BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition (Supplementary Materials)

In the supplementary materials, we provide more experimental results and analyses of our proposed BBN model, including:

1. Additional experiments of different manners for representation and classifier learning (cf. Section 3 and Figure 2 of the paper) on large-scale datasets iNaturalist 2017 and iNaturalist 2018;

2. Affects of re-balancing strategies on the compactness of learned features;

3. Comparisons between the BBN model and ensemble methods;

4. Coordinate graph of different adaptor strategies for generating α;

5. Learning algorithm of our proposed BBN model.

1. Additional experiments of different manners for representation and classifier learning (cf. Section 3 and Figure 2 of the paper) on large-scale datasets iNaturalist 2017 and iNaturalist 2018

In this section, following Section 3 of our paper, we conduct experiments on large-scale datasets, *i.e.*, iNaturalist 2017 [3] and iNaturalist 2018, to further justify our conjecture (*i.e.*, the working mechanism of these class re-balancing strategies is to promote classifier learning significantly but might damage the universal representative ability of the learned deep features due to distorting original distributions.) Specifically, the representation learning stages are conducted on iNaturalist 2017. Then, to also evaluate the generalization ability for learned representations, classifier learning stages are performed on not only iNaturalist 2017 but also iNaturalist 2018.

As shown in Figure 1 of the supplementary materials, we can also have the observations from two perspectives on these large-scale long-tailed datasets:

- **Classifiers:** When we apply the same representation learning manner (comparing error rates of three blocks in the vertical direction), it can be reasonably found that RW/RS always achieve lower classification error rates than CE, which owes to their re-balancing operations adjusting the classifier weights updating to match test distributions.
- **Representations:** When applying the same classifier learning manner (comparing error rates of three blocks in the horizontal direction), it is a bit of surprise to see that error rates of CE blocks are consistently lower than error rates of RW/RS blocks. The findings indicate that training with CE achieves better classification results since it obtains better features. The worse results of RW/RS reveal that they lead to inferior discriminative ability of the learned deep features.

These observations are consistent with those on long-tailed CIFAR datasets, which can further demonstrate our discovery of Section 3 in the paper.

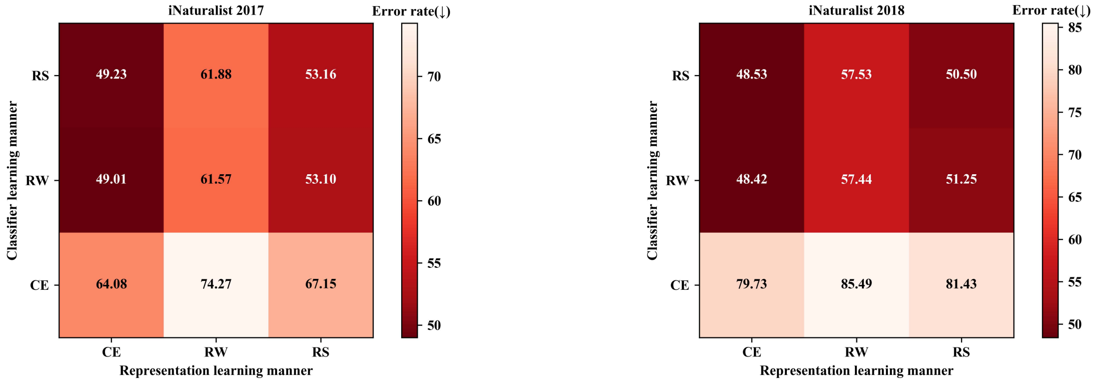

Figure 1. Top-1 error rates of different manners for representation learning and classifier learning on two large-scale long-tailed datasets iNaturalist 2017 and iNaturalist 2018. "CE" (Cross-Entropy), "RW" (Re-Weighting) and "RS" (Re-Sampling) are the conducted learning manners.

2. Affects of re-balancing strategies on the compactness of learned features

To further prove our conjecture that re-balancing strategies could damage the universal representations, we measure the compactness of intra-class representations on CIFAR-10-IR50 [1] for verification.

Concretely, for each class, we firstly calculate a centroid vector by averaging representations of this class. Then, ℓ_2 distances between these representations and their centroid are computed and then averaged as a measurement for the compactness of intra-class representations. If the averaged distance of a class is small, it implies that representations of this class gather closely in the feature space. We normalize the ℓ_2-norm of representations to 1 in the training stage for avoiding the impact of feature scales. We report results based on representations learned with Cross-Entropy (CE), Re-Weighting (RW) and Re-Sampling (RS), respectively.

As shown in Figure 2 of the supplementary materials, the averaged distances of re-balancing strategies are obviously larger than conventional training, especially for the head classes. That is to say, the compactness of learned features of re-balancing strategies are significantly worse than conventional training. These observations can further validate the statements in Figure 1 of the paper (*i.e.*, for re-balancing strategies, "the intra-class distribution of each class becomes more separable") and also

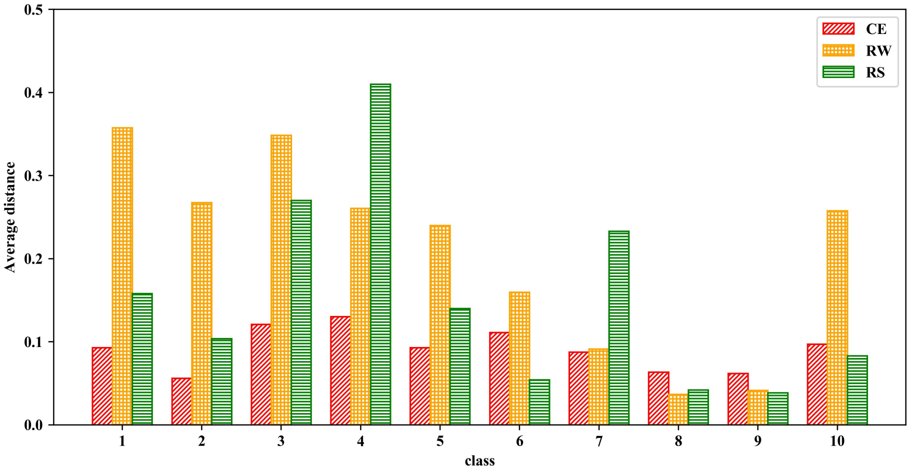

Figure 2. Histogram of the measurement for the compactness of intra-class representations on the CIFAR-10-IR50 dataset. Especially for head classes, representations trained with CE gather more closely than those trained with RW/RS, since the representations of each class are closer to their centroid. The vertical axis is the averaged distance between learned features of each class and their corresponding centroid (The smaller, the better).

Table 1. Top-1 error rates of our proposed BBN model and ensemble methods.

Methods	CIFAR-10-IR50	CIFAR-100-IR50	iNaturalist 2017	iNaturalist 2018
Uniform sampler + Balanced sampler	19.41	55.10	39.53	36.20
Uniform sampler + Reversed sampler	19.38	54.93	40.02	36.66
BBN (Ours)	**17.82**	**52.98**	**36.61**	**33.74**

the discovery of Section 3 in the paper (*i.e.*, re-balancing strategies "might damage the universal representative ability of the learned deep features to some extent").

3. Comparisons between the BBN model and ensemble methods

In the following, we compare our BBN model with ensemble methods to prove the effectiveness of our proposed model. Results on CIFAR-10-IR50 [1], CIFAR-100-IR50 [1], iNaturalist 2017 [3] and iNaturalist 2018 are provided in Table 1 for comprehensiveness.

As known, ensemble techniques are frequently utilized to boost performances of machine learning tasks. We train three classification models with uniform data sampler, balanced data sampler and reversed data sampler, respectively. For mimicking our bilateral-branch network design and considering fair comparisons, we provide classification error rates of (1) an ensemble of models learned with a uniform sampler and a balanced sampler, as well as (2) another ensemble of models learned with a uniform sampler and a reversed sampler.

As shown in Table 1 of the supplementary materials, our BBN model achieves consistently lower error rates than ensemble models on all datasets. Additionally, compared to ensemble models, our proposed BBN model can yield better performance with limited increase of network parameters thanks to its sharing weights design (cf. Sec. 4.2 of the paper).

4. Coordinate graph of different adaptor strategies for generating α

As shown in Figure 3 of the supplementary materials, we provide a coordinate graph to present how α varies with the progress of network training. The adaptor strategies shown in the figure are the same as those in Table 4 of the paper except for the β-distribution for its randomness.

Furthermore, as discussed in Sec. 5.5.2 of the paper, these decay strategies yield better results than the other non-decay strategies. When α decreasing, the learning focus of our BBN gradually changes from representation learning to classifier learning, which fits our motivation stated in Sec. 4.3 of the paper. Among these decay strategies, our proposed parabolic decay is the best. Specifically, we can intuitively regard $\alpha > 0.5$ as the learning focus emphasizing representation learning, as well as $\alpha \leq 0.5$ as the learning focus emphasizing classifier learning. As shown in Figure 3 of the supplementary materials,

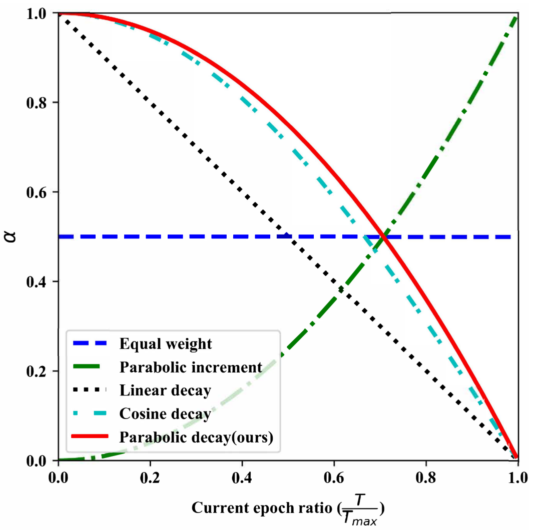

Figure 3. Different kinds of adaptor strategies for generating α. The horizontal axis indicates current epoch ratio $\frac{T}{T_{max}}$ and the vertical axis denotes the value of α. (Best viewed in color)

compared with other decay strategies, our parabolic decay *with the maximum degree* prolongs the epochs of the learning focus upon representation learning. As analyzed by theoretical understanding of learning dynamics in networks [2], network convergence speed is highly correlated with the number of layers. That is to say, the representation learning part (former layers) of networks requires more epochs to sufficiently converge, while the classifier learning part (later layers) requires relatively less epochs until sufficient convergence. In fact, our parabolic decay ensures that BBN could have enough epochs to fully update the representation learning part, *i.e.*, learning better universal features, which is the crucial foundation for learning robust classifiers. That is why our parabolic decay is the best.

5. Learning algorithm of our proposed BBN model

In the following, we provide the detailed learning algorithm of our proposed BBN. In Algorithm 1 of the supplementary materials, for each training epoch T, we firstly assign a value to α by the adaptor proposed in Eq. (5) of the paper. Then, we sample training samples by the uniform sampler and reversed sampler, respectively. After feeding samples into our network, we can obtain two independent feature vectors $\boldsymbol{f}_c$ and $\boldsymbol{f}_r$. Then, we calculate the output logits $\boldsymbol{z}$ and the prediction possibility $\hat{\boldsymbol{p}}$ according to Eq. (1) and Eq. (2) of the paper. Finally, the classification loss function is calculated based on Eq. (3) of the paper and we update model parameters by optimizing this loss function.

Algorithm 1 Learning algorithm of our proposed BBN

Require : Training Dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$; `UniformSampler`$(\cdot)$ denotes obtaining a sample from $\mathcal{D}$ selected by a uniform sampler; `ReversedSampler`$(\cdot)$ denotes obtaining a sample by a reversed sampler; $\mathcal{F}_{cnn}(\cdot;\cdot)$ denotes extracting the feature representation from a CNN; θ_c and θ_r denote the model parameters of the conventional learning and re-balancing branch; $\boldsymbol{W}_c$ and $\boldsymbol{W}_r$ present the classifiers' weights (*i.e.*, last fully connected layers) of the conventional learning and re-balancing branch.

```
for $T = 1$ to $T_{max}$ do
    $\alpha \leftarrow 1 - \left(\frac{T}{T_{max}}\right)^2$
    $(\mathbf{x}_c, y_c) \leftarrow \texttt{UniformSampler}(\mathcal{D})$
    $(\mathbf{x}_r, y_r) \leftarrow \texttt{ReversedSampler}(\mathcal{D})$
    $\boldsymbol{f}_c \leftarrow \mathcal{F}_{cnn}(\mathbf{x}_c; \theta_c)$
    $\boldsymbol{f}_r \leftarrow \mathcal{F}_{cnn}(\mathbf{x}_r; \theta_r)$
    $\mathbf{z} \leftarrow \alpha \boldsymbol{W}_c^{\top} \boldsymbol{f}_c + (1-\alpha) \boldsymbol{W}_r^{\top} \boldsymbol{f}_r$
    $\hat{\boldsymbol{p}} \leftarrow \texttt{Softmax}(\boldsymbol{z})$
    $\mathcal{L} \leftarrow \alpha E(\hat{\boldsymbol{p}}, y_c) + (1-\alpha) E(\hat{\boldsymbol{p}}, y_r)$
    Update model parameters by minimizing $\mathcal{L}$
end for
```

References

[1] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 2, 3

[2] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*, pages 1–14, 2014. 4

[3] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 2, 3