Mutual Learning for Long-Tailed Recognition

Changhwa Park
42dot Inc.
changhwa.park@42dot.ai

Junho Yim*
AI Technology Division, LG Energy Solution
junho.yim@lgensol.com

Eunji Jun*
Institute of Advanced Technology Development, Hyundai Motor Group
ejjun@hyundai.com

Abstract

Deep neural networks perform well in artificially-balanced datasets, but real-world data often has a long-tailed distribution. Recent studies have focused on developing unbiased classifiers to improve tail class performance. Despite the efforts to learn a fine classifier, we cannot guarantee a solid performance if the representations are of poor quality. However, learning high-quality representations in a long-tailed setting is difficult because the features of tail classes easily overfit the training dataset. In this work, we propose a mutual learning framework that generates high-quality representations in long-tailed settings by exchanging information between networks. We show that the proposed method can improve representation quality and establish a new state-of-the-art record on several long-tailed recognition benchmark datasets, including CIFAR100-LT, ImageNet-LT, and iNaturalist 2018.

1. Introduction

Deep neural networks show high recognition accuracy in artificially-balanced datasets, such as ImageNet [8], COCO [21], and Places [35]. However, it is difficult to obtain delicately manipulated data rather than long-tail distributed data in practice [25]. Under the imbalanced circumstance, a naively learned neural network is easily dominated by head classes and performs disastrously in tail classes [1].

The apparent issue is that classifier predictions are entangled with the long-tailed distribution [15]. In this respect, several recent studies have focused on learning an unbiased classifier by properly calibrating the classifier boundary [17, 15, 33]. For example, instead of instance-balanced (natural) sampling, class-balanced sampling, which samples uniformly across classes, has been used. These strategies enhanced overall performance by increasing the performance of the tail classes, but this was often suffered by a decrease in the performance of the head classes. More importantly, the performance of classifier-focused approaches is limited by the quality of the learned representations.

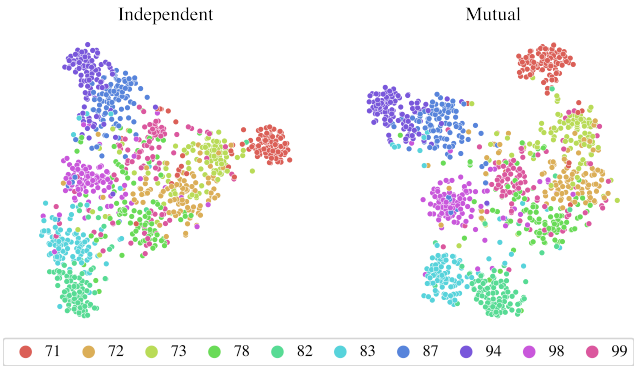


Figure 1: t-SNE visualization of representations trained with independent learning and mutual learning, respectively, on CIFAR100-LT. To compare the quality of representations in tail classes, ten random tail classes are chosen.

It is difficult to learn high-quality representations in a long-tailed environment because it overfits a limited number of tail class samples. If samples from different classes are mixed in the learned representation space, performance will be limited regardless of how finely we tune the classifier boundary. For example, Figure 1 (left) shows tail class representations learned using cross-entropy loss with instance-balanced sampling, and we cannot expect satisfactory performance from the tangled representations. To quantify the quality of learned representations, we freeze the feature encoder and train a linear classifier using the test dataset. Then, the accuracy of the linear classifier can be thought of as the upper bound accuracy, and we define it as feature accuracy. Although instance-balanced sampling learns better representations than other sampling strategies [17], its fea-

*Work done at AIRS Company, Hyundai Motor Group

ture accuracy is 58.3%. In other words, any classifier cannot get an accuracy better than 58.3% with the learned representations, regardless of how we finetune it. Accordingly, it is crucial to learn high-quality representations under a long-tailed environment to achieve higher performance.

In this study, we propose a framework for improving the quality of representations under a long-tailed circumstance and combining an unbiased classifier. We exploit decoupled learning of representations and a classifier since the optimal strategy for representation learning differs from that for classifier training. For example, instance-balanced sampling learns better representations than class-balanced sampling, whereas class-balanced sampling learns a better classifier than instance-balanced sampling [17]. A simple method to improve the generalization ability in a conventional classification problem is to ensemble multiple networks. Long-tailed recognition, on the other hand, possesses the aforementioned inherent issues, thus we focus on facilitating representation learning. To improve generalization ability at the feature level, we propose a simple yet effective framework using mutual learning technique [32], which trains multiple networks together and penalizes divergence between their outputs.

We are motivated by the property of mutual learning, which corrects what models have not seen of each other. The information from peer models can alleviate the model's tendency to overfit tail class samples, allowing them to learn better representations. Using the feature accuracy and other classifier fine-tuning methods [17], we empirically demonstrate that mutual learning improves the quality of representations under the long-tailed condition. For instance, representations learned using mutual learning, which is shown in Figure 1 (right), have a feature accuracy of 61.4% (+3.1%). In addition, we find that the sampling strategy is important in collaborative learning that instance-balanced sampling outperforms class-balanced sampling. Upon the learned high-quality representations, we can apply any unbiased classifier. In this paper, we apply a simple classifier, Post-Compensated softmax (PC softmax) [15], to disentangle the training data distribution from the model prediction.

We make the following observations and contributions.

- We focus on learning better representations and suggest mutual learning to achieve it. We empirically show that mutual learning can learn more generalizable features than independent learning in a long-tailed setting.
- Sampling strategy matters: even in mutual learning, instance-balanced sampling learns more generalizable representations than class-balanced sampling.
- We propose an effective framework that combines the proposed feature extraction method with a simple disentangling classifier, PC softmax.
- We extensively evaluate the proposed framework on various long-tailed benchmark datasets, including CIFAR100-LT [3] (+0.1∼1.7%), ImageNet-LT [22] (+2.8∼2.9%), and iNaturalist 2018 [27] (+2.0%), and achieve state-of-the-art performance.

2. Related Works

2.1. Re-balancing

Classical re-sampling methods include under-sampling high-frequency instances [9], over-sampling low-frequency instances [4, 11], and class-balanced sampling [26, 23]. Another line of work to compensate for the imbalanced distribution is cost-sensitive learning which gives more weight to minor classes [20, 3, 7, 29]. However, these approaches are pruned to overfit minor classes or underfit major classes, resulting in unsatisfactory overall performance.

2.2. Multi-expert Networks

Recently, multi-experts-based methods have led to significant performance improvements in both head and tail classes by pursuing expertise in each expert. BBN [34] dynamically combines a conventional learning branch that employs an instance-balanced sampler and a re-balancing branch that employs a class-balanced sampler. RIDE [28] trains multiple experts independently while penalizing inter-expert correlation to encourage diversity between them. ACE [2] introduces complementary experts, in which each expert is assigned a diverse but overlapping class subset. In contrast to these methods, we pursue collaborative learning among experts to learn better representations. NCL [18] also uses the concept of collaborative learning in combination with hard category mining to stimulate learning from a partial perspective. However, because it combines multiple optimization functions, it suffers from hyper-parameter tuning, and its algorithm design to use independent networks as experts necessitates a large amount of computing power. In this paper, we propose a simple but effective method that is easily adaptable to other approaches.

2.3. Knowledge Distillation and Mutual Learning

Knowledge distillation [14] from a teacher model to a student model has been recently introduced to the long-tailed recognition area. LFME [30] divides the entire long-tailed dataset into subsets with a smaller imbalance to train expert models and then distill knowledge into a unified student model. RIDE [28] applies knowledge distillation from a model with more experts to a model with fewer experts for further advancements. SSD [19] and DIVE [13] exploit self-supervision and power normalization, respectively, to obtain a flatter label distribution as teacher signals.

In contrast to one-way knowledge distillation, deep mutual learning [32] proposes collaboratively learning an en-

semble of students to teach each other throughout the training process. The secondary class probabilities of experts act as the salient cue to each other, and the learned model finds a much wider minimum than an independent model. Similar to our work, [10] also develops the concept of collaborative learning with instance-balanced sampling and class-balanced sampling. However, we make an intriguing observation that sampling strategy is important in mutual learning as it is in independent training. We empirically show that using only instance-balanced sampling learns better representations and achieves higher classification accuracy.

3. Method

3.1. Overall Framework

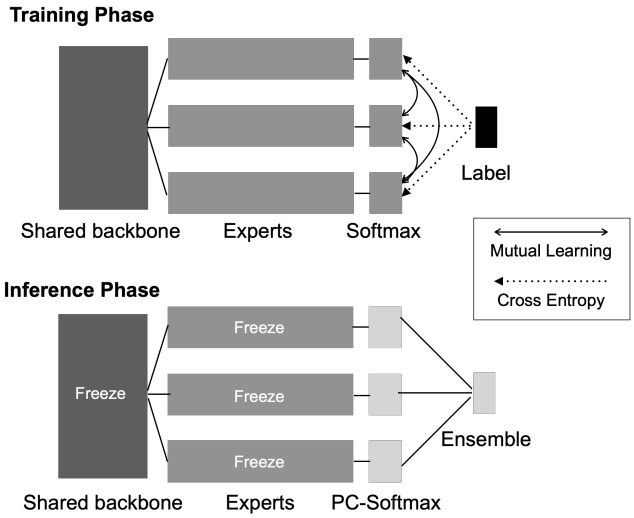


Figure 2: Overall framework. In the training phase, we employ mutual learning loss and classification loss to learn high-quality representations. In the inference phase, PC softmax is used to balance the biased predictions.

The overall framework is shown in Figure 2. We use a shared backbone architecture to reduce the computational complexity, similar to [28]. To illustrate, f_θ is shared by all experts, and each expert is denoted as g_{θ_k}, where $k \in [1, K]$, and K is the number of experts. The output of f_θ is fed into each expert, which includes the fully connected layer, and the outputs of experts are subject to classification loss, $\mathcal{L}_{\text{Classify}}$, and mutual learning loss, $\mathcal{L}_{\text{Mutual}}$. All the experts use an instance-balanced sampler and are co-trained throughout the training process. After learning high-quality representations through mutual training, we employ PC softmax as a classifier.

3.2. Mutual Learning Loss

Mutual learning loss [32] is proposed to distill knowledge between cohort models. Cohort models' inter-class correlation information helps each other avoid falling into local minima. In the long-tailed recognition problem, an independent model is especially prone to falling into sharp minima since the cardinalities of tail classes are limited. We exploit the mutual learning loss to find more robust minima under the long-tailed circumstance and acquire better representations. When the number of experts is two, the mutual learning loss is defined as follows [32]:

$$\mathcal{L}^1_{\text{Mutual}} = D_{\text{KL}}(\mathbf{p}_2 \| \mathbf{p}_1), \tag{1}$$

$$\mathcal{L}^2_{\text{Mutual}} = D_{\text{KL}}(\mathbf{p}_1 \| \mathbf{p}_2), \tag{2}$$

where $\mathbf{p}_k = \sigma(g_{\theta_k}(f_\theta(\mathbf{x})))$, and $\sigma(\cdot)$ represents the softmax function.

When there are more than two experts, we can either use each cohort individually or their ensemble as a teacher. If we use each cohort as a teacher, the mutual learning loss for each expert becomes

$$\mathcal{L}^k_{\text{Mutual}} = \frac{1}{K-1} \sum_{l=1, l\neq k}^{K} D_{\text{KL}}(\mathbf{p}_l \| \mathbf{p}_k). \tag{3}$$

When using the ensemble of cohort models as a teacher, the mutual learning loss for each expert is defined as

$$\mathcal{L}^k_{\text{Mutual}} = D_{\text{KL}}(\mathbf{p}_{avg} \| \mathbf{p}_k), \tag{4}$$

$$\mathbf{p}_{avg} = \frac{1}{K-1} \sum_{l=1, l\neq k}^{K} \mathbf{p}_l. \tag{5}$$

We empirically find that there is no significant difference in performance between using the ensemble signal as a teacher and using each cohort as a teacher. In experiments, we apply the ensemble teacher for the mutual learning loss.

Each expert is supervised by cross-entropy loss in addition to the mutual learning loss, as shown below.

$$\mathcal{L}^k_{\text{Classify}} = -\log \mathbf{p}_k(y). \tag{6}$$

In summary, the overall objective is formulated as

$$\mathcal{L}_{\text{Total}} = \sum_{k=1}^{K} \left(\mathcal{L}^k_{\text{Classify}} + \mathcal{L}^k_{\text{Mutual}} \right). \tag{7}$$

Using the above objective, the shared backbone network and all experts are trained together.

3.3. Post-Compensated Softmax

We utilize mutual learning to obtain more generalizable representations than independent learning. Although the proposed framework can be applied with any other classifier-focused approach to train the classifier, we adopt

the post-compensation (PC) strategy [24, 1, 16, 15] in the inference phase, which is simple and almost cost-free. Under the long-tailed condition, the label distribution of the training dataset often does not match that of the test dataset. As the learned model is strongly entangled with the label distribution of the training dataset, it performs poorly on the test dataset. The PC softmax adjusts the model's output to match the arbitrary test label distribution as follows [15],

$$h_\theta^{\mathrm{PC}}(\mathbf{x})[y] = h_\theta(\mathbf{x})[y] - \log p_s(y) + \log p_t(y), \qquad (8)$$

where $h_\theta(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} g_{\theta_k}(f_\theta(\mathbf{x}))$, the ensemble of expert outputs, and p_s and p_t are the label distribution of the training dataset and the test dataset, respectively. We use the ensemble of expert outputs in the inference phase.

3.4. Feature Generalizability

Many long-tailed recognition approaches learn the features and the classifier at the same time. As [17] pointed out, it's unclear whether the performance is achieved by learning good representations or by shifting the classifier's boundary. Suppose the feature generalizability of a learned model by one method is better than the other. In that case, we can expect the former to outperform the latter when the same classification approach is used. In this regard, we emphasize the importance of learning better representations, which can be used in conjunction with other classifier-focused methods.

To quantify the quality of the learned representations, we measure linear classification accuracy on the frozen features of the test dataset. To illustrate, after training, we infer the features on the test dataset and train a single fully-connected classifier using the inferred test features and test labels. Then, the accuracy of the classifier learned using the test dataset serves as the upper bound for classification accuracy, which we referred to as feature accuracy. We evaluate the quality of learned representations using mutual learning and independent learning by this measure. The feature accuracy of mutual learning is higher than that of independent learning, demonstrating the effectiveness of mutual learning in generating high-quality representations. More results and discussions are presented in Section 4.3.

4. Experiments

4.1. Experimental Setup

4.1.1 Datasets

ImageNet-LT [22] is sampled from ImageNet-2012 dataset [8] using a Pareto distribution [25] with the power value $\alpha = 6$. It has 115.8K images from 1000 classes in total, with a maximum of 1280 images per class and a minimum of 5 images per class.

CIFAR100-LT [3] is a long-tailed version of CIFAR-100 with fewer training samples per class. The imbalance ratio is defined as the ratio between the maximum and minimum size of the classes. In CIFAR100-LT, the rest size of the classes decays exponentially. We experiment on three imbalance ratios, 10, 50, and 100.

iNaturalist 2018 [27] is a large scale real-world dataset for a species classification. It has 437.5K images from 8142 classes in total and has a high imbalance ratio of 500.

4.1.2 Implementation Details

We use ResNet-32 [12] for CIFAR100-LT, ResNet-50 and ResNeXt-50 [31] for ImageNet-LT, and ResNet-50 for iNaturalist 2018 as our backbone. Following [28], the first two stages of the network serve as a shared backbone, while the later stages serve as the experts' architecture. The number of filters in each expert is reduced by $\frac{1}{4}$ to reduce the computational cost, which is the same as the setting of [28]. For all experiments, we employ a three-experts model whose computational cost is comparable to that of the baseline model.

We use a cosine classifier and SGD optimizer with a momentum of 0.9 for all datasets. For CIFAR100-LT, we primarily follow the experimental settings of [3]. The training epoch is 200, and the multistep learning rate schedule is employed, which reduces the learning rate by 0.1 at the 160^{th} and 180^{th} epochs. For ImageNet-LT and iNaturalist 2018, we mainly follow the protocol of [33]. The training epochs for ImageNet-LT and iNaturalist 2018 are 180 and 200, respectively, and the cosine learning rate scheduler is used.

To assess the linear classification accuracy on the test dataset, we first infer all the test samples with the learned model and create a test features dataset. Then we initialize a single-layer fully connected network, which takes features as input, and train it with cross-entropy loss. We use SGD with a momentum of 0.9 for 200 epochs. Initial learning rate is 0.1 and it is decayed by 0.1 at 160^{th} and 180^{th} epochs. The top-1 accuracy of the learned classifier is reported.

4.1.3 Competing Methods

We compare the proposed method with recent state-of-the-art approaches: two-stage based method (MiSLAS [33]), logit-adjusted training (LADE [15]), knowledge distillation (LFME [30], SSD [19], DIVE [13]), and multi-experts (BBN [34], RIDE [28], ACE [2], NCL [18]). For multi-experts-based methods, except BBN, the results of their three-experts model are borrowed for a fair comparison.

4.1.4 Evaluation Metric

We evaluate the trained models on the corresponding test datasets and report the top-1 accuracy across all classes. To investigate the accuracy of each class and analyze how the model performs as the cardinality of class differs, we also report the average accuracy on three subsets of the entire

Method	Res-50	ResX-50	ResX-50 GFlops
Cross Entropy	47.9	49.0	4.29 (1.0x)
PC Softmax	52.7	53.5	4.29 (1.0x)
LADE [15]	-	53.0	4.29 (1.0x)
MiSLAS [33]	52.7	-	4.29 (1.0x)
SSD [19]	-	56.0	-
DIVE [13]	-	53.1	-
RIDE [28]	54.9	56.4	4.69 (1.1x)
ACE [2]	54.7	56.6	6.03 (1.4x)
Ours	**57.7**	**59.5**	6.12 (1.4x)
PaCo* [6]	57.0	58.2	-
NCL* [18]	**59.5**	60.5	12.86 (3.0x)
Ours*	59.4	**60.8**	6.12 (1.4x)

Table 1: Top-1 accuracy on ImageNet-LT with ResNet-50 and ResNeXt-50. * denotes models trained with RandAugment [5] for 400 epochs. GFlops are mainly based on [28]

Method	Many	Med.	Few
Cross Entropy	68.9	43.2	12.6
PC Softmax	64.8	50.6	31.9
LADE [15]	65.1	48.9	33.4
SSD [19]	66.8	53.1	35.4
DIVE [13]	64.1	50.4	31.5
RIDE [28]	67.6	53.5	35.9
Ours	**70.2**	**56.7**	**39.1**

Table 2: Class-wise top-1 accuracy comparison with state-of-the-arts on ImageNet-LT with ResNeXt-50.

classes following [22]: *Many-shot* (contains over 100 samples), *Medium-shot* (contains 20 to 100 samples), and *Few-shot* (contains under 20 samples) classes. The average accuracy over three independent runs is reported.

4.2. Comparison with State-of-the-arts

4.2.1 Results on ImageNet-LT

Table 1 shows that the proposed method outperforms state-of-the-art methods that do not use additional augmentation by a large margin on ImageNet-LT with various backbone networks, ResNet-50 and ResNeXt-50. RIDE, in particular, penalizes inter-expert correlation, whereas our method encourages collaborative learning among experts and outperforms the current best method on ResNet-50, RIDE, by 2.8%. Moreover, the performance improvement over the state-of-the-art method on ResNeXt-50, ACE, is 2.9%. When RandAugment [5] is used and trained with longer epochs, the proposed method performs as well as or better than NCL which uses three independent networks as experts and thus requires far more GFlops than our method.

Imbalance Ratio	10	50	100
Cross Entropy †	59.0	45.5	41.0
PC Softmax †	61.2	49.5	45.3
BBN [34]	59.1	47.0	42.6
LADE [15]	61.7	50.5	45.4
MiSLAS [33]	63.2	52.3	47.0
SSD [19]	62.3	50.5	46.0
DIVE [13]	62.0	51.1	45.4
RIDE ‡ [28]	58.0	51.9	48.0
ACE [2]	-	50.7	49.4
Ours	**63.3**	**54.0**	**49.6**

Table 3: Top-1 accuracy on CIFAR100-LT with an imbalance ratio of 10, 50, and 100. Rows with † denote results directly borrowed from [15]. ‡ denotes our reproduced results with the released code.

Method	Top-1 accuracy
Cross Entropy †	65.0
PC Softmax †	69.3
BBN [34]	69.6
LADE [15]	70.0
MiSLAS [33]	71.6
SSD [19]	71.5
DIVE [13]	71.7
ACE [2]	72.9
Ours	**74.9**

Table 4: Top-1 accuracy on iNaturalist 2018. Rows with † denote results directly borrowed from [15].

To further evaluate the proposed method, we also report the average accuracy of each category subset in Table 2. Introducing PC softmax to the model learned with cross-entropy loss improves the overall accuracy as in Table 1, but it sacrifices the performance of the many-shot classes as in Table 2. In comparison to using only PC softmax, our method with mutual learning technique even improves the performance of the many-shot subset and surpasses all other methods on all category subsets.

4.2.2 Results on CIFAR100-LT

Extensive experiments are carried out on CIFAR100-LT with imbalance ratios of 10, 50, and 100, and the results are provided in Table 3. In contrast to previous methods that show the best accuracy on specific imbalance ratio circumstances, the proposed method yields new state-of-the-art results for all imbalance ratio settings. It is worth noting that our method outperforms state-of-the-art methods that use mixup augmentation, such as MiSLAS and ACE.

Method	Independent				Mutual			
	Many	Med.	Few	All	Many	Med.	Few	All
Joint	67.5	38.5	7.9	39.5	70.3	40.2	7.2	**40.9 (+1.4)**
cRT [17]	61.0	43.5	19.9	42.6	63.2	46.9	22.4	**45.2 (+2.6)**
NCM [17]	58.2	44.7	23.4	43.1	59.1	46.6	25.4	**44.6 (+1.5)**
τ-norm [17]	64.0	42.2	17.9	42.5	67.7	43.9	15.9	**43.8 (+1.3)**
LWS [17]	61.0	43.9	21.5	43.2	63.2	46.9	24.0	**45.7 (+2.5)**
PC Softmax	60.7	44.6	23.6	43.9	63.1	48.1	25.3	**46.5 (+2.6)**
Feature	65.7	57.5	50.6	58.3	68.3	60.9	53.9	**61.4 (+3.1)**

Table 5: Comparison with independently learning a network. Top-1 accuracy on CIFAR100-LT with an imbalance ratio of 100 is reported. The mutual learning model is trained on a network of two experts, and the performance of each individual model is evaluated. "Feature" denotes the feature accuracy on the test features.

4.2.3 Results on iNaturalist 2018

To evaluate the effectiveness of the proposed method on real-world long-tailed circumstances, we conduct experiments on iNaturalist 2018. Results are presented in Table 4. Our method outperforms other methods with a large margin, demonstrating its effectiveness on fine-grained datasets with high imbalance ratios. To illustrate, we can observe a 2.0% performance improvement over ACE on iNaturalist 2018, which has an imbalance ratio of 500.

4.3. Effectiveness of Mutual Learning

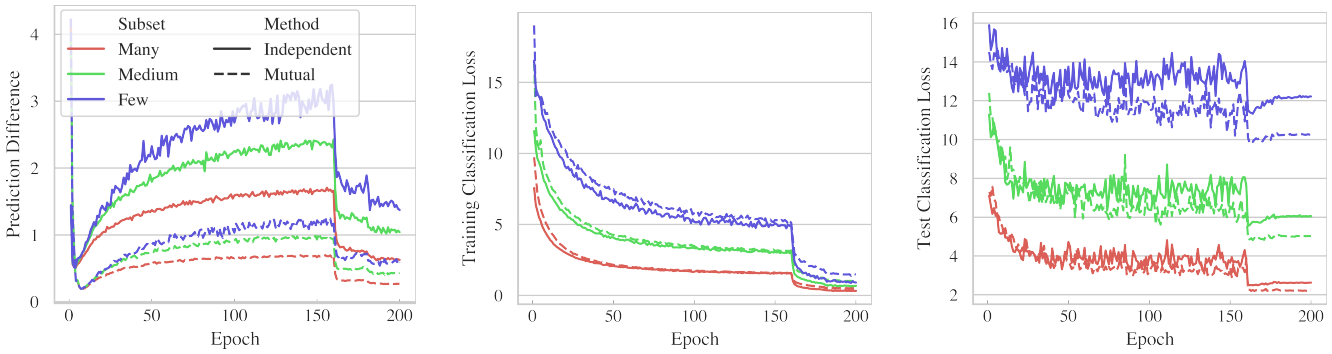
We further evaluate the effectiveness of mutual learning in obtaining high-quality representations under long-tailed circumstances. To investigate it, we train a model independently as well as a model with the mutual learning loss for comparison. After training the models, we consider the following classifiers to probe the quality of the learned representations: the classifier learned jointly with the representations (Joint), Classifier Re-training (cRT), Nearest Class Mean classifier (NCM), τ-normalized classifier (τ-norm), Learnable weight scaling (LWS) [17], PC softmax, and the classifier trained on the test dataset. The results are provided in Table 5. We can observe that the representations learned with mutual learning achieve higher accuracy than those learned independently on all kinds of classifiers. In particular, the performance improvement in test dataset linear classification is 3.1%. These results indicate that mutual learning has a positive impact on the learning of high-quality representations in long-tailed recognition.

Improving the quality of representations through mutual learning is easily adaptable to other methods. To demonstrate, we applied the two-experts mutual learning model to baseline methods, Focal loss [20], LDAM [3], and MiSLAS [33]. As shown in Table 6, all of the methods' accuracies have improved by a significant margin (2.1%$\sim$3.6%).

To better understand the role of mutual learning in long-tailed conditions, we demonstrate how the difference in prediction between two experts, training and test classification loss change as training progresses for many-shot, medium-shot, and few-shot class subsets. Figure 3a shows that few-shot classes exhibit higher prediction differences than the many-shot classes throughout the training process of independent learning. This suggests that the training result for tail class samples is highly stochastic and that it is easy to converge to a less generalizable solution. In Figure 3b and 3c, we can observe that independent learning converges to a lower training classification loss but a higher test classification loss than mutual learning. This implies that mutual learning has a regularization effect, which leads to better generalization. Furthermore, the difference in training classification loss between independent and mutual learning, as well as the test classification loss, are significantly greater in tail classes than in head classes. This demonstrates that the regularization effect of mutual learning is stronger in tail class samples than head class samples. We conjecture that mutual learning allows experts to exchange secondary class probability, making them less likely to fall in sharp minima even for tail class samples. Experts receive the same supervision signal, but because their initial states differ, they take different learning paths, allowing them to transfer correla-

Method	Many	Med.	Few	All
Focal [20]	65.0	35.1	8.0	37.4
Focal+ML	69.4	38.5	6.3	**39.6 (+2.2)**
LDAM [3]	61.4	43.4	19.6	42.6
LDAM+ML	66.6	46.4	22.2	**46.2 (+3.6)**
MiSLAS [33]	63.3	46.7	22.7	46.8
MiSLAS+ML	63.3	49.4	26.5	**48.9 (+2.1)**

Table 6: Effectiveness of applying mutual learning to other approaches. Top-1 accuracy on CIFAR100-LT with an imbalance ratio of 100 is reported. "ML" denotes mutual learning.



(a) Prediction difference between experts. (b) Training classification loss (c) Test classification loss

Figure 3: The prediction difference and classification loss trajectory of independent and mutual learning of the two-experts model on CIFAR100-LT. The average loss of each class subset is reported.

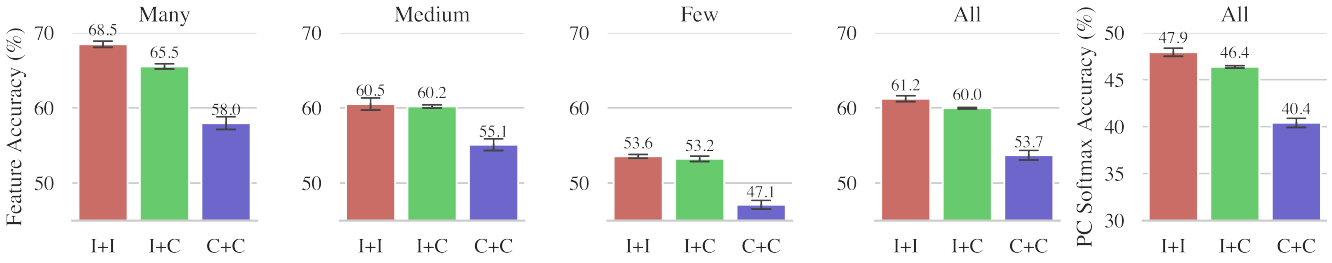


Figure 4: Feature accuracy and PC softmax accuracy of different sampling strategies for each subset on CIFAR100-LT. "I" and "C" refer to instance-balanced sampling and class-balanced sampling, respectively. For example, "I+C" trains a two-experts model with mutual learning, where each expert utilizes instance-balanced sampling and class-balanced sampling.

tions between classes that they have not seen before.

In Figure 1, we generate t-SNE embeddings of tail classes to show how features learned with mutual learning differ from those learned with independent learning. We can see that the features learned with mutual learning are more linearly separable than the other in few-shot circumstances.

4.4. Importance of Sampling Strategy

In independent training, it has been shown that using instance-balanced sampling produces better representations than using class-balanced sampling [17]. We discover that it also holds for the collaborative learning framework. To illustrate, we experiment three variations of sampling strategy for the collaborative learning of the two-experts model: instance-balanced sampling + instance-balance sampling (our setting, I+I), instance-balanced sampling + class-balanced sampling (I+C) [10], and class-balanced sampling + class-balanced sampling (C+C). In Figure 4, we measure the feature accuracy to determine how well they learned representations. We can observe that the sampling strategy which only used instance-balanced sampling achieves higher overall feature accuracy than those used class-balanced sampling. Using only class-balanced samplers, in particular, yields significantly poorer results than other sampling strategies. The PC softmax accuracy results show a similar trend. These findings suggest that instance-balanced sampling outperforms class-balanced sampling when learning linearly separable representations even under the collaborative learning scheme.

4.5. Larger Expert Number

Figure 5 shows how the proposed method scales with more experts. We can observe that as the number of experts increases, the advantage of mutual learning over a mere ensemble prediction of experts grows more in medium-shot and few-shot classes than in many-shot classes. This demonstrates that poor generalization on tail classes under long-tailed circumstances can be enhanced further with mutual learning with a large number of experts.

We can also observe that utilizing an ensemble teacher for mutual learning produces similar results as using each individual cohort as a teacher. It is in contrast with when mutual learning is applied to uniformly distributed datasets; using individual cohort teachers yields better performance than using an ensemble teacher [32]. Our reasoning is that predictions on few-shot classes are more likely to be inaccurate, so the ensemble teacher is a better source of information in the long-tailed setting than in the uniform setting.

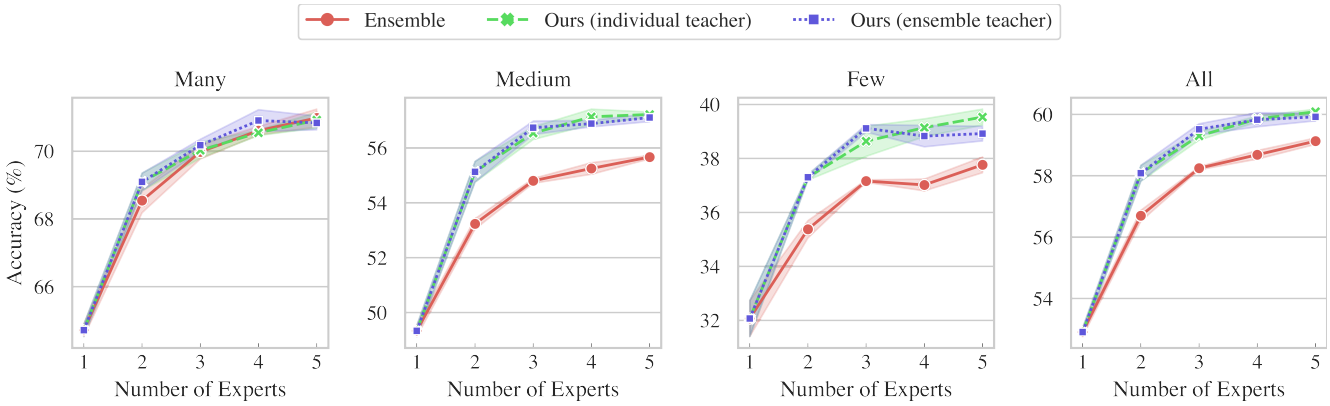


Figure 5: Top-1 accuracy of different numbers of experts on ImageNet-LT with ResNeXt-50. "Ensemble" denotes the ensemble accuracy of the multi-experts network learned with a standard cross-entropy loss. Individual teacher and ensemble teacher represent the two variations of mutual learning loss described in Section 3.2.

Number of Experts	$\mathcal{L}_{\text{Classify}}$	$\mathcal{L}_{\text{Mutual}}$	PC Softmax	Many	Med.	Few	All
1	✓			68.9	43.2	12.6	49.0
1	✓		✓	64.8	50.6	31.9	53.5
2	✓	✓	✓	69.1	55.1	37.3	58.1
3	✓	✓		**73.9**	47.8	18.4	53.8
3	✓		✓	70.0	54.8	37.2	58.2
3	✓	✓	✓	70.2	**56.7**	**39.1**	**59.5**

Table 7: Ablation study on the efficacy of each component. Top-1 accuracy on ImageNet-LT with ResNeXt-50 is reported.

4.6. Ablation study

We provide the results of ablated models to investigate the contribution of each component of the proposed framework in Table 7. Table 7 shows that PC softmax increases the accuracy of the baseline model by 4.5%. Employing three experts with the reduced dimension advances the performance by 4.7%, and it is comparable to that of other competing multi-experts-based methods. By applying $\mathcal{L}_{\text{Mutual}}$, we can improve the model's performance even further with 1.3% of advancement. These results demonstrate that the proposed framework has an advantage over merely ensembling a group of experts.

Although PC softmax has a noticeable effect on model performance in this framework, applying it to other existing methods may be ineffective because they already use balanced classifiers. Table 8 shows the results of applying PC softmax to baseline approaches. Because the other methods already employ their own balancing techniques, applying PC softmax to them overcompensates the tail classes while penalizing the head classes, resulting in lower performance.

Method	Many	Med.	Few	All
Cross Entropy	68.7	41.8	10.3	47.9
Cross Entropy+PC	64.1	50.3	29.0	**52.7 (+4.8)**
MiSLAS [33]	61.7	51.3	35.8	52.7
MiSLAS+PC	43.3	45.9	52.3	**45.8 (−6.9)**
RIDE [28]	66.2	51.7	34.9	54.9
RIDE+PC	59.7	51.7	45.3	**53.9 (−1.0)**

Table 8: Applying PC softmax to other approaches. Top-1 accuracy on ImageNet-LT with ResNet-50 is reported. "PC" denotes PC softmax.

5. Conclusion

In this paper, we revisit the mutual learning strategy to foster better representations in long-tailed recognition. We empirically show that mutual learning can help us learn more generalizable features than independent learning. Furthermore, we highlight the significance of sampling strategy in mutual learning by demonstrating that instance-balanced sampling performs best. We extensively evaluate the efficacy of mutual learning on several long-tailed recognition benchmarks, including CIFAR100-LT, ImageNet-LT, and iNaturalist 2018, and achieve state-of-the-art performance. Last but not least, the mutual learning framework is simple and easily adaptable to other cutting-edge approaches.

References

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

[2] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 112–121, October 2021.

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[6] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021.

[7] 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*, pages 9268–9277, 2019.

[8] 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*, pages 248–255. Ieee, 2009.

[9] Chris Drummond and Robert Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*, 01 2003.

[10] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and rebalanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15089–15098, 2021.

[11] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 235–244, October 2021.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[15] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636, 2021.

[16] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[17] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

[18] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2022.

[19] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 630–639, 2021.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[22] 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*, pages 2537–2546, 2019.

[23] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[24] Dragos Margineantu. When does imbalanced data require more than cost-sensitive learning. In *Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets*, pages 47–50, 2000.

[25] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001.

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

[27] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and

Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[28] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

[29] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020.

[30] 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*, pages 247–263. Springer, 2020.

[31] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[32] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[33] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16489–16498, 2021.

[34] 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*, pages 9719–9728, 2020.

[35] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. 2014.