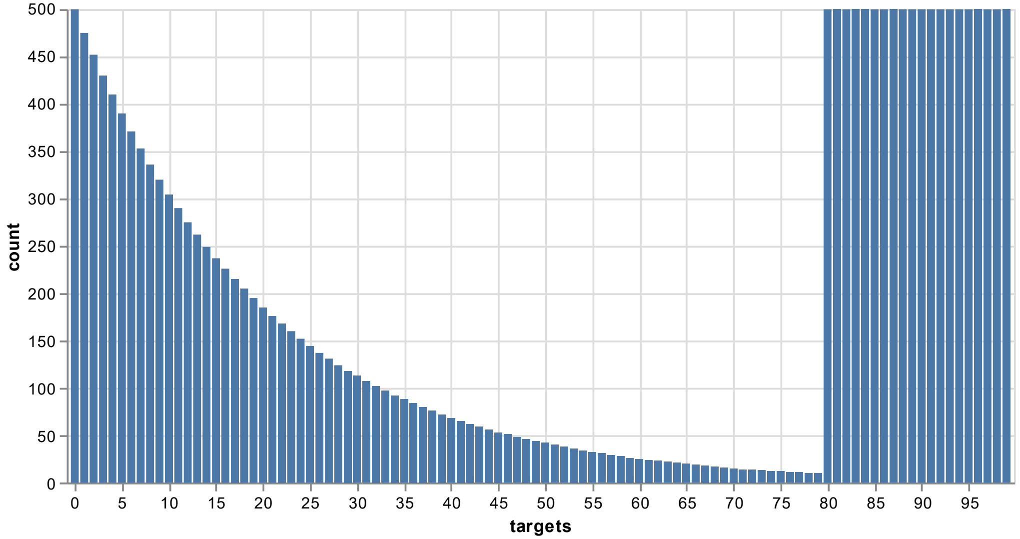


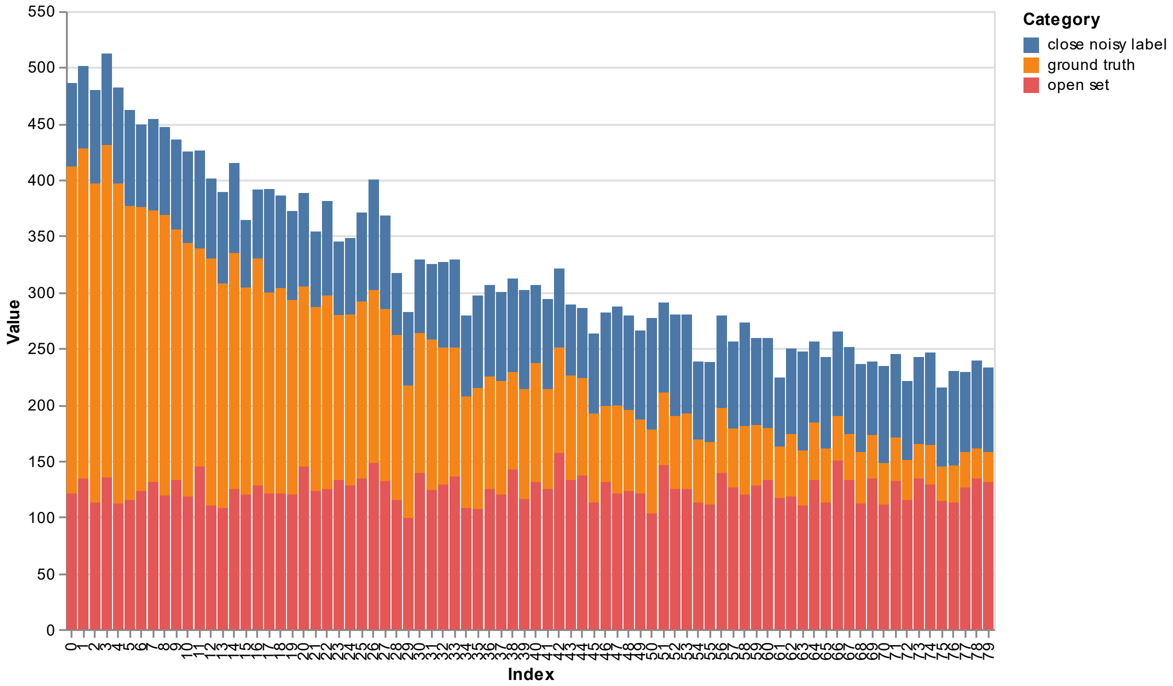
(a) Standard (b) CB-Focal (c) CB-Resampling

(d) LDAM-RW (e) Balanced Softmax (f) Ours

Figure 5. t-SNE visualization of test set on long-tailed CIFAR-10 with imbalance ratio 100. We can observe that LDAM and our method appear to learn more separable representations than Standard training and the other algorithms.

Close set noise ratio: 0.4
Open set noise ratio: 0.2
Imbalance ratio: 50
Close samples: Open samples = 10116: 10000





Thank you