Pattern Recognition 141 (2023) 109608

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Margin-aware rectified augmentation for long-tailed recognition

Liuyu Xiang[a], Jungong Han[b], Guiguang Ding[c,*]

[a] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[b] Department of Computer Science, Aberystwyth University, UK
[c] Beijing National Research Center for Information Science and Technology (BNRist), School of Software, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 7 September 2022
Revised 15 March 2023
Accepted 11 April 2023
Available online 18 April 2023

Keywords:
Long-tailed recognition
Data augmentation
Mixup

ABSTRACT

The long-tailed data distribution is prevalent in real world and it poses great challenge on deep neural network training. In this paper, we propose Margin-aware Rectified Augmentation (MRA) to tackle this problem. Specifically, the MRA consists of two parts. From the data perspective, we analyze that data imbalance will cause the decision boundary be biased, and we propose a novel Margin-aware Rectified mixup (MR-mixup) that adaptively rectifies the biased decision boundary. Furthermore, from the model perspective, we analyze that the imbalance will also lead to consistent 'gradient suppression' on minority class logits. Then we propose Reweighted Mutual Learning (RML) that provides extra 'soft target' as supervision signal and augments the 'encouraging gradients' on the minority classes. We conduct extensive experiments on benchmark datasets CIFAR-LT, ImageNet-LT and iNaturalist18. The results demonstrate that the proposed MRA not only achieves state-of-the-art performance, but also yields a better-calibrated prediction.

1. Introduction

In recent years, the development of deep neural networks has achieved great success in various applications, such as object recognition, object detection and so on. To train a well-performed deep neural network, one has to collect a dataset with sufficient training samples for all categories such as ImageNet [1]. However, it is often costly to collect such a balanced dataset in real-world applications. For example, in applications such as large-scale object recognition with thousands of categories, it is easy to collect a large number of images of sparrows, but it would cost much more effort to collect sufficient samples of, let's say, armadillo. Moreover, in some cases, it is even infeasible to collect a balanced training set as some samples are so rare to be collected. For example, in medical image analysis, some types of diseases are so rare that there could be only a few samples available.

The long-tailed phenomenon greatly limits the robustness and generalizability of current deep learning algorithms in various applications [2]. Since most deep learning algorithms are designed with the assumption of a balanced training set, they tend to be severely biased when training with long-tailed distributed dataset. In such cases, the deep neural networks tend to be overconfident on the majority categories, and perform poorly on minority categories.

Current approaches for long-tailed recognition can be broadly categorized into two types: one-stage and multi-stage. One-stage methods do not change the original training procedure and use reweighting or resampling to alleviate the dominance of the majority classes. Specifically, reweighting methods design novel loss functions to reweigh instances from different categories, so that the impact from the majority classes will be downgraded and the dominance is alleviated. The resampling methods, on the other hand, adopt majority undersampling or minority oversampling, so that a more balanced training distribution can be obtained. The multi-stage methods usually adopt multiple training phases to obtain a more robust network, including decoupling methods, mixture of experts and meta/transfer learning methods, etc. The decoupling methods [3] decouples the representation learning and classifier and train them separately in a two-stage stage manner. The mixture of experts methods [4,5] learn an ensemble of classifiers to debias the data imbalance. The meta learning and transfer learning methods [6], however, tackle the long-tail problem by transferring the rich knowledge from the majority classes to tail classes.

Orthogonal to the aforementioned methods, we propose a augmentation method termed Margin-aware Rectified Augmentation (MRA), to tackle the long-tailed classification problem. The MRA stems from the following observations: firstly, from the data perspective, the poor performance of minority classes derives from

* Corresponding author.
E-mail addresses: xiangly@bupt.edu.cn (L. Xiang), dinggg@tsinghua.edu.cn (G. Ding).

https://doi.org/10.1016/j.patcog.2023.109608
0031-3203/

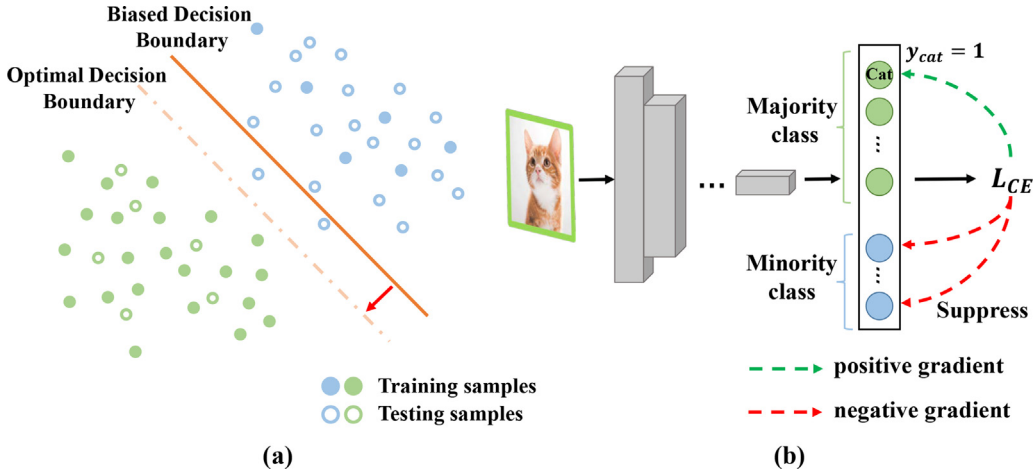

Fig. 1. The long-tailed distribution brings up two problems: (a) a biased decision boundary and (b) consistent gradient suppression on minority categories, both of which we aim to address in this paper.

insufficient training data, or in other words, the lack of variations and diversity of minority categories, which can be alleviated by the data synthesis or augmentation methods such as Mixup [7] and SMOTE [8]. However, current augmentation approaches are not particularly designed for (e.g., mixup), or neglect the characteristics of long-tailed problem (e.g., SMOTE), that is, a biased decision boundary. As shown in Fig. 1(a), the observed minority training samples (solid circle) **can not well represent the underlying distribution** due to the lack of sufficient data. When trained with standard empirical risk minimization, the territory of minority classes will be narrowed and the decision boundary is biased, in which case the minority testing samples (hollow circle) could easily fall on the other side of the decision boundary and be misclassified. To rectify the resulting biased decision boundary, we borrow the theoratical analysis of optimal margin from Cao et al. [9], which indicates that that for two classes a, b, the optimal margin that minimizes the generalization error bound γ_a, γ_b satisfies: $\frac{\gamma_b}{\gamma_a} = \frac{N_a^{1/4}}{N_b^{1/4}}$. When mixing up two samples from different categories, we rectify the mixed up label accordingly so that the decision boundary can also be rectified through data augmentation.

Secondly, from the model perspective, we analyze that the data imbalance also leads to an imbalanced gradient distribution. As shown in Fig. 1(b), during deep neural network training, the minority class logits will receive much more 'discouraging gradients' (the red dashed lines) than 'encouraging gradients' (the green dashed lines) from the majority categories (e.g., the majority class *cat*), due to their insufficient training data. In other words, they are consistently suppressed by the majority classes during training. The resulting consequence is that majority predictors tend to be overconfident and minority samples tend to be misclassified as majority ones. To address the gradient imbalance issue, we propose Reweighted Mutual Learning (RML) to provide an extra 'soft target' as supervision and augment the encouraging positive gradients on the minority classes. To be more specific, we train two identical peer networks and teach each other mutually via knowledge distillation [10]. Due to the stochastic characteristic of deep neural network training, the peer network will learn in a complementary way and promote each other [11]. Moreover, we design a reweighting scheme so that the augmented positive gradients of minority samples will be emphasized.

We conduct extensive experiments on the long-tailed benchmark datasets including CIFAR-LT, ImageNet-LT and iNaturalist18. We show that the proposed MRA can be easily combined with one-stage and multi-stage methods and achieve state-of-the-art performances. We also analyze the Expected Calibration Error (ECE) and show that the MRA can also alleviate the misaligned confidence in long-tailed classification.

The key contributions of this paper are as follows:

- From the data aspect, we analyze the necessity of debiasing decision boundary and propose Margin-aware Rectified mixup, which enlarges the tail class margins by rectifying the labels of the augmented samples.
- From the model aspect, we analyze the gradient suppression on minority classes and propose Reweighted Mutual Learning to augment 'encouraging gradients' for the minority classes.
- We conduct extensive experiments on three benchmark datasets: CIFAR-LT, ImageNet-LT and iNaturalist18 to verify the effectiveness of MRA. We also demonstrate through visualizations of decision boundary, ablation studies and calibration analysis that the proposed MRA not only achieves state-of-the-art performance, but is also efficient, flexible and reliable.

The rest of this paper is organized as follows: related works on long-tailed recognition are reviewed in Section 2. Section 3 describes the proposed MRA method in detail. Then the experimental results are demonstrated in Section 4 and finally we draw the conclusion in Section 5.

2. Related work

The long-tailed classification problem, especially in the context of deep learning, has gained increasing attention in recent years. Current long-tailed classification approaches can be broadly divided into one-stage and multi-stage methods based on their training stages.

2.1. One-stage methods

One-stage methods maintain the common deep neural network training procedure and use techniques such as resampling and reweighting to tackle the data imbalance.

Resampling methods either under-sampling majority classes or over-sampling minority classes. For undersampling methods, one common approach is to randomly discard majority class samples so that a more balanced and uniform training distribution is sampled [12,13].

For oversampling methods, one representative is the Synthetic Minority Oversampling TEchnique (SMOTE) [8], where minority samples are oversampled by interpolating synthetic minority instances. Han et al. [14] propose borderline-SMOTE that oversamples the borderline instances. Maldonado et al. [15] propose a

feature-weighted oversampling approach to tackle the data imbalance. While resampling could alleviate the imbalance during training, it may also lead to information loss (under-sampling) or minority class overfitting (over-sampling). More recently, Balanced softmax loss [16] is proposed to optimize a parameterized sampling strategy via meta-learning.

Reweighting methods usually design cost-sensitive loss functions to alleviate the majority class dominance during training. This is usually achieved by down-weighting majority class losses. Focal Loss [17] proposes to downweigh the loss of well-classified examples. Label-Distribution-Aware Margin Loss (LDAM) [9] provide theoretical analysis on the optimal margin between minority and majority classes, and encourage the minority categories to have larger margins. Class-balanced Effective Number Loss proposes [18] effective number of samples for reweighting. Equalization Loss and its variants [19,20] analyze that the tail classes are consistently suppressed in terms of gradient in imbalanced object detection, and propose reweighting to alleviate the gradient suppression issue. Du et al. [21] propose parameter-free loss which requires no hyper-parameter tuning. There are also efforts trying to automatically learn a reweighting function via meta learning [22–24].

2.2. Multi-stage methods

Multi-stage methods modify the original training procedure and involve multiple training stages, including decoupling methods, mixture of experts, head-to-tail transfer learning, etc.

Decoupling methods is first proposed in Kang et al. [3] where representation network is first learned with ordinary instance-balanced sampling during the first stage and the classifier is re-trained in the second stage with class-balanced resampling. There are also variants including DisAlign [25] and LogitAdjust [26]. The DisAlign argues that the representation network is already well-trained in the first stage of Decouple and proposes a generalized alignment strategy in the second stage. LogitAdjust adjusts the logits with a label-dependent offset and can be applied either post-hoc or during training.

Mixture of experts methods [4,5,27] usually use multiple networks in an ensemble manner to promote the performance. LFME [4] trains multiple expert networks on different subsets and distills a unified network while RIDE [5] learns a dynamic routing scheme between multiple expert networks with shared parameters.

Transfer learning or meta learning methods transfer the knowledge from majority to minority classes so that the performance on minority categories is promoted. MetaModelNet [28] proposes to progressively learn a transformation mapping from head to tail classifiers/regressors while OLTR [6] proposes to learn a meta embedding with a memory module for head-to-tail knowledge transfer.

2.3. Data augmentation methods

Apart from common one-stage and multi-stage methods, various data augmentation methods [29–31] are also proposed to promote the long-tailed performance. Among these methods, LEAP [29] models the feature distribution as Gaussian distribution and transfers the majority distribution to minority. M2m [32] generates minority samples by transforming from majority samples in an adversarial-like strategy. DFG (Discriminative Feature Generation). Suh et al. [33] is trained to generate discriminative feature via attention maps. MetaSAug [31] on the other hand, adopt meta learning to automatically learn an implicit semantic data augmentation. While these methods have greatly promoted the performance, they usually require sophisticated augmentation calculation and lacks flexibility. The closest method to our proposed MRA is Remix [30], which modifies mixup [7] to cope with the imbalance. Both Remix and the proposed MR-mixup aim to modify the labels of the augmented samples, so that the classifier tends to predict more of tail classes. However, remix is designed with hand-crafted rules with several extra hyperparameters. In contrast, the proposed MR-mixup rectifies the margin continuously with theoretical support and with no hyperparameters. Moreover, we also introduce RML that augments the tail class gradient and further improves the long-tailed recognition performance.

However, it is designed with hand-crafted discrete rectification and requires several extra hyperparameters. By contrast, the proposed MR-mixup rectifies the margin continuously and requires no extra hyperparameters and can be regarded as a generalization of Remix.

Compared with previous augmentations methods, the proposed MRA can not only be efficiently plugged in to any one or multi-stage methods, but are also particularly designed to cope with the characteristics of long-tailed recognition, i.e., biased decision boundary and consistent minority gradient suppression.

3. Proposed method

3.1. Overview and problem setup

As illustrated in Fig. 1, the proposed MRA mainly focuses on addressing two issues: (1) rectifying biased decision boundary through data augmentation, (2) augmenting positive gradients on minority classes via reweighted mutual learning. We first briefly introduce the problem setting, then elaborate on details of MR-mixup and RML.

Problem setup. Suppose we have a training dataset with long-tailed distribution: $\mathcal{D}_{train} = \{x_i, y_i\}$, $i \in \{1, \dots, N\}$ where x_i is the i-th data point with label y_i, and N is the total number of training samples. We denote C as the total number of classes and n_c to be the number of samples for class c where $\sum_i n_i = N$. We use imbalance ratio to indicate the ratio between the largest and smallest n_i, i.e., $\max\{n_i\}/\min\{n_i\}$

Without loss of generality, we assume that the classes are sorted by their cardinality in decreasing order, such that $n_1 \geq n_2 \geq \dots \geq n_C$. Since the training set is long-tailed, we have $n_1 \gg n_C$. For test-time evaluation, we have a balanced test set $\mathcal{D}_{test}$, such that $N_1^{test} \simeq N_C^{test}$.

3.2. Margin-aware rectified mixup

Firstly, we briefly introduce preliminaries including mixup [7] augmentation and optimal margin analysis [9]. Then we will describe how to design MR-mixup with theoretical support.

mixup. is a well-known data augmentation method which proves to be beneficial for neural network generalization. Given two samples $(x_1, y_1), (x_2, y_2)$, the augmented data point is formulated as follows:

$$\tilde{x} = \lambda x_1 + (1-\lambda)x_2 \qquad \tilde{y} = \lambda y_1 + (1-\lambda)y_2 \tag{1}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$

Optimal margin. is introduced as follows:

Theorem 1 ([9]). *For binary classification, let $\mathcal{F}$ be a hypothesis class of neural networks with Rademacher complexity upper bound $\mathfrak{R}_j \leq \sqrt{\frac{C(\mathcal{F})}{n_j}}$ where n_j denotes number of samples in class j. Suppose some classifier $f \in \mathcal{F}$ can achieve a total sum of margins $\gamma_1' + \gamma_2' = \beta$ with margins $\gamma_1', \gamma_2' > 0$. Then there exists a classifier $f^* \in \mathcal{F}$ with margins*

$$\gamma_1^* = \frac{\beta n_2^{1/4}}{n_1^{1/4} + n_2^{1/4}}, \quad \gamma_2^* = \frac{\beta n_1^{1/4}}{n_1^{1/4} + n_2^{1/4}} \tag{2}$$

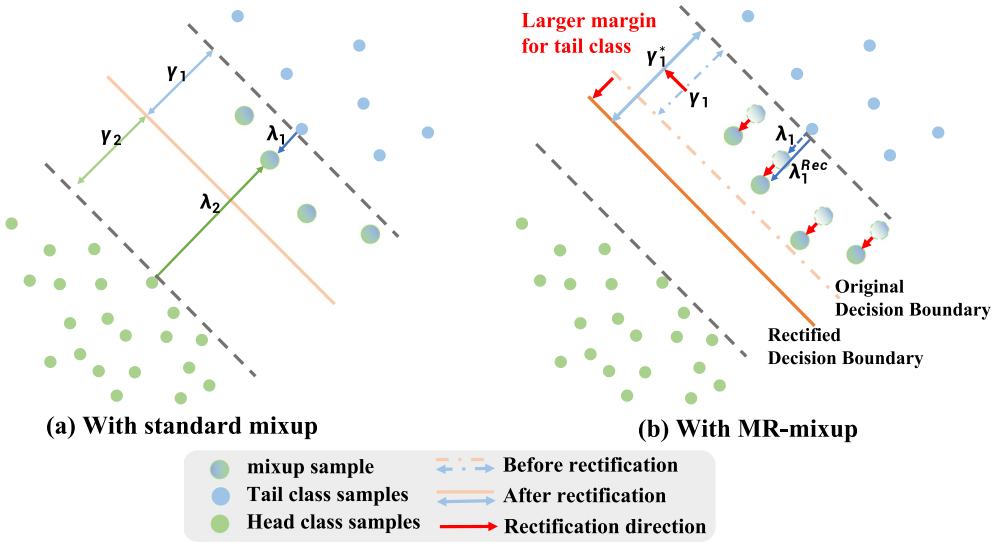

Fig. 2. Illstration of MR-mixup. Since our goal is to rectify the decision boundary and let margin γ becomes γ^*, we propose to rectify the labels of the augmented samples where λ is rectified to λ^{Rec} proportionally.

that obtains the optimal generalization error.

While mixup improves the neural network's generalizability, we further incorporate the optimal margin into its formulation and propose MR-mixup. Specifically, we propose to 'shift' the label of the mixed up sample according to the optimal margin (see the red arrow in Fig. 2), so that the margin is adjusted and the decision boundary is rectified. We adopt to linearly rectify the mixed up label for calculation simplicity.

First, we consider the case where $\lambda < 0.5$, that is, the mixed up sample falls on the minority side of the decision boundary, as illustrated in Fig. 2. Suppose we have a mixed up sample $\tilde{x} = \lambda x_1 + (1-\lambda)x_2$ with label $\tilde{y} = \lambda y_1 + (1-\lambda)y_2$. When trained with ordinary empirical risk minimization, the decision boundary with margin $\gamma_1 = \gamma_2$ is biased due to the under-represented minority distribution. Then we wish to rectify the decision boundary (orange solid line) to the one with optimal margin (orange dashed line). Recall the optimal margin in Theorem 1, the minority class margin should be enlarged from $\gamma_1 = (\gamma_1+\gamma_2)/2$ to

$$\gamma_1^* = \frac{n_2^{1/4}(\gamma_1+\gamma_2)}{n_1^{1/4}+n_2^{1/4}} \tag{3}$$

In order to rectify the decision boundary to reach the optimal margin γ_1^*, we propose to rectify the label $\tilde{y}$ of the mixed up sample proportionally. Concretely, we linearly 'shift' the mixing factor λ_1 so that the rectified λ_1^{Rec} satisfies:

$$\frac{\lambda_1}{\lambda_1^{Rec}} = \frac{\gamma_1}{\gamma_1^*} = \frac{(\gamma_1+\gamma_2)/2}{\frac{n_2^{1/4}(\gamma_1+\gamma_2)}{n_1^{1/4}+n_2^{1/4}}} \tag{4}$$

Then we have:

$$\lambda_1^{Rec} = \frac{2\lambda n_2^{1/4}}{n_1^{1/4}+n_2^{1/4}} \tag{5}$$

Then the MR-mixup sample's label is rectified as:

$$\tilde{y}^{Rec} = \lambda_1^{Rec} y_1 + (1-\lambda_1^{Rec})y_2 \tag{6}$$

Similarly, when $\lambda \geq 0.5$, that is, the mixed up sample falls on the majority side of the decision boundary. In this case, the optimal majority class margin satisfies:

$$\gamma_2^* = \frac{n_1^{1/4}(\gamma_1+\gamma_2)}{n_1^{1/4}+n_2^{1/4}} \tag{7}$$

Then we consider linearly rectifying λ_2 to λ_2^{Rec} so that:

$$\frac{\lambda_2}{\lambda_2^{Rec}} = \frac{\gamma_2}{\gamma_2^*} = \frac{(\gamma_1+\gamma_2)/2}{\frac{n_1^{1/4}(\gamma_1+\gamma_2)}{n_1^{1/4}+n_2^{1/4}}} \tag{8}$$

Similarly, we have:

$$\lambda_2^{Rec} = \frac{2\lambda n_1^{1/4}}{n_1^{1/4}+n_2^{1/4}} \tag{9}$$

Then the MR-mixup sample's label is rectified as:

$$\tilde{y}^{Rec} = (1-\lambda_2^{Rec})y_1 + \lambda_2^{Rec} y_2 \tag{10}$$

Note that while the MR-mixup rectifies the labels of the augmented samples $\tilde{y}$, the original augmented input $\tilde{x}$ is unchanged. We wish the prediction of the classifier leaning towards the tail class given the same augmented sample $\tilde{x}$. In this way, the tail class margin is enlarged accordingly.

Toy examples To give an intuitive understanding of how MR-mixup rectifies decision boundary, we conduct experiments on two toy imbalanced datasets *two moon* and *circle*. We train a one hidden layer MLP with imbalance ratio $= 5$. Then we plot $p(y|x)$ as well as test samples shown in Fig. 3. We compare MR-mixup with (1) without any augmentation, (2) with regular mixup, (3) with Remix, (4) with MR-mixup. From the visualization, we come with the following observations: first, the training data imbalance will cause the majority class (red) to have a larger territory or margin than minority class (blue), which will lead to misclassification and is in accordance with our previous analysis. Second, all mixup-based methods enlarge the margin of the tail class (blue area) and narrow the margin of the head class (red area). Finally, compared to mixup and Remix, the proposed MR-mixup yields a clearer and better-rectified boundary. When compared to mixup, MR-mixup results in a larger minority class margin (see the top-right in *two-moon* case). When compared to remix, its boundary is clearer as remix tends to 'over-smoothen' the boundary. For example, on the *circle* dataset, the leftmost blue point could be misclassified as head class by remix. The result in Fig. 3 illustrates the superiority of MR-mixup over mixup and remix.

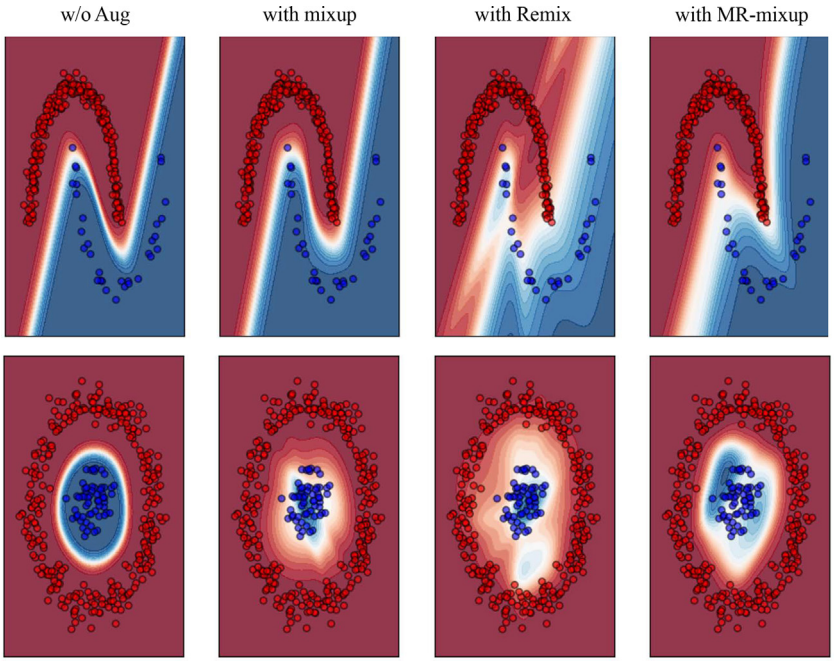

Fig. 3. Visualization on toy datasets to illustrate how MR-mixup rectifies the decision boundary. It can be observed that compared to other baselines, the proposed MR-mixup enlarges the minority margin while also maintaining a clearer boundary.

3.3. Reweighted mutual learning

The MR-mixup augments the minority classes to alleviate the data imbalance, we further propose RML to augment the minority gradients to alleviate the gradient imbalance issue. First, we give a brief review of the gradient suppression on minority classes, which has also been similarly analyzed in previous works [19,20].

Consider a neural network with pre-softmax logits $\mathbf{z} = [z_1, \ldots, z_C]$ and $\mathbf{y} = [y_1, \ldots, y_C]$ to be the one-hot ground-truth vector. The neural network is trained with cross entropy loss:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(\sigma_i), \quad \sigma_i = \text{Softmax}(z_i) \tag{11}$$

Then the gradient of logit z_i with respect to $\mathcal{L}_{CE}$ is:

$$\frac{\partial \mathcal{L}_{CE}}{\partial z_i} = \begin{cases} \sigma_i - 1 & y_i = 1 \\ \sigma_i & y_i = 0 \end{cases} \tag{12}$$

From Eq. (12), it can be observed that for a minority category c_i, it only receives positive gradient $\sigma_i - 1$ when the current sample belongs to c_i ($y_i = 1$), which is rare. Otherwise, it will suffer from a negative gradient σ_i since $y_i = 0$, which is much more frequent.

To mitigate the minority gradient suppression, we resort to mutual learning [11] where an identical peer network f' is trained as well. Denote its post-softmax output as p_i, the mutual learning loss with respect to f is calculated as:

$$\mathcal{L}_{ML}(\sigma, p) = -\sum_{i=1}^{C} p_i \log(\frac{p_i}{\sigma_i}) \tag{13}$$

Similarly, since KL-divergence is asymmetric, we have mutual loss for f' as well:

$$\mathcal{L}'_{ML}(p, \sigma) = -\sum_{i=1}^{C} \sigma_i \log(\frac{\sigma_i}{p_i}) \tag{14}$$

Then the gradient of z_i with respect to $\mathcal{L}_{ML}$ is:

$$\frac{\partial \mathcal{L}_{ML}(\sigma, p)}{\partial z_i} = p_i(\sigma_i - 1) + \sum_{k \neq i} p_k \sigma_i \tag{15}$$

In this case, z_i receives both positive gradients $p_i(\sigma_i - 1)$ and negative gradients $\sum_{k \neq i} p_k \sigma_i$. Compared to the one-hot 'hard target' in cross-entropy, mutual learning provides a 'soft target' p_i as supervision signal and alleviates the over-confidence of majority classifier. Furthermore, we propose to reweigh the mutual learning in an instance-wise manner to augment the minority gradients:

$$\mathcal{L}_{RML}(\sigma, p) = -w_i \sum_{i=1}^{C} p_i \log(\frac{p_i}{\sigma_i}) \tag{16}$$

where w_i is the instance-wise weight. Intuitively, if the current sample i belongs to minority class c, i.e., $y_i = c$, it is expected to have higher p_i and should be emphasized. In practice, we adopt class-balanced effective number [18]:

$$w_i = \frac{1 - \eta}{1 - \eta^{n_c}} \tag{17}$$

where η is the hyperparameter. In this way, the RML guides the peer networks to learn from each other, with special focus on minority samples. With the RML, the long-lasting gradient suppression on minority classes is compensated and mitigated.

During training, we make a copy of the original training batch for MR-mixup augmentation and the RML loss is computed on the original data. The final loss is computed as:

$$\mathcal{L} = \mathcal{L}_{CE}(x, y) + \mathcal{L}_{CE}(\tilde{x}, \tilde{y}^{Rec}) + \mathcal{L}_{RML}(\sigma, p) + \mathcal{L}'_{RML}(p, \sigma) \tag{18}$$

During testing, we discard peer network f' and use f for inference, so that RML will not bring any extra computational cost at inference time. The whole training pipeline is shown in Algorithm 1.

4. Experiments

4.1. Datasets

We conduct experiments on three benchmark datasets: CIFAR-LT, ImageNet-LT and iNaturalist18.

- CIFAR-LT is a long-tailed version of CIFAR dataset introduced in Cao et al. [9]. It is created with exponential decay imbalance

Algorithm 1: Pseudo code for MRA training.

Input : Imbalanced training set D_{train}, main network f and peer network f, total epochs T, Batch size B.
Output: Learned main network f.
for *epoch = 1 to T* **do**
  Sample mini-batch $\{(x_i, y_i)\}_{i=1}^{B}$ from D_{train}
  Randomly sample B pairs of $\{(x_j^1, y_j^1, x_j^2, y_j^2)\}_{j=1}^{B}$ from mini-batch
  Sample mixup factor λ for each pair, then rectify λ to λ^{Rec} according to Eq. 5 or Eq. 9
  Apply MR-mixup according to Eq. 6 or Eq. 10
  where$\tilde{x}_j = \lambda x_j^1 + (1-\lambda)x_j^2 \quad \tilde{y}_j^{Rec} = \lambda_1^{Rec} y_j^1 + (1-\lambda_1^{Rec})y_j^2$
  Forward pass $\{(x_i, y_i)\}$ through main network f and get logits σ_i
  Forward pass $\{(x_i, y_i)\}$ through peer network f and get logits p_i
  Forward pass $(\{\tilde{x}_j, \tilde{y}_j^{Rec}\})$ through main network f and get logits $\tilde{\sigma}_j$
  Calculate L_{CE} with $\{(x_i, y_i)\}, \{(\tilde{x}_j, \tilde{y}_j^{Rec})\}$ according to Eq. 11
  Calculate L_{RML} and L_{RML} with σ, p according to Eq. 16
  Calculate total loss according to Eq. 18
  $L_{CE}(x, y) + L_{CE}(\tilde{x}, \tilde{y}^{Rec}) + L_{RML}(\sigma, p) + L_{RML}(p, \sigma)$
  Loss backward, update main network and peer network.
return network f

and controllable imbalance ratio. In our experiments, we evaluate with imbalance ration 10 and 100 following [9].

- ImageNet-LT is a long-tailed version of ImageNet introduced in Liu et al. [6]. It is sampled from the original ImageNet dataset following the Pareto distribution with power value $\alpha = 6$. It has 1000 categories, 115 K training samples, 50 K testing samples and the imbalance ratio is 1280/5.
- iNaturalist18 is a large-scale dataset collected from real-world and is extremely long-tailed. It has 8142 categories, 437 K training samples, 24K testing samples and the imbalance ratio is 1000/2.

4.2. Baselines

We compare with one-stage methods, multi-stage methods and data augmentation methods in the following experiments. We briefly introduce our baseline methods.

- One-stage methods include (1) vanilla Empirical Risk Minimization (ERM), which is the commonly adopted training strategy for most deel learning algorithms, (2) Focal loss [17], which is a re-weighting method, (3) Class-balanced resampling (CB RS), where each class is sampled with equal probability, (4) Class-balanced reweighting loss based on effetive number of samples (CB RW), (5) Class-balanced effective number based Focal loss (CB Focal) [18], (6) Deferred resampling (DRS) [9], where instance-balanced sampling is first used and then switch to class-balanced sampling, (7) Label-Distribution-Aware Margin loss (LDAM) [9], which is a state-of-the-art method consisting of a re-weighting loss and DRS, (8) Domain Adaptation class-balanced reweighting (DA-RW) [24], which is a state-of-the-art adaptive re-weighting loss, (9) Range loss [34], which is also a re-weighting method, and (10) Balanced Meta Softmax [16], which consists of a meta sampler and balanced softmax.
- Multi-stage methods include (1) Bilateral-Branch Network (BBN) [35], which is a two-branch network with different sampling strategies, (2) De-confound-TDE [36], which direct causal effect of an input is calculated, (3) Few-shot meta learning (FSLwF) [37], which is a few-shot learning method, (4) OLTR [6], where a meta-embedding is adopted, (5) Multiple Experts (LFME) [4], where several teacher networks are trained to distill a unified student model, (6) Decoupling [3], where the representation network and the classifier are trained in a two-stage manner.
- Data augmentation methods include (1) mixup [7], (2) Remix with Class-balanced Resampling (Remix-CB-RS), with Deferred Resampling (Remix-DRS) and with Deferred re-weighting (Remix-DRW), which is an improved version of mixup, (3) Majority to minority transfer with LDAM loss (M2m-LDAM) [32], where minority samples are synthesis in a adversarial-like method, and (4) SMOTE [8], which is a traditional data synthesis method.

4.3. Implementation details

All experiments are conducted with PyTorch on NVIDIA GeForce 2080 Ti GPUs. For CIFAR-LT experiments, we follow the training rules in Cao et al. [9] for a fair comparison. We train ResNet-32 for 200 epochs with batch size 128, SGD with momentum 0.9, weight decay 2×10^{-4}. The initial learning rate is 0.1 and decays by 0.1 at 160 and 180 epoch. For ImageNet-LT experiments, we follow the setting in Kang et al. [3], Liu et al. [6] and train ResNet-10 for 90 epochs with batch size 256, SGD with momentum 0.9, with initial learning rate 0.2 with cosine annealing learning rate schedule. For iNaturalist18 experiments, we use random cropping to 224×224 and train ResNet-50 from scratch for 90 epochs with batch size 64, SGD with momentum 0.9 and learning rate 0.1 with cosine annealing schedule. For all experiments η is set to 0.9999 following [18] and mixup hyperparameter $\alpha = 1$ following the original implementation [7].

4.4. Comparison with state-of-the-art methods

Performance on CIFAR-LT The result is shown in Table 1. We build MRA upon two baselines: (1) CB-RS, which is a commonly used re-sampling strategy in one-stage methods, (2) DRS, an improved version of resampling, where instance-balanced sampling is first used, then switch to class-balanced sampling after certain number of epochs.

The result in Table 1 shows that when combined with deferred resampling, MRA outperforms all baselines by a large margin and achieves state-of-the-art performance in accuracy. Meanwhile, MRA-CB-RS also exhibits competitive performance. Moreover, if we compare MRA with its own baseline, i.e., CB-RS and DRS, it can be observed that MRA brings significant improvement. The MRA-CB-RS even improves CB-RS by 9.83% on CIFAR100-LT with imbalance ratio 100. The MRA-DRS also improves the DRS with large performance gain. Notably, we also observe that the improvement brought by MRA becomes larger as the imbalance ratio grows. The performance gain increases from 2.32(MRA-CB-RS)/1.68(MRS-DRS) to 7.78(MRA-CB-RS)/7.27(MRS-DRS) for CIFAR-10-LT when the imbalance ratio increases from 10 to 100. Similarly, the performance gain increases from 4.44(MRA-CB-RS)/4.54(MRS-DRS) to 9.83(MRA-CB-RS)/7.24(MRS-DRS) for CIFAR-100-LT when the imbalance ratio increases from 10 to 100. The result demonstrates that MRA is beneficial for the long-tailed recognition, especially in the severe imbalance case.

Performance on ImageNet-LT The result is shown in Table 2. We build MRA upon two baselines: (1) CB-RS, a commonly used re-sampling strategy in one-stage methods, (2) Decouple, which is the cutting edge of multi-stage methods. We show that MRA can be easily combined with either one-stage or multi-stage methods and achieves promising performance.

The result in Table 2 shows that when combined with Decouple [3], MRA achieves the highest accuracy and outperforms all

Table 1
Results on CIFAR-LT. Baseline results are from Cao et al. [9] or their original papers.

Dataset	CIFAR10-LT		CIFAR100-LT	
Imbalance ratio	100	10	100	10
ERM	70.36	86.39	38.32	55.70
Focal Loss [17]	70.38	86.67	38.41	55.78
CB RS	70.55	86.79	33.44	55.06
CB RW [18]	72.37	86.54	33.99	57.12
CB Focal [18]	74.57	87.10	36.02	57.99
DRS [9]	75.07	87.52	40.86	57.75
LDAM [9]	77.03	88.16	42.04	58.71
DA-RW [24]	80.00	87.40	44.08	58.00
BBN [35]	79.82	88.32	42.56	59.12
De-confound-TDE [36]	80.60	88.50	44.10	59.60
mixup [7]	73.09	88.00	40.83	58.37
Remix-CB-RS [30]	76.23	87.70	41.13	58.62
Remix-DRS [30]	79.53	88.85	46.53	60.52
Remix-DRW [30]	79.76	89.02	46.77	61.23
M2m-LDAM [32]	79.10	87.50	43.50	57.60
MRA-CB-RS	78.33(+7.78)	89.11(+2.32)	43.27(+9.83)	59.50(+4.44)
MRA-DRS	**82.34**(+7.27)	**89.20**(+1.68)	**48.10**(+7.24)	**62.29**(+4.54)

Table 2
Results on ImageNet-LT. *denotes reproduced results. Other baseline results are from Liu et al. [6] or their original papers.

Many $N_c > 100$	Medium $20 < N_C \leq 100$	Few shot $N_c < 20$	Overall	
ERM	40.9	10.7	0.4	20.9
Lifted Loss [38]	35.8	30.4	17.9	30.8
Focal Loss [17]	36.4	29.9	16.0	30.5
Range Loss [34]	35.8	30.3	17.6	30.7
CB RS *	42.7	34.1	15.7	34.8
MetaSoftmax [16]	50.3	39.5	25.3	41.8
FSLwF [37]	40.9	22.1	15.0	28.4
OLTR [6]	43.2	35.1	18.5	35.6
LFME [4]	47.0	37.9	19.2	38.8
Decouple [3]*	51.9	38.1	21.0	41.0
SMOTE * [8]	41.9	33.4	15.3	34.1
mixup * [7]	40.2	35.5	20.2	35.1
MRA-CB-RS	44.2(+1.5)	37.8(+3.7)	19.1(+3.4)	37.6(+2.8)
MRA-Decouple	**53.0** (+1.1)	**40.8** (+2.7)	**27.3** (+6.3)	**43.6** (+2.6)

baselines including one-stage, multi-stage and data augmentation methods. Meanwhile, the MRA-CB-RS also demonstrates competitive performance. It outperforms other data augmentation methods and most one-stage methods.

More importantly, if we compare the improvement brought by MRA on each subset of ImageNet-LT (Many, Medium, Few-shot subsets), we find that most improvement comes from the Few subset. The performance gain increases from 1.5% to 3.4% as the subset changes from many to few-shot for MRA-CB-RS. Similarly, the performance gain increases from 1.1% to 6.3% as the subset changes from many to few-shot for MRA-Decouple. This result indicates that MRA is extremely effective on augmenting the minority classes.

Performance on iNaturalist18 We also conduct experiments on real-world dataset iNaturalist18, which is directly constructed without any data sampling. The result in Table 3 shows that MRA-Decouple also yields state-of-the-art performance on iNaturalist18. Compared with the Decouple baseline, the MRA-Decouple brings 1.6% improvement in accuracy. This result demonstrates the effectiveness of MRA for real-world long-tailed recognition.

Efficiency analysis While MRA achieves state-of-the-art performance on three benchmark datasets, we argue that it is also computationally efficient and tuning efficient. First, the MRA is an augmentation-based method and brings little extra computational cost. In contrast to many existing state-of-the-art methods (such as multi-stage methods) which improve the performance while suffering from extra high computational cost, the main component MR-mixup augments the training sample in an online manner which is computationally efficient. Second, it can be observed that MRA requires little hyperparameters. The α in MR-mixup is inherited from mixup and kept as default value. The η in RML can also be replaced by other hyperparameter-free reweighting methods. Thus, the MRA is tuning efficient. Finally, the MRA is not only lightweight but also flexible, as it can be easily combined with other methods.

Table 3
Results on iNaturalist18.

Method	Accuracy
ERM	57.1
CB Focal	61.1
DRW	63.7
DRS	63.6
LDAM	66.0
Decouple	65.6
BBN	66.3
MRA-Decouple	**67.2** (+1.6)

4.5. Ablation study of MRA

To verify the effectiveness of each component of MRA, we conduct ablation study and the result is shown in Table 4. From the result, we conclude that both MR-mixup and RML contribute a lot to the overall improvement. We observe that MR-mixup alone already outperforms all other data augmentation methods including mixup, Remix-CB-RS, Remix-DRS, Remix-DRW and

Table 4
Ablation study on CIFAR-LT.

Dataset	CIFAR10-LT		CIFAR100-LT	
Imbalance ratio	100	10	100	10
DRS	75.07	87.52	40.86	57.75
MR-mixup-DRS	82.18	89.13	47.78	61.26
RML-DRS	80.15	89.10	46.60	62.03
MRA-DRS	82.34	89.20	48.10	62.29

Table 5
Ablation study on RML.

Dataset	CIFAR10-LT		CIFAR100-LT	
Imbalance ratio	100	10	100	10
DRS	75.07	87.52	40.86	57.75
ML only	78.31	88.60	44.73	60.89
RML-Focal	79.72	89.05	46.14	61.63
RML-CB	**80.15**	**89.10**	**46.60**	**62.03**

M2m-LDAM. This demonstrates the superiority of MR-mixup over existing data augmentation methods. The RML, on the other hand, also produces competitive performance. Meanwhile, we observe that both MR-mixup and RML yield larger performance improvement over DRS when the imbalance ratio increases, indicating their effectiveness of dealing with severe imbalance. They also behave slightly differently under different imbalance ratios. The MR-mixup achieves higher accuracy when the imbalance ratio is 100, where RML yields superior performance when the imbalance ratio is 10. Finally, the combination of MR-mixup and RML, i.e., MRA, produces the highest performance.

4.6. Ablation study of RML

In order to verify the effectiveness of RML, we conduct a more detailed ablation study on RML. We compare with the following variants: (1) ML only, where only mutual learning is adopted without the reweighting factor. (2) RML-Focal, where focal loss is adopted for reweighting. (3) RML-CB, where class-balanced loss is adopted for reweighting. The result is shown in Table 5. The result on CIFAR-10 and CIFAR-100 shows that (1) the improvement of RML-CB mostly comes from the mutual learning, as the mutual learning alone yields large improvement on the DRS baseline. (2) The reweighting strategy is also effective, as both focal loss and CB loss reweighting factor bring further improvements upon the mutual learning. In comparison, the RML-CB yields slightly superior performance than the RML-Focal, and is adopted in the final version of MRA.

4.7. Comparison of different label shifting strategies

In order to verify the effectiveness of linear shifting, we compare with different shifting strategies. First, we formalize the shifting strategies as follows: consider the case where $\lambda_1 < 0.5$, and we wish to enlarge the tail class margin from γ_1 to γ_1^*, i.e., $\gamma_1 < \gamma_1^*$. Then we rectify the mixup factor from λ_1 to λ_1^{Rec} according to

$$\frac{\lambda_1}{\lambda_1^{Rec}} = (\frac{\gamma_1}{\gamma_1^*})^{\tau} \tag{19}$$

where $\tau \geq 0$ is the hyper-parameter controlling the scale of the enlargement. Then we compare the performance of different shifting strategies as shown in Table 6. The results show that (1) the general idea of shifting the augmented label is effective, as it consistently improves the performance with different τ. (2) The linear shifting ($\tau = 1$) is the most effective one among different strategies. When $\tau = 0$, there is no shifting at all, and the MR-mixup degenerates to ordinary mixup with significant performance drop. When $\tau = 0.5$, the tail class margins are enlarged, but not as large as the linear shifting strategy, the performance becomes superior to mixup, but still inferior to MR-mixup. This result verifies the effectiveness of shifting strategy and tail class margin enlargement. When $\tau = 2$, the tail class margins are further enlarged, and become larger than the linear strategy, the performance also drops slightly. This indicates that a larger tail class margin may not always ensure a higher performance, as it may hamper the overall performance. Finally, we conclude that the linear shifting strategy is effective, easy to implement and yields the best performance.

Table 6
Comparison of different label shifting strategies.

Dataset	CIFAR10-LT		CIFAR100-LT	
Imbalance ratio	100	10	100	10
DRS	75.07	87.52	40.86	57.75
$\tau = 0$ (mixup-DRS)	77.93	87.90	44.16	59.25
$\tau = 0.5$	79.85	88.49	45.10	60.09
$\tau = 1$ (MR-mixup-DRS)	**82.18**	**89.13**	**47.78**	**61.26**
$\tau = 2$	81.74	88.97	47.31	61.17

4.8. Comparison of class-wise accuracy

To further investigate where the overall improvement of MRA comes from, we plot the class-wise accuracy on CIFAR10-LT with imbalance ratio 100 and 10. The results is shown in Fig. 4. First, the result shows that the Empirical Risk Minimization (ERM) performs relatively well on majority classes, but performs poorly on the minority classes, indicating the dominance of majority classes during the imbalance training. Second, the previous state-of-the-art method LDAM improves the tail class accuracy by a large margin, but it also decreases the head class accuracy (e.g., class id = 3 in the left figure). The result shows that while LDAM improves the overall performance, it pays the price of majority class performance degradation. Finally, compared with LDAM, the proposed MRA alleviates the majority degradation (e.g., class id = 3 in both figures) while also producing higher tail class accuracy. This result demonstrates the effectiveness of the proposed MRA.

4.9. Comparison of confusion matrix

To investigate the effectiveness of MR-mixup, RML and MRA, we choose DRS as baseline and plot the corresponding confusion matrix. The result is shown in Fig. 5. From the result, we observe that the DRS still has confusion between class 0,1 (majority) and class 8,9 (minority). This may be due to the fact that majority class 0,1 tend to be dominating during training, thus the minority class 8,9 tend to be easily misclassified. Compared with DRS, the proposed MR-mixup, RML and MRA all improve the class confusion and yield superior performance.

4.10. Confidence calibration

While MRA achieves SOTA performances, we also wish its predictive probability $p(y|x)$ could reflect the true probability. To this end, we calculate the Expected Calibration Error (ECE) and plot the reliability diagrams with 15 bins [39] shown in Fig. 6. We expect a well-calibrated classifier's output to be close to the actual probability and produces low ECE. We compare the reliability diagram of the following variants: (a) Vanilla model, (b) With DRS, (c) With Remix, (d) With MR-mixup, (e) With MRA. All augmentations methods adopt DRS as the baseline. From the results in Fig. 6, we observe the long-tailed distribution would cause the ordinary neural network to be miscalibrated. In other words, it may cause

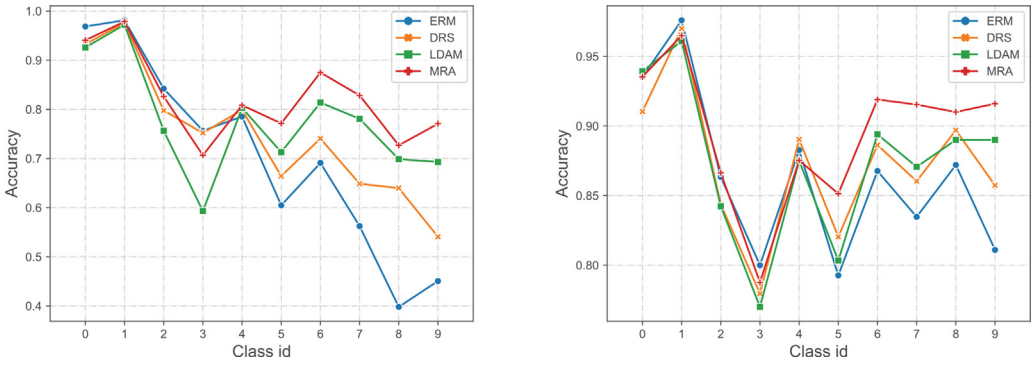

Fig. 4. Comparison of classwise accuracy on CIFAR10-LT with imbalance ratio=100 (left) and imbalance ratio=10 (right).

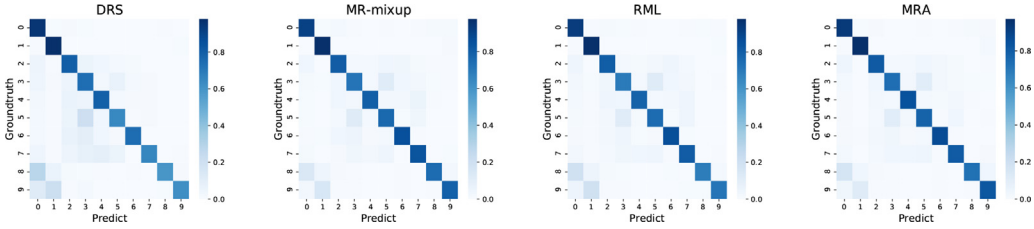

Fig. 5. Confusion matrix of DRS, MR-mixup, RML and MRA respectively on CIFAR10-LT with imbalance ratio = 100.

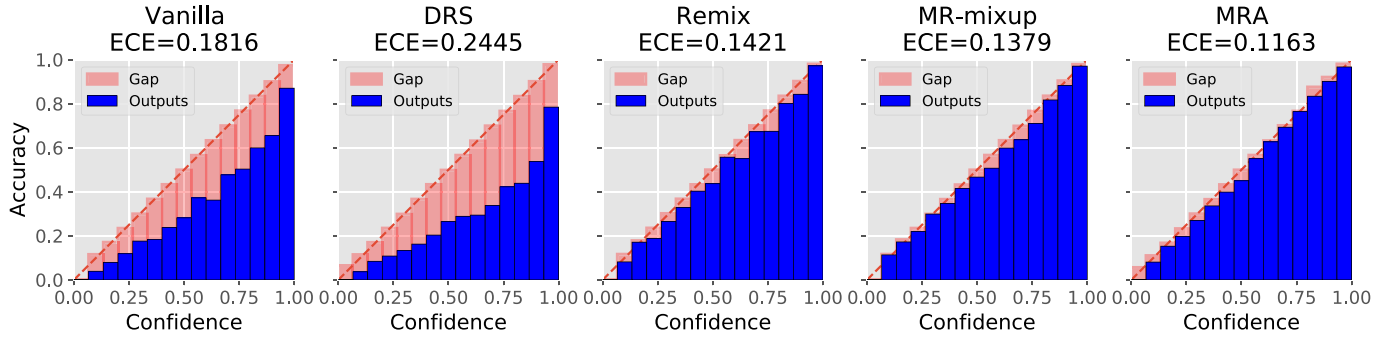

Fig. 6. Reliability diagrams of on CIFAR-100-LT with imbalance ratio 100. The result shows that both MR-mixup and RML significantly improves the miscalibration phenomenon.

the deep network to be overconfident on the majority classes. Moreover, while the DRW improves the accuracy, it worsens the ECE and leads to a severer miscalibrated classifier. Meanwhile, the mixup-based augmentation is overall beneficial for the model calibration. Compared with state-of-the-art mixup-based method Remix, the proposed MR-mixup yields superior calibration results. Finally, the MRA further improves the result, demonstrating the effectiveness of RML in improving model calibration.

5. Conclusion

In this paper, we propose a novel Margin-aware rectified augmentation method for long-tailed classification. We aim to address two issues: rectifying decision boundary through data augmentation, and mitigating minority gradient suppression through mutual learning. We first propose MR-mixup, which is derived from optimal margin analysis, to augment the minority classes as well as rectify the decision boundary. Moreover, we propose Reweighted Mutual Learning to provide an extra 'soft supervision signal' to augment the minority gradients and alleviate the overconfidence of majority classes. The proposed MRA is flexible and easy to implement, and thus can be easily combined with existing methods. We conduct extensive experiments on three benchmark datasets: CIFAR-LT, ImageNet-LT and iNaturalist18. The experimental results show that when combined with other one-stage or multi-stage methods, MRA brings significant improvement and outperforms baseline methods by a large margin. We also demonstrate through the results of class-wise accuracy and confusion matrix that MRA is especially beneficial for tail class improvement. Moreover, the confidence calibration experiment shows that MRA produces a better-calibrated classifier. The MRA aims to address the critical challenges in long-tailed recognition [40]. While many state-of-the-art methods, such as multi-stage methods, improve the performance while suffering from extra high computational costs, the MRA is designed to improve the overall performance with little extra computational cost. It is also flexible as it can be easily combined with other methods. The MRA also has its limitations. Currently, it is designed mainly for image classification, and we plan to generalize MRA to more visual understanding tasks such as object detection, and more modalities such as natural language.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by National Key R&D Program of China (Grant no. 2022YFF1202400), National Natural Science Foundation of China (Nos. 61925107, U1936202), the Science Fund for Creative Research Groups of the National Natural Science Funds of China (No. 62021002).

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[2] T. Wang, Y. Li, B. Kang, J. Li, J. Liew, S. Tang, S. Hoi, J. Feng, The devil is in classification: a simple framework for long-tail instance segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 728–744.
[3] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, Y. Kalantidis, Decoupling representation and classifier for long-tailed recognition, in: International Conference on Learning Representations, 2019.
[4] L. Xiang, G. Ding, J. Han, Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification, in: 16th European Conference on Computer Vision, ECCV 2020, Springer Nature, 2020, pp. 247–263.
[5] X. Wang, L. Lian, Z. Miao, Z. Liu, S.X. Yu, Long-tailed recognition by routing diverse distribution-aware experts, arXiv preprint arXiv:2010.01809 (2020).
[6] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, S.X. Yu, Large-scale long-tailed recognition in an open world, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2537–2546.
[7] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).
[8] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[9] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, in: Advances in Neural Information Processing Systems, 2019, pp. 1565–1576.
[10] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).
[11] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4320–4328.
[12] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst., Man, Cybern., Part B (Cybernetics) 39 (2) (2008) 539–550.
[13] M. Koziarski, Radial-based undersampling for imbalanced data classification, Pattern Recognit. 102 (2020) 107262.
[14] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International Conference on Intelligent Computing, Springer, 2005, pp. 878–887.
[15] S. Maldonado, C. Vairetti, A. Fernandez, F. Herrera, FW-smote: a feature-weighted oversampling approach for imbalanced classification, Pattern Recognit. 124 (2022) 108511.
[16] J. Ren, C. Yu, S. Sheng, X. Ma, H. Zhao, S. Yi, H. Li, Balanced meta-softmax for long-tailed visual recognition, arXiv preprint arXiv:2007.10740 (2020).
[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[18] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.
[19] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, J. Yan, Equalization loss for long-tailed object recognition, arXiv preprint arXiv:2003.05176 (2020).
[20] J. Tan, X. Lu, G. Zhang, C. Yin, Q. Li, Equalization loss v2: a new gradient balance approach for long-tailed object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1685–1694.
[21] J. Du, Y. Zhou, P. Liu, C.-M. Vong, T. Wang, Parameter-free loss for class-imbalanced deep learning in image classification, IEEE Transactions on Neural Networks and Learning Systems (Early Access) (2021) 1–7.
[22] M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning, arXiv preprint arXiv:1803.09050 (2018).
[23] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, D. Meng, Meta-weight-net: learning an explicit mapping for sample weighting, in: Advances in Neural Information Processing Systems, 2019, pp. 1919–1930.
[24] M.A. Jamal, M. Brown, M.-H. Yang, L. Wang, B. Gong, Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7610–7619.
[25] S. Zhang, Z. Li, S. Yan, X. He, J. Sun, Distribution alignment: a unified framework for long-tail visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2361–2370.
[26] A.K. Menon, S. Jayasumana, A.S. Rawat, H. Jain, A. Veit, S. Kumar, Long-tail learning via logit adjustment, arXiv preprint arXiv:2007.07314 (2020).
[27] Z. Chen, J. Duan, L. Kang, G. Qiu, in: Class-imbalanced deep learning via a class-balanced ensemble, IEEE Transactions on Neural Networks and Learning Systems (Volume: 33, Issue: 10, October 2022) (2021) 5626–5640.
[28] Y.-X. Wang, D. Ramanan, M. Hebert, Learning to model the tail, in: Advances in Neural Information Processing Systems, 2017, pp. 7029–7039.
[29] J. Liu, Y. Sun, C. Han, Z. Dou, W. Li, Deep representation learning on long–tailed data: alearnable embedding augmentation perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2970–2979.
[30] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, D.-C. Juan, Remix: rebalanced mixup, in: European Conference on Computer Vision, Springer, 2020, pp. 95–110.
[31] S. Li, K. Gong, C.H. Liu, Y. Wang, F. Qiao, X. Cheng, Metasaug: meta semantic augmentation for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5212–5221.
[32] J. Kim, J. Jeong, J. Shin, M2m: imbalanced classification via major-to-minor translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13896–13905.
[33] S. Suh, P. Lukowicz, Y.O. Lee, Discriminative feature generation for classification of imbalanced data, Pattern Recognit. 122 (2022) 108302.
[34] X. Zhang, Z. Fang, Y. Wen, Z. Li, Y. Qiao, Range loss for deep face recognition with long-tailed training data, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5409–5418.
[35] B. Zhou, Q. Cui, X.-S. Wei, Z. Chen, BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9716–9725.
[36] K. Tang, J. Huang, H. Zhang, Long-tailed classification by keeping the good and removing the bad momentum causal effect, arXiv preprint arXiv:2009.12991 (2020).
[37] S. Gidaris, N. Komodakis, Dynamic few-shot visual learning without forgetting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4367–4375.
[38] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4004–4012.
[39] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.
[40] Y. Fu, L. Xiang, Y. Zahid, G. Ding, T. Mei, Q. Shen, J. Han, Long-tailed visual recognition with deep models: amethodological survey and evaluation, Neurocomputing 509 (2022) 290– 309.

Liuyu Xiang is currently an Associate Research Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China. He received his B.Sc. degree from EE, University of Science and Technology of China in 2017 and Ph.D. degree from School of Software, Tsinghua University. His research interests include computer vision and machine learning.

Jungong Han is currently a Chair Professor and the Director of the Research of Computer Science, Aberystwyth University, U.K. He also holds an Honorary Professorship with the University of Warwick, U.K. His research interests include computer vision, artificial intelligence, and machine learning.

Guiguang Ding is currently an Associate Professor with the School of Software, Tsinghua University, China. His research interests include the areas of multimedia information retrieval, computer vision, and machine learning.