FEATURE-BALANCED LOSS FOR LONG-TAILED VISUAL RECOGNITION

Mengke Li   Yiu-ming Cheung   Juyong Jiang*

Department of Computer Science, Hong Kong Baptist University, Hong Kong
{csmkli, ymc, csjyjiang}@comp.hkbu.edu.hk

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9860003

ABSTRACT

Deep neural networks frequently suffer from performance degradation when the training data is long-tailed because several majority classes dominate the training, resulting in a biased model. Recent studies have made a great effort in solving this issue by obtaining good representations from data space, but few of them pay attention to the influence of feature norm on the predicted results. In this paper, we therefore address the long-tailed problem from feature space and thereby propose the feature-balanced loss. Specifically, we encourage larger feature norms of tail classes by giving them relatively stronger stimuli. Moreover, the stimuli intensity is gradually increased in the way of curriculum learning, which improves the generalization of the tail classes, meanwhile maintaining the performance of the head classes. Extensive experiments on multiple popular long-tailed recognition benchmarks demonstrate that the feature-balanced loss achieves superior performance gains compared with the state-of-the-art methods.

***Index Terms*—** Long-tailed recognition, class imbalance learning, feature-balanced loss, deep neural networks

1. INTRODUCTION

In classification problems, real-world data often exhibits a long-tailed distribution: a few majority classes have large amounts of samples, while numerous minority classes are with only a few samples. This extreme imbalance class distribution leads to the model training dominated by head classes. As a result, the model performance for tail classes is severely degraded. Nowadays, it is still challenging to effectively train a model on long-tailed data in visual recognition tasks.

To address the issue of extreme data imbalance caused by long-tailed distribution, an intuitive way is to re-balance the model via class-balanced sampling [1, 2] or loss function reweighting [3, 4]. However, these methods result in overfitting to the tail classes, which invariably inhibits the performance of the model. Most recently, Cui *et al.* [5] have proposed to re-weight the loss function or re-sample the data based on the "effective number" of each class, which has been shown empirically effective. This "effective number" strategy, on the other hand, does not truly address the issue of uneven feature distribution for long-tailed data. Subsequently, Cao *et al.* [6] utilized the label-distribution-aware margin (LDAM) to re-weight the loss, which can improve the generalization performance of tail classes. Nevertheless, it calculates the predicted logit through the cosine distance, which neglects the significant influence of the feature norm.

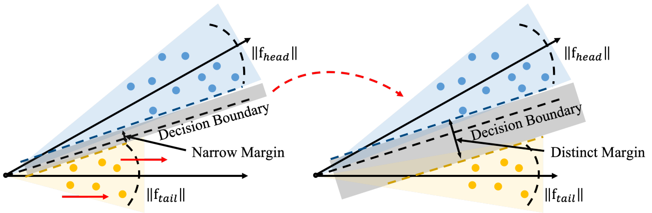

Fig. 1. A schematic diagram of the influence of feature norm on decision margin in embedding space. With the increase of the feature norm of the tail class samples, the margin becomes clear and the separability of the samples can be enhanced, which in turn improves the model generalization towards the samples.

In this paper, we address the long-tailed problem from a feature norm perspective and thereby proposing the feature-balanced loss (FBL). As shown in Fig. 1, it can be seen that training samples with less feature norm are difficult to classify because of the unclear margins between each class. The increase of feature norm can enlarge the margins between classes and enhance the separability of the samples. Based on this observation, we add a class-based stimulus to the predicted logit to encourage larger tail class feature norm to improve its generalization. Different from LDAM that utilizes hard margins to increase intra-class compactness, our FBL enlarges decision margin without compressing the embedding space distribution of each class. Furthermore, we adopt curriculum learning [7] strategy to gradually increase the class-based stimulus so that the network initially concentrates on the head classes, and then gradually shifts its attention to the tail classes as the training progresses. In this way, the classification accuracy of the tail classes can be improved while

* Corresponding author. We thank Ms. Hui Ji for her help on part of the experiments and literature review. This work was supported in part by NSFC/RGC JRS Grant: N_HKBU214/21, GRF Grant: 12201321, NSFC Grant: 61672444, and HKBU Grant: RC-FNRA-IG/18-19/SCI/03. Source code is available at https://github.com/juyongjiang/FBL.

maintaining the performance of the head classes. We validate the proposed FBL on five popular benchmark datasets, *i.e.*, CIFAT-10-LT, CIFAT-100-LT, ImageNet-LT, iNaturalist 2018 and Places-LT. We also conduct an additional experiment on feature norm visualization, which demonstrates that feature norm is one of the key factors for improving the accuracy of long-tailed data classification.

Our main contributions are summarized as follows:

- We propose the novel FBL for long-tailed visual recognition by adding an extra classes-based stimulus to the logit. The proposed FBL encourages larger feature norms for tail classes, thereby improving the generalization performance of these classes.

- We propose to gradually increase the intensity of stimulus in the way of curriculum learning. This robust training strategy not only enhances the classification accuracy of tail classes to a large extent, but also maintains the performance of head classes.

- We conduct extensive experiments on commonly used long-tailed datasets, which demonstrates the superiority of the proposed method in comparison with the state-of-the-art methods.

2. RELATED WORK

Long-tailed visual recognition has received increasing attention in computer vision because of the prevalence of data imbalance in the real world. This section will make an overview of the most related works.

2.1. Loss Modification

Loss modification aims to re-balance the importance of different classes by tuning the loss values. It addresses the class imbalance problem from two perspectives: sample-wise and class-wise. Sample-wise methods [8, 9] assign large relative weights to the difficult samples through the fine-grained parameters in the loss. For example, focal loss [9] utilizes the sample prediction hardness as the re-weighting coefficient of the loss function. However, the classification difficulty of a sample may not be directly related to its corresponding class size. Hence, the sample-wise method is incapable of handling the large-scale and severe imbalance data. Class-wise methods [4, 5, 10] assign the loss function with class-specific parameters that are negatively correlated to the label frequencies. For example, Cui *et al.* [5] proposed to re-weight the loss function by the “effective number” of each class instead of label frequency. Nevertheless, it does not completely alleviate the problem of biased feature distribution.

2.2. Logit Adjustment

Logit adjustment addresses the class imbalance problem by calibrating the logit to the prior during inference or training. Typically, a number of approaches adjust the loss during training. Most recently, Cao *et al.* [6] have proposed label-distribution-aware margin accompanied with the deferred scheme (LDAM-DRW), which enforces tail classes to have large relative margins to increase their classification accuracy. Furthermore, DisAlign [11] adaptively aligns the logit to a balanced class distribution to adjust the biased decision boundary, which can re-balance the classifier well. Besides, another kind of method post-hoc shifts the predicted logits. For example, Menon *et al.* [12] proposed logit adjustment (LA) to post-process the logit based on the label frequencies of training data. In contrast, Hong *et al.* [13] proposed LADE, which post-adjusts logits with the label frequencies of testing data, allowing the distribution of the test set to be arbitrary.

3. PROPOSED METHOD

To mitigate the training bias towards the head classes caused by long-tailed data, we propose the FBL as a more powerful supervised signal for optimizing deep neural networks (DNNs).

3.1. Analysis of Softmax Loss Function in Classification

Given a training sample x with the label y from the training set $\mathcal{T}$ with total C classes and N training samples. We use $\mathbf{f} \in \mathbb{R}^D$ to represent the feature of x obtained from the embedding layer with dimension D. $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_C\} \in \mathbb{R}^{D\times C}$ represents the weight matrix of the classifier, where $\mathbf{w}_i$ represents the weight vector of class i in the classifier. The predicted logit of class i is represented by z_i, thus, $z_i = \mathbf{w}_i^T\mathbf{f}$. We use the subscript y to represent the target class. That is, z_y indicates the target logit and z_i $(i \neq y)$ is the non-target logit. The original softmax loss function for the given sample x is:

$$L_{\text{softmax}}(x) = -\log \frac{e^{z_y}}{\sum_j e^{z_j}}. \tag{1}$$

The gradient of L_{softmax} w.r.t. z_i is:

$$\frac{\partial L_{\text{softmax}}}{\partial z_i} = \begin{cases} p_i - 1, & i = y \\ p_i, & i \neq y \end{cases}, \tag{2}$$

where $p_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$. In backward propagation, the gradients of the target class are negative, and those of the non-target classes are positive. Thus, the training samples punish the non-target class weights $\mathbf{w}_i$ $(i \neq y)$ by p_i. The weights of tail classes which have fewer training instances always receive punishment signals. As a result, the weight norm of the classifier for tail classes is always reduced. Therefore, we obtain the following properties:

Property 1. The weight norm $\|\mathbf{w}_i\|$ of the classifier for class i is correlated with the class size n_i.

In addition, we introduce additional property of softmax loss that was found by Yuan *et al.* [14]:

Property 2. By fixing the weight vector and direction of feature vectors, softmax loss is a function that monotonically decreases with the increasing of feature L_2-norm when features are correctly classified.

Property 1 indicates that the target logit $z_y = \mathbf{w}_y^T\mathbf{f}$ of tail class is usually suppressed because of the relatively small $\mathbf{w}_y^T$. Meanwhile, Property 2 shows that feature norm is an important factor to achieve a lower loss, so that the features can be more separable. To improve the performance on tail classes, we can encourage larger feature norm for tail classes to diminish the bias towards the head classes.

3.2. FBL with Curriculum Learning

To stimulate the large feature norm, we can add an additional constraint item to the original cross-entropy loss:

$$L' = -\log\frac{e^{z_y}}{\sum_i e^{z_i}} + \alpha\frac{\lambda_y}{\|\mathbf{f}\|}, \tag{3}$$

where α is the parameter used to adjust the strength of the constraint, and λ_y controls the stimulus intensity towards different classes. Since Property 1 in Sec. 3.1 states that the weight norm of classifier for tail classes is usually suppressed, the logits of tail classes will be unfairly reduced. To diminish this bias, we can encourage large feature norms for tail classes and thus assign them stronger stimulation. Therefore, λ_y is negatively correlated with the number of samples in class y.

For the sake of analysis of the loss function, we rewrite Eq. (3) as:

$$\begin{aligned} L' &= -\log\frac{e^{z_y}}{\sum_j e^{z_j}} + \log e^{\frac{\lambda_y}{\|\mathbf{f}\|}} \\ &= -\log\frac{e^{z_y - \frac{\lambda_y}{\|\mathbf{f}\|}}}{\sum_j e^{z_j}} \\ &= -\log p_y \end{aligned}, \tag{4}$$

where $p_y = \frac{e^{z_y - \frac{\lambda_y}{\|\mathbf{f}\|}}}{\sum_j e^{z_j}}$. As the sum of the probabilities of all classes obtained by Eq. (4) is not equal to 1, *i.e.*, $\sum_{y=1}^{C} p_y \neq 1$, we further modify the logit to ensure that the total predicted probabilities of all classes are equal to 1. The feature-balanced logit z_j^b of class j is introduced and is expressed as:

$$z_j^b = z_j - \alpha\frac{\lambda_j}{\|\mathbf{f}\|}. \tag{5}$$

In addition, λ_j controls the intensity of the stimulus, which should be weak for head classes and strong for tail classes. Subsequently, we set λ_j at:

$$\lambda_j = \log n_{max} - \log n_j, \tag{6}$$

so that it is zero for the most frequent class and is much stronger for tail classes.

Algorithm 1: FBL with curriculum learning

```
Input: Training dataset $\mathcal{S}$
Output: Predicted labels
Initialize the DNN model $\phi((x, y); \theta)$ randomly,
  where $\theta$ is the parameter of the model;
for $t = 1$ to $T$ do
    Sample mini-batch training samples $\mathcal{B}$ from the
      long-tailed data $\mathcal{S}$ with batch size of $b$;
    Obtain the constraint strength parameter $\alpha$:
      $\alpha \leftarrow \alpha(t)$;
    Obtain the stimulus intensity parameter $\lambda_j$:
      $\lambda_j \leftarrow \log n_{max} - \log n_j$;
    Calculate the loss by Eq. (8):
      $\mathcal{L}((x, y); \theta) = \frac{1}{b}\sum_{(x,y)\in\mathcal{B}} L_{\text{FBL}}(x, y)$;
    Update model parameters:
      $\theta \leftarrow \theta - \alpha'\nabla_\theta\mathcal{L}((x, y); \theta)$;
end
```

Furthermore, the stronger the constraint on feature (*i.e.*, $\frac{\lambda_j}{\|\mathbf{f}\|}$) is, the more the model focuses on the tail classes. We can adopt the idea of curriculum learning [7], which makes the model initially focus on easy samples (*i.e.*, head classes), and then gradually shift to learning difficult samples (*i.e.*, tail classes). To achieve this, we can choose the learning strategy that gradually increases α as the training progresses. Therefore, we replace α by $\alpha(t)$ which is related to the training epoch t. We empirically select the parabolic increase learning strategy, which is expressed as:

$$\alpha(t) \propto (\frac{t}{T})^2, \tag{7}$$

where t is the training epoch and T is the total number of epochs. Sec. 4.5 also provides experimental results for different learning strategies.

The final loss function L_{FBL} is expressed as:

$$L_{\text{FBL}} = -\frac{1}{N}\sum_i \log\frac{e^{z_{y_i}^b}}{\sum_j e^{z_j^b}}. \tag{8}$$

This loss function is named as FBL–feature-balanced loss, because it balances the logit of different classes based on feature norm. The algorithm of our proposed method is summarized in Algorithm 1.

4. EXPERIMENTS

4.1. Datasets

To demonstrate the effectiveness of our proposed FBL, we conduct the experiments on five benchmark datasets with the various scales.

Table 1. An Overview of Long-Tailed Datasets

Dataset	CIFAR-10-LT		CIFAR-100-LT		ImageNet-LT	Places-LT	iNat 2018
# Classes	10		100		1,000	365	8,142
IF	100	50	100	50	256	996	500
# Train. img.	12,406	13,996	10,847	12,608	115,846	62,500	437,513
Tail class size	50	100	5	10	5	5	2
Head class size	5,000	5,000	500	500	1,280	4,980	1,000
# Val. img.	-	-	-	-	20,000	7,300	24,426
# Test img.	10,000	10,000	10,000	10,000	50,000	36,500	-

CIFAR-10/100-LT [6] down-samples the original balanced version of CIFAR-10/100 [15] per class by an imbalanced factor $IF = \frac{N_{max}}{N_{min}}$ (where N_{max} and N_{min} are the numbers of training samples in the most and the least frequent classes, respectively). CIFAR-10-LT and CIFAR-100-LT have two typical variants, namely, with $IF = \{100, 50\}$.

ImageNet-LT [16] is a large-scale long-tailed dataset for object classification through sampling a subset following the Pareto distribution with the power value $\alpha = 6$ from ImageNet-2012 [17]. It includes 115.8K images with the class size ranging from 5 to 1,280, imitating the long-tailed distribution that regularly existed in the real world.

Places-LT is a long-tailed version of the large-scale scene classification dataset Places-365 [18]. There are 184.5K images with class sizes ranging from 5 to 4,980. Moreover, the gap between the sizes of tail and head classes of this dataset is larger than that of ImageNet-LT.

iNaturalist 2018 (iNat 2018) is the iNaturalist species classification and detection dataset [19], which is a massive real-world long-tailed dataset. In its 2018 version, iNaturalist comprises 437,513 training images from 8,142 classes. In the light of different classes, the numbers of the training samples follow an exponential decay.

Table 1 summarizes the details of the above datasets.

4.2. Implementation Details

We use Pytorch to implement and train all the backbones with stochastic gradient descent with momentum.

Backbone. Following the protocol of Cui *et al.* [5], ResNet-32 is adopted as the backbone for all CIFAR-10/100-LT datasets. For ImageNet-LT and iNat 2018, ResNet-50 is applied. For Places-LT, we follow Liu *et al.* [16] and start from a ResNet-152 pre-trained on the original balanced version of ImageNet. Except for ResNet-152, all the backbones are trained from scratch.

Training details. For CIFAR-10/100-LT, we train the backbone with 200 epochs and batch size of 64. The initial learning rate (lr) is set at 0.1, and we anneal lr by 100 at the 160-th and 180-th epoch, respectively. For the three large-scale datasets, backbone is trained with 180 epochs, batch size of 512, and initial $lr = 0.2$. We divide lr by 10 at 120-th and 160-th epochs.

Table 2. Comparison results on CIFAR-10/100-LT. Top-1 accuracy (%) are reported. The best results are shown in **underline bold**.

Dataset	CIFAR-10-LT		CIFAR-100-LT	
Backbone Net	ResNet-32			
IF	100	50	100	50
CE loss (baseline)	71.07	75.31	39.43	44.20
LDAM-DRW [6] (*NeurIPS* 2019)	77.03	81.03	42.04	47.62
BBN [20] (*CVPR* 2020)	79.82	81.18	42.56	47.02
LA [12](*ICLR* 2021)	80.92	-	43.89	-
FBL (ours)	**82.46**	**84.30**	**45.22**	**50.65**

Table 3. Comparison results on ImageNet-LT, iNaturalist 2018 and Places-LT. Top-1 accuracy (%) are reported. The best results are shown in **underline bold**.

Dataset	ImageNet-LT	iNat 2018	Places-LT
Backbone Net	ResNet-50	ResNet-50	ResNet-152
CE loss (baseline)	44.51	63.80	27.13
LDAM-DRW [6] (*NeurIPS* 2019)	48.80	68.00	-
Decoupling [2] (*ICLR* 2020)	47.70	69.49	37.62
LA [12] (*ICLR* 2021)	50.44	66.36	-
FBL (ours)	**50.70**	**69.90**	**38.66**

4.3. Comparison Methods

The vanilla training with cross-entropy (CE) loss is chosen as the baseline method. We compare the proposed method with the state-of-the-art ones, *i.e.*, the logit modification methods including: LDAM-DRW [6] and LA [12], the most recently proposed two-stage method–BBN [20] on the small-scale datasets (CIFAR-10/100-LT) and decoupling on the large-scale datasets (imageNet-LT, iNat 2018 and Places-LT).

4.4. Long-Tailed Recognition Results

Results on CIFAR-10/100-LT. We conduct the comparison experiments on CIFAR-10/100-LT with $IF = \{100, 50\}$. Table 2 summarizes the top-1 accuracy. Our FBL outperforms the other competing methods by noticeable margins across all the datasets. For example, FBL outperforms the state-of-the-art method – LA by 1.54% and 1.33% with $IF = 100$ on CIFAR-10-LT and CIFAR-100-LT, respectively.

Results on large-scale datasets. FBL yields good performance on all large-scale datasets, which is consistent with that CIFAR-10/100-LT. Table 3 shows the comparison results.

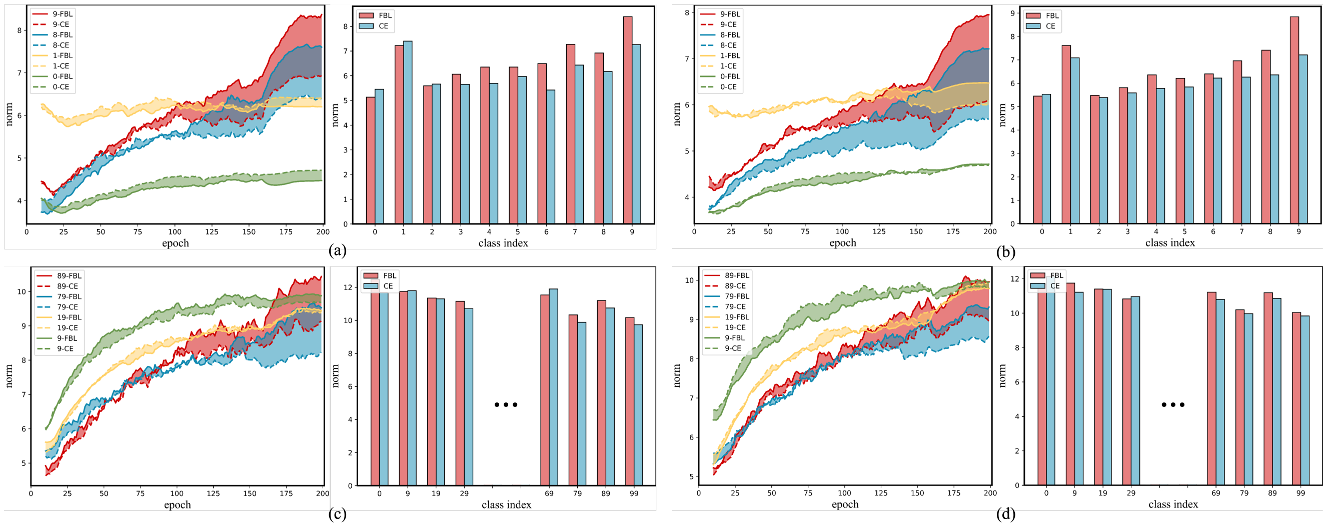

Fig. 2. **Top**: The changes of feature norm on *head classes* (class index-$\{0, 1\}$) and *tail classes* (class index-$\{8, 9\}$) with respect to training epochs (left) and the feature norm distribution of classes over test dataset (right) on CIFAR-10 with $IF = 100$ (a) and 50 (b). **Bottom**: The changes of feature norm on *head classes* (class index-$\{9, 19\}$) and *tail classes* (class index-$\{79, 89\}$) with respect to training epochs (left) and the feature norm distribution of classes over test dataset (right) on CIFAR-100 with $IF = 100$ (c) and 50 (d).

Table 4. Ablation experiment of different learning strategy on CIFAR-10-LT with $IF = 100$.

$\alpha(t)$	Representation	Acc.(%)
Linear decrease	$1 - t/T$	75.97
Linear increase	t/T	81.67
Sine increase	$\sin(t/T \cdot \pi/2)$	81.22
Cosine increase	$1 - \cos(t/T \cdot \pi/2)$	80.79
Parabolic increase	$(t/T)^2$	**82.46**

The proposed FBL that can be trained end-to-end not only achieves better results than LA, but also is superior to the two-stage method, *i.e.*, LDAM-DRW and decoupling. For example, on ImageNet-LT, FBL outperforms LDAM-DRW and decoupling by 1.90% and 3.00%, respectively.

4.5. Ablation Study

We conduct an ablation study to investigate the effectiveness of different learning strategies adopted by $\alpha(t)$. Table 4 summarizes their performance. It can be seen that the classification accuracy of the linear decrease strategy is 75.97%, which is only higher than that of the baseline method (71.07%). It is not as competitive as other learning strategies, because it makes the DNN model focus on hard samples (*i.e.*, tail classes) first. As the training progresses, the network gradually forgets what it has previously learned. Therefore, there is basically no improvement in the performance of the tail classes. Other strategies that increase α with the training epoch t gradually shift the network's attention from the head classes to the tail, which can avoid forgetting the tail classes and improve the overall performance.

Table 5. Per-class accuracy $(\%)$ of test set on CIFAR-10-LT.

Class index	0	1	2	3	4	5	6	7	8	9
IF	100									
CE loss	**91.0**	**98.2**	**83.2**	72.5	78.8	65.1	68.8	59.5	49.0	44.6
FBL	88.1	94.7	81.9	**73.0**	**83.6**	**75.1**	**86.3**	**77.3**	**82.7**	**81.9**
IF	50									
CE loss	**84.5**	**95.8**	68.5	**74.6**	81.1	72.7	82.9	67,5	59.1	66.4
FBL	83.7	92.1	**81.7**	73.9	**85.0**	**76.1**	**87.7**	**85.0**	**88.5**	**89.3**

Table 6. Per-class accuracy $(\%)$ of test set on CIFAR-100-LT.

Class index	0	9	19	29	$\cdots$	69	79	89	99
IF	100								
CE loss	**89.0**	72.0	**59.0**	**48.0**	$\cdots$	45.0	12.0	3.0	2.0
FBL	86.0	**77.0**	54.0	45.0	$\cdots$	**60.0**	**27.0**	**22.0**	**8.0**
IF	50								
CE loss	**88.0**	**79.0**	53.0	49.0	$\cdots$	53.0	10.0	19.0	13.0
FBL	87.0	77.0	**56.0**	**57.0**	$\cdots$	**62.0**	**48.0**	**38.0**	**17.0**

4.6. Feature-balanced Results

To further validate the effects of the proposed FBL, especially the tail classes, we visualize the changes of the feature norm (*i.e.*, $\|\mathbf{f}\|$) with respect to training epochs and feature norm distribution of classes over the test set on CIFAR-10/100-LT. The results are shown in Fig. 2. The corresponding per-class accuracy is presented in Table 5 and 6, respectively. The following phenomena can be seen:

- The capability of the model to learn from different classes of samples is diverse. Specifically, in Fig. 2 (a) and (b), the feature norms of head classes samples (class index-$\{0, 1\}$) reach stable in very early training epochs due to enough training samples. Differently, on CIFAR-100-LT (as shown in Fig. 2 (c) and (d)), the feature norms of the samples from all classes including head (class index-$\{9, 19\}$) and tail classes (class index-$\{79, 89\}$) are constantly changing as the epoch increases, which have a similar phenomenon to the tail classes in CIFAR-10-LT (class index-$\{8, 9\}$ in Fig. 2 (a) and (b)) because they all suffer from insufficient training samples.

- Compared with CE loss, our FBL encourages larger feature norms of tail class samples to eliminate representation bias towards head classes. The area (S^{*}_{area}(class index)) enclosed by the curve of CE loss and FBL becomes larger as the number of class samples decreases, e.g. $S^{tail}_{area}(9) > S^{tail}_{area}(8) > S^{head}_{area}(1) > S^{head}_{area}(0)$ in Fig. 2 (a), which is in line with our motivation.

These observations not only justify our intuition about the influence of feature norm on decision margin, but also offer a new promising way to investigate long-tailed visual recognition.

5. CONCLUSIONS

In this work, we have proposed a novel FBL to address the long-tailed classification from feature space. FBL encourages larger feature norms of tail classes by adding relatively stronger stimuli to the logits of tail classes, which can mitigate the representation bias towards head classes in the feature space. In addition, a curriculum learning strategy has been adopted to gradually increase the stimuli in training, which can keep the good accuracy of the model for the head classes and improve the performance of the tail classes. FBL allows DNNs to be trained end-to-end without the risk of a performance drop from head classes. Extensive experiments have demonstrated the superiority of the proposed FBL.

6. REFERENCES

[1] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *IJCNN*, 2008, pp. 1322–1328.

[2] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *ICLR*, 2020.

[3] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for imbalanced classification," in *CVPR*, 2016, pp. 5375–5384.

[4] Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous Ahmed Sohel, and Roberto Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *TNNLS*, vol. 29, no. 8, pp. 3573–3587, 2018.

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019, pp. 9268–9277.

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019, pp. 1567–1578.

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.

[8] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun, "Learning to reweight examples for robust deep learning," in *ICML*, 2018, vol. 80, pp. 4331–4340.

[9] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *TPAMI*, vol. 42, no. 2, pp. 318–327, 2020.

[10] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan, "Equalization loss for long-tailed object recognition," in *CVPR*, 2020, pp. 11662–11671.

[11] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *CVPR*, 2021, pp. 2361–2370.

[12] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar, "Long-tail learning via logit adjustment," in *ICLR*, 2021.

[13] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang, "Disentangling label distribution for long-tailed visual recognition," in *CVPR*, 2021, pp. 6626–6636.

[14] Yuhui Yuan, Kuiyuan Yang, Jianyuan Guo, Chao Zhang, and Jingdong Wang, "Feature incay for representation regularization," in *ICLR (workshop)*, 2018.

[15] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," *Tech Report*, 2009.

[16] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019, pp. 2537–2546.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[18] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million image database for scene recognition," *TPAMI*, vol. 40, no. 6, pp. 1452–1464, 2018.

[19] 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 *CVPR*, 2018, pp. 8769–8778.

[20] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *CVPR*, 2020, pp. 9719–9728.