Mutual Exclusive Modulator for Long-Tailed Recognition

Anonymous CVPR submission

Paper ID 27

Abstract

The long-tailed recognition (LTR) is the task of learning high-performance classifiers given extremely imbalanced training samples between categories. Most of the existing works address the problem by either enhancing the features of tail classes or re-balancing the classifiers to reduce the inductive bias. In this paper, we try to look into the root cause of the LTR task, i.e.*, training samples for each class are greatly imbalanced, and propose a straightforward solution. We split the categories into three groups,* i.e.*, many, medium and few, according to the number of training images. The three groups of categories are separately predicted to reduce the difficulty for classification. This idea naturally arises a new problem of how to assign a given sample to the right class groups? We introduce a mutual exclusive modulator which can estimate the probability of an image belonging to each group. Particularly, the modulator consists of a light-weight module and learned with a mutual exclusive objective. Hence, the output probabilities of the modulator encode the data volume clues of the training dataset. They are further utilized as prior information to guide the prediction of the classifier. We conduct extensive experiments on multiple datasets,* e.g.*, ImageNet-LT, Place-LT and iNaturalist 2018 to evaluate the proposed approach. Our method achieves competitive performance compared to the state-of-the-art benchmarks.*

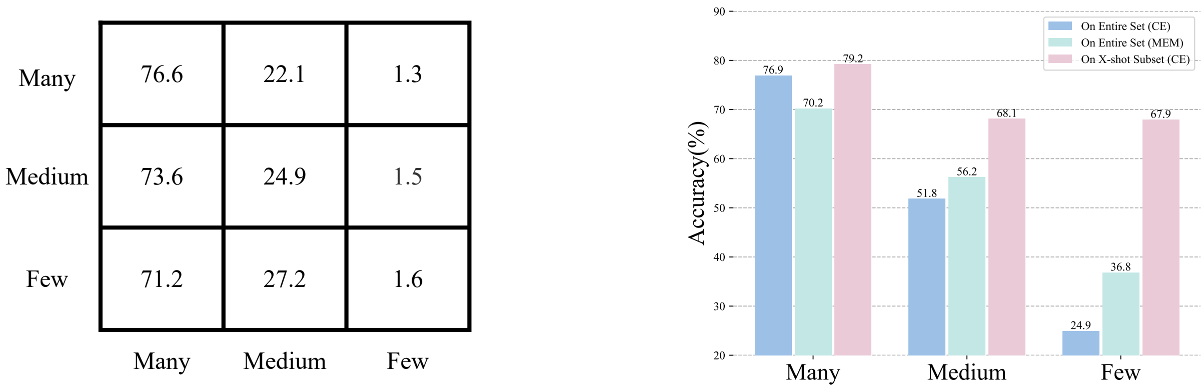
1. Introduction

In the past years, deep learning has achieved significant progress in computer vision [21]. The huge success in deep technologies is inseparable from the availability of high-quality large-scale datasets, *e.g.*, ILSVRC [34], MS COCO [27], and Places [46]. In contrast with these canonical datasets which are manually well-balanced across different categories *w.r.t* training data samples, the real-world data are always extremely skewed and exhibit long-tailed distribution. Most of the samples are congregated on a few of categories, *i.e.*, head classes, while the rest categories, *i.e.*, tail classes, possess very limited samples. Traditional models learned on such datasets perform very weak generalization ability and obtain poor recognition accuracy on tail classes.

To alleviate the skewness and increasing the performance of tail classes, previous works can be roughly summarized into two streams. The first stream aims to increase the representation capability of models and improve the extracted features for tail classes by seeking various constraints and network architectures [6,7,29,42]. Another line of works focuses on the adjustment of the decision boundaries of classifiers [7,14,32,35]. For example, [20] proposes to boost the recognition performance of tail classes by adjusting the cardinality of the classifier layer so as to adjust the separating hyperplane between categories.

In this paper, we first analyze the misclassified samples across different classes. For simplicity, we follow the same setting in [29] and divide the given dataset into three groups, *i.e.*, *Many*, *Medium* and *Few*, on the basis of the samples in each classes. Figure 1a depicts the proportion of misclassified images from one group to another. For the misclassified images in “Medium” group, 73.6% of them are predicted into the “Many” group. Similarly, for “Few” group, 71.2% of the wrongly predicted images are attributed into the “Many” group and 27.2% are assigned into the “Medium” group. Such an observation naturally arises a question: *Can reducing the misclassification rate between groups facilitate the accurate prediction?* To illustrate this, we hypothesize an ideal case that the ground-truth group that each sample belongs to is available. We then compute the precision on each group separately and plot it in the light pink column in Figure 1b. Compared to the case without group prior (light blue column), the dramatically improved performance of each group, for example, 67.9 *v.s.* 24.9 in “Few” group, answers yes to this question.

We introduce a Mutual Exclusive Modulator (MEM) to realize this purpose. MEM aims to learn and predict the mutually exclusive guidance of the groups for a given image. This prior-guidance assigns a set of selection probability on the three groups and enables the classifier to predict the right category within the chosen group. With the guide of MEM, the classifier can be re-adjusted and achieve



	Many	Medium	Few
Many	76.6	22.1	1.3
Medium	73.6	24.9	1.5
Few	71.2	27.2	1.6

(a) Misclassification percentage between groups.

(b) Comparision on ImageNet-LT.

Figure 1. (a) The digits indicate the percentage of images being misclassified to the wrong groups. For example, the digit $d_{i,j}$ on i row, j column means the misclassified images in group i has $d_{i,j}$ percent be misclassified to group j. Analysis is conducted on ImageNet-LT dataset with Vit-Base. (b) Evaluation accuracy compares between the entire set and the X-shot subset, where $X \in \{\text{Many, Medium, Few}\}$. For the entire set evaluation, which is also the common evaluation approach, consider all classes in this process. However, for X-shot evaluation, we only use the X-shot classes output rather than overall classes to calculate the X-shot accuracy..

a more accurate prediction from the guided label space. To learn the MEM guidance, we follow the same idea of the network activations, *i.e*., the output of MEM is activated if the group of an image is rightly predicted, otherwise it is depressed. Particularly, an embedding vector is employed to represent the group a given image belongs to. The magnitude of the embedding is optimized to be large w.r.t. the correct groups. Otherwise, the magnitude should be small and close to zero for the wrong groups. To achieve this goal of the feature magnitudes, we propose an objective function to learn the data-aware embedding in the training process. In detail, we first train a standard classification model by supervised learning. Based on the representation of the standard model, we train a mutual exclusive modulator to estimate the activation values for each group.

After harvesting the group guidance information, it is further utilized to obtain precise category predictions. We propose a data-aware classifier by fusing the group information into the estimation of the accurate labels in a soft manner. In particular, the data-aware classifier has the analogical property with the canonical classifier, so it can also be trained by backward propagation directly. In Figure 1b, we reveal that the proposed method increases the recall rate to facilitate the classification precision in each group (the light green column). Especially, the accuracy in the few group significantly improves by 11.9%. Our contribution can be summarized as follows:

- We analyze the relationship between the categories and the number of training samples by splitting them into different groups, and explore that the causes of low accuracy for tail classes are the low recall rates;
- We propose a novel mutual exclusive modulator to boost the recall of tail classes so as to improve the overall accuracy of long-tailed recognition;
- We conduct extensive experiments compared with the most relevant decoupled learning methods, *i.e*., τ-normalized, cRT, and LWS, and the state-of-the-art for long-tailed recognition. We show that our method achieves a more balanced performance between head and tail classes and outperforms the state-of-the-art by a nontrivial margin.

2. Related Work

Long-Tailed visual recognition. In a general way, methodologies in long-tailed recognition can be categorized as classes re-balancing, multi-stage learning, and ensemble learning. Classes re-balancing can be further divided into two types: re-sampling and re-weighting. For re-sampling, The most intuitive approach is under-sampling head classes or over-sampling tail classes to achieve more balance between head classes and tail classes [13, 14]. Another way for re-sampling is to apply a class-balanced sampling based on the cardinality of each class [32, 35]. For re-weighting, various methods propose to assign different losses to different instances [7, 26, 33]. Multi-stage based methods deal with long-tailed datasets by conducting a multi-steps schema [20, 24]. For instance, [20] decouples the learning procedure into representation learning and classifier learning. More recently, ensemble-based methods usually adopt multiple experts to reduce the model variance, *e.g*., RIDE [37], LFME [38], TADE [43], BBN [45], and NCL [23]. For example, LFME constructs three experts in the cardinality-adjacent subset to reduce the imbalance in training with the assumption that training on these subsets is better than jointly trained counterparts.

Self-supervised pretraining. Self-supervised learning unleashes the potential of vision transformers [3, 5, 15]. In the past few years, contrastive learning is very popular, which aims to learn invariances from different augmented views of images [4, 12, 16]. Recently, Masked Image Modeling (MIM) [11, 15, 28, 40] become more and more prevalent for vision transformers. MIM is the task that reconstructs image content from a masked image. Mask Autoencoders (MAE) [15] is a recent representative work. MAE builds an asymmetric encoder and decoder to reconstruct the corrupted input images in which most tokens are randomly masked.

Out-of-distribution detection. Out-of-distribution detection [1, 9, 18, 22, 25] is crucial to ensuring the reliability of the learning system [41]. It is a task to discriminate whether a sample in the inference is from a different distribution of the training data. Some approaches reject the out-of-distribution samples by setting a threshold on maximum softmax scores, they assume that the out-of-distribution samples will have a low maximum softmax score. However, different from the methods making a such assumption on out-of-distribution samples, [10] proposes two simple yet effective loss functions, the Entropic Open-Set loss, and Objectosphere loss, to jointly train the model with in-distribution data and out-of-distribution data. The Objectosphere loss attempts to increase the feature magnitude for in-distribution data and decrease it for other data. Motivated by [10], we design a regularization objective for the sub-network of MEM to activate the positive samples and depress the negative samples.

3. Method

For long-tailed recognition, various approaches of re-sampling [32, 35] and re-weighting [7, 33] aim to alleviate the bias towards the head classes and most of them adhere to joint learning representation and classification. [20] decouple the learning procedure into representation learning and classifier learning, which presents a great potential way from a different perspective. We follow this schema and propose a novel Mutual Exclusive Modulator (MEM) to receive more accurate predictions. Practically, we first train a standard classification model by supervised learning. Then, based on the representation of the standard model, we train an individual mutual exclusive modulator to generate a group of adaptive weights. By encoding the adaptive weights, the data-aware classifier act on logits in a soft-routing manner to boost the classification accuracy.

3.1. Mutual exclusive modulator

Motivated by the observation in Figure 1a that most of the misclassified samples are incorrectly recognized into the wrong group rather than the group they belong to. The performance can be significantly promoted if the samples were recalled in the right group, as illustrated in Figure 1b. Thus, we explore a new way of trying to recall the misclassified samples to the right group and propose a novel module *i.e.*, Mutual Exclusive Modulator (MEM). Without loss of generality, we split the classes into three mutual exclusive groups and denote them as $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$, following the previous practices [29]. The *mutual exclusive* indicates that a category can only be assigned into one specific group according to the number of training images.

Figure 2 depicts the overall structure of our approach. Given an image x_i, the encoder $\mathcal{F}$ transfers x_i to representation z_i, $z_i = \mathcal{F}(x_i)$. With input z_i, the MEM generates a group of adaptive weights via two procedures. First, the sub-networks inside the MEM map the representation z_i to latent vectors which imply the signal of the groups. To explicitly express the signal, we adopt the magnitude (L2-norm) of the latent vector to represent it. Such a magnitude represents the influence contributed by each group. In order to learn it, we design an objective to regularize the sub-networks. Afterward, we generate the adaptive weights by fusing the magnitude from each sub-network with the top K logits of each group through the designed *fusion module*. Finally, the data-aware classifier is given the final output. Our learning objective and the detail of the fusion module will be described as follows.

For each batch data in practice, we split the samples into three mutual exclusive groups based on their labels. We denote a batch input as $X = \{z_i, c_i\}, i \in \{1, \ldots, n\}$, where z_i is the input feature after encoder and c_i is the ground-truth label. Meanwhile, each sample can be further categorized into three mutual exclusive groups with group label $g \in \{1, 2, 3\}$. We try to capture the peculiarity of each sample and find the group label it belongs to. To realize this, we design three sub-networks $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3$ which parameterized with θ_1, θ_2 and θ_3 respectively, and a learning objective. Specifically, the learning objective can be formulated as:

$$r(z_i) = \lambda \begin{cases} \max(\xi - Q_g(z_i; \theta_g), 0)^2 & \text{if } c_i \in \mathcal{G}_g, \\ \max(Q_g(z_i; \theta_g) - \mu, 0)^2 & \text{otherwise.} \end{cases} \quad (1)$$

The learning objective aims to regularize the magnitude $\mathcal{Q}_g(z_i; \theta_g)$, which is the magnitude of the output of sub-network $\mathcal{Q}_g$, to be higher than ξ for a *positive class*, otherwise lower than μ for a *negative classes*. For each sub-network $\mathcal{Q}_g$, classes in the $\mathcal{G}_g$ are positive classes, other are negative classes. For example, sub-network $\mathcal{Q}_1$ treats samples in $\mathcal{G}_1$ as positive and the rest as negative. Intuitively,

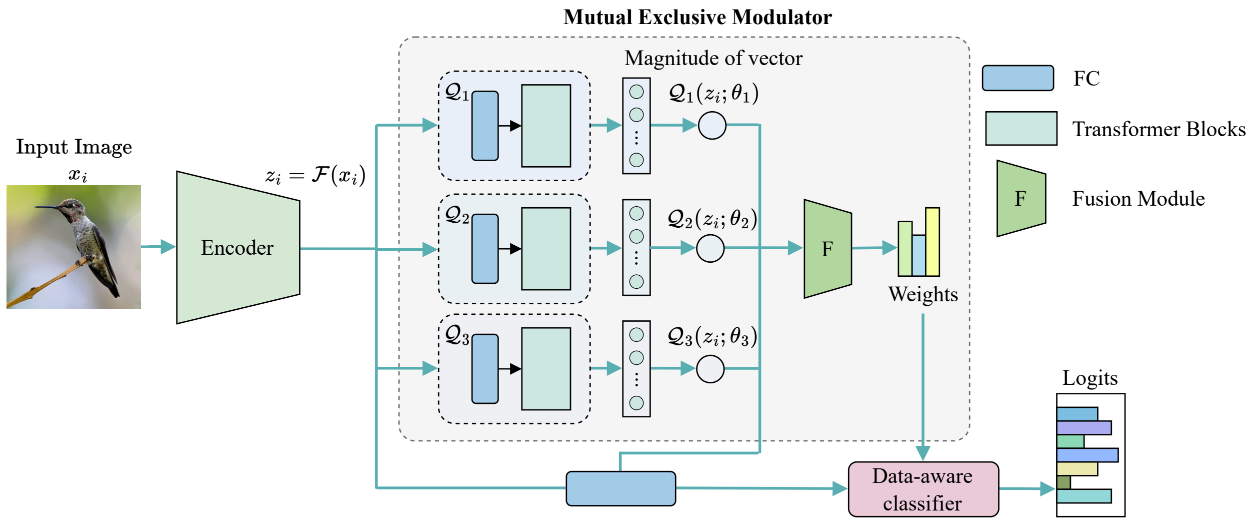


Figure 2. An illustration of our proposed Mutual Exclusive Modulator (MEM). MEM consists of three sub-networks and a fusion module. The sub-network maps the representation z_i to a latent vector that implies the signal of the group. To generate the adaptive weights, the fusion module encodes the magnitude from each sub-network and top K logits of each group from classifier.

each sub-network focuses on one group and tries to acquire the difference between groups. To elaborate the construction of the learning objective, we take sub-network $\mathcal{G}_1$ as an example. We leverage its output latent vector and compute the magnitude (L2-norm). We denote the magnitude as $\mathcal{Q}_1(z_i;\theta_1)$. For notation convenience, we first split X in each group g into positive part X_g^+ and negative part X_g^- . Formally, the positive part X_g^+ and negative part X_g^- can be expressed as:

$$\begin{cases} X_g^+ = \{(z_i, c_i) \mid c_i \in \mathcal{G}_g, (z_i, c_i) \in X\} \\ X_g^- = X \backslash X_g^+ \end{cases} \tag{2}$$

From another perspective, the X_g^+ and X_g^- can be also seen as ID (in-distribution) and OOD (out-of-distribution) samples for sub-network $\mathcal{Q}_g$ respectively. Based on X_g^+ and X_g^-, we formulate the regularization objective as:

$$\begin{cases} \mathcal{L}_{X_g^+} = \frac{1}{|X_g^+|} \sum_{z \in X_g^+} \max(\xi - \mathcal{Q}_g(z;\theta_g), 0)^2; \\ \mathcal{L}_{X_g^-} = \frac{1}{|X_g^-|} \sum_{z \in X_g^-} \max(\mathcal{Q}_g(z;\theta_g) - \mu, 0)^2 \end{cases} \tag{3}$$

where $|\cdot|$ stands for the cardinality in the set. The objective $\mathcal{L}_{X_g^+}$ is to encourage the output of $\mathcal{Q}_g$ greater than ξ for positive part X_g^+, however, objective $\mathcal{L}_{X_g^-}$ encourages the output lower than μ for negative part X_g^-. Averaging independently for the $\mathcal{L}_{X_g^+}$ and the $\mathcal{L}_{X_g^-}$ can equally contribute to the loss in a positive and negative part level. Otherwise, there will be a big difference in quantitative level between the loss of positive part and the negative part, especially when the distribution of the positive and negative part is extremely imbalanced in the batch. Then the regularization objective for sub-network Q_g can be summed up as:

$$\mathcal{L}_{Q_g} = \mathcal{L}_{X_g^+} + \mathcal{L}_{X_g^-} \tag{4}$$

Finally, the regularization objective for all three sub-networks can be written as:

$$\mathcal{L}_{REO} = \sum_{g \in 1,2,3} \mathcal{L}_{Q_g} \tag{5}$$

To get the final output of MEM, we take a fusion module to fuse the $\mathcal{Q}(z_i) = [\mathcal{Q}_1(z_i;\theta_1), \mathcal{Q}_2(z_i;\theta_2), \mathcal{Q}_3(z_i;\theta_3)]$ and top K logits of each group, where K is a hyper-parameter. The fusion module can be implemented as a simple multilayer perceptron (MLP). Suppose the logits are $\ell_i = \{\ell_i^1, \ell_i^2, \dots \ell_i^m\}$ for representation z_i, and the logits of group g can be write as:

$$\ell_{i,g} = \{\ell_i^j | j \in \{1, \dots, n\}, j \in \mathcal{G}_g\} \tag{6}$$

Suppose $\mathcal{T}_i = [\mathcal{T}_i^1, \mathcal{T}_i^2, \mathcal{T}_i^3]$ where $\mathcal{T}_i^g, g \in \{1, 2, 3\}$ is the set of top K values in $\ell_{i,g}$. Then the fusion operation can be formulated as:

$$w_i = \text{MLP}(\text{concat}(\ \mathcal{Q}(z_i),\ \mathcal{T}_i)), \tag{7}$$

where $w_i \in \mathbb{R}^3$ denotes the adaptive weights that output from the fusion module, and *concat* means the concatenate operation between the parameters.

3.2. Data-aware classifier

After harvesting the adaptive weights w_i, we construct a data-aware classifier (DAC) by utilizing the group information in a soft-routing manner. Together with the logits $\ell_i = \{\ell_i^1, \ell_i^2, \dots \ell_i^m\}$ of representation z_i, we can formulate our data-aware classifier as:

$$p_i = \text{softmax}(\{\ell_i^j \cdot w_i^{g(j)} \,|\, j \in \{1, \dots, m\}), \tag{8}$$

where $g(j) \in \{1, 2, 3\}$ denotes the mapping function from the class index j to the corresponding group.

The loss of data-aware classifier is thus:

$$L_{DAC} = -\frac{1}{n}\sum_{i=1}^{n} y_i \text{log}(p_i) \tag{9}$$

In the proposed individual mutual exclusive modulator, the two supervisions are employed together to achieve a comprehensive learning, and the final loss is written as:

$$\mathcal{L} = L_{REO} + L_{DAC} \tag{10}$$

4. Experiments

4.1. Experimental setup

Datasets. We conduct our experiments on three large-scale datasets (i.e., ImageNet-LT [29], Places-LT [29], iNaturalist 2018 [36]). ImageNet-LT and Places-LT are the long-tailed versions of ImageNet [8] and Places-365 [46] respectively. ImageNet-LT has 1,000 classes and contains 115.8k samples, with maximum of 1,280 samples and minimum 5 samples for a category. Places-LT contains 184.5K samples from 365 classes, with class samples ranging from 4,980 to 5. The iNaturalist 2018 is a large-scale species dataset collected in the natural world. It contains 437.5K samples for 8,142 classes.

Evaluation protocols. Following previous works [20, 29], the top-1 accuracy is adopted for evaluation. Moreover, we follow the setting in [29] to split the dataset into many-shot (with more than 100 samples), medium-shot (with 20∼100 samples), and few-shot (with less than 20 samples) and report the accuracy of each shot. All the results are reported as a percentage.

Implementation details. For ImageNet-LT and iNaturalist 2018, we report results based on ResNeXt-50 [39] and the transformer networks, i.e., Vit-Base [15], Vit-Large [15] and CVit-Base [11]. We apply MAE in [15] to pretrain the Vit-Base and CVit-Base for 400 epochs and Vit-Large for 800 epochs, while 100 epochs train from scratch for ResNeXt-50. For Places-LT, we report results based on ResNet-152 [17] and the transformer networks mentioned above. We start training from the finetuned classification model from [15] and [11] for transformer network, while the same setting follows [29] for ResNet-152 [17]. If not specified, for all experiments, we use Adamw [31] with betas=(0.9, 0.95), cosine learning rate schedule [30] with learning rate warmup to 1.5e-4 for 40 epochs, mask ratio 0.75 and weight decay 0.05, for MAE pre-training. For supervised finetuning, we train it for 100 epochs with 5 epochs warmup, cosine learning rate schedule, layer-wise learning decay following [15] and weight decay 0.05. For the training of MEM, we follow the training of decoupling methods in [20] to finetune 10 epochs while freezing the backbone. The value K in fusion operation 7 is 5 by default. And the details of the sub-network and the choose for hyper-parameters in the regularization objective can be referred in 4.3.

4.2. Results Comparisons

In this section, we make the comparison in two parts. The first part are about comparison with other decoupling methods which are most relevant to ours and the second part are more comprehensive results with previous methods as shown in Table 1. For the first part, we reproduce the corresponding methods for a fair comparison, and show our superior performances on all datasets mentioned above. For the second part, we compare with the state-of-the-art methods that based on convolutional networks.

4.2.1 Comparison with other decoupling methods

ImageNet-LT. Comparisons between different methods on ImageNet-LT are shown in Table 2. Most prior works present results based on CNN-based models, but results on recent prevalent structures are lacking. In this paper, we not only provide the comparisons on ResNeXt-50 but also the results on transformer based models. We reproduce the decoupling methods (τ-normalized, cRT and LWS) [20] on all backbones for a fair comparison. Compare with these methods, without bells and whistles, we achieve a new state-of-the-art for all backbones. Specifically, the MEM surpasses cRT by 0.82%, 0.81% and 0.66% on Vit-Base, Vit-Large and CVit-Base respectively, cRT is the best among all three decouple methods. And we found CVit-Base with fewer parameters but the accuracy is much higher than Vit-Base, e.g., 62.51% vs. 58.88%. A reasonable explanation is that the CVit-Base benefits from hybridizing convolutions and transformer blocks. Moreover, as shown in Figure 3, with the proposed mutual exclusive modulator, our method achieves more balance accuracy on each split than other decoupling methods. For example, MEM has higher medium-shot and few-shot accuracy with a slightly lower many-shot accuracy.

iNaturalist 2018. Comparisons on iNaturalist-2018 are shown in Table 3. Under a fair training setting, MEM surpasses τ-normalized, cRT and LWS consistently across

Method	Dataset	Many	Medium	Few	All
MiSLAS [44]	ImageNet-LT	62.0	49.1	32.8	51.4
Balanced Softmax [33]		64.1	48.2	33.4	52.3
LADE [19]		64.4	47.7	34.3	52.3
ACE [2]		-	-	-	56.6
RIDE [37]		68.2	53.8	36.0	56.9
PaCo [6]		68.2	58.7	41.0	60.0
NCL [23]		-	-	-	60.5
TADE [43]		68.6	61.2	47.0	62.1
MEM(ours)		76.0	62.0	43.7	**64.9**
LADE [19]	iNaturalist 2018	-	-	-	69.3
Balanced Softmax [33]		-	-	-	70.6
MiSLAS [44]		-	-	-	70.7
RIDE [37]		70.9	72.4	73.1	72.6
ACE [2]		-	-	-	72.9
PaCo [6]		70.3	73.2	73.6	73.2
NCL [23]		-	-	-	74.9
TADE [43]		78.3	77.0	76.7	77.0
MEM (ours)		83.6	83.0	81.3	**82.4**
MiSLAS [44]	Places-LT	-	-	-	38.3
LADE [19]		-	-	-	39.2
Balanced Softmax [33]		-	-	-	39.4
TADE [43]		40.4	43.2	36.8	40.9
PaCo [6]		36.1	47.9	35.3	41.2
NCL [23]		-	-	-	41.8
MEM (ours)		49.1	47.4	37.4	**46.0**

Table 1. Comparisons with previous works on dataset ImageNet-LT, iNaturalist 2018 and Places-LT. The results show that our proposed method (MEM) outperforms the state-of-the-art method by a large margin.

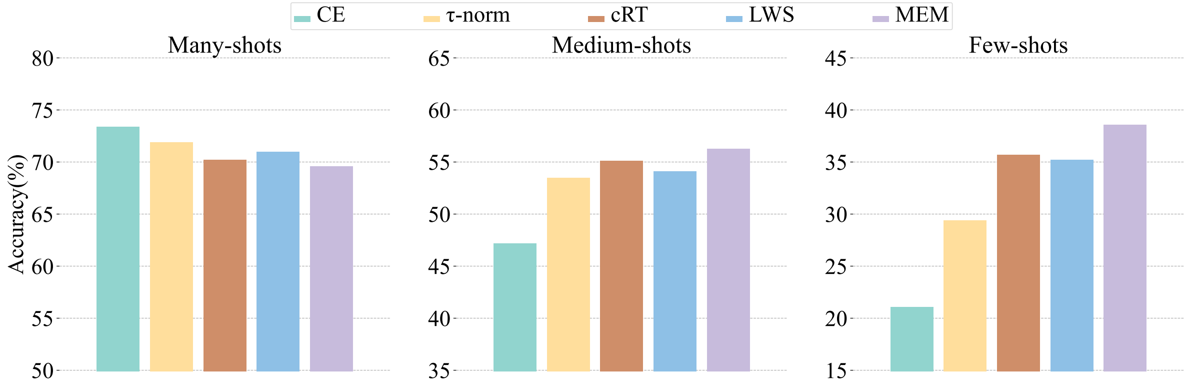


Figure 3. An illustration of the performance of each split with Vit-Base on ImageNet-LT. Different colors denote different methods.

all backbones (i.e., ResNeXt-50, Vit-Base, Vit-Large and CVit-Base). For example, the MEM outperforms cRT, which achieves the highest performance among all three decouple methods, on ResNeXt-50 (65.58% vs. 65.21%) and

Method	ResNeXt-50	Vit-Base	Vit-Large	CVit-Base
CE	45.17	53.48	59.67	57.54
τ-normalized [20]	48.99	57.07	62.74	60.56
cRT [20]	49.25	58.06	64.04	61.85
LWS [20]	48.16	57.87	63.30	61.34
MEM	**49.60**	**58.88**	**64.85**	**62.51**

Table 2. The accuracy on ImageNet-LT with Vision Transformer. We compare with the state-of-the-art decoupling methods on ResNeXt-50, Vit-Base, Vit-Large and CVit-Base.

Method	ResNeXt-50	Vit-Base	Vit-Large	CVit-Base
CE	60.14	72.59	79.47	78.18
τ-normalized [20]	61.85	76.27	81.96	81.11
cRT [20]	65.21	76.29	82.09	81.14
LWS [20]	63.94	76.24	82.11	81.19
MEM	**65.58**	**76.63**	**82.39**	**81.40**

Table 3. The accuracy on iNaturalist-2018 with Vision Transformer. We compare with the state-of-the-art decoupling methods on ResNeXt-50, Vit-Base, Vit-Large and CVit-Base.

Method	ResNet-152	Vit-Base	Vit-Large	CVit-Base
CE	30.74	35.98	37.68	36.95
τ-normalized [20]	31.18	36.12	37.83	37.23
cRT [20]	37.16	43.48	44.81	44.11
LWS [20]	36.73	42.84	44.80	43.47
MEM	**37.49**	**44.21**	**46.04**	**44.35**

Table 4. The accuracy on Places-LT with Vision Transformer. We compare with the state-of-the-art decoupling methods on ResNet-152, Vit-Base, Vit-Large and CVit-Base.

Vit-Base (76.63% vs. 76.29%). Once again, the CVit-Base with fewer parameters but has much higher accuracy than Vit-Base (81.40% vs. 76.63%). Moreover, we achieve a new state-of-the-art of 82.39% with Vit-Large that is outperforming the previous best result by 5.4%, see Table 1 for detail.

Places-LT. We further evaluate our MEM on Places-LT dataset. We follow the protocol of [29], initialize from a full ImageNet pre-trained model. We present the results after 30 epochs of fine-tuning, as shown in Table 4. Our MEM exceeds all other approaches, including CE, τ-normalized, cRT, and LWS. Moreover, we achieve the highest accuracy 46.04% on Vit-Large, which is 4.2% higher than the best results of previous works, see Table 1 for detail.

4.2.2 More comprehensive comparison

Here we make a more comprehensive comparison with the previous works on ImageNet-LT, iNaturalist 2018 and Places-LT, as shown in 1. The results of ours are based on the Vit-Large for all datasets. More specifically, our result on ImageNet-LT is 2.8% higher than the best, e.g., TADE [43] which is based on ResNeXt-152 [39]. On iNaturalist 2018, our result is 5.4% higher than TADE with ResNeXt-152. Moreover, for Places-LT, our MEM outperforms the best, i.e., NCL [23], by 4.2%. NCL is based on an ensemble of ResNeXt-50 models. By the comparison with previous CNN based works in Table 1 shows transformer networks could be a better choice for long-tailed recognition.

4.3. Ablation Study

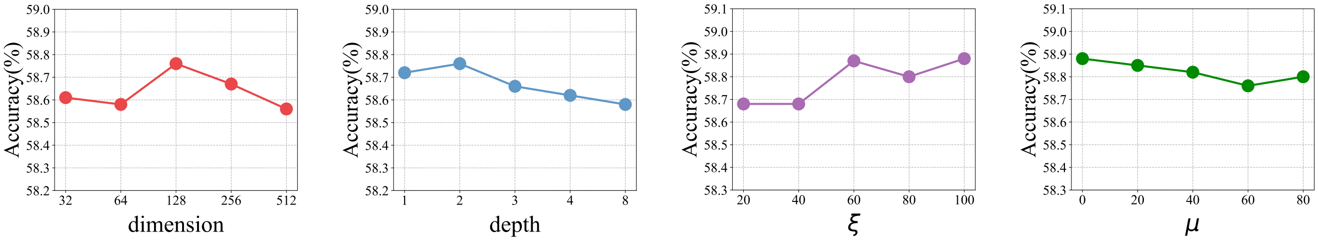
Pre-training for vision transformer. Ablations are conducted on ImageNet-LT with Vit-Base. As shown in Table 5, Vit-Base without MAE pre-training gets a very low accuracy, i.e., 27.72% for 100 epochs. It only attains 39.88% in accuracy when increasing the number of epoch form 100 to 500. However, with the training schedule that 400 epochs for MAE pre-training and 100 epochs for supervised fine-tuning, the accuracy is significantly improved from 39.9% up to 53.5%, which shows the necessity of MAE pre-training for transformer based models on long-tailed datasets. One explanation could be that there are insufficient samples in the long-tailed datasets for transformer based models to give play to its strength. From another point of view, the MAE could be regarded as providing with a fair initialization for the following supervised training.

Strategies for group partitioning. An intuitive way for group partition is based on the cardinality of each class. In long-tailed recognition, the classes are divided into {many, medium, few}-shot by default and we denote it as strategy (1) here. However, we try to explore whether there are more proper partition methods for mutual exclusive modulator learning. As shown in Table 5, we first conduct a random and even partition which is denoted as strategy (2). Afterwards, we split the classes evenly according to the cardinality of classes and denote it as strategy (3). We find that even in a random partition, our MEM can work well. But the strategies which utilize the information of the cardinality of each class, i.e., strategy (1) and strategy (3), work better.

The capacity of sub-network. Our sub-network inside MEM can be flexibly designed, we explore from two aspects (i.e., sub-network width and depth), as shown in Figure 4a and 4b. Figure 4a varies the width (the dimension of embedding) of sub-network, while we fix the sub-network

case	Partitioning strategy	epochs	Many	Medium	Few	All
w/o pre-training	/	100	46.5	19.5	4.6	27.72
w/o pre-training	/	500	61.0	31.6	10.9	39.88
w/o pre-training	/	1000	59.5	30.0	11.2	38.59
w/ pre-training	/	400+100	73.4	47.2	21.1	53.48
+MEM	Strategy (1)	+10	67.9	55.9	43.0	58.67
+MEM	Strategy (2)	+10	71.2	55.0	34.1	58.18
+MEM	Strategy (3)	+10	67.5	56.6	41.8	58.66

Table 5. Comparisons on pre-training and without pre-training on ImageNet-LT with Vit-Base. We present the results of without pre-training with supervised learning for 100, 500 and 1000 epochs, respectively. And the result of 400 epochs MAE pre-training and 100 epochs supervise learning is given for comparison. We also present the results of our MEM with different partitioning strategies.



(a) Sub-network width. (b) Sub-network depth. (c) Hyper-parameter ξ. (d) Hyper-parameter μ.

Figure 4. The ablation experiments with Vit-Base on ImageNet-LT. (a) and (b) are the ablation of the capacity of sub-network inside MEM. (c) and (d) are the ablation on the hyper-parameters which used for our regularization objective (see Equation 3).

depth. And Figure 4b varies the depth (number of transformer block) of sub-network, while we fix the sub-network width. We found that our sub-network is not sensitive to the width or depth. Interestingly, our sub-network with a single transformer block can yield strong performance (58.72%).

The hyper-parameters in regularization objective. We compare the different ξ and μ for our regularization objective (see Equation 3), as shown in Figure 4c and 4d. Figure 4c varies the parameter ξ and fix the parameter μ. When μ is 0, the accuracy has 0.2% improvement when ξ changes from 20 to 100. This can be explained as the regularization objective needs a large margin between positive classes and negative classes to optimize. And for Figure 4d, we vary the parameter μ and under a certain ξ. Figure 4d shows that the accuracy has minor adjustments under different μ. In general, our regularization objective is not sensitive to the hyper-parameters ξ and μ.

5. Conclusions

In this paper, we first focus on the behaviors of existing models on three separate groups, *i.e.*, Many, Medium, and Few, in the long-tailed datasets and reveal that the reason for the poor performance is the severe confusion between groups. The model tends to categorize samples in one group to another. Then, we investigate an ideal case that the images are first classified to the right group before the final label is predicted. The overall performance has seen huge promotion with this simple assumption. Motivated by this, we thus propose a straightforward structure, which is called Mutual Exclusive Modulator (MEM), to capture the characteristics of each group and the discrepancy among them. It perceives a set of adaptive weights for different groups. Together with original logits, the data-aware classifier make the final prediction by following a soft-routing manner. MEM achieves significantly better results on different backbones, including convolutional networks and transformer networks, compared with the state-of-the-art methods.

References

[1] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 3

[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*, pages 112–121, 2021. 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 3

[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. 1, 6

[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. 1, 2, 3

[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. 5

[9] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3

[10] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[11] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 3, 5

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[13] 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. 2

[14] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 1, 2

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3, 5

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[17] 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. 5

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 3

[19] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 6

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 1, 2, 3, 5, 7

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3

[23] 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. 2, 6, 7

[24] 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. 2

[25] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 3

[26] 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. 2

[27] 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. 1

[28] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*, 2022. 3

[29] 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. 1, 3, 5, 7

[30] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[32] 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. 1, 2, 3

[33] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 2, 3, 6

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[35] 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. 1, 2, 3

[36] 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. 5

[37] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020. 2, 6

[38] 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. 2

[39] 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. 5, 7

[40] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3

[41] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 3

[42] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020. 1

[43] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*, 2021. 2, 6, 7

[44] 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. 6

[45] 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. 2

[46] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1, 5