Local and Global Logit Adjustments for Long-Tailed Learning

Yingfan Tao\textsuperscript{1,*}, Jingna Sun\textsuperscript{2},†, Hao Yang\textsuperscript{2}, Li Chen\textsuperscript{2}, Xu Wang\textsuperscript{2}, Wenming Yang\textsuperscript{1,†}, Daniel Du\textsuperscript{2}, Min Zheng\textsuperscript{2}  
\textsuperscript{1}Tsinghua University, \textsuperscript{2}ByteDance Inc

Abstract

Multi-expert ensemble models for long-tailed learning typically either learn diverse generalists from the whole dataset or aggregate specialists on different subsets. However, the former is insufficient for tail classes due to the high imbalance factor of the entire dataset, while the latter may bring ambiguity in predicting unseen classes. To address these issues, we propose a novel Local and Global Logit Adjustments (LGLA) method that learns experts with full data covering all classes and enlarges the discrepancy among them by elaborated logit adjustments. LGLA consists of two core components: a Class-aware Logit Adjustment (CLA) strategy and an Adaptive Angular Weighted (AAW) loss. The CLA strategy trains multiple experts which excel at each subset using the Local Logit Adjustment (LLA). It also trains one expert specializing in an inversely long-tailed distribution through Global Logit Adjustment (GLA). Moreover, the AAW loss adopts adaptive hard sample mining with respect to different experts to further improve accuracy. Extensive experiments on popular long-tailed benchmarks manifest the superiority of LGLA over the SOTA methods.

1. Introduction

Deep learning has brought profound improvements to various vision tasks, including classification, detection, segmentation, etc. The success of deep learning is undoubtedly inseparable from large-scale well-designed datasets, such as ImageNet \textsuperscript{9}, COCO \textsuperscript{25} and Places \textsuperscript{51}, which usually exhibit approximately uniform distribution over different classes. However, constructing these artificially balanced datasets is extremely difficult: sufficient instances must be collected for the tail classes whose samples are few by nature. According to the inherently existing power law \textsuperscript{43}, most real-world data follows a long-tailed distribution: a few head classes occupy a large portion of samples, while most tail classes only have small portions. It is challenging to learn directly from these long-tailed data, because deep learning models tend to be dominated by the head classes that appear most during training, while resulting in poor performance on tail classes. This paper aims to mitigate such problems in model training on long-tailed data.

Among the existing literature coping with the long-tail problem, some design class re-balancing strategies for training, including re-sampling \textsuperscript{5,12,17,33} or cost-sensitive learning \textsuperscript{24,20,21,40,34}. In addition, decoupled learning proposes a two-stage training process that decouples the representation learning and the classifier learning \cite{13,25} to preserve the broken feature representation caused by the re-balancing methods. Most recent efforts rely on ensemble learning to achieve state-of-the-art performances on long-tail visual recognition, where multiple experts are trained in a complementary manner, then aggregated together for in-
Figure 2: (a) Some ensemble methods train generalists \((\psi_{\Omega_1,\ldots,\Omega_3})\) on the entire dataset \(S\) that is severely imbalanced. (b) Some others train specialists on individual subsets \((S_1,\ldots,S_3)\) that are less imbalanced, but these specialists may suffer from limited perceptions. (c) Our method trains all experts on the entire dataset \(S\), with a novel CLA strategy to ensure an adaptive local/global awareness for the experts. Assuming this is a three-expert model, the first two experts \(\psi_{\Omega_1,\Omega_2}\) controlled by LLA excel at different subsets (marked with smile symbols), while the last expert \(\psi_{\Omega_3}\) optimized by GLA will further boost the ensemble results and delivers superior performance.

The ensemble process involves averaging the predictions of each expert to make the final decision. These experts are commonly trained either using the entire dataset to create generalists [39, 21] (Figure 2a) or by combining specialists trained on different subsets [1, 6, 41] (Figure 2b). However, the former approach experiences a high level of data imbalance during training, while the latter restricts the vision of each model seeing only a subset of the training data, which impairs the capabilities of each specialist as well as the overall ensemble model since they only encounter a limited number of classes and lack effective collaboration during training.

Inspired by the above insights, we propose a novel framework for long-tailed recognition: the **Local and Global Logit Adjustments (LGLA)** framework. LGLA contains two core parts: the **Class-aware Logit Adjustment (CLA)** strategy and the **Adaptive Angular Weighted (AAW)** loss. CLA has two adjustment strategies, namely **Local Logit Adjustment (LLA)** and **Global Logit Adjustment (GLA)**. LLA trains multiple experts on the whole dataset but specializing in non-overlapping subsets and possessing the most comprehensive knowledge in their respective areas. Meanwhile, GLA trains one expert that achieves a global perspective and excels at handling inversely long-tailed distributions. AAW further improves classification performance by introducing adaptive hard sample mining, which enhances discriminative ability by capturing and re-weighting hard samples for each expert during training.

The core insight of LGLA is simple: *each skill-diverse expert should always have access to the whole data during training*. The various skills of each expert controlled by LLA within a certain subset, as well as the capability of the last expert learned by GLA, should be obtained through an adaptive approach, instead of arbitrarily restricting their training data. Besides, the sharing of the same vision for all experts would also eliminate possible ambiguity during their ensemble.

As illustrated in Figure 2, instead of training generalists on the full dataset (Figure 2a), or training specialists on partial subsets (Figure 2b), LGLA (Figure 2c) combines the advantages of both the generalist and the specialists by training experts all on the full data, leading to diversely skilled recognition ability w.r.t both fine-grained subsets and the whole data distribution.

To sum up, LGLA inherits the strengths of generalists and specialists, resulting in improved classification performance on long-tailed data. As illustrated by Figure 1 under a fair setting ensuring the same backbone and same training data, our approach achieves superior performance over existing methods on most public benchmarks. Overall, our contributions can be summarized as follows:

- **We propose a novel Local and Global Logit Adjustments (LGLA) method to boost long-tailed recognition tasks. LGLA possesses the merits of more fine-grained optimization, higher diversity among experts, and an entire feature space learning.**

- **We propose a Class-aware Logit Adjustment (CLA) strategy to instruct differentiated learning among experts, and an Adaptive Angular Weighted (AAW) loss for adaptive instance re-weighting, which better handles samples of different difficulties.**

- **Extensive experiments on popular long-tailed benchmarks, including CIFAR-10/100-LT [8], ImageNet-LT[27], iNaturalist 2018 [36] and Places-LT [51] have demonstrated the superiority of LGLA over the SOTA competitors, as shown by Figure 1.**
2. Related Work

This section introduces existing methods for long-tailed learning. We focus on training strategies and intentionally exclude methods that require a large amount of extra training or pre-training data.\textsuperscript{28}\textsuperscript{35}.

Class Re-balancing Class re-balancing handles long-tailed distribution problem by re-balancing the contribution of each class during training. This method can be divided into two types: re-sampling and cost-sensitive learning. The re-sampling methods re-balance data distribution by over-sampling the tail classes\textsuperscript{3} \textsuperscript{12} or under-sampling the head classes\textsuperscript{11} \textsuperscript{17} \textsuperscript{33}. However, over-sampling often causes tail category overfitting, while under-sampling inevitably neglects a large amount of information from head data. On the other hand, the cost-sensitive learning methods emphasize the tail classes by assigning larger weights on them\textsuperscript{24} \textsuperscript{20} \textsuperscript{40} or randomly neglect the gradients from head classes\textsuperscript{34} by assigning different weights in loss for different samples. Nevertheless, cost-sensitive learning often makes the network difficult to optimize, especially on large-scale data\textsuperscript{30}. It also brings more performance fluctuations on tail classes\textsuperscript{43}.

Decoupled Learning Although yielding great improvement in classifying long-tailed data, re-balancing methods often suffer from the degeneration of feature representation during the representation learning stage. To cope with this, two-stage decoupled methods are proposed\textsuperscript{18} \textsuperscript{4} \textsuperscript{16} \textsuperscript{44} \textsuperscript{50}, which first train a feature extractor with standard instance-balanced sampling, then fine-tune the classifier with class-balanced sampling. Previous works focusing on classifier learning can be fine-tune the classifier with class-balanced sampling.

Ensemble Learning Ensemble learning utilizes complementary knowledge by training and aggregating multiple experts, which can be further divided into two categories: The first category\textsuperscript{39} \textsuperscript{21} \textsuperscript{46} \textsuperscript{50} directly trains each expert on the whole dataset. For example, NCL\textsuperscript{21} highlights the importance of cooperation by conducting distillation between every two model pairs. RIDE\textsuperscript{39} brings about diversified experts by maximizing the KL-divergence between the predicted probabilities from diverse models. However, training on a highly imbalanced dataset inevitably causes negligence in the sample-few classes. The second category\textsuperscript{41} \textsuperscript{1} \textsuperscript{6} often trains each expert on a subset of training data, which leads to a less severe data imbalance problem and results in better performance. For instance, ResLT\textsuperscript{6} trains three experts on three subsets respectively corresponding to the all, the medium+tail, and the tail classes, and combines their outputs as the final result. However, the learned feature space of each expert in ResLT is incomplete with unseen classes, which would bring ambiguity to the ensemble results. Unlike these methods, our LGLA takes advantages of both strategies above. Through an adaptive training on the whole data for all experts, and by leveraging a novel logit adjustments strategy combining LLA and GLA, LGLA achieves advantageous performance on most public benchmarks over existing approaches.

3. Methodology

3.1. Preliminaries

As shown by Figure 3, LGLA consists of $K$ experts with a shared backbone $f_{Ω}$. Denote the experts as $Ψ = \{ψ_{Ω1}, ..., ψ_{Ω_{K-1}}, ψ_{ΩK}\}$, with $ψ_{Ω_{k-1}…Ω_{K}}$ being the $K-1$ experts controlled by LLA and the last expert $ψ_{Ω_{K}}$ learned by GLA; $Ω_{k}$ is the parameter of the $k$-th expert. We use $W_{k} = \{W_{k1}, W_{k2}, ..., W_{kC}\} \in \mathbb{R}^{k \times C}$ to represent the weight of the last FC layer within the classifier of the $k$-th expert (which is not shared across experts), where $C$ is the number of classes, $d$ denotes the dimension of features. Note that the rows $W_{1} ≤ C$ of $W_{k}$ can also be regarded as the center of the learned features w.r.t the $k$-th expert on the $j$-th class.

Let $S = \{x_{i}, y_{i}\}$ be the training data, with $x_{i}$ being the $i$-th image and $y_{i}$ the corresponding label. Given $x_{i}$, the backbone $f_{Ω}$ extracts features $z_{i} = f_{Ω}(x_{i})$ from $x_{i}$, then the classifier $ψ_{Ωk}$ obtains the logit $v_{i} = ψ_{Ωk}(z_{i})$ from $z_{i}$. It is straightforward to apply softmax onto the logit $v_{i}$ as:

$$p(x_{i}) = \frac{\exp(v_{i}^{y_{i}})}{\sum_{j=1}^{C} \exp(v_{i}^{j})},$$

with $v_{i}^{j}$ being the $j$-th value of $v_{i}$. Losses like cross-entropy are then imposed on $p(x_{i})$ for training.

3.2. Class-aware Logit Adjustment Strategy

The vanilla softmax suffers from the negligence of the discrepancy between the posterior distributions of training and test data, leading to biases under long-tailed scenarios. In order to encourage greater intra-class compactness and inter-class separability, logit adjusting – a strategy proposed by face recognition approaches\textsuperscript{10} \textsuperscript{37} \textsuperscript{26} – also reveals its potential in some long-tailed learning methods\textsuperscript{31} \textsuperscript{29} \textsuperscript{2} \textsuperscript{48}. Existing methods employ logit adjustment with the following formula:

$$p(x_{i}) = \frac{\exp(v_{i}^{y_{i}} + T(y_{i}))}{\sum_{j=1}^{C} \exp(v_{i}^{j} + T(j))},$$

where $T(y_{i})$ is the logit adjusting term.
where $T(\cdot)$ is a logit adjustment function conditioning only on class labels. Unlike such logit adjustment used by existing methods [31, 29, 2, 48], our proposed Class-aware Logit Adjustment (CLA) strategy cooperates with \(K\) experts presenting different adjustments for different experts, and the probability output of the \(k\)-th expert can be formulated as:

$$
p(x_i, k) = \frac{\exp \left( \frac{v^j_{i,k} + T(k, y_i)}{v^j_{i,k} + T(k, j)} \right)}{\sum_{j=1}^{C} \exp \left( \frac{v^j_{i,k} + T(k, j)}{v^j_{i,k} + T(k, j)} \right)}. \tag{3}
$$

Unlike the $T(j)$ in Eq. 2 which conditions only on the class label $j$, $T(k, j)$ in Eq. 3 is designed to reflect the awareness of the $k$-th expert w.r.t. the $j$-th class. In specific, we first sort all the classes in descending order according to their cardinality; then we divide all the $C$ classes into $K$ groups $C = \{C_1, C_2, ..., C_{K-1}\}$, ensuring that the number of samples $\|S_k\| := \{(x_i, y_i) : y_i \in C_k\}$ belonging to each group $C_k$ is approximately equal, i.e. $\|S_p\| \approx \|S_q\|, 1 \leq p, q \leq K - 1$. Through this way, $C_1$ contains mostly the head classes while $C_{K-1}$ includes mostly the tail classes. The entire training dataset can be denoted as $S = S_1 \cup S_2 \cup ... \cup S_{K-1}$.

Now, we list the designed CLA function $T(k, j)$ of the $k$-th expert model for the class $j$ as follows:

$$
T(k, j) = \begin{cases} 
\log(n_j), & j \in C_k, k \neq K \\
\log(n_{max}), & j \notin C_k, k \neq K \\
\tau \cdot \log(n_j), & k = K 
\end{cases} \tag{4}
$$

where $n_j$ is the frequency of class $j$, $n_{max}$ denotes the frequency of the largest class, and $\tau$ is a hyper-parameter.

To illustrate the intuition behind our design, we substitute Eq. 4 into Eq. 3 obtaining the following equation:

$$
P(x_i, k) = \frac{\exp \left( \frac{v^j_{i,k} + T(k, y_i)}{v^j_{i,k} + T(k, j)} \right)}{\sum_{j=1}^{C} \exp \left( \frac{v^j_{i,k} + T(k, j)}{v^j_{i,k} + T(k, j)} \right)}, \tag{5}
$$

with $M(y_i, j, k) := T(k, j) - T(k, y_i)$. $M(y_i, j, k)$ adjusts the decision boundaries between the ground truth class $y_i$ and all the other non ground truth classes $j$ for the $k$-th expert. Its formulation can be expanded as:

$$
M(y_i, j, k) = \begin{cases} 
\log\left(\frac{n_j}{n_{y_i}}\right), & y_i \in C_k, j \in C_k, k \neq K \\
\log\left(\frac{n_{max}}{n_{y_i}}\right) > 0, & y_i \in C_k, j \notin C_k, k \neq K \\
\log\left(\frac{n_{max}}{n_{y_i}}\right) < 0, & y_i \notin C_k, j \in C_k, k \neq K \\
\log\left(\frac{n_{max}}{n_{y_i}}\right) = 0, & y_i \notin C_k, j \notin C_k, k \neq K \\
\tau \cdot \log\left(\frac{n_j}{n_{y_i}}\right), & y_i \notin C_k, j \in C_k, k = K 
\end{cases} \tag{6}
$$

Local Logit Adjustment We propose Local Logit Adjustments (LLA) to meticulously explore subset representation ability and eliminate ambiguity from unseen classes. LLA, denoted by the first four lines ($k \neq K$) in Eq. 6 trains the first $K - 1$ experts with full data and restricts their specialization to the sample groups $S_1, ..., S_{K-1}$. In specific, when $y_i \in C_k$ and $j \in C_k$, indicating that both the two classes $y_i$ and $j$ belong to the group that the $k$-th expert is skilled in, then $M(y_i, j, k) = \log(n_j/n_{y_i})$ which is positive when $n_{y_i} < n_j$ while negative when $n_j < n_{y_i}$. As
Global Logit Adjustment (LLA) enables the first $K - 1$ experts to acquire specialized knowledge on certain subsets. However, the non-specialized subsets trained with original Softmax are still susceptible to the long-tailed distribution. To address this issue and further promote the ensemble performance, we introduce Global Logit Adjustment (GLA), presented in the last line ($k = K$) of Eq. 5. GLA trains another expert to learn an inversely long-tailed distribution, which offsets the long-tailed distribution of the non-specialized subsets in the first $K - 1$ experts. The decision boundary for this expert has an offset of $\tau \cdot \log(n_j/n_i)$, with a scaler $\tau > 1$ to amplify the margin and adjust the ensemble results.

Figure 3 also shows the L2 norms of per-class weights from the classifiers of a three-expert LGLA model. Higher norms in a classifier usually contribute more to the performance [50]. The first two experts excel at the first and last subsets, respectively, while the last expert is skilled in the inversely long-tailed distribution. The ensemble weights are obtained by averaging the weights of all experts, resulting in an approximately uniform among various categories, which shows the effectiveness of our proposed LGLA on a uniform test set. It is noteworthy that LGLA enlarges the variety among all the experts meanwhile ensures the integrity of each learned feature space.

3.3. Adaptive Angular Weighted Loss

Hard sample mining emphasizes hard instances during training by giving them higher weights in loss. During the process of training, tail classes iterate fewer times than head classes and gradually become hard to score, so the instance-level re-weighting approach is important and effective for long-tailed learning. However, it is strenuous to find suitable weights to fit the model. Existing approaches like OHem [32] score hard examples with high weights, yet it ignores the optimization of easy samples; Focal loss [24] reduces the weights of simple samples, while paying more attention to hard samples, but it does not consider the nature of distribution within long-tailed learning. Different from these methods, we propose a novel Adaptive Angular Weighted (AAW) loss to address hard instances for long-tailed data and re-weight them adaptively during training.

We measure the difficulty of the $i$-th sample w.r.t the $k$-th expert using the angle between the input feature ($f_k^y$) of the last FC layer and its corresponding class centers of the $k$-th expert ($W_k^y$) in the cosine space, which is formulated as:

$$\theta_k^y = \arccos \left( \frac{W_k^y \cdot f_k^y}{||W_k^y|| \cdot ||f_k^y||} \right).$$

(7)

For each iteration, we compute $K - 1$ reference angles using $\hat{\theta}_k^{ref} = \text{mean}(\theta_{k}^y), y_i \in C_k$ to guide the optimization of samples belonging to different groups. Based on the fact that features with larger cosine distances are more difficult to learn, we define an instance in $C_k$ with the angle greater than $\hat{\theta}_k^{ref}$ as the hard sample and make its weight increase to $1 + \theta_k^y - \hat{\theta}_k^{ref}$. Meanwhile, the samples whose angles are smaller than the reference angles will not be neglected by keeping their weights unchanged. Therefore, AAW loss can adaptively re-weight those instances away from the class center during different training phases, and the re-weighting function is formulated as:

$$g(\theta_k^y, \theta_k^{ref}) = \begin{cases} 1, & \theta_k^y \leq \theta_k^{ref} \\ 1 + \theta_k^y - \theta_k^{ref}, & \theta_k^y > \theta_k^{ref} \end{cases}$$

(8)

Cooperated with CLA, the overall loss of our proposed LGLA is formulated as:

$$L = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} g(\theta_k^y, \theta_k^{ref}) \log \left( \frac{\exp(v_{i,k}^y + T(k, y_i))}{\sum_{j=1}^{C} \exp(v_{i,j}^y + T(k, j))} \right)$$

(9)
### 3.4. Advantages over Previous Methods

In general, four appealing properties of LGLA make it stand out among previous multi-expert ensemble methods: First, LLA makes experts extremely explore the representation ability in each subset. Preserving the merits of previous ensemble specialists methods [1, 41], we also shrink down the specialized area to a subset with a lower imbalance factor, which is more helpful to extract the discrimination among long-tailed classes. Different from them, all the experts in LGLA possess the same feature space, which is more reasonable to fuse the logits for obtaining stunning ensemble performance. Secondly, the number of experts in LGLA depends on the number of subsets, which can be easily expanded for a more fine-grained optimization, and experiments in Figure 7 verify that more experts can get better results. Thirdly, AAW loss can adaptively re-weight hard samples in training to further improve the recognition ability. Finally, as shown in Figure 6, the ensemble model can bias to a specific region, like many, medium, or few split, by increasing the weights of the logit from the $k$-th expert, which makes our method adapt to test sets with different class distributions.

### 4. Experiments

#### 4.1. Datasets and Protocols

We validate the effectiveness of our proposed method on five major long-tailed datasets: CIFAR-10/100-LT [8], ImageNet-LT [27], iNaturalist 2018 [36] and Places-LT [51]. CIFAR-10/100-LT [8] are the different long-tailed versions of the original CIFAR datasets [20], with 10 and 100 classes, respectively. Both of them follow an exponential decay in the sample sizes across different classes, which is controlled by an Imbalance Factor (IF). IF evaluates how imbalanced a dataset is, denoted as the ratio of the number of the largest class to that of the smallest class. Two values of IF, namely 50 and 100 are used for each dataset. ImageNet-LT [27] is created from the ImageNet [9] dataset following the Pareto distribution with power value $\alpha = 6$. It obtains 115K images from 1,000 classes with the IF=256. Places-LT is a terribly imbalanced version of Places [51], containing 62.5K samples of 365 categories with the IF=996. iNaturalist 2018 [50] is currently the largest long-tailed visual dataset, possessing 437.5K images from 8,142 classes with the IF=512.

We evaluate our proposed method on balanced test sets after long-tailed training and report the top-1 accuracy on all categories. We follow [27], to divide a validation dataset into three subsets: many (with more than 100 instances), medium (with 20 to 100 instances), and few (with less than 20 instances) splits.

#### 4.2. Implementation Details

Following [2, 21, 39], we adopt ResNet-32 [13] as the backbone for CIFAR-10/100-LT, ResNet-50/ResNeXt-50 [42] for ImageNet, ResNet-50 for iNaturalist 2018 and pretrained ResNet-152 for Places-LT, respectively. In addition, current ensemble methods [46, 21] constitute independent convolution layers and a fully-connected layer as an expert. However, we find that group convolution provides a better alternative, which can slightly promote classification accuracy on many test sets without introducing additional parameters. The effect of group convolution will be discussed in 4.4. Moreover, following [21, 7], we use RandAugment [5] on all the benchmarks except Places-LT, and NVIDIA Tesla V100 GPUs are employed for training. We exclude methods that rely on a large amount of extra training or pre-training data [28, 35] for fair comparison.

---

**Table 1:** Top-1 accuracy (%) on CIFAR-100-LT and CIFAR-10-LT with IF=100/50. All the methods use the ResNet-32 backbone. The best and the secondary results are marked in **bold** and *underline*, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR-100-LT 100</th>
<th>CIFAR-100-LT 50</th>
<th>CIFAR-10-LT 100</th>
<th>CIFAR-10-LT 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB Focal loss [8]</td>
<td>38.7 46.2</td>
<td>74.6 79.3</td>
<td>42.0 50.0</td>
<td>87.5 91.0</td>
</tr>
<tr>
<td>LDAM+DRW [9]</td>
<td>42.0 45.1</td>
<td>77.0 79.3</td>
<td>44.1 49.2</td>
<td>81.4 82.2</td>
</tr>
<tr>
<td>LDAM+DAP [9]</td>
<td>43.5 47.9</td>
<td>- -</td>
<td>79.1 -</td>
<td>- -</td>
</tr>
<tr>
<td>BBN [50]</td>
<td>39.4 47.0</td>
<td>79.8 82.2</td>
<td>42.3 -</td>
<td>60.6 -</td>
</tr>
<tr>
<td>LFME [41]</td>
<td>42.3 -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>CAM [47]</td>
<td>47.8</td>
<td>51.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RIDE [9]</td>
<td>49.1 -</td>
<td>- -</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Logit Adj. [29]</td>
<td>43.9 -</td>
<td>-</td>
<td>77.7 -</td>
<td>-</td>
</tr>
<tr>
<td>LADe [15]</td>
<td>45.4</td>
<td>50.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MISLAS [49]</td>
<td>47.0</td>
<td>52.3</td>
<td>82.1 85.7</td>
<td>-</td>
</tr>
<tr>
<td>Hybrid-SC [38]</td>
<td>46.7</td>
<td>51.9</td>
<td>81.4 85.4</td>
<td>-</td>
</tr>
<tr>
<td>DiVe [14]</td>
<td>45.4</td>
<td>51.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SSD [22]</td>
<td>46.0</td>
<td>50.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ACE [11]</td>
<td>49.6</td>
<td>51.9</td>
<td>81.4 84.9</td>
<td>-</td>
</tr>
<tr>
<td>PaCo [7]</td>
<td>52.0</td>
<td>56.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ResLT [9]</td>
<td>49.7</td>
<td>54.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NCL [21]</td>
<td>54.2</td>
<td>58.2</td>
<td>85.5 87.3</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>56.5</td>
<td>60.6</td>
<td>87.8 90.2</td>
<td>-</td>
</tr>
<tr>
<td>Ours (GC)</td>
<td>57.2</td>
<td>61.6</td>
<td>87.5 89.8</td>
<td>-</td>
</tr>
</tbody>
</table>

**Figure 5:** Accuracy gain comparisons between our method and SOTA on all the many-/medium-/few-shot. Experiments are conducted on CIFAR-100-LT dataset (IF=100).
## 4.3. Comparing with Existing Methods

Experimental results on the major long-tailed benchmarks, including CIFAR-10/100-LT [8], ImageNet-LT [27], iNaturalist 2018 [36] and Places-LT [51] are listed in Table 1, 2, 3, respectively, demonstrating the superiority of our method over the state-of-the-art. To ensure the fairness of comparisons, we list the results of LGLA with both the original network and \( K \)-group convolution (denoted by GC).

It is worth noting that without additional explanation, the number of experts in LGLA is set to 3 for fair comparisons with other three-expert ensemble modules [46, 21, 39, 1].

### CIFAR-10/100-LT

Table 1 suggests that our proposed LGLA consistently outperforms the state-of-the-art by a large margin on CIFAR-100-LT and CIFAR-10-LT with IF=50 and IF=100. Concretely, compared with the current SOTA [21], obvious improvements are gained by 2.3% (IF=100) and 2.4% (IF=50) on CIFAR-100-LT, and 2.3% (IF=100) and 2.9% (IF=50) on CIFAR-10-LT. When using “GC”, the accuracies are further improved to 57.2% (IF=100) and 61.6% (IF=50) on CIFAR-100-LT without introducing additional parameters. Class-wise accuracy gain with NCL is compared in Figure 5, which shows the effectiveness of LGLA on medium-/few-shot.

### ImageNet-LT

We report the results on ImageNet-LT with different backbones in Table 2. It can observe that the performance is further improved to 59.7% with backbone ResNet-50 and 61.1% (GC) with backbone ResNeXt-50.

### iNaturalist 2018

Table 3 presents the performance comparison of LGLA and other SOTA competitors on iNaturalist 2018. We find that LGLA results in around 3% better performance on few-shot, showing the effectiveness of our method on tail classes. Finally, there is around 1.3% (GC) performance gain over SOTA in all categories.

### Places-LT

Following previous work [7, 6, 21], we use a pretrained model on ImageNet and fine-tune 30 epochs on Places-LT. So that we only provide results with the original expert network. As shown in Table 2, we can observe that...
our proposed LGLA still yields a state-of-the-art result on the extremely imbalanced benchmarks Places-LT at 42.0%.

4.4. Component Analysis

Data augmentation. Data augmentation techniques have proved to be useful in long-tailed learning. Many works adopt data augmentation for obtaining richer feature representations. For example, [1][47][49] use Mixup and [21][7] employ RandAugment [5]. Following PaCo [7] and NCL [21], RandAugment is used in our method. In order to investigate the effect of data augmentation, we conduct experiments of different ensemble methods with and without RandAugment. From Table 4, we can observe that RandAugment brings significant improvement to classification accuracies. However, our method sees a major performance increase over the current SOTA methods irrespective of whether RandAugment is used or not.

CLA. We evaluate the effectiveness of CLA on CIFAR-100-LT. The first three rows of Table 5 show the classification accuracies of a three-expert model trained with Softmax, BALSM, and CLA respectively. It observes that compared with Softmax and BALSM, CLA considerably improves the performance by 8.56% and 1.29% respectively.

To validate that LLA is actually making experts excel at a specific data region, we train a 4-expert model on CIFAR-100-LT, with each expert specializing in Many-/Medium-/Few-shot, respectively. The weight of the last expert controlled by CLA is fixed, we increase the fusion weight of each expert from 0.1 to 0.9 and evaluate the ensemble model on each split. Figure 6 shows that when we raise the fusion weight of an expert, the corresponding accuracy on the ensemble model simultaneously increases, which makes LGLA adapt to different class distributions.

AAW loss. As shown in Table 5, we apply AAW to different classification losses, namely BALMS, and CLA, to explore the effects of the proposed Adaptive Angular Weighted (AAW) loss. There can be seen obvious increases in accuracy by adding AAW to BALMS (55.24% to 56.51%), and CLA (56.53% to 57.15%). This indicates that AAW is helpful to improve performance on long-tailed data.

Number of experts. LGLA is a multi-expert model, and we evaluate the performance under different numbers of experts. From Figure 7, we can observe that increasing the number of experts from 2 to 7 brings obvious improvements to “All” classes. In specific, The performance is greatly improved when only using a small number of experts, however, when $K > 3$, the performance gains tend to be stable.

Independent Convolution vs. Group Convolution. The current ensemble methods [39][21][46][31] construct each expert with Independent Convolution (IC) layers and a fully-connected layer, but we find that using Group Convolution (GC) instead of IC can achieve better results without introducing extra parameters. When using GC, representations should be firstly concatenated $K$ times and then $K$ groups of convolution are utilized for representation learning. Then the obtained $K$ groups of features are sent to $K$ independent fully-connected layers, respectively. Figure 8 shows the results of LGLA using IC and GC with current SOTA [21] which employs IC for comparison. It observes that GC results in relatively higher accuracies and whether IC or GC is used, our method achieves better performance.

5. Conclusion

This paper presents a novel Local and Global Logit Adjustments (LGLA) method for long-tailed learning. LGLA consists of two main components: CLA and AAW. The CLA strategy, applies LLA to extract more discriminative features for each subset and applies GLA to further build superior ensemble performance. The AAW loss also boosts performance by capturing and adaptively re-weighting hard instances according to the deviations between features and class centers. Extensive experiments on popular long-tailed benchmarks highlight the effectiveness and superiority of our LGLA over SOTA methods.
References


on computer vision and pattern recognition, pages 212–220, 2017.


