Received: 23 October 2022 | Revised: 19 December 2023 | Accepted: 21 March 2024

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12374

ORIGINAL RESEARCH

Improving long-tail classification via decoupling and regularisation

Shuzheng Gao[1] | Chaozheng Wang[1] | Cuiyun Gao[1,2,3] | Wenjian Luo[1] | Peiyi Han[1] | Qing Liao[1,2] | Guandong Xu[4]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

[2]Peng Cheng Laboratory, Shenzhen, China

[3]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, Harbin Institute of Technology, Shenzhen, China

[4]Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, New South Wales, Australia

Correspondence

Cuiyun Gao.
Email: gaocuiyun@hit.edu.cn

Funding information

National Key Research and Development Program of China, Grant/Award Numbers: 2022YFB3103900, 2023YFB3106504; Major Key Project of PCL, Grant/Award Numbers: PCL2022A03, PCL2023A09; Shenzhen Basic Research, Grant/Award Number: JCYJ20220531095214031; Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, Grant/Award Number: 2022B1212010005; Shenzhen International Science and Technology Cooperation Project, Grant/Award Number: GJHZ20220913143008015; Natural Science Foundation of Guangdong Province, Grant/Award Number: 2023A1515011959; Shenzhen-Hong Kong Jointly Funded Project, Grant/Award Number: SGDX20230116091246007; Shenzhen Science and Technology Program, Grant/Award Numbers: RCBS20221008093131089, ZDSYS20210623091809029

Abstract

Real-world data always exhibit an imbalanced and long-tailed distribution, which leads to poor performance for neural network-based classification. Existing methods mainly tackle this problem by reweighting the loss function or rebalancing the classifier. However, one crucial aspect overlooked by previous research studies is the imbalanced feature space problem caused by the imbalanced angle distribution. In this paper, the authors shed light on the significance of the angle distribution in achieving a balanced feature space, which is essential for improving model performance under long-tailed distributions. Nevertheless, it is challenging to effectively balance both the classifier norms and angle distribution due to problems such as the low feature norm. To tackle these challenges, the authors first thoroughly analyse the classifier and feature space by decoupling the classification logits into three key components: classifier norm (i.e. the magnitude of the classifier vector), feature norm (i.e. the magnitude of the feature vector), and cosine similarity between the classifier vector and feature vector. In this way, the authors analyse the change of each component in the training process and reveal three critical problems that should be solved, that is, the imbalanced angle distribution, the lack of feature discrimination, and the low feature norm. Drawing from this analysis, the authors propose a novel loss function that incorporates hyperspherical uniformity, additive angular margin, and feature norm regularisation. Each component of the loss function addresses a specific problem and synergistically contributes to achieving a balanced classifier and feature space. The authors conduct extensive experiments on three popular benchmark datasets including CIFAR-10/100-LT, ImageNet-LT, and iNaturalist 2018. The experimental results demonstrate that the authors' loss function outperforms several previous state-of-the-art methods in addressing the challenges posed by imbalanced and long-tailed datasets, that is, by improving upon the best-performing baselines on CIFAR-100-LT by 1.34, 1.41, 1.41 and 1.33, respectively.

KEYWORDS

artificial intelligence, neural network

1 | INTRODUCTION

Deep learning has achieved significant progress in many tasks with large-scale datasets [1–5]. In traditional classification setting [6–8], the instance numbers of different classes are rebalanced artificially for facilitating evaluation. However, data of some classes are hard to collect in the real world, which makes the collected datasets generally exhibit a long-tailed distribution, where head classes occupy most of the data while tail classes have rarely few instances. Models trained on

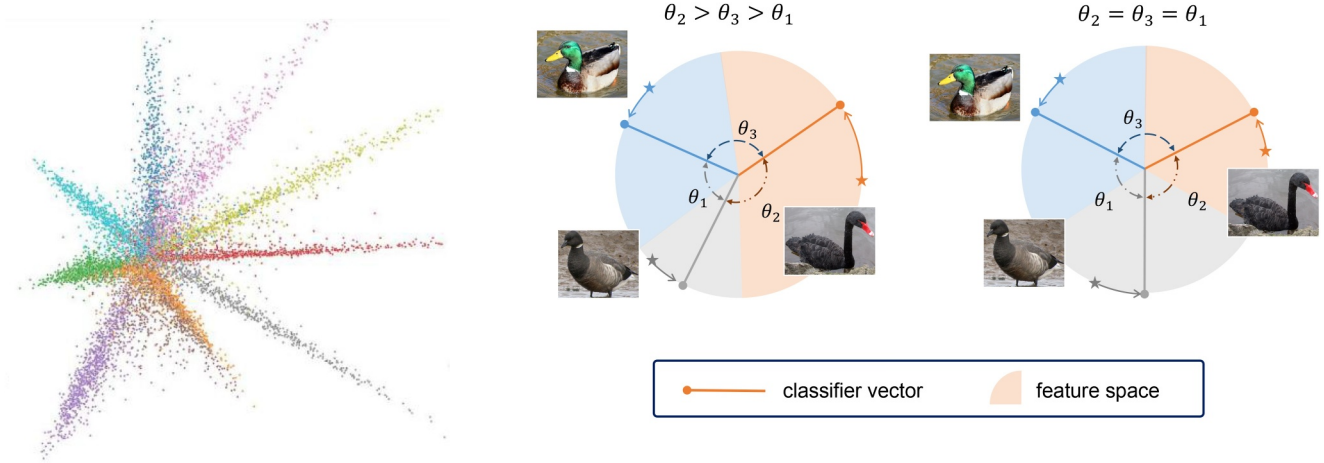
The first two authors contributed equally and are joint first authors.

imbalanced data are likely to be biased towards the head class and perform poorly on the tail classes, which degrades the overall performance. It is a great challenge to train an unbiased network under the data-biased scenario [9, 10].

To address this problem, many techniques have been proposed to re-balance the data distribution in order to be less dominated by the head classes. Early techniques [9, 11] on long-tailed classification focus on re-weighting and re-sampling the data. The core ideas are to rectify the classifier by assigning a larger weight to tail classes or sample more data from them. However, the re-balancing strategies damage the representative ability of the learnt deep features and raise the risk of over-fitting on tail class [12]. To improve the generalisation on tail classes, Cao et al. [13] proposed the label distribution aware margins (LDAM) from the perspective of generalisation error bound. Ref. [14] empirically shows that the imbalance of classifier weight norms caused by imbalanced data distribution degrades the model's performance. Then they proposed a two-stage method that trains the network with vanilla CE at the first stage and retrains the classifier with different classifier correction methods. Another work [15] analysed the class imbalance problem from the perspective of Bayes probabilistic and proposes a novel loss function Balanced Softmax. Menon et al. [16] analysed long-tailed distribution from the same perspective, and further explored the post-hoc logit adjustment and parameter for Balance Softmax loss function. A recent work [17] proposed LAbel distribution DisEntangling loss (LADE) to better involve the label disentanglement into the training stage. Besides working on loss function, other studies proposed data augmentation techniques [18–21] to improve the performance in data imbalance scenarios. Some recent works also investigated the effect of different backbones models for long-tail classification [22, 23].

Despite the notable achievements of previous studies, they mainly focus on the classifier rebalancing process while ignoring the issue of an imbalanced feature space learnt under long-tailed distributions, which can also negatively impact model generalisation. Figure 1a presents the feature space learnt by previous classifier rebalancing methods. We can find that although these methods can solve the imbalanced classifier norm problem, the learnt feature space is still highly imbalanced, that is, the feature space occupied by head classes is much larger than that of tail classes. Specifically, the space occupied by head classes significantly surpasses that of tail classes. Consequently, the features of instances from tail classes (i.e. orange and brown) become indistinguishable as they cluster together, while the features of instances from head classes (i.e. purple and green) maintain considerable distances between each other, making them easily distinguishable. This will largely hamper the model performance on the tail classes. To gain deeper insights into this problem, we offer an example in Figure 1b to elucidate why prior approaches struggle to achieve a balanced feature space. In the left part of Figure 1b, we observe that previous methods solely address the magnitudes of classifier vectors without balancing the angle distribution between them, ultimately leading to an imbalanced feature space. In contrast, the right part of Figure 1b shows that a balanced feature space can only be achieved when both the magnitudes of classifier vectors and the angles between them are balanced. Thus, it becomes imperative to tackle the issue of an imbalanced feature space for long-tailed classification. However, it is challenging to effectively balance the classifier norms and angle distribution at the same time due to problems such low feature norm, as we illustrated in Section 3.1.

To address these challenges, we conduct an in-depth analysis of the last fully connected layer in a classification model by decoupling the classification logits into three distinct components: classifier norm (magnitude of the classifier vector), feature norm (magnitude of the feature vector), and cosine similarity between the classifier and feature



(a) Imbalanced feature space learnt by weight normalization-based methods

(b) Imbalanced feature space caused by imbalanced angle distribution and a balanced feature space.

FIGURE 1 An illustration of imbalanced feature space problem.

vectors. In this way, we quantitatively validate and evaluate the imbalanced angle distribution problem. We further analyse the change of each component in the training process and reveal three critical problems that should be solved, that is, the imbalanced angle distribution, the lack of feature discrimination, and the low feature norm. Based on the analysis, we propose a new loss function to effectively regularise the classifier norms, angle distribution, and feature norms to achieve a balanced feature space. Specifically, to get a balanced angle distribution, we distribute the classifier vectors as uniformly as possible on a hypersphere by employing the maximisation of minimal pairwise angles regularisation. Then in the additive angular margin regularisation part, we propose to fix the classifier norms to balance the classifier and add the label-aware angular margin to improve the discrimination of feature space. Furthermore, to address the low feature norm problem caused by fixed classifier norms and prevent overfitting, we propose an additional regularisation term to penalise low feature norms. Eventually, we combine all the regularisation terms together to a unified training loss. We evaluate the effectiveness of our proposed loss function on three popular benchmark datasets including CIFAR-10/100-LT, ImageNet-LT, and iNaturalist 2018. The experimental results show that our loss function outperforms a couple of previous state-of-the-art methods.

In summary, the main contributions of this work are as follows:

1. We figure out the imbalanced angle distribution problem which is harmful to the learnt feature space from long-tailed distribution but ignored by previous work.
2. To obtain a more balanced and discriminative feature space, we propose a novel loss function that contains the hyperspherical uniformity, additive angular margin, and feature norm regularisation.
3. Extensive experiments show that our method can outperform other methods on long-tailed classification datasets including CIFAR-10/100-LT, ImageNet-LT, and iNaturalist 2018.

Paper structure. Section 2 illustrates the background knowledge and related work of this paper. Section 3 presents our proposed methodology for long-tail classification. Section 4 introduces the experimental setup and describes the evaluation results. Finally, Section 5 concludes the paper.

2 | RELATED WORK

2.1 | Long-tail classification

The traditional classification has the balanced distribution over all classes [24] but the long-tailed distribution problem will largely hurt the performance [25]. To tackle the long-tailed classification problem, prior methods for imbalance learning include re-sampling by over-sampling the minority or under-sampling the majority and re-weighting by applying a larger weight for minor categories. Compared with directly re-weighting by the number of different classes, ref. [26] considered the effective samples of each class and designed a class-balanced cross-entropy loss function. Cao et al. [13] proposed to introduce label distribution aware margin into loss function to improve the generalisation of tail class. With the observation that the re-weighting and re-sampling methods benefit the classifier but harm the representation learning, Kang et al. [14] proposed some classifier rebalance methods to learn the representation and classifier at two stages to avoid this problem in previous work. Another work [12] proposed to transit from representation learning stage to classifier learning stage with an annealing factor which is more smoothing. Refs. [15, 16] focused on the classifier revision part and proposed a better classifier rebalance methods which is the optimal solution from statistical perspective and ref. [17] further proposed to improve the generalisation of this method by regularising the Donsker–Varadhan representation. Apart from accuracy, a recent work [27] focused on the calibration problem and proposed label-aware smoothing to deal with miscalibration for models trained on long-tailed datasets. However, these methods mainly focused on the classifier rebalancing process while ignoring the imbalanced feature space learnt under long-tailed distribution, which is also harmful for model generalisation. Different from them, our method aims at rebalancing both the classifier and feature space and can achieve better performance.

Apart from these, many methods have been proposed to help further improve the performance by combining them with the above optimising objective. Many work [18, 20, 28] leverage data augmentation to help the network learn a better representation on tail class. Ref. [19] improved the mixup strategy under imbalanced distribution and proposed remix. Ref. [29] utilised self-supervised learning to further boost the performance of many loss functions. Recent approaches leverage causal inference [30], meta-learning [31], multi-experts [32] and contrastive learning [33] and achieve better performance. These methods are orthographic to our research and can be combined with our method to further boost the performance.

2.2 | Hyperspherical learning

Hyperspherical uniformity encourages vectors to be spaced apart with an angle as large as possible such that these vectors can be uniformly distributed over the hypersphere [34] which is equivalent to the Thompson problem [35] in physics. Inspired by this, Liu et al. [36] proposed a minimum hyperspherical energy (MHE) objective as generic regularisation for neural networks. Since this objective function is hard to optimise due to highly non-linear and non-convex optimisation as the space dimensionality becomes higher, ref. [37] proposed CoMHE which utilises projection mappings to reduce the dimensionality. Ref. [38] proposed hyperspherical prototype networks in which the prototype of each class is fixed in advance. Wang et al. [39] proposed maximising the minimal

pairwise angles (MMA) from the Tammes problem and achieved closer solutions to the optimal solutions.

3 | PROPOSED APPROACH

In this section, we start with an empirical study to analyse the classification layer of a traditional neural network. By conducting several experiments using the decoupling framework, we reveal some problems under the long-tail distribution. Inspired by the investigation, we then propose a set of strategies to regularise each component individually. Finally, we present a comparison with existing typical methods from the decision boundary perspective.

3.1 | Empirical study

To analyse why vanilla CE training suffers a large performance drop under long-tail distribution, we first revisit the classification layer of a traditional neural network. We decouple the logits into classifier norm, feature norm, and cosine similarity of classifier vector and feature vectors. We then propose a novel objective metric, *classification space*, to evaluate the quality of balance among the classifiers of trained neural networks. Through experiments and analysis, we reveal several problems under the decoupling framework. These problems will be proven to be critical in improving the long-tail classification performance, as addressed in the following sections.

3.1.1 | Classification space

The classification layer of a traditional neural network is given as the following equation:

$$p(y=i|x)=\frac{e^{\phi_i}}{\sum_{j=1}^{N}e^{\phi_j}} \tag{1}$$

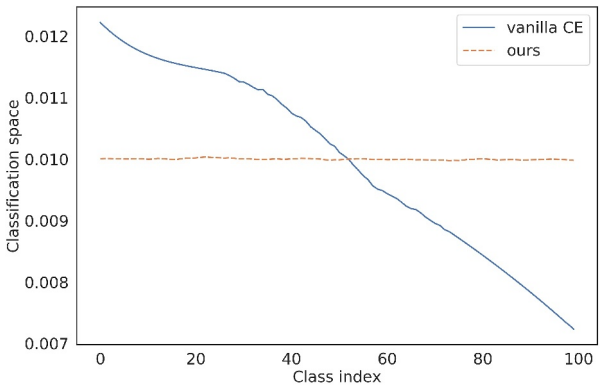
where $\phi_i = w_i x$. Following the decoupling framework [40], ϕ_i can also be expressed as $\phi_i = \|w_i\|\|x\|\cos(\theta_{w_i,x})$. The classification boundary of each pair of classes is determined by $\|w_i\|\cos(\theta_{w_i,x}) = \|w_j\|\cos(\theta_{w_j,x})$. On the one hand, the larger the classifier norm of one class, the larger the space it occupied; on the other hand, with equal norms, the occupied space of one class can be approximately quantified by the total pair-wise angular distance with the other classifiers. Therefore, we propose *classification space* (CS) as an objective metric to quantify the relative space occupied by each class such that,

$$CS_i = \|w_i\|\sum_{j=1}^{C}\arccos(\hat{w}_i\cdot\hat{w}_j) \tag{2}$$

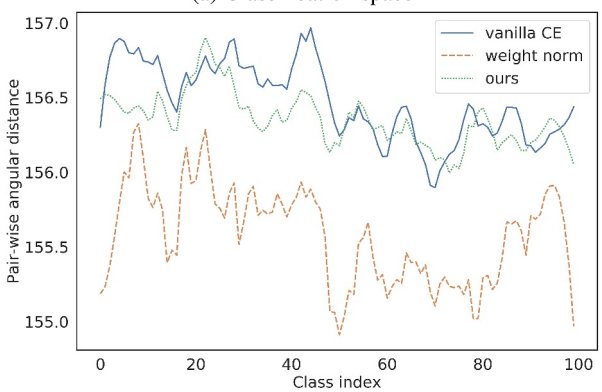
where C is the number of classes and $\hat{w}_i = \frac{w_i}{\|w_i\|}$ is the norm of each prototype.

3.1.2 | Investigation

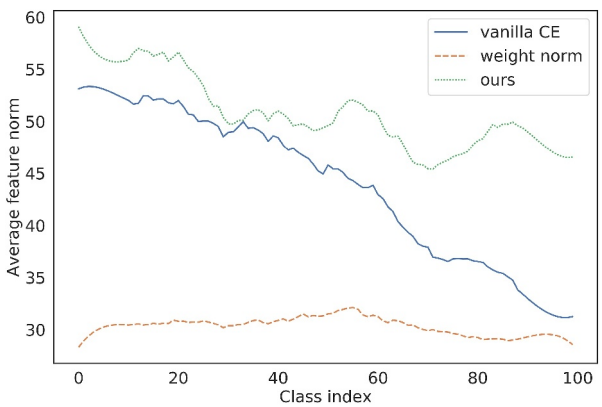
Equipped with *CS*, we visualise the classifiers resulting from a traditional cross-entropy training loss in Figure 2. As shown in Figure 2a, the classifiers exhibit a severe phenomenon of imbalance. A quick and attempted modification might be to fix all classifier norms to be a constant during training. In fact, the classifier normalisation has been proposed by ref. [14], but only



(a) Classification space



(b) Pair-wise angular distance



(c) Average feature norm

FIGURE 2 Comparison of CE, weight norm, and our proposed method on CIFAR100-100. Figure (a) shows the classification space per class for cross entropy and our method. Figure (b) shows the pair-wise angular distance per class for cross entropy, weight norm, and our method. Figure (c) shows the average feature norm of each class for cross entropy, weight norm, and our method.

in the inference stage to improve the result. One might be surprised by the observation that classifier normalisation during training leads to worse performance under the long-tail training setting, since classifier normalisation has been reported to accelerate and improve the training progress in a variety of other settings [40, 41]. In an attempt to shed light on this phenomenon, we investigate the change in the distribution of *CS* and feature norm across different classes, and the results are shown below.

Figure 2b displays the total pair-wise angular distance (the second term of Eq. 2) as a function of class index, with and without classifier normalisation. As we can see, the classifier normalisation results in a lower total pair-wise angular distance which indicates that the prototype of each class is closer than vanilla CE. This makes the classification procedure harder.

Figure 2c plots the average feature norm for each class, with and without classifier normalisation. Once again, the presence of classifier normalisation during training harms the training process from the perspective of feature norm. We can find that the average feature norm for each class is much lower and raise the risk of overfitting.

3.2 | Decoupling regularisation

As discussed above, training with long-tail data results in a set of imbalanced classifiers. To address these issues, we propose a set of strategies that aim to regularise each component individually and eventually combine them into a unified training loss. Specifically, the hyperspherical uniformity regularisation term and the normalisation of classification weight vector are proposed for a more balanced feature space. The additive angular margin part is to improve the discrimination of feature space. The feature norm regularisation is proposed to deal with the low feature norm problem observed from the weight norm before (Figure 2c) which can easily lead to overfitting. The rest of the strategies are described as follows.

3.2.1 | Hyperspherical uniformity

To achieve a balanced distribution of the total pairwise angular distance among each class, we propose to introduce a regularisation item that aims to uniformly distribute the normalised classifier vectors on a hypersphere. This is equivalent to a well-known problem in physics—Thomson problem [35] where one seeks to find a solution that distributes N electrons on a unit sphere as evenly as possible with minimum potential energy [36]. Several algorithms have been proposed for solving this problem in the literature [36, 37], with the latest and most successful method proposed by ref. [39]. This method treats the problem as a Tammes problem and tries to maximise the minimal pairwise angles (MMA). Formally, the MMA objective function can be expressed as the following equation:

$$L_{mma} = -\frac{1}{C}\sum_{i=1}^{C} \min_{j\neq i} \arccos\left(\hat{w}_i \cdot \hat{w}_j\right) \tag{3}$$

where $\hat{w}_i = \frac{w_i}{\|w_i\|}$ is the normalised classifier vector. Under a long-tail class distribution setting, it is beneficial to distribute the classifier vectors such that the tail classes have larger minimal pairwise angles, as shown in the ablation study 4.5. Therefore, we modify the MMA to become label frequency aware such that

$$L_{hu} = -\sum_{i=1}^{C} weight_i \cdot \min_{j\neq i} \arccos\left(\hat{w}_i \cdot \hat{w}_j\right) \tag{4}$$

$$weight_i = \frac{n_i^{-\frac{1}{k}}}{\sum_{j=1}^{C} n_j^{-\frac{1}{k}}} \tag{5}$$

where n_i is the number of training instances of class i and the parameter k is set to 4 by referencing previous work [13].

3.2.2 | Additive angular margin

Adding margins to the classifier boundary is known to be effective to improve the generalisation of neural networks. This regularisation strategy can be naturally integrated with the strategy of regularising the angles between the classifiers [13, 42–44]. A classical approach to account for the issue of long-tail distribution is to introduce a label-aware margin [43]. Inspired by the previous works, we propose a novel label-aware additive angular margin such that

$$p(y|x) = \frac{e^{\|x\|\cos\left(\theta_y + \theta_{W_y,x}\right)}}{e^{\|x\|\cos\left(\theta_y + \theta_{W_y,x}\right)} + \sum_{c\neq y} e^{\|x\|\cos\left(\theta_{W_c,x}\right)}} \tag{6}$$

$$L_{am} = -\frac{1}{N}\sum_{i=1}^{N} \log p\left(y_i|x_i\right) \tag{7}$$

where $\|x\|$ is the feature norm, $\theta_{W_c,x}$ is the angle between feature and classifier vectors, θ_y is the angular margin for class y and s is a learnable parameter. We set a larger margin for tail classes in the following way:

$$\theta_y = \frac{\pi}{m} \cdot n_y^{-\frac{1}{k}} \tag{8}$$

where n_y is the number of each class and m is a hyper-parameter to control the amount of margin. The parameter k is also set to 4 following previous work [13]. It is worth noting that the previous work of label-aware margin acts on the cosine similarity, which has been shown to suffer critical optimisation issues, while our regularisation strategy works directly on the angle and therefore facilitates the training process as discussed in ref. [43].

3.2.3 | Feature norm regularisation

As discussed above regarding feature norm, we propose to regularise it to be not too small by minimising,

$$L_{fr} = \frac{1}{N}\sum_{i}^{N} g(\|x_i\|) \tag{9}$$

where $g(x) = \frac{1}{x}$. As we can see, this term encourages larger feature norms for every instance to avoid a negative impact on accuracy. It is worth noting the feature normalisation has been shown to facilitate the training of neural networks in the literature [13, 43]. But our proposed feature norm regularisation is more flexible and can potentially bring additional benefits to the training of neural networks.

3.2.4 | Final loss function

Finally, the proposed loss function is defined in Eq. 10, which is a combination of hyperspherical uniformity regularisation, feature norm regularisation, and angular margin loss. λ_a and λ_b are a non-negative weighting coefficient:

$$Loss = L_{am} + \lambda_a L_{hu} + \lambda_b L_{fr} \tag{10}$$

Similar to LDAM, we also use the deferred re-balancing training strategy which first trains the network with standard loss function and re-weighting each class after annealing the learning rate.

3.3 | Comparison with other method

From the geometry perspective, we compare our loss function with the previous methods in Figure 3. From Figure 3a, we can see that τ Norm [14] improves the model performance on the balanced test set by allocating the classification space of head class to tail class at test time. The experimental results in ref. [14] also indicated that τ Norm performs well on tail class but sacrifices the performance on the head class. The same issue also exists in Balanced Softmax [15]; in Figure 3b, we can see that the inconsistent decision boundary at training and testing time makes the head class sample between the boundary misclassified. As for LDAM in Figure 3c, the margin loss can ensure the training boundary is contained in the testing boundary, but the cosine classifier makes the classification space of the tail class much lower than the head class. Our proposed loss function can solve the above problems and achieve better performance.

From the theoretical perspective, our method distinguish from previous work as it can better deal with the collapse problem [45, 46]. Neural collapse [45] is a phenomenon where the embeddings of each class collapse to their class mean. Based on this, Fang et al. [46] further investigated the imbalanced scenario and proposed the minority collapse, denoting

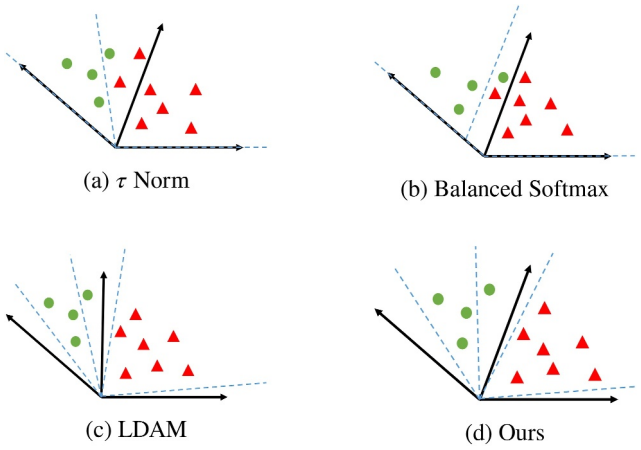


FIGURE 3 The classification boundary under different loss function. The red triangle and the green circle represent head class and tail class respectively. The blue dash and the black line represent the training and testing classification boundary.

that the classifier vectors of the minorities (i.e. tail classes) collapse to the same vector. Recent works have proved that dealing with this will lead to a better generalisation performance [46, 47]. Therefore, based on this, we can find that our proposed hyperspherical uniformity regularisation term is helpful for avoiding the collapse problem and leads to better generation performance.

4 | EXPERIMENTS

4.1 | Datasets

We evaluate the proposed method on four long-tailed datasets, including the CIFAR-10/100-LT [13], ImageNet-LT [48] and iNaturalist 2018 [49].

CIFAR-10/100-LT: CIFAR-10-LT and CIFAR-100-LT are both created by sampling from the original CIFAR dataset [50]. For a fair comparison, we do not create a new dataset from original CIFAR dataset and use the long-tailed versions of CIFAR datasets with the same setting as those used in ref. [13]. The data imbalance ratio is denoted by a factor β, $\beta = \frac{N_{\max}}{N_{\min}}$ where $N_{\max}$ and $N_{\min}$ are the numbers of training samples for the most and least frequent classes, respectively. We conduct experiments on four imbalance factors 10, 50, 100, and 200.

ImageNet-LT: ImageNet-LT is proposed in ref. [48]. ImageNet-LT is a long-tailed classification dataset by sampling a subset from ImageNet [51] following the Pareto distribution with power value $\alpha = 6$. It contains 115,846 training examples in 1000 classes, with class cardinality ranging from 5 to 1280.

iNaturalist 2018: The iNaturalist 2018 [49] is a large-scale real-world species classification dataset with severe long-tail distribution. The iNaturalist 2018 dataset contains 437,513 training images from 8142 classes, with an imbalance ratio of 500. For fair comparisons, we use the official training and validation splits in our experiments.

4.2 | Baseline

Since the method we proposed is a loss function, we mainly compare our model with state-of-the-art long-tail loss functions. Other methods such as refs. [18–20] can also be combined with our method and further boost the performance. Specifically, we mainly compare our method with focal loss [9], ClassBalance loss [26], C-SMOTE [52], LDAM [13], LDAM-DRW [13], decouple methods (LWS, τ Norm and cRT) [14], and Balanced Softmax [15]. Besides, we also compare with some advanced methods proposed in the past 3 years, including LADE [17], KCL [53], CDB [54] and MixGradient [55].

4.3 | Implementation details

For all the experiments over different datasets, we follow previous works [13, 56] and use the SGD optimiser with momentum $\gamma = 0.9$ to train the network. The initial learning rate is set to 0.2 and the weight decay is set to $2 \cdot 10^{-4}$ if not specific. All experiments are conducted with fixed random seed similar to ref. [56]. For CIFAR-10/100-LT, we use ResNet-32 as the backbone and use the multistep learning rate schedule, which decreases the learning rate by 0.01 at the 160^{th} and 180^{th} epoch. For ImageNet-LT, we mainly follow refs. [14, 56] and use ResNet-10 and ResNet-50 as backbones. ResNet and its variants are widely used in the literature [44, 57, 57]. We train the network with 90 epoch and follow ref. [56] to anneal the learning rate by 0.1 at epoch 60 and 80. For iNaturalist 2018, we use ResNet-50 as the backbone and train the network with 90 epochs. The learning rate decay is used at 60^{th}, 80^{th} epoch for 90 epochs. As for the hyper-parameters, we set λ_a and λ_b to 0.15 and 200, respectively. All the experiments are conducted on a server with 4 Nvidia Tesla V100 GPUs.

4.4 | Experimental results

We show our experimental results on the three datasets, respectively.

Results on CIFAR-10/100-LT: The results of four imbalance ratios on CIFAR-10/100-LT are shown in Table 1. The results are reproduced based on the code released by the authors. From the result, we can see that Balanced Softmax [15], LADE [17], LDAM-DRW [13], KCL [53] and cRT [14] achieve second-best results on different datasets and imbalance ratios. But our proposed method consistently achieves the best results on all datasets and imbalance ratios and outperforms the second-best results by at least 1 point in most settings. Specifically, on CIFAR-10, except CIFAR10-100 our method gets an obvious improvement compared with all baselines and achieves the highest improvement 1.33 points on CIFAR10-50. On CIFAR-100, compared with the strongest baseline Balanced Softmax, our method can improve it with at least 1.41 points and the highest improvement is 2.01 points on CIFAR100-100. These results demonstrate the effectiveness of our proposed regularisation methods. Besides, compared with KCL which also aims to create a more balanced feature space for long-tailed classification, our method consistently outperforms it on different settings. Specifically, on CIFAR-100, our method improves it by a large margin, for example, 2.48 points and 2.70 points on CIFAR100-10 and CIFAR-100-50, respectively. This shows that the proposed hyperspherical uniformity regularisation term is more effective in learning a balance and uniform feature space, especially when the number of classes become very large.

TABLE 1 Experimental results on CIFAR-10/100-LT.

Approach	CIFAR-10				CIFAR-100			
	im10	im50	im100	im200	im10	im50	im100	im200
CE	86.92	76.55	73.13	66.40	56.84	42.81	39.13	35.28
Focal	86.62	77.79	72.27	63.97	56.40	42.90	37.08	34.30
ClassBalance	86.90	78.13	72.68	68.77	57.57	44.79	38.77	35.56
C-SMOTE	81.50	75.02	70.17	64.89	50.52	43.69	39.44	35.23
LDAM	87.32	78.83	73.55	66.75	57.29	46.16	40.60	36.53
LDAM-DRW	87.32	81.03	77.03	71.94	57.40	47.29	<u>42.48</u>	37.52
Balanced softmax	<u>88.31</u>	80.74	77.96	72.35	58.33	<u>47.59</u>	41.88	38.09
LADE	87.98	80.82	<u>78.78</u>	72.65	57.82	46.80	41.56	37.93
LWS	87.41	77.77	70.04	65.74	58.08	45.40	41.57	37.52
τ norm	87.03	78.70	72.82	70.21	57.48	45.40	42.40	37.43
cRT	87.90	79.56	73.02	69.48	57.86	45.37	41.57	<u>38.16</u>
KCL	88.00	<u>81.70</u>	77.60	–	57.60	46.30	42.80	–
CDB	–	–	–	–	<u>58.74</u>	46.82	42.59	37.72
MixGradient	84.40	79.38	75.52	<u>72.74</u>	53.57	47.43	42.46	38.14
Ours	**88.94**	**82.36**	**78.87**	**73.54**	**60.08**	**49.00**	**43.89**	**39.49**

Note: Best and second-best results are marked in bold and underline, respectively.

Results on ImageNet-LT: The experimental comparison to baselines on ImageNet-LT is presented in Table 2. We compare our method with many strong baselines. The results show that our method can achieve good performance with both ResNet-10 and ResNet-50 as backbones. Specifically, our method yields 42.5% top-1 overall accuracy with ResNet-10 as the backbone, which improves the strongest baseline Balanced Softmax (combined with meta sampler) by 0.7 points. This indicates that our proposed loss function can help weak backbone get much improvement and even surpass the strong backbone model. Besides, our method can also achieve good performance when using ResNet-50 as the backbone. These results show that our proposed method is also effective on large-scale datasets.

Results on iNaturalist 2018: Table 3 presents the experimental results on the naturally long-tailed datasets iNaturalist 2018. We conduct experiments for 90 epochs. Our method reaches the best accuracy of 68.1% compared with

TABLE 2 Experimental results on ImageNet-LT.

Approach	ResNet-10	ResNet-50
CE	36.5	41.7
OLTR	35.6	41.9
Remix	37.6	46.2
τ_norm	40.6	46.7
cRT	41.8	47.3
LWS	41.4	47.7
LDAM-DRW	40.7	48.8
Balanced Softmax	41.8*	**50.0***
Ours	**42.5**	48.9

Note: The bold figures indicates the best results.
* denotes combined with meta sampler.

TABLE 3 Experimental results on iNaturalist 2018.

Approach	ResNet-50
CE	61.7
τ_norm	65.6
cRT	65.2
LWS	65.9
LDAM-DRW	66.1
BBN	66.3
Balanced Softmax	66.4
Ours	**68.1**

Note: The bold figures indicates the best results.

previous methods. Specifically, compared with the strongest baseline Balanced Softmax, our method significantly outperforms it by 1.7 points. Compared with other classification boundary adjustment methods including τ Norm, LDAM-DRW, and Balanced Softmax, our method improves them by a large margin, which demonstrates that the classification boundary obtained by our method has better generalisation for the test set. This also validate the analysis in Section 3.3.

4.5 | Further evaluation

Ablation study: To verify the effectiveness of each component in our method, we also conduct an ablation study. As shown in Table 4, removing the margin will cause more than 1 point performance drop on three imbalance ratios, which indicates that the generalisation improvement brought by margins is essential for long-tail classification. We can also find that our model suffers from an obvious performance drop after eliminating L_{hu} and L_{fr} . Specifically, when removing L_{hu}, the performance drop 2.19 points on CIFAR100-100. As for L_{fr}, the performance of our method also drops significantly on the three imbalance ratios. This suggests that the model can learn a better feature space by regularising the shortcoming from training with weight normalisation.

TABLE 4 Ablation study on CIFAR-100.

Imbalance factor	50	100	200
Ours	**49.00**	**43.89**	**60.08**
Ours w/o margin	47.42	42.43	58.82
Ours w/o L_{hu}	48.48	41.70	59.46
Ours w/o L_{fr}	47.04	41.51	59.20

Note: The bold figures indicates the best results.

Evaluation on the classification space: As we discussed in our empirical study, we use classification space, pair-wise angular distance, and the average feature norm to evaluate the balanced degree of classifier and feature space. Thus, we also conduct experiments to validate whether our proposed method achieves the purpose above. As shown in Figure 2a, our method can clearly give a more balanced classification space over all classes. As for the pair-wise angular distance, we can see that our method which is combined with weight normalisation and hyperspherical uniformity can avoid the problem and reach a higher pair-wise angular distance. This property can help improve the generalisation of neural networks. As for the average feature norm, we can also see that our method can generate a more balanced and larger feature norm for each class.

Visualisation of feature space: We also visualise the feature space of our method and each component of our method in Figure 4 to show how each component helps long-tail classification. The visualisation is conducted by adding another full-connected layer (64×2) before the classifier. In vanilla CE, we can find that the feature norm of tail classes (red and orange) is much lower than other classes which makes it hard to be classified. When using weight normalisation, the feature norm of each class is more balanced but we can also find that the angle between some classes (orange and brown) is very close. This hinders the model's ability to distinguish them. After adding L_{hu} regularisation, we can find that the angle between different classes is more balanced. Eventually, adding margin helps the feature of different classes distributed away from each other. This helps the long-tail classification.

Parameter analysis: In the section, we study the effect of the parameters on the performance of the proposed method in our loss function. We analyse two main parameters λ_a and λ_b in Equ. 10 that could greatly affect the balance of different losses. The analysis results are shown in Figure 5. As can be seen, with the growth of λ_a and λ_b, the model performance first increases to the peak and suffers from an obvious degradation with further improvement. This shows the importance of balancing the contribution of three regularisation terms. Besides, we can also find that our method achieves the highest accuracy when $\lambda_a = 0.15$ and $\lambda_b = 200$. So we set them to these values in our experiments.

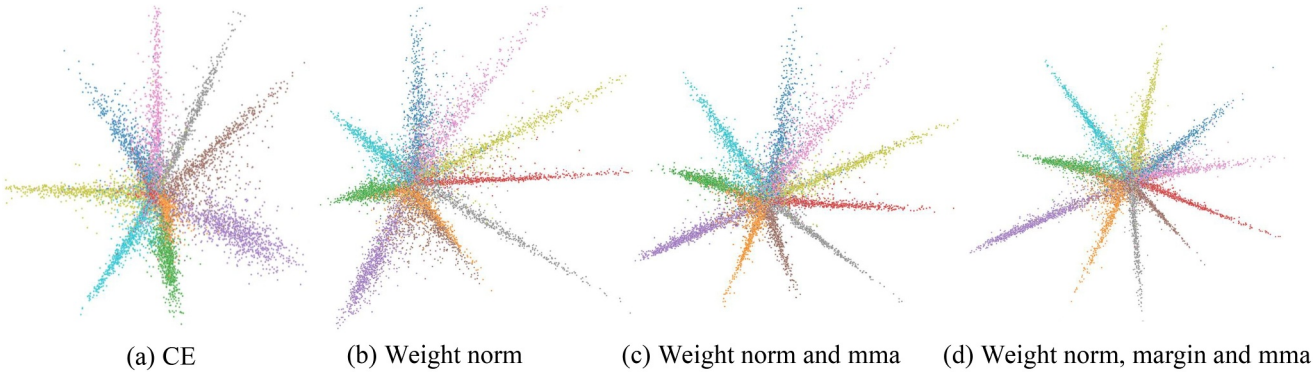


FIGURE 4 Feature space visualisation of different methods on CIFAR10-50 test set.

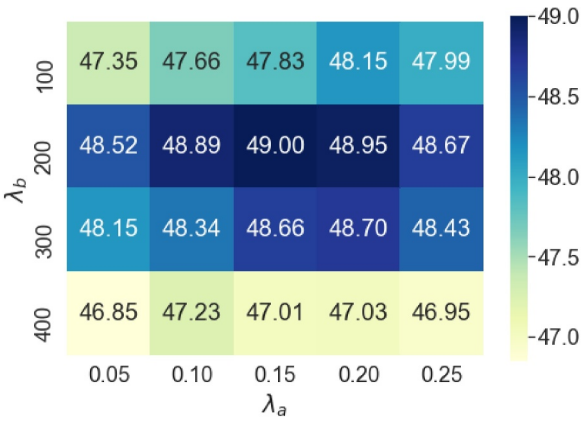


FIGURE 5 Analysis of two parameters λ_a and λ_b on the performance of our loss function. Heat map visualisation is based on CIFAR100-50.

5 | CONCLUSION

In this paper, we conducted a thorough analysis of the classifier and feature space under long-tail distribution with a decoupling approach. With the analysis, we developed a set of strategies that aim to regularise each component individually and eventually combined them into a unified training loss. Extensive experiments show that our proposed method outperforms other methods on several long-tailed classification benchmarks.

ACKNOWLEDGEMENTS

This research is supported National Key R&D Program of China (No. 2022YFB3103900), Natural Science Foundation of Guangdong Province (Project No. 2023A1515011959), Shenzhen-Hong Kong Jointly Funded Project (Category A, No. SGDX20230116091246007), Shenzhen Basic Research (General Project No. JCYJ20220531095214031), Shenzhen International Science and Technology Cooperation Project (No. GJHZ20220913143008015), the Major Key Project of PCL (Grant No. PCL2022A03), National Key Research and Development Program of China under Grant 2023YFB3106504, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant 2022B1212010005, the Major Key Project of PCL under Grant PCL2023A09, Shenzhen Science and Technology Program under Grant ZDSYS 20210623091809029 and RCBS20221008093131089.

CONFLICT OF INTEREST STATEMENT

The authors certify that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

We use public datasets in our paper. All datasets are available at https://github.com/zhangyongshun/BagofTricks-LT.

ORCID

Shuzheng Gao https://orcid.org/0000-0002-8102-480X
Qing Liao https://orcid.org/0000-0003-1012-5301

REFERENCES

1. Tadepalli, Y., et al.: Content-based image retrieval using Gaussian-hermite moments and firefly and grey wolf optimization. CAAI Trans. Intell. Technol. 6(2), 135–146 (2021). https://doi.org/10.1049/cit2.12040
2. Onan, A.: Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. Journal of King Saud University-Computer and Information Sciences 34(5), 2098–2117 (2022). https://doi.org/10.1016/j.jksuci.2022.02.025
3. Yang, J., et al.: A two-branch network with pyramid-based local and spatial attention global feature learning for vehicle re-identification. CAAI Trans. Intell. Technol. 6(1), 46–54 (2021). https://doi.org/10.1049/cit2.12001
4. Gao, S., et al.: Code structure-guided transformer for source code summarization. ACM Trans. Software Eng. Methodol. 32(23), 1–23:32 (2023). https://doi.org/10.1145/3522674
5. Luo, F., et al.: Multiscale diff-changed feature fusion network for hyperspectral image change detection. IEEE Trans. Geosci. Rem. Sens. 61, 1–13 (2023). https://doi.org/10.1109/tgrs.2023.3241097
6. Onan, A.: An ensemble scheme based on language function analysis and feature engineering for text genre classification. J. Inf. Sci. 44(1), 28–47 (2018). https://doi.org/10.1177/0165551516677911
7. Xing, Y., Zhu, J.: Deep learning-based action recognition with 3d skeleton: a survey. CAAI Trans. Intell. Technol. 6(1), 80–92 (2021). https://doi.org/10.1049/cit2.12014
8. Duan, Y., et al.: Classification via structure-preserved hypergraph convolution network for hyperspectral image. IEEE Trans. Geosci. Rem. Sens. 61, 1–13 (2023). https://doi.org/10.1109/tgrs.2023.3258977
9. Lin, T., et al.: Focal loss for dense object detection. In: ICCV, pp. 2999–3007. IEEE Computer Society (2017)
10. Thabtah, F.A., et al.: Data imbalance in classification: experimental evaluation. Inf. Sci. 513, 429–441 (2020). https://doi.org/10.1016/j.ins.2019.11.004

11. Drumnond, C., Holte, R.: Class imbalance and cost sensitivity: why undersampling beats oversampling. In: ICML-KDD 2003 Workshop: Learning from Imbalanced Datasets (2003)
12. Zhou, B., et al.: BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In: CVPR, pp. 9716–9725. Computer Vision Foundation / IEEE (2020)
13. Cao, K., et al.: Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss, pp. 1565–1576. NeurIPS (2019)
14. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. In: ICLR. OpenReview.net (2020)
15. Ren, J., et al.: Balanced Meta-Softmax for Long-Tailed Visual Recognition. NeurIPS (2020)
16. Menon, A.K., et al.: Long-tail learning via logit adjustment. In: ICLR. OpenReview.net (2021)
17. Hong, Y., et al.: Disentangling label distribution for long-tailed visual recognition. In: CVPR, pp. 6626–6636. Computer Vision Foundation / IEEE (2021)
18. Kim, J., Jeong, J., Shin, J.: M2m: imbalanced classification via major-to-minor translation. In: CVPR, pp. 13893–13902. Computer Vision Foundation / IEEE (2020)
19. Chou, H., et al.: Remix: rebalanced mixup. In: ECCV Workshops (6), pp. 95–110. Springer (2020)
20. andf Kaixiong Gong, S.L., et al.: Metasaug: meta semantic augmentation for long-tailed visual recognition(2021)
21. Wang, C., et al.: Label-aware distribution calibration for long-tailed classification. CoRR abs/2111, 04901 (2021)
22. Xu, Z., et al.: Learning imbalanced data with vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, pp. 15793–15803 (2023)
23. Long, A., et al.: Retrieval augmented classification for long-tail visual recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, pp. 6949–6959 (2022)
24. Onan, A., Korukoğlu, S., Bulut, H.: A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. Inf. Process. Manag. 53(4), 814–833 (2017). https://doi.org/10.1016/j.ipm.2017.02.008
25. Onan, A., et al.: Consensus Clustering-Based Undersampling Approach to Imbalanced Learning. Scientific Programming (2019). 2019
26. Cui, Y., et al.: Class-balanced loss based on effective number of samples. In: CVPR, pp. 9268–9277. Computer Vision Foundation / IEEE (2019)
27. Zhong, Z., et al.: Improving calibration for long-tailed recognition. In: CVPR, pp. 16489–16498. Computer Vision Foundation / IEEE (2021)
28. Chu, P., et al.: Feature space augmentation for long-tailed data. ECCV 29, 694–710 (2020). Springer. https://doi.org/10.1007/978-3-030-58526-6_41
29. Yang, Y., Xu, Z.: Rethinking the Value of Labels for Improving Class-Imbalanced Learning. NeurIPS (2020)
30. Tang, K., Huang, J., Zhang, H.: Long-tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. NeurIPS (2020)
31. Jamal, M.A., et al.: Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In: CVPR, pp. 7607–7616. Computer Vision Foundation / IEEE (2020)
32. Wang, X., et al.: Long-tailed recognition by routing diverse distribution-aware experts. In: ICLR. OpenReview.net (2021)
33. Wang, P., et al.: Contrastive learning based hybrid networks for long-tailed image classification. In: CVPR, pp. 943–952. Computer Vision Foundation / IEEE (2021)
34. Liu, W., et al.: Learning with hyperspherical uniformity. In: AISTATS, pp. 1180–1188. PMLR (2021)
35. Thomson, J.J.: Xxiv. on the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. London, Edinburgh Dublin Phil. Mag. J. Sci. 7(39), 237–265 (1904). https://doi.org/10.1080/14786440409463107
36. Liu, W., et al.: Learning towards minimum hyperspherical energy. NeurIPS, 6225–6236 (2018)
37. Lin, R., et al.: Regularizing neural networks via minimizing hyperspherical energy. In: CVPR, pp. 6916–6925. Computer Vision Foundation / IEEE (2020)
38. Mettes, P., Vander Pol, E., Snoek, C.: Hyperspherical Prototype Networks, pp. 1485–1495. NeurIPS (2019)
39. Wang, Z., et al.: MMA Regularization: Decorrelating Weights of Neural Networks by Maximizing the Minimal Angles. NeurIPS (2020)
40. Liu, W., et al.: Decoupled networks. In: CVPR, pp. 2771–2779. Computer Vision Foundation / IEEE Computer Society (2018)
41. Liu, W., et al.: Sphereface: deep hypersphere embedding for face recognition. In: CVPR, pp. 6738–6746. IEEE Computer Society (2017)
42. Liu, W., et al.: Large-margin softmax loss for convolutional neural networks. In: ICML, pp. 507–516. JMLR.org (2016)
43. Deng, J., et al.: Arcface: additive angular margin loss for deep face recognition. In: CVPR, pp. 4690–4699. Computer Vision Foundation / IEEE (2019)
44. Xiao, J., et al.: Learning discriminative representation with global and fine-grained features for cross-view gait recognition. CAAI Trans. Intell. Technol. 7(2), 187–199 (2022). https://doi.org/10.1049/cit2.12051
45. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proc. Natl. Acad. Sci. USA 117(40), 24652–24663 (2020). https://doi.org/10.1073/pnas.2015509117
46. Fang, C., et al.: Exploring deep neural networks via layer-peeled model: minority collapse in imbalanced training. Proc. Natl. Acad. Sci. USA 118(43), e2103091118 (2021). https://doi.org/10.1073/pnas.2103091118
47. Thrampoulidis, C., et al.: Imbalance trouble: revisiting neural-collapse geometry. NeurIPS (2022)
48. Liu, Z., et al.: Large-scale long-tailed recognition in an open world. In: CVPR, pp. 2537–2546. Computer Vision Foundation / IEEE (2019)
49. Horn, G.V., et al.: The inaturalist species classification and detection dataset. In: CVPR, pp. 8769–8778. Computer Vision Foundation / IEEE Computer Society (2018)
50. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images(2009)
51. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
52. Bernardo, A., Valle, E.D.: An extensive study of c-smote, a continuous synthetic minority oversampling technique for evolving data streams. Expert Syst. Appl. 196, 116630 (2022)
53. Kang, B., et al.: Exploring balanced feature spaces for representation learning. In: ICLR. OpenReview.net (2021)
54. Sinha, S., Ohashi, H., Nakamura, K.: Class-difficulty based methods for long-tailed visual recognition. IJCV 130(10), 2517–2531 (2022). https://doi.org/10.1007/s11263-022-01643-3
55. Peng, X., Wang, F., Li, L.: Mixgradient: a gradient-based re-weighting scheme with mixup for imbalanced data streams. Neural Network. 161, 525–534 (2023). https://doi.org/10.1016/j.neunet.2023.02.017
56. Zhang, Y., et al.: Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In: AAAI, pp. 3447–3455. AAAI Press (2021)
57. Jafarbigloo, S.K., Danyali, H.: Nuclear atypia grading in breast cancer histopathological images based on CNN feature extraction and LSTM classification. CAAI Trans. Intell. Technol. 6, 426–439 (2021)

How to cite this article: Gao, S., et al.: Improving long-tail classification via decoupling and regularisation. CAAI Trans. Intell. Technol. 1–10 (2024). https://doi.org/10.1049/cit2.12374