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

How Re-sampling Helps for Long-Tail Learning?

Jiang-Xin Shi[1*] Tong Wei[2*] Yuke Xiang[3] Yu-Feng Li[1†]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

[2]School of Computer Science and Engineering, Southeast University, Nanjing, China

[3]Consumer BG, Huawei Technologies, Shenzhen, China

{shijx,liyf}@lamda.nju.edu.cn, weit@seu.edu.cn, yuke.xiang@huawei.com

NeurlPS 2023

NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Related Work

1. Re-sampling and Re-weighting:

[1] denotes that Class-balance re-sampling can help gain for classifier learning but hurts representation learning.
Two-stage adopt in order not to impact the representation.

2. Head-to-tail Knowledge Transfer:

Assume that the head classes and the tail classes share some common knowledge
such as the same intra-class variances, the same model parameters, or the same semantic features

3. Data Augmentation :

- Mixup can have a positive effect on representation learning but a negative or negligible effect on classifier learning.
- CAM-based methods separate the features, then augments(flipping, rotating) or combines with the components from the tail classes. These methods neglect that the learned model has limited generalization ability on tail classes.

[1] Decoupling representation and classifier for long-tailed recognition. In ICLR, 2020.

Motivation

Two-stage:
Step1: uniform sampler: uniform sampling is beneficial to representation learning.

Step2: class-balanced sampler: class-balanced sampling can be used to fine-tune the linear classifier.

Question: Can re-sampling benefit long-tail learning in the single-stage framework?

Hypothesis: Class-balanced sampler overfits the oversampled irrelevant contexts and learns unexpected spurious correlations

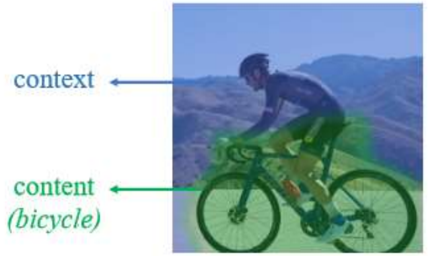


Figure 6: Illustration of context and content. Taking a photo of the bicycle as an example.

[1] Decoupling representation and classifier for long-tailed recognition. In ICLR, 2020.

Motivation/Re-sampling can learn discriminative representations

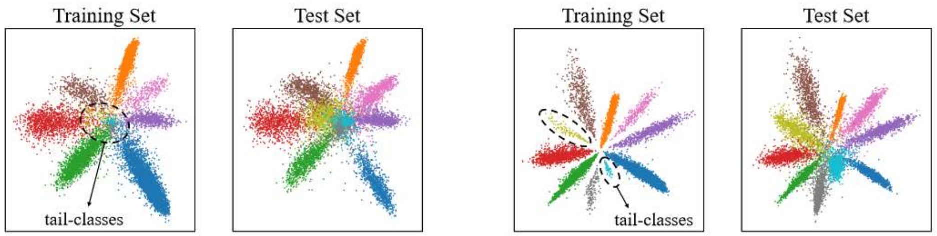
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 1: Test accuracy (%) of CE with uniform sampling, classifier re-training (cRT), and class-balanced re-sampling (CB-RS) on four long-tail benchmarks. We report the accuracy in terms of all, many-shot, medium-shot, and few-shot classes.

	MNIST-LT				Fashion-LT				CIFAR100-LT				ImageNet-LT			
	All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few
CE	65.8	**99.1**	89.9	0.0	45.6	**94.7**	43.1	0.0	39.1	**65.8**	36.8	8.8	35.0	**57.7**	26.5	4.7
cRT	82.5	96.6	89.4	58.8	60.3	77.1	61.4	42.1	**41.6**	63.0	**40.4**	**16.5**	**41.9**	52.9	**39.2**	**23.6**
CB-RS	**90.8**	98.7	**94.4**	**77.7**	**80.5**	86.6	**74.3**	**82.8**	34.1	59.5	31.1	6.2	37.6	47.5	36.5	16.7

In MNIST-LT and Fashion-LT:

- CE vs. cRT(same representation): re-sampling can help for classifier learning.
- cRT vs. CB-RS(same classifier): CB-RS learns batter representation than uniform sampling



(a) Uniform sampling. (b) Class-balanced re-sampling.

Figure 2: Visualization of learned representation of training and test set on MNIST-LT. Using class-balanced re-sampling yields more discriminative and balanced representations.

Motivation/Re-sampling is sensitive to irrelevant contexts

Table 1: Test accuracy (%) of CE with uniform sampling, classifier re-training (cRT), and class-balanced re-sampling (CB-RS) on four long-tail benchmarks. We report the accuracy in terms of all, many-shot, medium-shot, and few-shot classes.

	MNIST-LT				Fashion-LT				CIFAR100-LT				ImageNet-LT			
	All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few
CE	65.8	**99.1**	89.9	0.0	45.6	**94.7**	43.1	0.0	39.1	**65.8**	36.8	8.8	35.0	**57.7**	26.5	4.7
cRT	82.5	96.6	89.4	58.8	60.3	77.1	61.4	42.1	**41.6**	63.0	**40.4**	**16.5**	**41.9**	52.9	**39.2**	**23.6**
CB-RS	**90.8**	98.7	**94.4**	**77.7**	**80.5**	86.6	**74.3**	**82.8**	34.1	59.5	31.1	6.2	37.6	47.5	36.5	16.7

- MNIST and Fashion are highly semantically correlated
- samples on CIFAR and ImageNet contain complex contexts

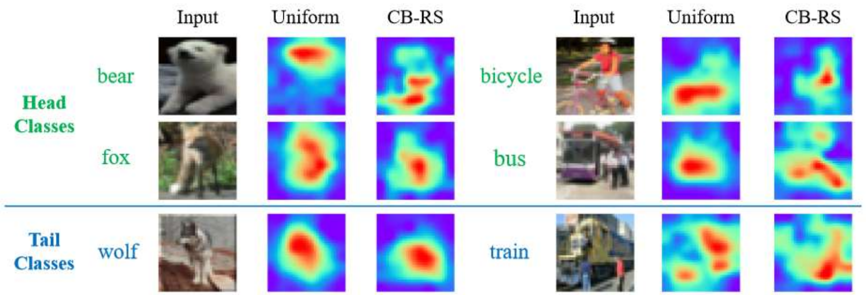


Figure 3: Visualization of features with Grad-CAM [17] on CIFAR100-LT. Uniform sampling mainly learns label-relevant features, while re-sampling overfits the label-irrelevant features.

Motivation/Re-sampling is sensitive to irrelevant contexts

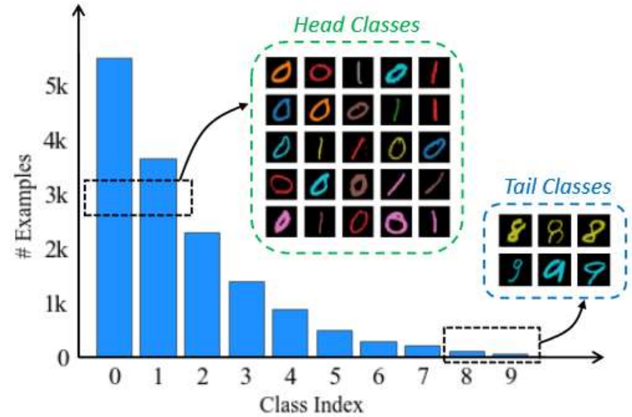
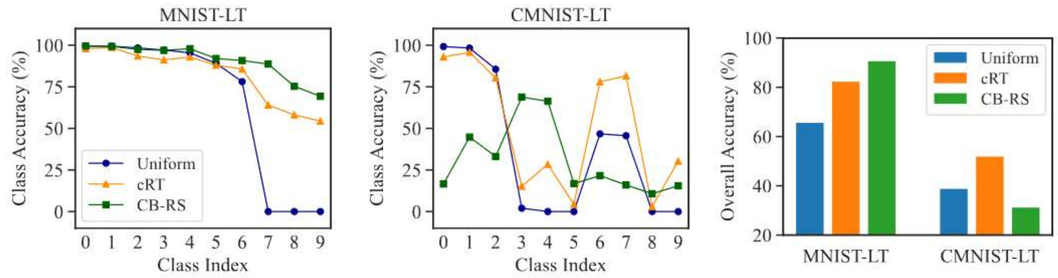


Figure 7: Illustration of the CMNIST-LT benchmark.



(a) Comparison of class accuracy. (b) Overall accuracy.

Figure 4: Comparison of Uniform sampling, cRT, and CB-RS on MNIST-LT and CMNIST-LT.

Method

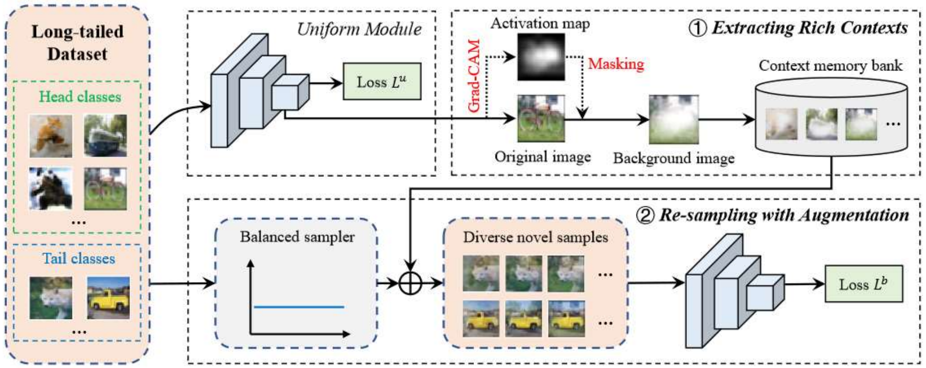


Figure 5: An overview of the proposed method.

Algorithm 1 Training procedure of context-shift augmentation

Input: training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$; context memory bank Q, maximum volume size V; model parameters ϕ, f^u, f^b; loss functions ℓ^u, ℓ^b;

Procedure:

Initialize model parameters ϕ, f^u, f^b;
Re-sampling a class-balanced dataset $\widetilde{\mathcal{D}} = \{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)\}_{i=1}^N$;
Empty memory bank Q;
for epoch $= 1, \ldots, T$ **do**
  repeat
    Draw a mini-batch $(\boldsymbol{x}_i, y_i)_{i=1}^B$ from $\mathcal{D}$;
    Draw a mini-batch $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)_{i=1}^B$ from $\widetilde{\mathcal{D}}$;
    // uniform module
    for $i = 1, \ldots, B$ **do**
      Calculate $\boldsymbol{z}_i^u = f^u(\phi(\boldsymbol{x}_i))$ and $\mathcal{L}_i^u = \ell^u(\boldsymbol{z}_i^u, y_i)$;
      if $p(y = y_i \mid \boldsymbol{x}_i, \phi, f^u) \geq \delta$ **then**
        Calculate background mask $\boldsymbol{M}_i$ of $\boldsymbol{x}_i$;
        Push $(\boldsymbol{x}_i, \boldsymbol{M}_i)$ into Q;
      end if
    end for
    Calculate $\mathcal{L}^u = \frac{1}{B}\sum_{i=1}^B \mathcal{L}_i^u$;
    // balanced re-sampling module
    if Size of Q reaches V **then**
      Obtain contexts $(\check{\tilde{\boldsymbol{x}}}_i, \boldsymbol{M}_i)_{i=1}^B$ from Q;
      $\lambda \sim \mathrm{Uniform}(0, 1)$;
      for $i = 1, \ldots, B$ **do**
        $\tilde{\boldsymbol{x}}_i = \lambda \boldsymbol{M}_i \odot \check{\tilde{\boldsymbol{x}}}_i + (1 - \lambda \boldsymbol{M}_i) \odot \tilde{\boldsymbol{x}}_i$;
        Calculate $\boldsymbol{z}_i^b = f^b(\phi(\tilde{\boldsymbol{x}}_i))$ and $\mathcal{L}_i^b = \ell^b(\boldsymbol{z}_i^b, \tilde{y}_i)$;
      end for
      Calculate $\mathcal{L}^b = \frac{1}{B}\sum_{i=1}^B \mathcal{L}_i^b$;
    else
      Assign $\mathcal{L}^b = 0$;
    end if
    // total objective function
    Calculate $\mathcal{L} = \mathcal{L}^u + \mathcal{L}^b$;
    Update model parameters ϕ, f^u, f^b with $\mathcal{L}$;
  until all training data are traversed.
end for

Experiments

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

Method	CIFAR-10-LT			CIFAR-100-LT		
	100	50	10	100	50	10
CE	70.4	74.8	86.4	38.4	43.9	55.8
mixup [37]	73.1	77.8	87.1	39.6	45.0	58.2
LDAM+DRW [4]	77.1	81.1	88.4	42.1	46.7	58.8
BBN(include mixup) [39]	79.9	82.2	88.4	42.6	47.1	59.2
Remix+DRW(300 epochs) [5]	79.8	-	89.1	46.8	-	61.3
cRT+mixup	79.1 / 10.6	84.2 / 6.89	89.8 / 3.92	45.1 / 13.8	50.9 / 10.8	62.1 / 6.83
LWS+mixup	76.3 / 15.6	82.6 / 11.0	89.6 / 5.41	44.2 / 22.5	50.7 / 19.2	62.3 / 13.4
MiSLAS	**82.1 / 3.70**	**85.7 / 2.17**	**90.0 / 1.20**	**47.0 / 4.83**	**52.3 / 2.25**	**63.2 / 1.73**

Table 4: Top-1 accuracy (%) / ECE (%) for ResNet-32 based models trained on CIFAR-10-LT and CIFAR-100-LT.

[1] Improving calibration for long-tailed recognition. In CVPR. 2021.

NeurIPS 2019

Table 2: Test accuracy (%) on CIFAR datasets with various imbalanced ratios.

Dataset	CIFAR100-LT			CIFAR10-LT		
Imbalance Ratio	100	50	10	100	50	10
CE	38.3	43.9	55.7	70.4	74.8	86.4
Focal Loss [31]	38.4	44.3	55.8	70.4	76.7	86.7
CB-Focal [7]	39.6	45.2	58.0	74.6	79.3	87.1
CE-DRS [15]	41.6	45.5	58.1	75.6	79.8	87.4
CE-DRW [15]	41.5	45.3	58.1	76.3	80.0	87.6
LDAM-DRW [15]	42.0	46.6	58.7	77.0	81.0	88.2
cRT [6]	42.3	46.8	58.1	75.7	80.4	88.3
LWS [6]	42.3	46.4	58.1	73.0	78.5	87.7
BBN [14]	42.6	47.0	59.1	79.8	82.2	88.3
mixup [29]	39.5	45.0	58.0	73.1	77.8	87.1
Remix [33]	41.9	-	59.4	75.4	-	88.2
M2m [32]	43.5	-	57.6	79.1	-	87.5
CAM-BS [13]	41.7	46.0	-	75.4	81.4	-
CMO [27]	43.9	48.3	59.5	-	-	-
cRT+mixup [34]	45.1	50.9	62.1	79.1	84.2	89.8
LWS+mixup [34]	44.2	50.7	62.3	76.3	82.6	89.6
CSA (ours)	45.8	49.6	61.3	80.6	84.3	89.8
CSA + mixup (ours)	**46.6**	**51.9**	**62.6**	**82.5**	**86.0**	**90.8**

Experiments/Ablation

Table 4: Ablation study on the context bank Q.

	All	Many	Med.	Few
Ours	45.8	64.3	49.7	18.2
Ours w/o Q	41.2 (-4.6)	65.1 (+0.8)	41.9 (-7.8)	10.7 (-7.5)

Table 5: Influence of the threshold δ.

δ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Accuracy	45.47	45.37	45.55	44.83	45.52	45.59	45.42	45.08	**45.83**	44.93

the training process progresses, most samples will fit well, so our method is not sensitive to δ

Table 6: Influence of the bank volume size (compared with the mini-batch size B).

Volume	$\times 1$	$\times 2$	$\times 4$	$\times 8$	$\times 16$	$\times 32$	$\times 64$
Accuracy	**45.83**	45.60	45.57	45.32	45.55	45.37	45.26

the latest incoming contexts are more convincing

Experiments/Ablation

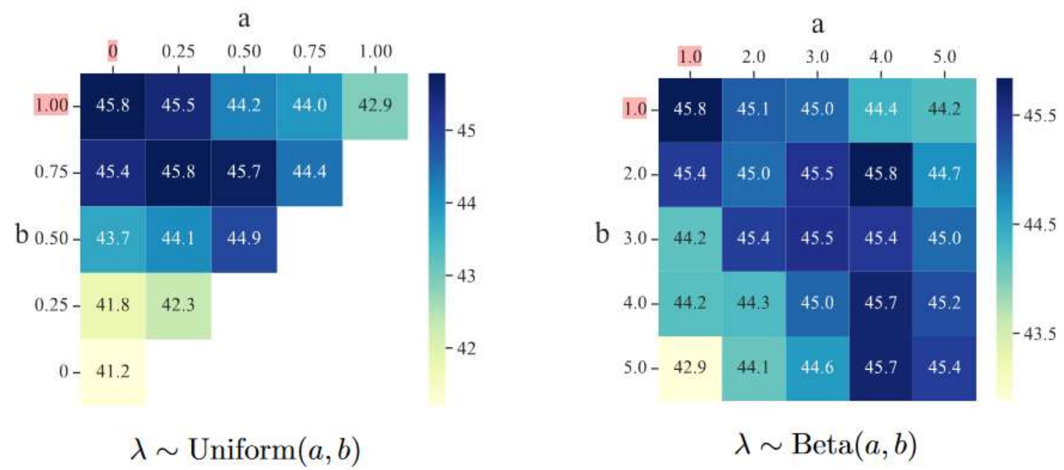


Figure 8: Influence of variants λ.

$$\tilde{\boldsymbol{x}}_i = \lambda \boldsymbol{M}_i \odot \check{\check{\boldsymbol{x}}}_i + (1 - \lambda \boldsymbol{M}_i) \odot \tilde{\boldsymbol{x}}_i$$

[1] Improving calibration for long-tailed recognition. In CVPR. 2021. Beta(1, 1)

Experiments/Ablation

Table 7: Comparison between the uniform module and the re-sampling module in *context-shift augmentation*.

	All	Many	Med.	Few
Uniform module	39.4	**68.3**	37.3	6.1
Re-sampling module	**45.8**	64.3	**49.7**	**18.2**
Ensemble results	43.0	67.5	44.1	11.4

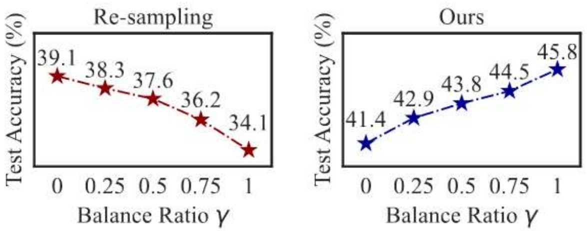


Figure 9: Comparison of re-sampling and our method under different balance ratios γ.

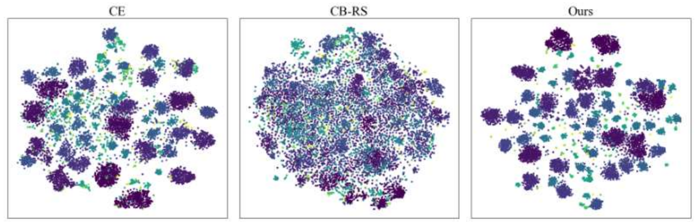


Figure 10: Visualization of learned representation on CIFAR100-LT.

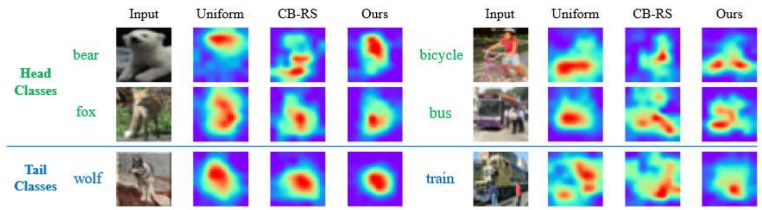


Figure 11: Visualization of features with Grad-CAM on CIFAR100-LT. Our method can alleviate the negative impact on head-class samples caused by the overfitting problem.

Thank you