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

The 37th Conference on Neural Information Processing Systems (NeurIPS 2023)

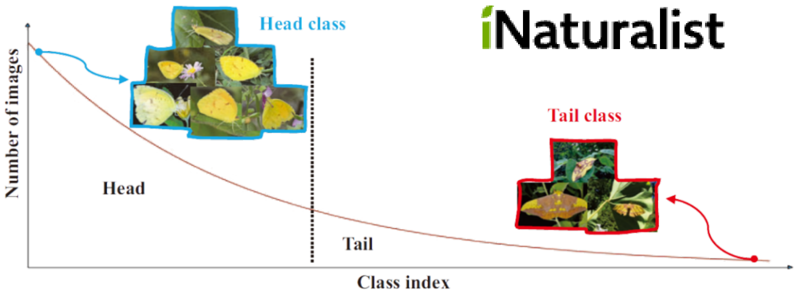
Outline

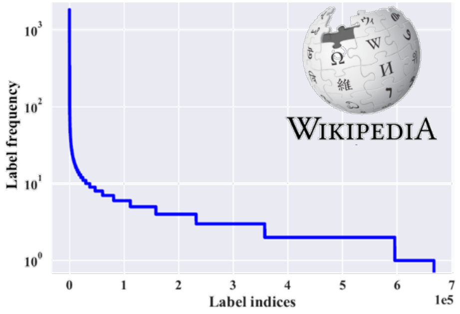
- Background
- Motivation
- Method
- Conclusion

Background

- DNNs have achieved great success by applying well-designed models on large-scale elaborated datasets

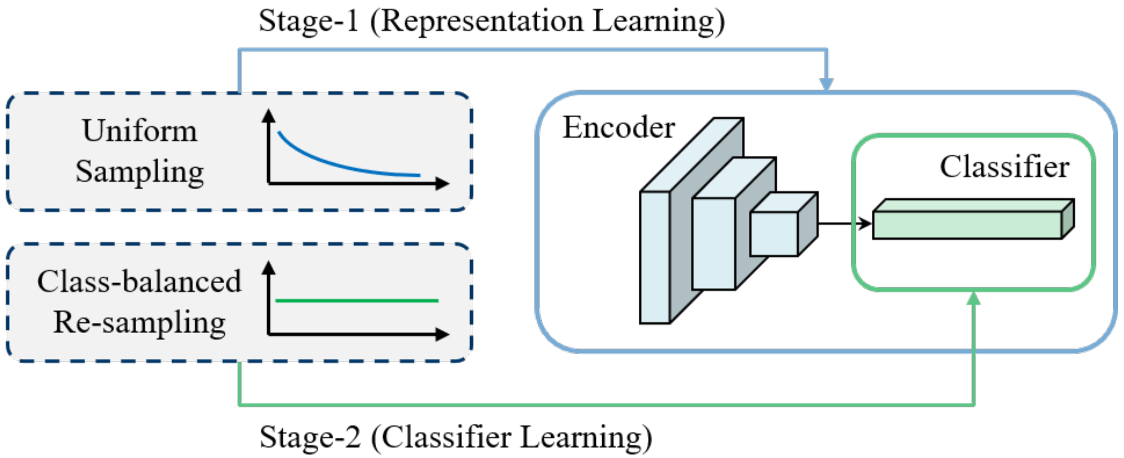
- However, real-world data often exhibits a **long-tail class distribution**





Two-stage Learning

- Stage-1:
  a) Adopt uniform sampling
  b) Jointly train the feature encoder & the classifier



- Stage-2:
  a) Adopt class-balanced re-sampling
  b) Fix the feature encoder
  c) Re-train the classifier

- Representative methods: cRT, DRS, BBN, ……

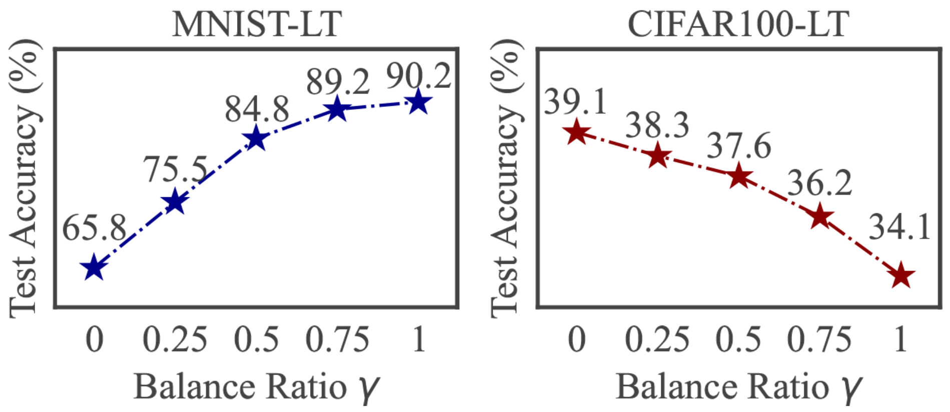
Outline

- ☐ Background
- ☐ Motivation —— *Can re-sampling benefit long-tail learning in the single-stage framework?*
- ☐ Method
- ☐ Conclusion

Motivation

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

- **Re-sampling leads to opposite effects on long-tail datasets**
  - On MNIST-LT dataset,
    Re-sampling **helps** long-tail learning
    (More balanced, more helps).
  - On CIFAR100-LT dataset,
    Re-sampling **harms** long-tail learning
    (More balanced, more harms).



Success/Failure of Re-sampling

- Comparing CE, cRT, CB-RS on four long-tail datasets

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

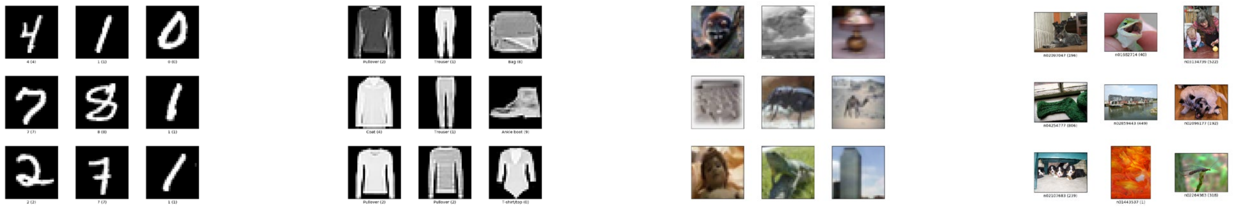
- cRT performs best on CIFAR100-LT and ImageNet-LT, indicating that CB-RS can help for classifier learning, while harms representation learning.
- CE-RS outperforms cRT on MNIST-LT and Fashion-LT, indicating that CB-RS learns better representations than uniform sampling on these two datasets.

Hypothesize

- We hypothesize that **re-sampling is sensitive to the contexts in the samples**

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

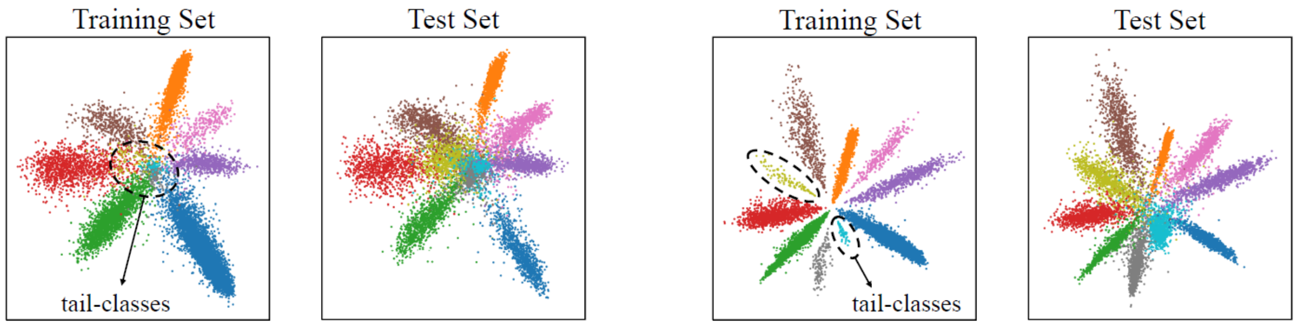


Highly semantically correlated

Contain irrelevant contexts

A Closer Look at Re-sampling

- **Re-sampling can learn discriminative representations**



(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.

- With uniform sampling on MNIST-LT, the representation space is dominated by head classes
- By applying class-balanced re-sampling (CB-RS), both head and tail classes are discriminative.

A Closer Look at Re-sampling

- **Re-sampling is sensitive to irrelevant contexts**

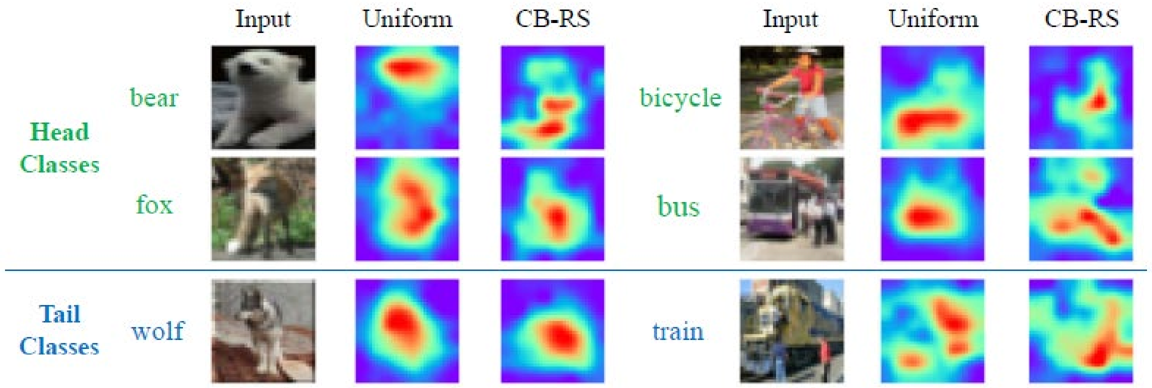


Figure 3: Visualization of features with Grad-CAM [17] on CIFAR100-LT. Uniform sampling mainly learns label-relevant features, while re-sampling overfits the label-irrelevant features.

- On CIFAR100-LT, class-balanced re-sampling (CB-RS) leads to overfitting on the irrelevant contexts from tail classes, and unexpectedly affects the representation of head classes.

Proposed benchmark

- We design Colored-MNIST-LT (CMNIST-LT) by injecting colors into MNIST-LT to artificially construct irrelevant contexts, and compare cRT and CB-RS on these two datasets.

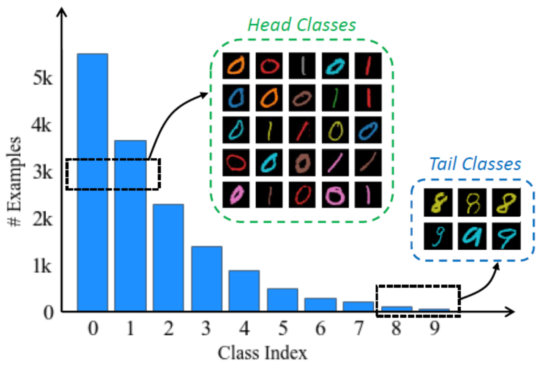
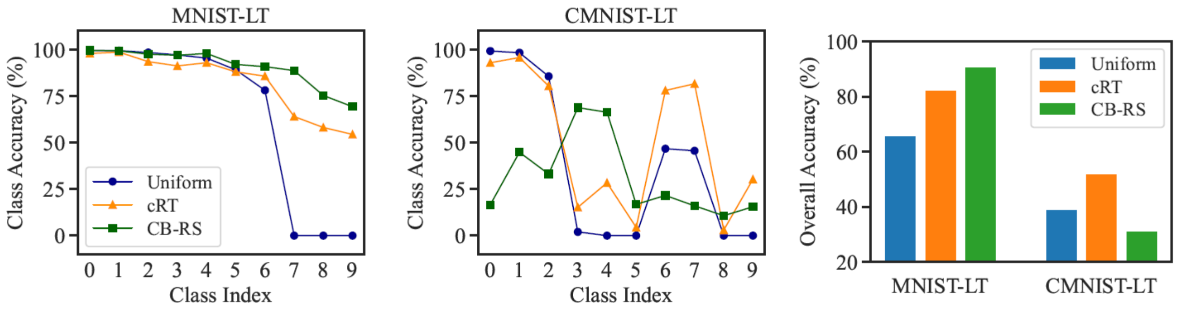


Figure 7: Illustration of the CMNIST-LT benchmark.



(a) Comparison of class accuracy.

(b) Overall accuracy.

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

- The results show that CB-RS succeeds on MNIST-LT and fails on CMNIST-LT, thus validating the negative impact of irrelevant contexts on re-sampling.

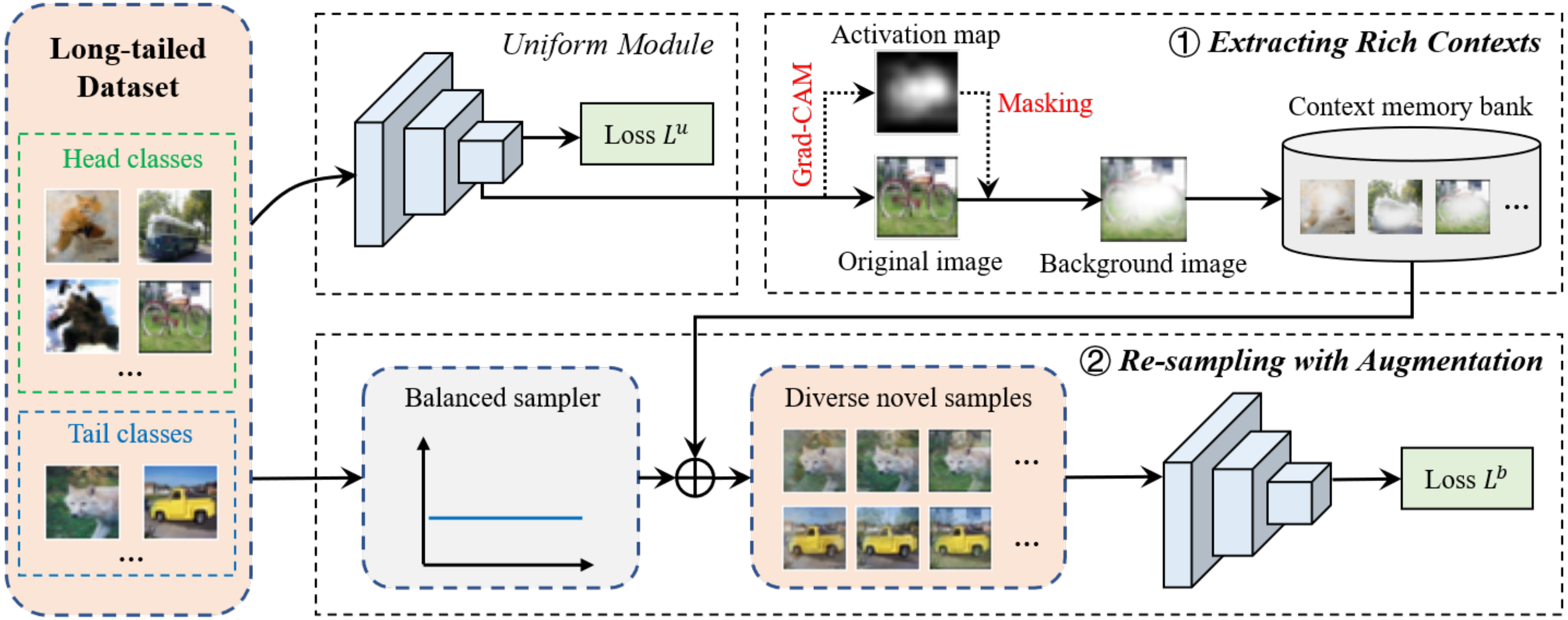
Outline

- Background
- Motivation
- Method —— *How to avoid the irrelevant contexts?*
- Conclusion

Method

- **Context-Shift Augmentation (CSA)**

—— a simple approach to make re-sampling robust to context-shift



Experiments

- Results on long-tail datasets

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**

Table 3: Test accuracy (%) on ImageNet-LT dataset.

	ResNet-10 (All)	ResNet-50			
		All	Many	Med.	Few
CE	34.8	41.6	64.0	33.8	5.8
Focal Loss [31]	30.5	-	-	-	-
OLTR [5]	35.6	-	-	-	-
FSA [28]	35.2	-	-	-	-
cRT [6]	41.8	47.3	58.8	44.0	26.1
LWS [6]	41.4	47.7	57.1	45.2	29.3
BBN [14]	-	48.3	-	-	-
CMO [27]†	-	49.1	67.0	42.3	20.5
CSA (ours)	42.7	49.1	62.5	46.6	24.1
CSA† (ours)	**43.2**	**49.7**	63.6	47.0	23.8

† denotes a longer training of 100 epochs.

✓ CSA outperforms re-sampling/re-weighting, head-to-tail knowledge transfer, and data augmentation methods

Experiments

✓ CSA remedies class-balanced re-sampling

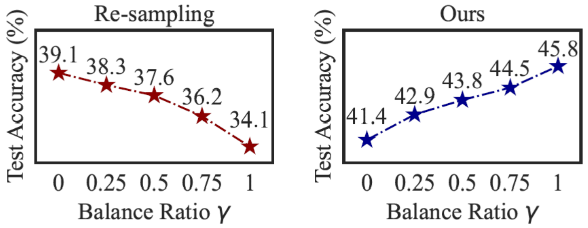


Figure 9: Comparison of re-sampling and our method under different balance ratios γ.

✓ CSA yields better representations

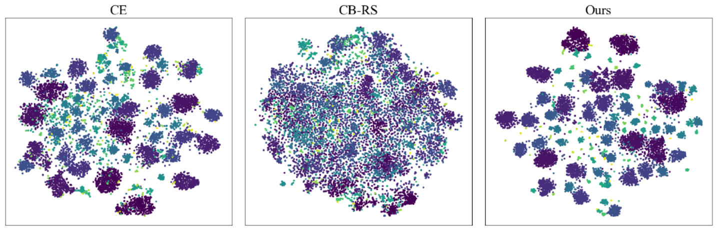


Figure 10: Visualization of learned representation on CIFAR100-LT.

✓ CSA can be integrated with SOTA

Table 11: Accuracy (%) on CIFAR100-LT by integrating the proposed CSA into BCL

Imbalance Ratio	100	50	10
BCL	51.9	56.6	64.9
BCL w/ CSA	**52.6**	**57.1**	**65.8**

✓ CSA does not lead to much overhead

Table 12: Training time cost per epoch on CIFAR100-LT.

	w/ CE	w/ BCL
Single-Branch	2.04 s	4.76 s
Dual-Branch	2.38 s	4.85 s
Ours	2.98 s	5.10 s

Outline

- ☐ Background
- ☐ Motivation
- ☐ Method
- ☐ Conclusion

Conclusion

- This paper investigates the reasons behind the success/failure of re-sampling approaches in long-tail learning

- This paper proposes a new context-shift augmentation module.

Code is available:

Thanks!