LABEL-OCCURRENCE-BALANCED MIXUP FOR LONG-TAILED RECOGNITION

Shaoyu Zhang[1,2], Chen Chen[1,2], Xiujuan Zhang[3], Silong Peng[1,2]*

[1]Institute of Automation, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China
[3] Inner Mongolia Key Laboratory of Molecular Biology on Featured Plants, China

ABSTRACT

Mixup is a popular data augmentation method, with many variants subsequently proposed. These methods mainly create new examples via convex combination of random data pairs and their corresponding one-hot labels. However, most of them adhere to a random sampling and mixing strategy, without considering the frequency of label occurrence in the mixing process. When applying mixup to long-tailed data, a *label suppression* issue arises, where the frequency of label occurrence for each class is imbalanced and most of the new examples will be completely or partially assigned with head labels. The suppression effect may further aggravate the problem of data imbalance and lead to a poor performance on tail classes. To address this problem, we propose *Label-Occurrence-Balanced Mixup* to augment data while keeping the label occurrence for each class statistically balanced. In a word, we employ two independent class-balanced samplers to select data pairs and mix them to generate new data. We test our method on several long-tailed vision and sound recognition benchmarks. Experimental results show that our method significantly promotes the adaptability of mixup method to imbalanced data and achieves superior performance compared with state-of-the-art long-tailed learning methods.

Index Terms— Long-tailed learning, mixup, data augmentation, class-balanced sampler, vision and sound recognition

1. INTRODUCTION

Deep convolutional neural networks have led to a series of breakthroughs for visual and sound recognition. The training of such networks often needs a rich supply of data to improve the generalization ability. In this regard, a number of data augmentation and regularization techniques have been recently proposed, including mixup-based methods [1, 2].

Mixup [1], which generates new examples by combining random data pairs and their labels, has shown promising
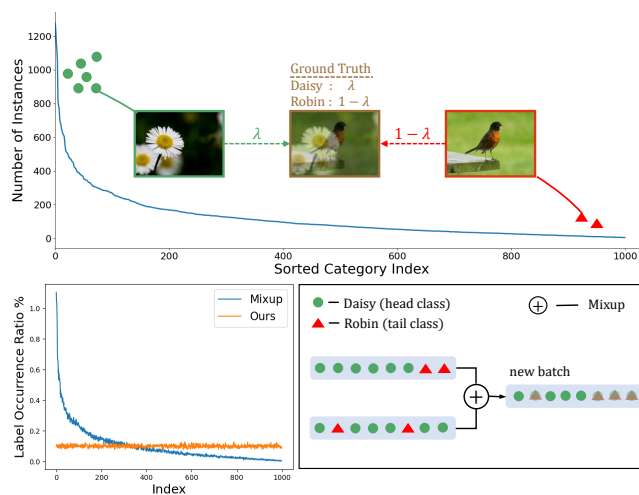
Fig. 1. *Top*: Illustration of the number of instances per class in ImageNet-LT, the distribution of which follows a long-tailed distribution. Mixup creates examples by convex combinations of data pairs and their labels. *Bottom*: Applying mixup to long-tailed data leads to *label suppression*, where head labels occupy the main position and most of the created data will be assigned with head labels (*right*). Label occurrence ratio for each class is highly imbalanced in the mixing process, while our method re-balance this ratio among classes (*left*).

performance on model generalization [3] and calibration [4]. Motivated by this idea, many follow-up methods [5, 6] were proposed and have proved to be effective on commonly used datasets, e.g., CIFAR and ImageNet ILSVRC 2012 [7]. However, these datasets are often collected with relatively balanced data distribution among classes. In contrast, real-world data often follow a long-tailed distribution, where head classes occupy a significantly larger number of data than tail classes. When handling such imbalanced data, mixup may not be an effective method to improve performance. Instead, directly mixing random data pairs and labels may cause a problem that label occurrence among classes is imbalanced and most of the mixed examples will be embedded with features and labels from head classes, which we called *label suppression*.

In this paper, we introduce the concept of *label occurrence ratio* to demonstrate the phenomenon of label suppression. As training data pairs and their labels are mixed with random mixing ratios, each new example may belong to multiple classes proportionally. After mixup, the expected volume of a class in new data can be represented by the sum of the mixing ratios from this class. We therefore define the *label occurrence ratio* as the proportion of the expected volume of a class among all newly created data. As shown in Fig. 1, the label occurrence ratios for different classes from ImageNet-LT are highly imbalanced after mixup. In addition, about 94.7% of new examples will be completely or partially assigned with head labels. The suppression effect actually introduces noise to tail data and further increases difficulty in learning tail classes.

To address this problem, a natural idea is to balance the label occurrence, either by applying class-conditional mixing ratio or class-conditional sampling. Considering that the former may lead to very small mixing ratio for head classes and fail to combine informative features, we apply the latter and adjust sampling strategy to re-balance the distribution of label occurrence. In this regard, class-balanced sampling [8] is a commonly used re-balancing method to learn imbalanced data. Although effective, recent study [9, 10, 11] find it may lead to overfitting on tail classes and hurt the representation learning. However, in long-tailed scenarios, we observe a natural complementarity of class-balanced sampling method and mixup method: mixup method increases the diversity of sampled data and alleviates risk of overfitting on tail classes, while class-balanced sampling helps to keep the mixed label occurrence relatively balanced to alleviate label suppression and learn unbiased classifier. Motivated by this observation, we propose label-occurrence-balanced mixup, which employs two independent class-balanced samplers to generate data pairs with balanced label occurrence among classes (see bottom left of Fig. 1), and then mixes the data pairs to create new data. Despite its simplicity, our method effectively generalizes the mixup-based methods to real-world long-tailed data.

The main contributions of this paper are as follows:

- We explore mixup-based methods in long-tailed scenarios and analyze their failure to achieve expected performance due to label suppression. We further define the label occurrence ratio to demonstrate this phenomenon.

- We discuss the merits/demerits of mixup and class-balanced sampling, and discover a complementary property of these two methods.

- We propose label-occurrence-balanced mixup to alleviate label suppression and significantly improve the performance of mixup in long-tailed scenarios.

2. RELATED WOKS

Mixup. Mixup training [1], which shares the same idea with between-class learning in sound recognition [2], has been shown to substantially improve model generalization and robustness [3]. Manifold mixup [6] extends the idea of mixing data pairs from input space to feature space. Recently, Yun et al. [5] propose CutMix via regional replacement of random data pairs. Besides supervised settings, mixup-based methods also prove to be effective in semi-supervised learning [12].

Long-tailed Recognition. Recent advances in tackling long-tailed challenges are mainly based on re-balancing methods and meta-learning methods. Re-balancing methods can be divided into two regimes: re-sampling [13] and cost-sensitive re-weighting [14, 9]. Re-sampling methods create a relatively balanced data distribution by over-sampling, under-sampling, or class-balanced sampling [8], while re-weighting methods design class-wise weights to adjust learning focus on different classes. In addition, some meta-learning based methods [15] facilitates learning tail data by transferring the information from head classes to tail classes. Beyond that, Remix [16], which applies mixup to long-tailed scenarios, adjusts label distribution by mixing some head-tail data pairs while keeping the ground truth only being the tail label.

3. METHOD

In this section, we firstly provide an overview of mixup method and introduce a metric for evaluating label suppression, and then describe the proposed label-occurrence-balanced mixup.

3.1. Ovreview of Mixup Method

Instead of Empirical Risk Minimization (ERM), mixup method creates new examples in the vicinity of original data and trains model based on the principle of Vicinal Risk Minimization (VRM) [17]. To be specific, given training data x and its label y, mixup creates new example $(\tilde{x}, \tilde{y})$ by combining two random training data (x_i, y_i) and (x_j, y_j) linearly

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \qquad (1)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j, \qquad (2)$$

the combination ratio $\lambda \in (0, 1)$ between the two data points is sampled from the beta distribution $\text{Beta}(\alpha, \alpha)$.

Due to the randomness in selecting data pairs and combination ratio, mixup relaxes the constraint of finite training datasets and encourages model to learn on diverse in-between examples. However, the randomness in sampling data will lead to an imbalanced class distribution for long-tailed data.

3.2. Label-Occurrence-Balanced Mixup

Here, we first introduce *label occurrence ratio* to quantitatively demonstrate the phenomenon of label suppression. Formally, for a dataset consisting of N data points from C classes, mixup shuffles the dataset twice and mixes the $2N$ data points to generate N new examples. For convenience, we double the index and let $A \equiv \{1, \dots, 2N\}$ be the set of indexes of data
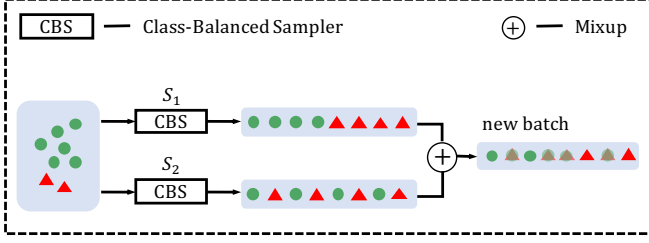
Fig. 2. Framework of our label-occurrence-balanced mixup. We apply two class-balanced samplers, where head classes (green) and tail classes (red) have an equal probability of been selected. Each of the sampler works independently and generates a data batch with uniform distribution among classes. Then a new batch with balanced label occurrence is created by mixing the two batches.

points from the two shuffles and $I(k) \equiv \{i \mid y_i = k, i \in A\}$ be the set of indexes of data points from class k. In addition, the number of data points in class k is denoted by $n_k = |I(k)|$. Assuming the probability of data point (x_i, y_i) being selected is P_i, we define the *label occurrence ratio* for class k as

$$\gamma_k = \frac{\sum_{i \in I(k)} P_i \lambda_i}{\sum_{j \in A} P_j \lambda_j}, \tag{3}$$

where λ_i is the corresponding mixing ratio of (x_i, y_i). In mixup, each data point has the same probability of being selected, i.e., $P_i = 1/(2N)$ for $i \in A$. Because the number of data points n_k for each class is imbalanced, the distribution of γ_k is also imbalanced, which leads to label suppression.

To address the problem of label suppression, the key idea is to balance the distribution of γ_k, either by adjusting λ_i or P_i. We find the former method of adjusting λ_i is sub-optimal, as the mixing ratio for all the head examples will be too small to provide informative feature. Therefore, we re-balance the label occurrence by adjusting P_i. As shown in Fig. 2, we employ two independent class-balanced samplers S_1 and S_2 to generate data pairs. In this case, each class has an equal probability of being selected by the two samplers. The probability of sampling an example with label k is

$$p_k = \frac{1}{C}. \tag{4}$$

Accordingly, the probability of an example (x_i, y_i) of being selected is class-conditional:

$$P_{i|i \in I(k)} = \frac{1}{n_k} p_k = \frac{1}{n_k C}. \tag{5}$$

The above process can be seen as a two-stage sampling operation from one class list and C per-class sample lists, which first selects a class k from the class list and then samples an example from the per-class sample list of class k uniformly.

During training, the samplers generate two data points, $(x_i^{S_1}, y_i^{S_1})$ from S_1 and $(x_j^{S_2}, y_j^{S_2})$ from S_2, respectively. For exhaustively taking advantage of the data diversity, both of the data points are sampled from the whole dataset independently, which are not limited in the same mini-batch. Then we perform mixup on the data pair to create new examples:

$$\tilde{x}_{\text{LOB}} = \lambda x_i^{S_1} + (1 - \lambda) x_j^{S_2} \tag{6}$$

$$\tilde{y}_{\text{LOB}} = \lambda y_i^{S_1} + (1 - \lambda) y_j^{S_2}. \tag{7}$$

Due to the dual class-balanced samplers, the new batch composed of $(\tilde{x}_{\text{LOB}}, \tilde{y}_{\text{LOB}})$ have a balanced distribution of γ_k.

To further improve the training performance, we employ a deferred re-balancing training strategy [9], which first trains with vanilla mixup before annealing the learning rate, and then uses the proposed label-occurrence-balanced mixup.

4. EXPERIMENTS

4.1. Datasets

ESC-50-LT. ESC-50 [18] contains a total of 2000 environmental recordings equally balanced between 50 classes. We select 8 examples per class to form validation set, and sample the rest of examples following Pareto distribution to form a long-tailed training set. The imbalance ratio ρ denotes the ratio between the number of examples of the most frequent class and the least frequent class. We set $\rho = 10$ for ESC-50-LT.
CIFAR-10-LT and CIFAR-100-LT. Following the prior work [14, 9], we use the long-tailed version of the CIFAR-10 and CIFAR-100 datasets with $\rho = 10, 100$.
ImageNet-LT. ImageNet-LT is constructed by sampling a long-tailed subset of ImageNet-2012 [7]. It has 115.8K images from 1000 categories, with the number of images per class ranging from 1280 to 5.

4.2. Implementation Settings

For sound recognition on ESC-50-LT, we use EnvNet [19] and EnvNet-v2 [2] as backnone networks and follow the training settings of [2]. For visual recognition on CIFAR-10/CIFAR-100-LT, we train a ResNet-32 backnone network for 200 epochs, following [9]. For ImageNet-LT, we choose ResNet-10 as the backbone network and train the model for 90 epochs, following [11]. The mixing ratio $\lambda \sim \text{Beta}(\alpha, \alpha)$, where we set $\alpha = 0.2$ for ImageNet-LT, and $\alpha = 1$ for other datasets.

4.3. Experimental Results

Competing methods. We compare the proposed label-occurrence-balanced mixup with ERM baseline and three groups of methods: 1) mixup-based methods, including mixup training [1], manifold mixup [6] and Remix [16], where mixup is replaced by between-class (BC) learning [2] for sound recognition on ESC-50-LT; 2) sampling-based methods,

Dataset	CIFAR-10-LT		CIFAR-100-LT		ImageNet-LT
Imbalance ratio	100	10	100	10	256
ERM	29.7	12.9	60.7	43.4	65.4
Mixup [1]	28.4	11.5	59.1	41.6	67.2
Manifold Mixup [6]	30.2	13.3	60.4	42.6	67.5
Remix [16]	27.0	11.5	58.6	40.5	66.6
Ours	25.8	10.6	58.5	40.1	63.0
CB Samp. [8]	31.6	13.1	68.1	45.0	64.4
CB Samp.* [9]	26.5	12.3	58.5	42.4	60.1
LDAM-DRW [9]	23.0	11.8	58.0	41.3	64.0
BBN [10]	22.0	12.7	57.4	40.9	-
Logit Adj. [20]	22.3	11.8	56.1	42.3	-
LFME [21]	-	-	57.7	-	62.8
Ours*	**21.3**	**10.4**	**53.8**	**38.9**	**59.6**

Table 1. Top-1 validation error rates on CIFAR-10-LT/CIFAR-100-LT and ImageNet-LT. ∗ denotes the deferred re-balancing version of corresponding methods.

Backbone	EnvNet [19]		EnvNet-v2 [2]	
Error rates (%)	Top-1 error	Top-5 error	Top-1 error	Top-5 error
ERM	54.7	24.2	52.2	23.7
CB Samp. [8]	55.7	25.3	54.6	29.3
BC$_{(mixup)}$ [2]	47.0	25.2	45.2	20.8
Ours	**44.9**	**21.4**	**42.6**	**19.2**

**Table 2**. Top-1/Top-5 validation error rates on ESC-50-LT for EnvNet and EnvNet-v2.

including class-balanced sampling (CB Samp.) [8], deferred class-balanced sampling (CB Samp.*) [9]; 3) state-of-the-art methods, including recently proposed LDAM-DRW [9], BBN [10], logit adjustment loss [20] and LFME [21]. Our method is denoted as *Ours*, and the deferred re-balancing version of our method is denoted as *Ours**.

Results for visual recognition are reported in Table 1. Compared with mixup, our method obtains 4.2% relative improvement on ImageNet-LT. Our method also outperforms other mixup-based methods like manifold mixup and Remix. It is worth noting that class-balanced sampling method leads to even worse performance on some datasets, e.g., CIFAR-100-LT, while our method shows consistent performance gains on all the reported benchmarks. Furthermore, by integrating with deferred re-balancing training strategy, our method achieves lower error rates than most of the state-of-the-art methods, such as logit adjustment [20] and BBN [10].

Results for sound recognition show similar trends. As shown in Table 2, class-balanced sampling gets worse performance, probably due to overfitting on repeated samples. BC learning outperforms ERM, while our method further improves the performance for both EnvNet and EnvNet-v2 backbones.

4.4. Ablation Study

Alternating two samplers. The key idea of our method is to balance the label occurrence ratio by employing two class-balanced (CB) samplers. Here we discuss the effect of alternating the two samplers. Mixup could be seen as using
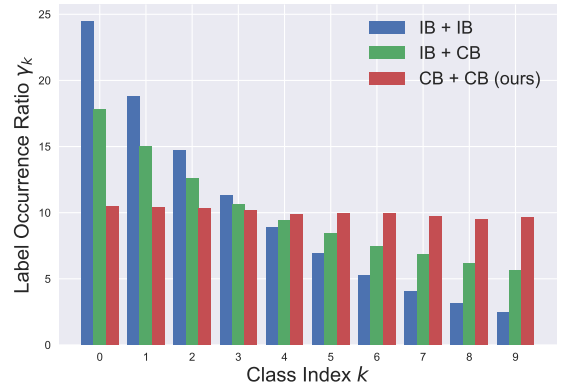
(a) γ_k for alternating samplers on CIFAR-10-LT, $\rho = 10$.

Combination of samplers	$\gamma_{max}/\gamma_{min}$	Top-1 error (%)
IB Sampler + IB Sampler (mixup)	9.91	11.5
IB Sampler + CB Sampler	3.16	11.4
CB Sampler + CB Sampler (ours)	**1.10**	**10.6**

(b) The correlation between performance and balance of γ_k.

Fig. 3. Ablation study for alternating the two samplers on CIFAR-10-LT, with imbalance ratio $\rho = 10$.

two instance-balanced (IB) samplers, where each example has the same probability of being selected. Beyond that, there is another case that an instance-balanced sampler and a class-balanced sampler are both used. We analyze the correlation between model performance and the balance of γ_k. Fig. 3(a) shows the distribution of γ_k for the three cases of combining samplers. We find that both the cases of IB Sampler + IB Sampler and IB Sampler + CB Sampler lead to an imbalanced distribution of γ_k, while our method achieves a balanced distribution. From Fig. 3(b) we can see that a more balanced γ_k, i.e., $(\gamma_{max}/\gamma_{min} \to 1)$, leads to a better model performance. **Adjusting λ or adjusting P.** In Equation 3, γ_k could be adjusted either by adjusting λ or P. For the former, to balance the distribution of γ_k, one may reduce each mixing ratio λ for head examples to a very small value, which is impracticable to provide informative features. Instead of constraining the mixing ratio λ directly, our method balances the summation of λ by controlling the probability P of sampling example from different classes. In experiments, adjusting λ achieves top-1 error rates of 26.2% and 15.9% for CIFAR-10-LT with $\rho = 100$ and $\rho = 10$, respectively, while our method achieves 25.8% and 10.6%, which shows more robust performance.

5. CONCLUSION

In this paper, we propose label-occurrence-balanced mixup, which addresses the problem of label suppression and generalizes mixup-based methods to real-world long-tailed scenarios. Label-occurrence-balanced mixup is a simple and effective method that shows consistent improvements on several vision and sound recognition benchmarks.

6. REFERENCES

[1] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[2] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada, "Learning from between-class examples for deep sound recognition," in *International Conference on Learning Representations*, 2018.

[3] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou, "How does mixup help with robustness and generalization?," in *International Conference on Learning Representations*, 2020.

[4] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[6] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6438–6447.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[8] Li Shen, Zhouchen Lin, and Qingming Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 467–482.

[9] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems*, 2019.

[10] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9719–9728.

[11] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *Eighth International Conference on Learning Representations (ICLR)*, 2020.

[12] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *International Joint Conference on Artificial Intelligence*, 2019, pp. 3635–3641.

[13] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[14] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[15] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2537–2546.

[16] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan, "Remix: Rebalanced mixup," in *European Conference on Computer Vision*. Springer, 2020, pp. 95–110.

[17] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik, "Vicinal risk minimization," *Advances in neural information processing systems*, pp. 416–422, 2001.

[18] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[19] Yuji Tokozume and Tatsuya Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2721–2725.

[20] Aditya Krishna Menon, Andreas Veit, Ankit Singh Rawat, Himanshu Jain, Sadeep Jayasumana, and Sanjiv Kumar, "Long-tail learning via logit adjustment," in *International Conference on Learning Representations (ICLR) 2021*, 2021.

[21] Liuyu Xiang, Guiguang Ding, and Jungong Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *European Conference on Computer Vision*. Springer, 2020, pp. 247–263.