BCE3S: Binary Cross-Entropy Based Tripartite Synergistic Learning for Long-tailed Recognition

Weijia Fan[1, 2, 3, 4], **Qiufu Li**[2, 3*], **Jiajun Wen**[1, 3], **Xiaoyang Peng**[5]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]School of Artificial Intelligence, Shenzhen University
[3]Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University
[4]CV:HCI Lab, Karlsruhe Institute of Technology
[5]School of Software Engineering, Sun Yat-sen University
weijia.fan@partner.kit.edu, liqiufu@szu.edu.cn, wenjiajun@szu.edu.cn, pengxy77@mail2.sysu.edu.cn

Abstract

For long-tailed recognition (LTR) tasks, high intra-class compactness and inter-class separability in both head and tail classes, as well as balanced separability among all the classifier vectors, are preferred. The existing LTR methods based on cross-entropy (CE) loss not only struggle to learn features with desirable properties but also couple imbalanced classifier vectors in the denominator of its Softmax, amplifying the imbalance effects in LTR. In this paper, for the LTR, we propose a binary cross-entropy (BCE)-based tripartite synergistic learning, termed BCE3S, which consists of three components: (1) BCE-based joint learning optimizes both the classifier and sample features, which achieves better compactness and separability among features than the CE-based joint learning, by decoupling the metrics between feature and the imbalanced classifier vectors in multiple Sigmoid; (2) BCE-based contrastive learning further improves the intra-class compactness of features; (3) BCE-based uniform learning balances the separability among classifier vectors and interactively enhances the feature properties by combining with the joint learning. The extensive experiments show that the LTR model trained by BCE3S not only achieves higher compactness and separability among sample features, but also balances the classifier's separability, achieving SOTA performance on various long-tailed datasets such as CIFAR10-LT, CIFAR100-LT, ImageNet-LT, and iNaturalist2018.

Code — https://github.com/wakinghours-github/BCE3S

Introduction

Though deep models trained on balanced datasets can surpass humans in visual recognition, their performance is still unsatisfactory on the imbalanced datasets. In the long-tailed datasets, the sample number decline significantly from the head to tail classes, which are more aligned with the data distribution in the real world, and the training of deep models on such datasets is easily dominated by the head classes, leading to imbalanced features and classifiers from the head to the tail classes, which consequently decreases the overall performance of the final models (Kang et al. 2020; Alshammari et al. 2022; Xuan and Zhang 2024).

*Corresponding author.

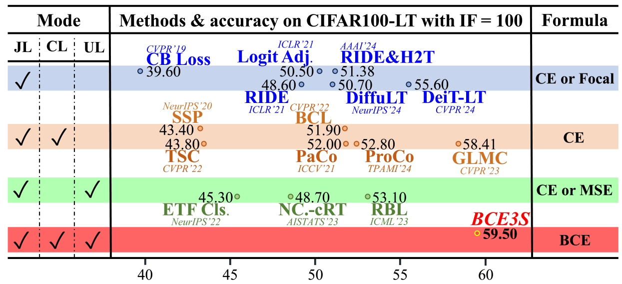


Table 1: Comparison of BCE3S with previous LTR methods in terms of learning mode and performance. "JL", "CL", and "UL" are short for joint learning between sample features and classifier vectors, contrastive learning among features, and classifier's uniform separability learning. Our Only BCE3S integrates all the three learning modes and achieves the highest accuracy (i.e., 59.50%) on CIFAR100-LT with IF = 100. More results can be found in Table 3.

In model training with cross-entropy (CE) loss, the classifier and sample features are jointly learned, while, on the long-tailed datasets, this learning mode is likely to fail to learn both well sample features and classifier. Currently, the re-balancing techniques such as re-sampling (Wallace et al. 2011), re-weighting (Lin et al. 2017; Cui et al. 2019; Alshammari et al. 2022), and logit adjustment (Menon et al. 2021; Zhao et al. 2022; Li, Cheung, and Lu 2022) have been developed and applied to improve the CE loss to balance the models' attention among different classes during the training. Combined with the joint learning, contrastive learning (Chen et al. 2020; He et al. 2020) has also been used to enhance the feature properties and improve LTR (Cui et al. 2021; Kang et al. 2021; Du et al. 2023; Xuan and Zhang 2024), by pulling the features of the same class closer and pushing those from different classes apart.

In LTR, in addition to enhancing sample features, it is also necessary to improve the discriminative capability of the classifiers learned on long-tailed datasets. Currently, there is no published work directly focusing on learning a well classifier for LTR. However, according to neural collapse (Papyan, Han, and Donoho 2020; Zhu et al. 2021), the optimal classifier that CE can learn is an equiangular tight frame (ETF) classifier where any two classifier vectors exhibit uniform and maximum separability. Then, customized ETF classifiers (Yang et al. 2022; Kasarla et al. 2022; Liu et al. 2023) are designed before the training, while they may not align well with the final learned features.

To improve LTR, a reasonable approach is to design a uniform separability learning framework for the classifier vectors and integrate it with joint learning and contrastive learning, creating a tripartite synergistic learning (TSL) paradigm to rebalance the features and classifiers during the model training. However, in the LTR tasks, the commonly used CE loss couples imbalanced metrics from different classes on its Softmax's denominator, making it challenging to effectively suppress the imbalance effect by combining the various CE-based learning modes. As BCE (binary CE) decouples the metrics from different classes by adopting multiple Sigmoid, it can flexibly adjust these metrics and effectively suppress the imbalance effect. In fact, Cui et al. (2019) have experimentally demonstrated that BCE (i.e., Sigmoid cross-entropy in the reference) achieves better LTR than CE, yet the potential of BCE in LTR has not been fully explored.

Therefore, in this paper, we design a BCE-based tripartite synergistic learning approach for LTR, termed BCE3S, which integrates: (1) BCE-based joint learning between sample features and classifiers, (2) BCE-based contrastive learning among features, and (3) BCE-based uniform separability learning for the classifier vectors. As Table 1 shows, with these three learning modes working in concert, BCE3S achieves the best LTR performance on CIFAR100-LT.

The main contributions of this paper are as follows:

1. We introduce the concept of tripartite synergistic learning (TSL) and propose BCE3S, a BCE-based LTR method integrating the joint learning between features and classifier, contrastive learning among features, and uniform learning for classifier vectors.
2. We explain in-depth that BCE-based uniform separability learning helps to train balanced classifier vectors, and analyze the advantage of BCE over CE in LTR.
3. We conduct extensive experiments to evaluate the performance of BCE3S on LTR and find that compared to CE-based methods, BCE3S achieves better intra-class compactness and inter-class separability of sample features, and the BCE-based uniform learning can effectively balance the classifier's separability. BCE3S achieves SOTA LTR results on four long-tailed datasets.

Related Work

Joint learning for both feature and classifier

Various techniques have been proposed to improve the LTR performance of the popular CE-based joint learning.

The re-sampling technique (Wallace et al. 2011) adjusts the sample distributions via over-sampling the tail classes or under-sampling the head classes to mitigate the class imbalance of the long-tailed datasets, which complicates the model training. The class-balanced loss (Cui et al. 2019) introduces a re-weighting factor inversely proportional to the sample numbers, which is adaptive to the CE, BCE, and focal losses based on joint learning. In (Kang et al. 2020), the authors decoupled the learning of sample features and classifier, and they rectified the recognition decision boundaries on the head and tail classes by fine-tuning with different rebalancing strategies, including the classifier re-weighting. DisAlign (Zhang et al. 2021) developed an adaptive calibration function to re-weight the learning of classifier vectors. The methods of BaLMS (Ren et al. 2020) and Logit Adj. (Menon et al. 2021) re-balance the model training by adjusting the logits according to the sample numbers. Besides these techniques, distillation (Rangwani et al. 2024), diffusion (Shao et al. 2024), and collaborative learning (Xu et al. 2023a) have also enhanced the LTR performance.

Existing LTR work primarily uses CE loss. However, research (Cui et al. 2019; Wightman, Touvron, and Jegou 2021; Touvron, Cord, and Jégou 2022) indicates that BCE loss also performs well in image recognition and shows potential for LTR (Cui et al. 2019), exemplified by LiVT (Xu et al. 2023b). Despite this, the advantages of BCE over CE in LTR remain underexplored and unexplained in detail.

Contrastive learning on sample features

Under the long-tailed scenario, PaCo (Cui et al. 2021) incorporates a set of learnable class centers into contrastive learning, re-balancing the feature learning from a perspective of optimization. KCL (Kang et al. 2021) creates balanced positive pairs by using the same number of samples per class. SSD (Li, Wang, and Wu 2021) employs a self-distillation framework to enhance the information exploitation in the tail class. BCL (Zhu et al. 2022) introduces class-averaging and class-complement to balance the gradient contribution in the contrastive learning. GLMC (Du et al. 2023) and DSCL (Xuan and Zhang 2024) mitigate head class bias using re-weighting loss and decoupled contrastive loss. ProCo (Du et al. 2024) models the feature space using a von Mises-Fisher distribution to eliminate the limitation of contrastive learning in requiring a large amount of samples in LTR tasks. These contrastive learning methods are designed on Softmax, measuring the relative value of feature similarity and amplifying the imbalance effect in LTR.

Uniform separability learning for classifier

The works of neural collapse (Papyan, Han, and Donoho 2020; Zhu et al. 2021) show that when training a recognition model on a balanced dataset and the CE loss reaches its minimum, the classifier vectors form an equiangular tight frame (ETF), where uniform and maximum separability exists between any two classifier vectors. However, on an imbalanced dataset, if the imbalance factor (IF) is too large, the classifier vectors of the tail classes will converge and collapse to the same vector (Fang et al. 2021), losing their separability.

To keep the classifier's uniform separability, Yang et al. (2022), Liu et al. (2023), and Kasarla et al. (2022) customize ETF classifiers before the LTR model training, which do not participate in the training and maintain the uniform separability; RBL (Peifeng et al. 2023) adopts a learnable orthogonal matrix to rotate the ETF classifier, adjusting its direction based on the sample features. However, these customized ETF classifiers struggle to align with the learned features. In this paper, we propose BCE-based uniform learning, helping to learn balanced classifiers aligned with features.

Method

Preliminary

Let $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$ be a dataset from K classes, where $\mathcal{D}_k$ contains the samples from the k-th class, and $n_k = |\mathcal{D}_k|$ denotes its sample number. Suppose $\mathcal{D}$ is an imbalanced and long-tailed dataset, and its K subsets have been sorted by the sample numbers in descending order, i.e., $n_k \geq n_\ell$ for $\forall k < \ell$, then n_1/n_K denotes the imbalance factor (IF).

In recognition task, for any sample $\boldsymbol{X} \in \mathcal{D}$, an deep model $\mathcal{M}(\cdot)$ first extracts its feature $\boldsymbol{x} = \mathcal{M}(\boldsymbol{X}) \in \mathbb{R}^d$, then, a linear classifier $\mathcal{C} = \{(\boldsymbol{w}_j, b_j)\}_{j=1}^{K}$ converts it into K metrics $\{\boldsymbol{w}_j^T \boldsymbol{x} + b_j\}_{j=1}^{K}$, which are applied to predict the sample's label, $\hat{k} = \arg\max_j \{\boldsymbol{w}_j^T \boldsymbol{x} + b_j\}_{j=1}^{K}$.

In training, for a sample from the k-th subset $\mathcal{D}_k, \forall k$, we denote it as $\boldsymbol{X}^{(k)}$ and its feature as $\boldsymbol{x}^{(k)} = \mathcal{M}(\boldsymbol{X}^{(k)})$. In the previous works (Ren et al. 2020; Li, Cheung, and Lu 2022; Alshammari et al. 2022; Du et al. 2023; Xuan and Zhang 2024; Wang et al. 2021; Chen et al. 2023), to learn the better model $\mathcal{M}$ and classifier $\mathcal{C} = \{(\boldsymbol{w}_j, b_j)\}_{j=1}^{K}$, Softmax was first applied to compute the probabilities $\Big\{\frac{\exp(\boldsymbol{w}_j^T \boldsymbol{x}^{(k)} + b_j)}{\sum_{\ell=1}^{K} \exp(\boldsymbol{w}_\ell^T \boldsymbol{x}^{(k)} + b_\ell)}\Big\}_{j=1}^{K}$ that $\boldsymbol{X}^{(k)}$ belongs to each class, then, cross-entropy was used to compute the CE loss,

$$L_{\text{ce}}^{(\text{sc})}(\boldsymbol{x}^{(k)}) = -\log\Big(\frac{\exp\big(\boldsymbol{w}_k^T \boldsymbol{x}^{(k)} + b_k\big)}{\sum_{\ell=1}^{K} \exp\big(\boldsymbol{w}_\ell^T \boldsymbol{x}^{(k)} + b_\ell\big)}\Big). \quad (1)$$

When applying CE loss to train models, the classifier vectors $\{\boldsymbol{w}_k\}$ and the sample features $\{\boldsymbol{x}^{(k)}\}$ are jointly trained and learned. The re-balancing techniques, such as re-sampling and re-weighting, have been used to improve CE loss and enhance its performance on long-tailed recognition (LTR).

Contrastive learning based on Softmax has also been introduced into LTR, comparing features in a projection space,

$$L_{\text{ce}}^{(\text{ss})}(\boldsymbol{x}^{(k)}) = -\log\Big(\frac{\exp\big(\frac{1}{\tau}\cos(\boldsymbol{z}^{(k)}, \boldsymbol{z}_*^{(k)})\big)}{\sum_{\ell=1}^{N} \exp\big(\frac{1}{\tau}\cos(\boldsymbol{z}^{(k)}, \boldsymbol{z}_*^{(\ell)})\big)}\Big), \quad (2)$$

where $\boldsymbol{z}^{(k)} = \mathcal{P}(\boldsymbol{x}^{(k)}), \boldsymbol{z}_*^{(k)} = \mathcal{P}(\boldsymbol{x}_*^{(k)})$, and $\mathcal{P}$ is a non-linear projection operator (Chen et al. 2020). In Eq. (2), τ is a temperature factor, $\cos(\boldsymbol{z}, \boldsymbol{z}_*) = \frac{\langle \boldsymbol{z}, \boldsymbol{z}_* \rangle}{\|\boldsymbol{z}\| \|\boldsymbol{z}_*\|}$ denotes the cosine similarity of any two vectors $\boldsymbol{z}, \boldsymbol{z}_* \in \mathbb{R}^{d'}$, and $\{\boldsymbol{z}_*^{(k)}\}_{k=1}^{K}$ are temporary features from the K classes, which could be saved in memory bank during the training.

According to neural collapse (Papyan, Han, and Donoho 2020; Zhu et al. 2021), the optimal classifier that CE can learn is the one whose classifier vectors $\{\boldsymbol{w}_k\}_{k=1}^{K}$ form an ETF which has uniform and maximum separability, i.e., satisfying $\cos(\boldsymbol{w}_k, \boldsymbol{w}_{k'}) = -\frac{1}{K-1}$ and $\|\boldsymbol{w}_k\| = \|\boldsymbol{w}_{k'}\|$ for $\forall k \neq k'$. The customized ETF classifiers (Yang et al. 2022; Liu et al. 2023) have been used to design LTR methods.

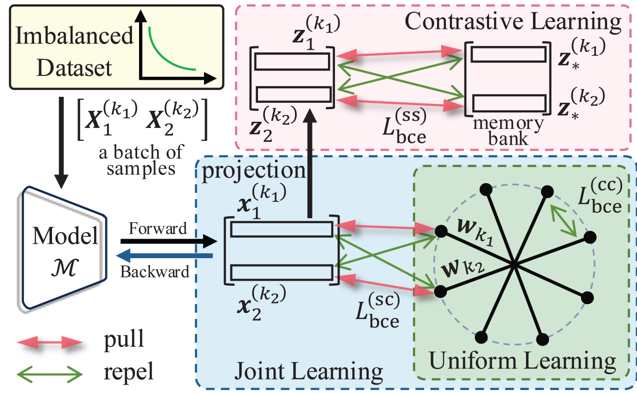


Figure 1: Pipeline of BCE3S, integrating all the three learning modes, i.e., joint learning (Eq. (3)) between sample feature and classifier, contrastive learning (Eq. (4)) among features, and classifier's uniform separability learning (Eq. (5)).

BCE-based tripartite synergistic learning

We propose the BCE-based tripartite synergistic learning (TSL), i.e., BCE3S, to integrate the above three learning modes. BCE3S consists of BCE-based joint learning $L_{\text{bce}}^{(\text{sc})}$, BCE-based contrastive learning $L_{\text{bce}}^{(\text{ss})}$, and BCE-based uniform learning $L_{\text{bce}}^{(\text{cc})}$. Fig. 1 shows its pipeline.

Joint learning. For a sample feature $\boldsymbol{x}^{(k)}$ from the k-th class, the joint learning $L_{\text{bce}}^{(\text{sc})}$ measures its similarities with the normalized classifier vectors using BCE formula,

$$L_{\text{bce}}^{(\text{sc})}(\boldsymbol{x}^{(k)}) = \log\big(1 + \exp(-\boldsymbol{w}_k^T \boldsymbol{x}^{(k)} - b_k)\big) + \sum_{\substack{j=1, j \neq k \\ p_j < r}}^{K} \log\big(1 + \exp(\boldsymbol{w}_j^T \boldsymbol{x}^{(k)} + b_j)\big), \quad (3)$$

where $\|\boldsymbol{w}_j\| = 1, \forall j$. In Eq. (3), the normalization of the classifier vectors could prevent their gradients from being dominated by the head class in training. which helps to learn more uniform classifier vectors in norm (Kang et al. 2020).

In $L_{\text{bce}}^{(\text{sc})}$, $r \in (0, 1]$ is a re-sampling parameter for the negative sample-to-class metrics $\{\boldsymbol{w}_j^T \boldsymbol{x}^{(k)}\}_{j \neq k}$, and p_j is randomly sampled from the uniform distribution $U(0, 1)$.

Contrastive learning. For any sample feature $\boldsymbol{x}^{(k)}$ from class k, similar to (Chen et al. 2020), $L_{\text{bce}}^{(\text{ss})}$ implements the contrastive learning in a projection space. A non-linear projection operator $\mathcal{P}$ is first applied to transform the sample feature into $\boldsymbol{z}^{(k)} = \mathcal{P}(\boldsymbol{x}^{(k)}) \in \mathbb{R}^{d'}$, and

$$L_{\text{bce}}^{(\text{ss})}(\boldsymbol{x}^{(k)}) = \log\Big(1 + \exp\big(-\frac{1}{\tau}\cos\big(\boldsymbol{z}^{(k)}, \boldsymbol{z}_*^{(k)}\big)\big)\Big) + \sum_{\substack{j=1 \\ j \neq k}}^{K} \log\Big(1 + \exp\big(\frac{1}{\tau}\cos\big(\boldsymbol{z}^{(k)}, \boldsymbol{z}_*^{(j)}\big)\big)\Big), \quad (4)$$

where $\big\{\boldsymbol{z}_*^{(j)} = \mathcal{P}(\boldsymbol{x}_*^{(j)})\big\}_{j=1}^{K}$ are projections of K sample features $\big\{\boldsymbol{x}_*^{(j)}\big\}_{j=1}^{K}$ stored in a memory bank.

Uniform separability learning for classifier. In the LTR tasks, the joint learning $L_{\text{bce}}^{(\text{sc})}$ tends to learn more discriminative classifier vectors for the head classes but less discriminative ones for the tail classes, and the contrastive learning $L_{\text{bce}}^{(\text{ss})}$ focuses to the sample features, which would aggravate the imbalanced effect. To balance the discriminative property of the classifier vectors for the different classes, we design a uniform separability learning for the classifier,

$$L_{\text{bce}}^{(\text{cc})}(\boldsymbol{w}_k) = \log\Big(1 + \exp\big(-\boldsymbol{w}_k^T\boldsymbol{w}_k\big)\Big) + \sum_{\substack{j=1\\j\neq k}}^{K}\log\Big(1+\exp\big(\boldsymbol{w}_k^T\boldsymbol{w}_j\big)\Big), \quad (5)$$

where the positive term equals to $\log(1+\mathrm{e}^{-1})$ as $\|\boldsymbol{w}_k\|^2 = 1$, which is a constant and was omitted in our experiments. The uniform learning $L_{\text{bce}}^{(\text{cc})}$ tries to maximize the separability among the classifier vectors $\{\boldsymbol{w}_k\}_{k=1}^K$, which helps to learn classifier with uniform and maximum separability (i.e., ETF classifier) and aligning with the final sample features by working together with the joint learning in BCE3S.

Training pipeline. In the experiments, we comprehensively learn the sample features and classifier vectors using BCE3S. As Fig. 1 shows, in each iteration during the training, a batch of samples $[\boldsymbol{X}_i^{(k_i)}]_{i=1}^B$ are fed into the model $\mathcal{M}$ to extract their features $[\boldsymbol{x}_i^{(k_i)}]_{i=1}^B$, where B is batch size. A tripartite synergistic loss is computed to train the model,

$$L_{\text{bce}}^{(\text{tri})}\big([\boldsymbol{x}_i^{(k_i)}]\big) = \frac{1}{B}\sum_{i=1}^{B}L_{\text{bce}}^{(\text{sc})}(\boldsymbol{x}_i^{(k_i)}) + \frac{\lambda_{\text{ss}}}{B}\sum_{i=1}^{B}L_{\text{bce}}^{(\text{ss})}(\boldsymbol{x}_i^{(k_i)}) + \frac{\lambda_{\text{cc}}}{K}\sum_{k=1}^{K}L_{\text{bce}}^{(\text{cc})}(\boldsymbol{w}_k), \quad (6)$$

where λ_{ss} and λ_{cc} are weight factors.

Although we design BCE3S, it should be noted that other losses, such as CE and MSE, can also be integrated within the TSL framework. However, for the LTR tasks, BCE not only inherits the advantage of fast convergence from CE but also suppresses the imbalance effects.

Analysis for tripartite synergistic learning (TSL)

We analyze BCE3S in balancing the separability among the different classes and enhancing the feature's properties.

Classifier during the training. Batch algorithm and back propagation are commonly used in the model training. In a iteration, with a batch of features $[\boldsymbol{x}_i^{(k_i)}]_{i=1}^B$, the normalized classifier vector $\boldsymbol{w}_k$ is updated via its gradient, i.e.,

$$\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k - \eta\frac{\partial L_{\text{bce}}^{(\text{tri})}\big([\boldsymbol{x}_i^{(k_i)}]\big)}{\partial \boldsymbol{w}_k}, \ \forall k, \quad (7)$$

where η is the learning rate, and

$$\frac{\partial L_{\text{bce}}^{(\text{tri})}\big([\boldsymbol{x}_i^{(k_i)}]\big)}{\partial \boldsymbol{w}_k} = -\frac{1}{B}\Bigg(\sum_{\substack{i=1\\k_i=k}}^{B}\underbrace{\sigma(-\boldsymbol{w}_k^T\boldsymbol{x}_i^{(k_i)})\boldsymbol{x}_i^{(k_i)}}_{\text{pulling term}} - \sum_{\substack{i=1\\k_i\neq k}}^{B}\underbrace{\sigma(\boldsymbol{w}_k^T\boldsymbol{x}_i^{(k_i)})\boldsymbol{x}_i^{(k_i)}}_{\text{repelling term}}\Bigg) + \frac{1}{K}\sum_{\substack{j=1\\j\neq k}}^{K}\underbrace{\sigma(\boldsymbol{w}_k^T\boldsymbol{w}_j)\boldsymbol{w}_j}_{\text{interactive term}}, \quad (8)$$

where σ is Sigmoid. In Eq. (8), the pulling terms and repelling terms are derived from the joint learning $L_{\text{bce}}^{(\text{sc})}$, and they pull the classifier vector $\boldsymbol{w}_k$ using the sample features $\boldsymbol{x}_i^{(k_i=k)}$ from the same class and repel it using the ones $\boldsymbol{x}_i^{(k_i\neq k)}$ from the different classes. Due to the higher probability that the batch contains samples from the head classes, the classifier vectors of the head classes are prone to be repelled by the samples of other head classes in different directions, while the tail classifier vectors are not easily to be repelled by the samples from the other tail classes, which results in high separability among the head classifier vectors and the low ones among the tail classifier vectors. The tail classifier vectors are also easily repelled by the head class samples, which would lead to higher separability between classifier vectors of the head and tail classes, because the repulsion from the head class is more concentrated on the tail classifier vectors. Totally, the joint learning $L_{\text{bce}}^{(\text{sc})}$ will learn a imbalanced classifier on LTR, which will affect the feature learning and result in imbalanced features across classes.

In contrast, the uniform learning $L_{\text{bce}}^{(\text{cc})}$ leads to the interactive term in Eq. (8). For each $\boldsymbol{w}_k, \forall k$, in every batch, the interactive term provides $K-1$ repelling forces from other classifier vectors, which directly, uniformly, and consistently maximize the separability among the K classifier vectors. Thereby, $L_{\text{bce}}^{(\text{cc})}$ helps to learn a balanced classifier on LTR. Combined with $L_{\text{bce}}^{(\text{sc})}$, the balanced classifier will align with the feature learning, and re-balance the features.

The contrastive learning $L_{\text{bce}}^{(\text{ss})}$ enhances the feature learning in the training, which does not directly update the classifier. However, when combined with joint learning $L_{\text{bce}}^{(\text{sc})}$, it can further exacerbate the imbalance of classifier separability through imbalanced sample features.

CE vs. BCE in LTR. On the LTR, both CE-based and BCE-based joint learning, $L_{\text{ce}}^{(\text{sc})}$ and $L_{\text{bce}}^{(\text{sc})}$, result in imbalanced classifier vectors among the head and tail classes, while they perform diversely on the sample features.

For a sample feature $\boldsymbol{x}^{(k)}$, it is updated by

$$\boldsymbol{x}^{(k)} \leftarrow \boldsymbol{x}^{(k)} - \eta\frac{\partial L_{\mu}^{(\text{sc})}\big(\boldsymbol{x}^{(k)}\big)}{\partial \boldsymbol{x}^{(k)}}, \quad (9)$$

where $\mu \in \{\text{ce}, \text{bce}\}$ denotes CE- or BCE-based joint learning, and η is the learning rate. The gradients $\frac{\partial L_{\mu}^{(\text{sc})}(\boldsymbol{x}^{(k)})}{\partial \boldsymbol{x}^{(k)}}$ have the similar form in the CE or BCE,

$$\underbrace{-\big(1-\texttt{Act}_{\mu}(\boldsymbol{w}_k^T\boldsymbol{x}^{(k)})\big)\boldsymbol{w}_k}_{\text{pulling}} + \sum_{\substack{j=1\\j\neq k}}^{K}\underbrace{\texttt{Act}_{\mu}(\boldsymbol{w}_j^T\boldsymbol{x}^{(k)})\boldsymbol{w}_j}_{\text{repelling}} \quad (10)$$

where, for $\forall j$,

$$\mathtt{Act}_{\mathrm{ce}}(\boldsymbol{w}_j^T \boldsymbol{x}^{(k)}) = \frac{\exp(\boldsymbol{w}_j^T \boldsymbol{x}^{(k)} + b_j)}{\sum_{\ell=1}^{K} \exp(\boldsymbol{w}_\ell^T \boldsymbol{x}^{(k)} + b_\ell)}, \tag{11}$$

$$\mathtt{Act}_{\mathrm{bce}}(\boldsymbol{w}_j^T \boldsymbol{x}^{(k)}) = \sigma(\boldsymbol{w}_j^T \boldsymbol{x}^{(k)}). \tag{12}$$

In the feature updating, $L_{\mathrm{ce}}^{(\mathrm{sc})}$ and $L_{\mathrm{bce}}^{(\mathrm{sc})}$ utilize the exponential inner products of the feature $\boldsymbol{x}^{(k)}$ with K classifier vectors $\{\boldsymbol{w}_j\}_{j=1}^{K}$ to compute one pulling term and $K-1$ repelling terms, while the imbalanced classifier vectors $\{\boldsymbol{w}_j\}_{j=1}^{K}$ will slow down the feature learning.

For CE-based joint learning, $L_{\mathrm{ce}}^{(\mathrm{sc})}$, in computing each pulling or repelling term, the K imbalanced exponential inner products are coupling on the denominator of Softmax $\mathtt{Act}_{\mathrm{ce}}$, thereby, the imbalanced effect is **injected again** into the feature learning. In contrast, BCE decouples the metrics between the feature with K classifier vectors, and only one classifier vector is applied in computing one pulling or repelling term by Sigmoid $\mathtt{Act}_{\mathrm{bce}}$. Therefore, BCE will achieve more balanced features across classes.

Similarly, BCE-based contractive learning and uniform separability learning, $L_{\mathrm{bce}}^{(\mathrm{ss})}$ and $L_{\mathrm{bce}}^{(\mathrm{cc})}$, could respectively result in better classifier and features than the CE-based ones, $L_{\mathrm{ce}}^{(\mathrm{ss})}$ and $L_{\mathrm{ce}}^{(\mathrm{cc})}$ (see supplementary for their formulas).

During the training, compared to CE, BCE adds at most $K-1$ logarithmic operations, which is negligible. In the testing or inference, it does not incur any additional cost.

Experiments and Results

We evaluate BCE3S on four long-tailed datasets: CIFAR10-LT/CIFAR100-LT (Yue et al. 2016), ImageNet-LT (Liu et al. 2022), and iNaturalist2018 (Van Horn et al. 2018). The CIFAR variants have IF of 100, 50, and 10, while ImageNet-LT and iNaturalist2018 have IF of 256 and 500, respectively. We train ResNets on their imbalance training set and evaluate them on the balance test set. Following (Kang et al. 2020; Alshammari et al. 2022; Du et al. 2024), we report total accuracy on the entire test set and three subsets: `Many` (>100 samples per class), `Medium` (≤ 100 and ≥ 20 samples per class), and `Few` (< 20 samples per class). More details about datasets and training can be found in supplementary.

Ablation Study

LTR results. We conduct ablation study for the proposed BCE3S on CIFAR100-LT with IF $= 100$ using ResNet32, and Table 2 presents the results. We first take the CE-based joint learning $L_{\mathrm{ce}}^{(\mathrm{sc})}$ as baseline. When training the model using BCE-based joint learning $L_{\mathrm{bce}}^{(\mathrm{sc})}$, though it reduces the accuracy on the `Many` subset, it improves the accuracy on the `Medium` and `Few` subsets, and increasing the overall accuracy from 51.48% to 52.88%, indicating that BCE pay more attention to the tail classes and improves the total accuracy at the cost of reduced accuracy on the head classes.

Based on $L_{\mathrm{ce}}^{(\mathrm{sc})}$, we compare CE- and BCE-based contrastive learning, $L_{\mathrm{ce}}^{(\mathrm{ss})}$ and $L_{\mathrm{bce}}^{(\mathrm{ss})}$. As Table 2 shows, $L_{\mathrm{bce}}^{(\mathrm{ss})}$ improves accuracy across all subsets, increasing total accuracy

Methods						Many	Med.	Few	All
$L_{\mathrm{ce}}^{(\mathrm{sc})}$	$L_{\mathrm{ce}}^{(\mathrm{ss})}$	$L_{\mathrm{ce}}^{(\mathrm{cc})}$	$L_{\mathrm{bce}}^{(\mathrm{sc})}$	$L_{\mathrm{bce}}^{(\mathrm{ss})}$	$L_{\mathrm{bce}}^{(\mathrm{cc})}$				
✓						82.29	51.37	15.67	51.48
			✓			81.11	55.06	17.40	52.88
✓	✓					84.09	54.80	17.53	53.87
✓				✓		**84.17**	55.20	17.90	54.15
	✓		✓			83.37	55.83	19.87	54.68
			✓	✓		82.74	56.57	20.63	54.95
✓		✓				82.31	51.83	17.20	52.11
✓					✓	81.23	53.69	18.17	52.67
		✓	✓			81.57	55.51	17.93	53.36
			✓		✓	81.03	56.51	19.20	53.90
✓	✓	✓				83.97	54.54	18.87	54.14
✓				✓	✓	83.77	54.49	18.67	53.99
	✓	✓	✓			83.77	56.17	21.40	55.40
			✓	✓	✓	83.34	**57.09**	**22.80**	**55.99**

Table 2: Ablation study for the proposed BCE3S and the CE based counterparts on CIFAR100-LT (IF $= 100$). The formulas of CE-based contrastive learning and uniform learning, $L_{\mathrm{ce}}^{(\mathrm{ss})}$ and $L_{\mathrm{ce}}^{(\mathrm{cc})}$, are presented in supplementary.

from 53.87% to 54.15%, and further to 54.95% when combined with $L_{\mathrm{bce}}^{(\mathrm{sc})}$. For uniform learning, $L_{\mathrm{bce}}^{(\mathrm{cc})}$ enhances the performance on `Medium` and `Few` despite reducing that on `Many`, yielding better overall performance (52.67%), which increases to 53.90% when combined with $L_{\mathrm{bce}}^{(\mathrm{sc})}$; meanwhile, it always perform better than CE-based one $L_{\mathrm{ce}}^{(\mathrm{cc})}$.

We finally compare different tripartite synergistic learning approaches based on CE and BCE, and achieve the best total LTR accuracy (55.99%) using the proposed BCE3S.

Separability and compactness. To analyze the models trained with various learning methods, for any class k, we define intra-class compactness $\mathcal{E}_k^{(\mathrm{x,com})}$, inter-class separability $\mathcal{E}_k^{(\mathrm{x,sep})}$ among features, and separability $\mathcal{E}_k^{(\mathrm{w,sep})}$ among the classifier vectors; as their names imply, $\mathcal{E}_k^{(\mathrm{x,com})}$ measures the compactness of features within the k-th class, $\mathcal{E}_k^{(\mathrm{x,sep})}$ measures the separability between the features from k-th class and that from other classes, and $\mathcal{E}_k^{(\mathrm{w,sep})}$ measures the separability between the k-th classifier vector with other ones, seeing supplementary for their definitions. The higher values suggest the better separability or compactness of the classifier vectors or features. Fig. 2 shows the results of these metrics for CE- and BCE-based methods, as well as their mean and standard deviation across different classes.

As the first row in the figure shows, the four CE-based methods achieve similar intra-class compactness, with their average values all approximately 82, and a noticeable imbalance is shown from the head to tail classes. In contrast, for the BCE-bsed methods, the average compactness of $L_{\mathrm{bce}}^{(\mathrm{sc})}$ reaches to 86.02. With the addition of $L_{\mathrm{bce}}^{(\mathrm{ss})}$ and $L_{\mathrm{bce}}^{(\mathrm{cc})}$, the compactness gradually increase to 95.47, while the imbalance across different classes gradually decreases and standard deviation of the compactness decreases from 5.55 to 1.81. As the second row shows, compared to the CE-based methods, the BCE-based ones achieve higher inter-class separability for sample features, with the highest mean

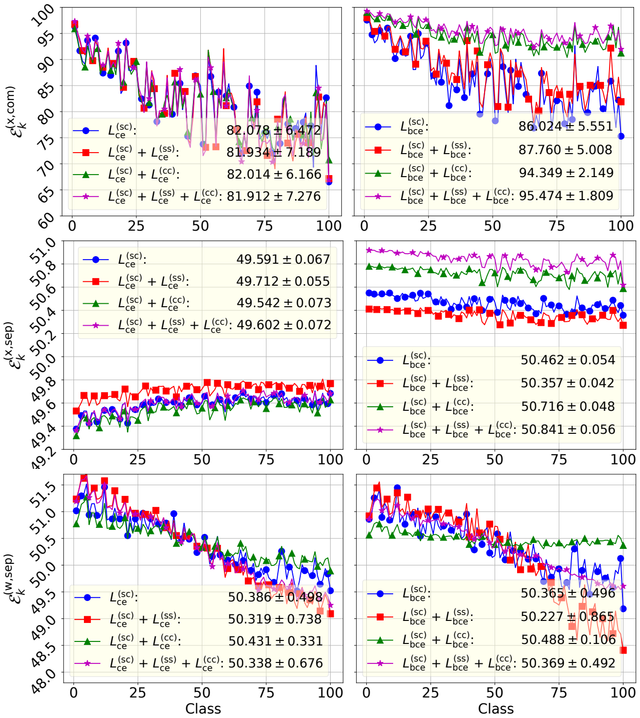


Figure 2: The intra-class compactness (top), inter-class separability (middle) of sample features, and separability (bottom) of classifier vectors on the training set of CIFAR100-LT (IF $= 100$), with the model trained using different CE- (left) and BCE-based (right) methods.

of separability increases from 49.71 ($L_{\mathrm{ce}}^{(\mathrm{sc})} + L_{\mathrm{ce}}^{(\mathrm{ss})}$) to 50.84 ($L_{\mathrm{bce}}^{(\mathrm{sc})} + L_{\mathrm{bce}}^{(\mathrm{ss})} + L_{\mathrm{bce}}^{(\mathrm{cc})}$). Totally, compared to CE-based methods, BCE3S enhances the feature properties and balances the property difference on the head and tail classes.

As the third row of Fig. 2 shows, the average separability of classifier vectors for the various methods is similar, close to 50.50, while they have significant differences across the different classes. The classifier separability of the two joint learning $L_{\mathrm{ce}}^{(\mathrm{sc})}$ and $L_{\mathrm{bce}}^{(\mathrm{sc})}$ has significant imbalance from the head to tail classes. When the CE- and BCE-based contrastive learning are added, the imbalance is respectively amplified, while the two kinds of uniform learning can reduce the imbalance effect. In particular, the BCE one significantly suppresses the separability imbalance, resulting in a standard deviation of only 0.106 for $L_{\mathrm{bce}}^{(\mathrm{sc})} + L_{\mathrm{bce}}^{(\mathrm{cc})}$.

These results indicate that BCE3S can not only enhance the feature properties, but also effectively improve the balance of features and classifier for the LTR models.

Feature distribution in t-SNE. To intuitively compare the feature properties between our BCE3S with CE-based methods, for ResNet32 trained on CIFAR10-LT with IF $=100$. We visually show the feature distributions of the 10 classes on the test set in Fig. 3 using t-SNE. As the figure shows, the feature clusters of class 3 and 5 (i.e., "cat" and "dog") of CE-based joint learning $L_{\mathrm{ce}}^{(\mathrm{sc})}$ overlap with each other; though the intersection decreases with the addition of $L_{\mathrm{ce}}^{(\mathrm{ss})}$ and $L_{\mathrm{ce}}^{(\mathrm{cc})}$, the final feature clusters of the two classes are not completely separated, indicating unsatisfactory inter-class separability. Meanwhile, the features of the tail classes (the 8th and 9th classes) are spreading over relatively elongated regions, indicating unsatisfactory intra-class compactness. In contrast, for BCE, the ten feature clusters of $L_{\mathrm{bce}}^{(\mathrm{sc})}$ are distributed in relatively independent regions, while with the addition of $L_{\mathrm{bce}}^{(\mathrm{ss})}$ and $L_{\mathrm{bce}}^{(\mathrm{cc})}$, the clusters of the tail classes becomes increasingly compact, indicating higher separability and compactness among the features.

Methods	CIFAR10-LT			CIFAR100-LT		
	100	50	10	100	50	10
CB-Focal CVPR'19	74.60	79.30	87.10	39.60	45.20	58.00
TSC CVPR'22	79.70	82.90	88.70	43.80	47.40	59.00
ETF+DR NeurIPS'22	76.50	-	87.70	45.30	-	-
NC-DRW AISTATS'23	81.90	-	89.80	48.60	-	63.10
RIDE (3 exp.) ICLR'21	-	-	-	48.60	51.40	59.80
BCL CVPR'22	84.50	87.20	91.10	51.90	56.40	64.60
NC.-cRT AISTATS'23	82.60	-	90.20	48.70	-	63.60
ResLT+H2T AAAI'24	81.77	84.99	-	49.60	54.39	-
RIDE+CMO CVPR'22	-	-	-	50.00	53.00	60.20
Logit Adj. ICLR'21	84.30	87.10	90.90	50.50	54.90	64.00
BCL+DODA ICLR'24	-	-	-	51.00	53.60	62.70
RIDE+H2T AAAI'24	-	-	-	51.38	55.54	-
DiffuLT NeurIPS'24	84.70	86.90	90.70	51.50	56.30	63.80
DiffuLT+RIDE NeurIPS'24	85.30	87.30	90.90	52.40	56.90	64.20
ProCo TPAMI'24	85.90	88.20	91.90	52.80	57.10	65.50
RBL ICML'23	84.70	87.60	-	53.10	57.20	-
DeiT-LT CVPR'24	87.50	89.80	-	55.60	60.50	-
GLMC CVPR'23	88.50	91.04	94.87	57.97	63.78	73.40
GLMC + MN	89.58	92.04	94.87	58.41	64.57	74.28
BCE3S	**90.08**	**92.55**	**95.71**	**59.50**	**65.23**	**76.13**

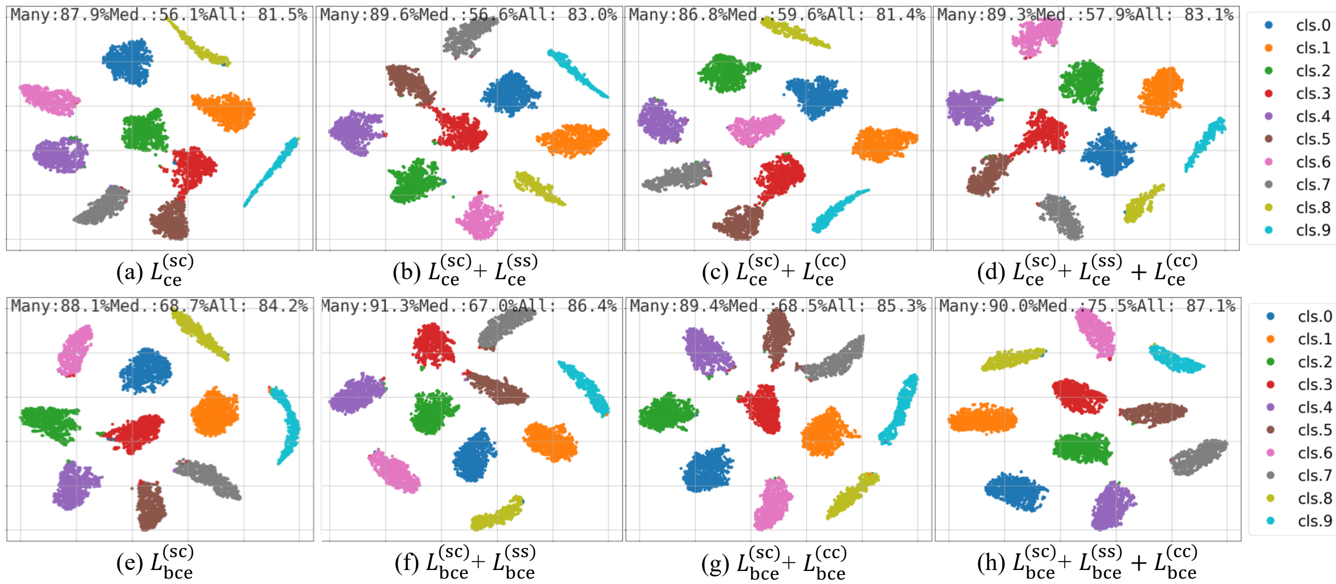
Table 3: LTR on CIFAR10-LT and CIFAR100-LT, IF $= 100$, 50, and 10. BCE3S consistently achieves the best results.

Parameter study. BCE3S applied hyper-parameters λ_{ss}, λ_{cc}, and τ to balance the impact of three learning components, and we have conducted a comprehensive study using ResNet32 on CIFAR100-LT with IF $= 100$. The detailed experimental results can be found in the supplementary.

Comparison with SOTA

We compare our BCE3S with a series of SOTA LTR methods on CIFAR10-LT, CIFAR100-LT, ImageNet-LT, and iNaturalist2018 with different backbones. To further explore the potential of BCE3S, we boost its performance by combining it with re-balancing techniques (Alshammari et al. 2022; Ren et al. 2020). The detailed experimental settings and more results can be found in the supplementary.

On **CIFAR10-LT and CIFAR100-LT**, we adopted a similar training strategy used in MaxNorm (MN) of Alshammari et al. (2022). As Table 3 shows, our BCE3S achieves the best top-1 accuracy on both CIFAR10-LT and CIFAR100-LT with different IFs. For example, on CIFAR100-LT, BCE3S achieves accuracies of 76.13%, 65.23%, and 59.50% with the three IFs, which respectively surpass the previous best results by 1.85%, 0.66%, and 1.09%, being new SOTA. On CIFAR10-LT dataset, BCE3S has also achieved SOTA results. Those results highlight the potential of a model trained using BCE3S.



(a) $L_{\text{ce}}^{(\text{sc})}$ (b) $L_{\text{ce}}^{(\text{sc})}+L_{\text{ce}}^{(\text{ss})}$ (c) $L_{\text{ce}}^{(\text{sc})}+L_{\text{ce}}^{(\text{cc})}$ (d) $L_{\text{ce}}^{(\text{sc})}+L_{\text{ce}}^{(\text{ss})}+L_{\text{ce}}^{(\text{cc})}$

(e) $L_{\text{bce}}^{(\text{sc})}$ (f) $L_{\text{bce}}^{(\text{sc})}+L_{\text{bce}}^{(\text{ss})}$ (g) $L_{\text{bce}}^{(\text{sc})}+L_{\text{bce}}^{(\text{cc})}$ (h) $L_{\text{bce}}^{(\text{sc})}+L_{\text{bce}}^{(\text{ss})}+L_{\text{bce}}^{(\text{cc})}$

Figure 3: Feature distribution on the CIFAR10-LT test set with CE (top) and BCE (bottom) learning methods. Compared to CE methods, features extracted using BCE-based joint learning $L_{\text{bce}}^{(\text{sc})}$ show improved intra-class compactness and inter-class separability. The contrastive learning $L_{\text{bce}}^{(\text{ss})}$ and uniform learning $L_{\text{bce}}^{(\text{cc})}$ further enhance these properties.

Methods	iNaturalist 2018			
	`Many`	`Med.`	`Few`	All
TSC CVPR'22	72.60	70.60	67.80	69.70
DR ICLR'20	72.88	71.15	69.24	70.49
GCL CVPR'22	66.43	71.66	72.47	71.47
GCL+H2T AAAI'24	67.74	71.92	72.22	71.62
BCL CVPR'22	69.40	72.40	71.80	71.80
DR+H2T AAAI'24	71.73	72.32	71.30	71.81
DSCL AAAI'24	74.20	72.90	70.30	72.00
RIDE ICLR'21	76.52	74.23	70.45	72.80
RIDE+H2T AAAI'24	75.69	74.22	71.36	73.11
PaCo (400 ep.) ICCV'21	70.30	73.20	73.60	73.20
ProCo (90 ep.) TPAMI'24	-	-	-	73.50
BCL+DODA ICLR'24	71.20	73.20	73.40	73.70
BCE3S (180 ep.)	77.16	74.45	72.78	73.99
DeiT-LT(ViT-B, 1K ep.) CVPR'24	70.30	75.20	76.20	75.10
ProCo(400 ep.) TPAMI'24	74.00	76.00	**76.80**	75.80
BCE3S (400 ep.)	**79.10**	**76.08**	74.28	**75.91**

Table 4: LTR results on iNaturalist2018 using ResNet50.

Methods	$\mathcal{M}$	ImageNet-LT			
		`Many`	`Med.`	`Few`	All
CB-Focal CVPR'19	R50	39.60	32.70	16.80	33.20
LDAM NeurIPS'19		60.40	46.90	30.70	49.80
KCL ICLR'21		61.80	49.40	30.90	51.50
TSC CVPR'22		63.50	49.40	30.40	52.40
GCL CVPR'22		62.24	48.62	52.12	54.51
GCL+H2T AAAI'24		62.36	48.75	52.15	54.62
BCL CVPR'22		65.30	53.50	36.30	55.60
DiffuLT NeurIPS'24		63.20	55.40	39.20	56.20
BCL+DODA ICLR'24		66.90	54.10	37.40	56.90
DiffuLT+RIDE NeurIPS'24		64.10	55.80	39.90	56.90
DSCL AAAI'24		**68.50**	55.20	35.40	57.70
ProCo TPAMI'24		68.20	55.10	38.10	57.80
BCE3S		68.14	**55.90**	**40.56**	**57.85**

Table 5: LTR results on ImageNet-LT using ResNet50 (R50). BCE3S achieves the best results on the test set.

On **iNaturalist2018**, as Table 4 shows, we apply BCE3S to train ResNet50 using a strategy similar to that of Du et al. (2024). When trained for 180 epochs, BCE3S achieves a competitive accuracy of 73.99% on the overall test set, outperforming most previous methods. When extended to 400 epochs, BCE3S achieves the best performance on test subsets of `Many` and `Med.`, and get an optimal overall accuracy of 75.91%, surpassing ProCo and DeiT-LT, though DeiT-LT uses ViT-B trained for 1,000 epochs.

On **ImageNet-LT**, we evaluate BCE3S using ResNet50. As Table 5 shows, BCE3S achieves the best performance on `Medium`, and `Few`, with accuracy of 55.90% and 40.56%, and obtains the best overall accuracy of 57.85%.

Conclusion

For long-tailed recognition (LTR) on imbalanced datasets, this paper proposes the BCE-based tripartite synergistic learning method, i.e., BCE3S, by integrating the joint learning between sample features and classifier vectors, contrastive learning among the features, and uniform separability learning for the classifier vectors. Compared with CE-based methods, BCE3S suppresses the imbalance effect among the classifier vectors of the head and tail classes, improves the models' focus on the feature learning in the tail classes, and enhances the intra-class compactness and inter-class separability of the features. The extensive experiments demonstrate that our BCE3S achieves the optimal LTR performance on various long-tailed datasets.

Acknowledgements

The research was supported by National Natural Science Foundation of China under Grant No. 62576217 and 8226113862, Natural Science Foundation of Ningxia Province under Grant No. 2025AAC020002, Guangdong Provincial Key Laboratory under Grant No. 2023B1212060076, and Scientific Foundation for Youth Scholars of Shenzhen University under Grant No. 868-000001032180.

References

Alshammari, S.; Wang, Y.-X.; Ramanan, D.; and Kong, S. 2022. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6897–6907.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.

Chen, X.; Zhou, Y.; Wu, D.; Yang, C.; Li, B.; Hu, Q.; and Wang, W. 2023. Area: adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19277–19287.

Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; and Jia, J. 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 715–724.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Du, C.; Wang, Y.; Song, S.; and Huang, G. 2024. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Du, F.; Yang, P.; Jia, Q.; Nan, F.; Chen, X.; and Yang, Y. 2023. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15814–15823.

Fang, C.; He, H.; Long, Q.; and Su, W. J. 2021. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43): e2103091118.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Kang, B.; Li, Y.; Xie, S.; Yuan, Z.; and Feng, J. 2021. Exploring Balanced Feature Spaces for Representation Learning. In *International Conference on Learning Representations*.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations*.

Kasarla, T.; Burghouts, G.; Van Spengler, M.; Van Der Pol, E.; Cucchiara, R.; and Mettes, P. 2022. Maximum class separation as inductive bias in one matrix. *Advances in neural information processing systems*, 35: 19553–19566.

Li, M.; Cheung, Y.-m.; and Lu, Y. 2022. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6929–6938.

Li, T.; Wang, L.; and Wu, G. 2021. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 630–639.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, X.; Zhang, J.; Hu, T.; Cao, H.; Yao, Y.; and Pan, L. 2023. Inducing neural collapse in deep long-tailed learning. In *International Conference on Artificial Intelligence and Statistics*, 11534–11544. PMLR.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Stella, X. Y. 2022. Open long-tailed recognition in a dynamic world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3): 1836–1851.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Papyan, V.; Han, X. Y.; and Donoho, D. L. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Peifeng, G.; Xu, Q.; Wen, P.; Yang, Z.; Shao, H.; and Huang, Q. 2023. Feature Directions Matter: Long-Tailed Learning via Rotated Balanced Representation. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 27542–27563. PMLR.

Rangwani, H.; Mondal, P.; Mishra, M.; Asokan, A. R.; and Babu, R. V. 2024. DeiT-LT: Distillation Strikes Back for Vision Transformer Training on Long-Tailed Datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23396–23406.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33: 4175–4186.

Shao, J.; Zhu, K.; Zhang, H.; and Wu, J. 2024. DiffuLT: Diffusion for Long-tail RecognitionWithout External Knowledge. *Advances in neural information processing systems*, 37.

Touvron, H.; Cord, M.; and Jégou, H. 2022. Deit iii: Revenge of the vit. In *European conference on computer vision*, 516–533. Springer.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2011. Class imbalance, redux. In *2011 IEEE 11th international conference on data mining*, 754–763. Ieee.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. 2021. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. In *International Conference on Learning Representations*.

Wightman, R.; Touvron, H.; and Jegou, H. 2021. ResNet strikes back: An improved training procedure in timm. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*.

Xu, Z.; Chai, Z.; Xu, C.; Yuan, C.; and Yang, H. 2023a. Towards Effective Collaborative Learning in Long-Tailed Recognition. *IEEE Transactions on Multimedia*.

Xu, Z.; Liu, R.; Yang, S.; Chai, Z.; and Yuan, C. 2023b. Learning imbalanced data with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15793–15803.

Xuan, S.; and Zhang, S. 2024. Decoupled Contrastive Learning for Long-Tailed Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6396–6403.

Yang, Y.; Chen, S.; Li, X.; Xie, L.; Lin, Z.; and Tao, D. 2022. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35: 37991–38002.

Yue, C.; Long, M.; Wang, J.; Han, Z.; and Wen, Q. 2016. Deep quantization network for efficient image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3457–3463.

Zhang, S.; Li, Z.; Yan, S.; He, X.; and Sun, J. 2021. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2361–2370.

Zhao, Y.; Chen, W.; Tan, X.; Huang, K.; and Zhu, J. 2022. Adaptive logit adjustment loss for long-tailed visual recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3472–3480.

Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.-P. P.; and Jiang, Y.-G. 2022. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6908–6917.

Zhu, Z.; Ding, T.; Zhou, J.; Li, X.; You, C.; Sulam, J.; and Qu, Q. 2021. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34: 29820–29834.