COVARIANCE-CORRECTED WHITENING ALLEVIATES NETWORK DEGENERATION ON IMBALANCED CLASSIFICATION

Anonymous authors
Paper under double-blind review

In this appendix, we first provide the detailed steps of our proposed Whitening-Net in Algrithm 1. Next, we present more visualization results in Section B to prove that the model does have network degeneration phenomenon on imbalanced data, where the experiments come from different datasets (CIFAR-10-LT, CIFAR-100-LT and iNaturalist-LT), different backbones (ResNet-10, ResNet-32, ResNet-110, EfficientNet-B0 and DenseNet121). Then, Section C give a simple example for understanding the superiority of our GRBS sampler. Finally, in Section D, we present the ablation studies on the hyper-parameters of GRBS and BET.

A ALGORITHM

Algorithm 1 Whitening-Net: An End-to-End Training Method for Imbalanced Classification

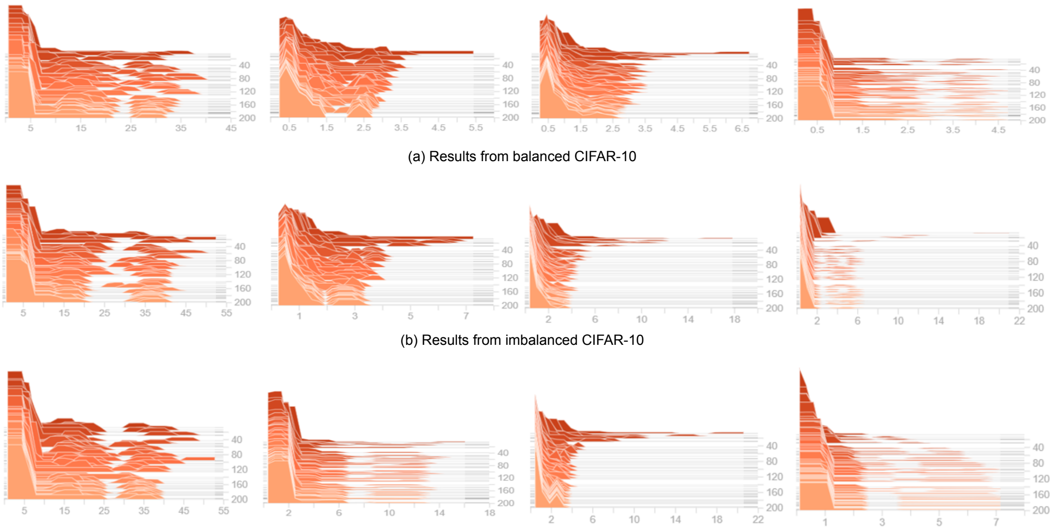
Required Samplers: Random Sampler with iterations T_1, GRBS Sampler with iterations T_2
Required Models: Initialized Backbone $f_{\theta 1}$ and Classifier $f_{\theta 2}$
Required: Inputs $\mathbf{X}$, Features $\mathbf{Z}$, Iteration T, Channel Whitening ϕ, Whitened features $\hat{\mathbf{Z}}$

```
for $t_1$=1 to $T_1$ do
    Extract features from random sampler $\mathbf{Z} = f_{\theta 1}(\mathbf{X})$
    Channel whitening $\hat{\mathbf{Z}} = \phi(\mathbf{Z})$
    Output logits $\hat{\mathbf{Y}} = f_{\theta 2}(\hat{\mathbf{Z}})$
    if $t_1/T = 0$ then
        for $t_2$=1 to $T_2$ do
            Extract features from GRBS sampler $\mathbf{Z} = f_{\theta 1}(\mathbf{X})$
            Channel whitening $\hat{\mathbf{Z}} = \phi(\mathbf{Z})$
            Output logits $\hat{\mathbf{Y}} = f_{\theta 2}(\hat{\mathbf{Z}})$
            Compute cross-entropy loss
            Update $f_{\theta 1}$ and $f_{\theta 2}$ by back-propagation
        end for
    end if
    Compute cross-entropy loss
    Update $f_{\theta 1}$ and $f_{\theta 2}$ by back-propagation
end for
```

B VISUALIZATION FOR NETWORK DEGENERATION

B.1 VISUALIZATION ON ALL HIDDEN LAYERS

In Figure 1, we visualize the channel-wised singular value distributions on different layers of ResNet-32. From Figure 1 (a) and (b), we observe that the main difference between the features learned in the imbalanced and balanced datasets is that the last intermediate hidden layer, i.e., the features fed into the classifier, learned on the imbalanced dataset have a significantly larger amount of nearly zero-valued singular values, which implies that these features are highly correlated. This feature representation collapse would make the training intractable and finally leads to degenerated solutions.

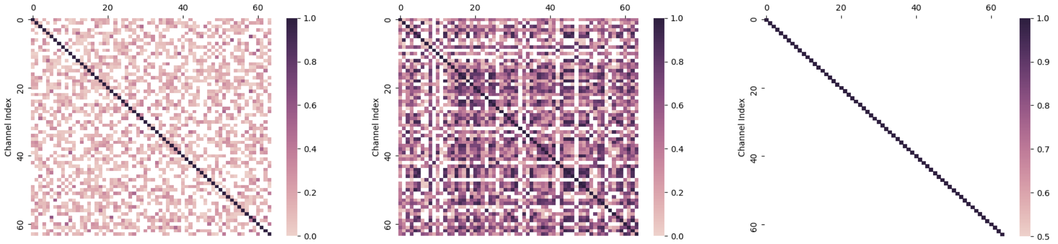


(a) Results from balanced CIFAR-10

(b) Results from imbalanced CIFAR-10

(c) Results from Imbalanced CIFAR-10 with channel whitening

Figure 1: Singular value histograms of features on different layers (The sub-figures of (a), (b) from left to right are: Layer_1, Layer_2, Layer_3 and Layer_p, where "p" denotes pooling. The last sub-figure on (c) is Layer_p after whitening transformation.) of ResNet-32 using end-to-end training. The first, middle and bottom rows present the results on balanced CIFAR-10, imbalanced CIFAR-10 and imbalanced CIFAR-10 with whitening, respectively. The vertical axis in each figure stands for the training epoch. We can see that the main difference between the balanced and imbalanced tasks is that a large amount of the singular values of the features fed into classifier (i.e., the last column) in the imbalanced task are nearly zero, which implies that these features are highly correlated. The bottom row demonstrates that our whitening can effectively decorrelate these features since the features have more large singular values.



(a) Balanced CIFAR-100

(b) Imbalanced CIFAR-100

(c) Imbalanced CIFAR-100 with channel whitening

Figure 2: The correlation coefficients between channel-wised features fed into the classifier at the last epochs. The experiments are constructed on CIFAR-100-LT datasets using ResNet-32.

B.2 Visualization on CIFAR-100-LT

As shown in Figure 2, we provide a visualization results on CIFAR-100-LT with imbalanced factor 200, in which we can draw the same conclusion as in the paper, i.e. the features trained on an imbalanced dataset have larger correlation coefficients, and the proposed ZCA whitening approach can alleviate this problem.

B.3 Visualization on More Network Architectures

We also construct experiments on CIFAR-10-LT dataset using more different backbones, e.g., ResNet-110, EfficientNet-B0 and DenseNet121, to prove the conclusion in the main paper. As shown in Figure 3, 4 and 5, the correlation coefficient values of the last layer features increases with the imbalance ratio of the dataset.

C A Simple Example for Understanding GRBS

we give a simple example to illustrate the advantage of our proposed sampler GRBS.

To be precise, we consider the simple imbalanced case that we are given a Gaussian mixture distribution comprised of nine Gaussian data distributions $\mathcal{N}(\mu, \sigma_i^2)$ with probability p_i, where $\sigma_1 > \sigma_2 > \ldots > \sigma_9$, $p_1 \approx p_2 \approx p_3 \gg p_4 \approx \ldots \approx p_9 \approx 0$ and $\sum_{i=1}^{9} p_i = 1$. The following remark shows such setting for variance is reasonable as existing works Khan et al. (2019) argue that in the imbalanced classification task, the variances of head classes are larger than the tail classes.

We denote the density of $\mathcal{N}(\mu, \sigma_i^2)$ as f_i. If we use the random sampler to independently sample N samples, i.e., $\mathbf{x}_i \sim \sum_{j=1}^{9} p_j f_j$. We consider the mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$. We denote the variance of $\bar{\mathbf{x}}$ of the random sampler as Var-RS and we can obtain that

$$\text{Var-RS} = \frac{1}{N}\sum_{i=1}^{9} p_i \sigma_i^2 \approx \frac{1}{N}\left(p_1\sigma_1^2 + p_2\sigma_2^2 + p_3\sigma_3^2\right) \approx \frac{1}{3N}\left(\sigma_1^2 + \sigma_2^2 + \sigma_3^2\right).$$

If we use our sampler and divide the categories into 3 groups, then $\bar{\mathbf{x}}$ takes the form of

$$\bar{\mathbf{x}} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i, \quad \mathbf{x}_i \sim \frac{p_1 f_1 + \tilde{p}_4 f_4 + \tilde{p}_7 f_7}{p_1 + \tilde{p}_4 + \tilde{p}_7} \text{ with probability} = \frac{1}{3}; \\ \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i, \quad \mathbf{x}_i \sim \frac{p_2 f_2 + \tilde{p}_5 f_5 + \tilde{p}_8 f_8}{p_2 + \tilde{p}_5 + \tilde{p}_8} \text{ with probability} = \frac{1}{3}; \\ \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i, \quad \mathbf{x}_i \sim \frac{p_3 f_3 + \tilde{p}_6 f_6 + \tilde{p}_9 f_9}{p_3 + \tilde{p}_6 + \tilde{p}_9} \text{ with probability} = \frac{1}{3}; \end{cases}$$

where we increase $p_4, p_5, \ldots, p_9$ to $\tilde{p}_4, \tilde{p}_5, \ldots, \tilde{p}_9$ in order to increase the importance the tail classes, i.e., $\tilde{p}_4 \gg p_4, \tilde{p}_5 \gg p_5, \ldots, \tilde{p}_9 \gg p_9$. Then, we denote the variance of our $\bar{\mathbf{x}}$ from GRBS as Var-GRBS and we can obtain that

$$\begin{aligned} \text{Var-GRBS} &= \frac{1}{3N}\left(\frac{p_1\sigma_1^2 + \tilde{p}_4\sigma_4^2 + \tilde{p}_7\sigma_7^2}{p_1 + \tilde{p}_4 + \tilde{p}_7} + \frac{p_2\sigma_2^2 + \tilde{p}_5\sigma_5^2 + \tilde{p}_8\sigma_8^2}{p_2 + \tilde{p}_5 + \tilde{p}_8} + \frac{p_3\sigma_3^2 + \tilde{p}_6\sigma_6^2 + \tilde{p}_9\sigma_9^2}{p_3 + \tilde{p}_6 + \tilde{p}_9}\right) \\ &< \frac{1}{3N}\left(\sigma_1^2 + \sigma_2^2 + \sigma_3^2\right) \approx \text{Var-RS}. \end{aligned}$$

Hence, our sampler can effectively reduce the variance of the sample mean without increasing the batch size. Moreover, simply increasing the batch size would not work, since the proportion of tail class is so small that we need a huge mini-batch, which cannot fit into the memory.

C.1 Visualization on iNaturalist-LT

We also present visualization results on large scaled iNaturalist-LT dataset to show the effectiveness of our proposed ZCA whitening. As shown in Figure 6, the result in top figure shows that when trained with ERM model, more than 95% of the singular values are smaller than 10. In contrast, we can see that when trained with Weighting-Net, the learned features have more large-valued singular values, implying that the features are effectively decorrelated.

D Ablation Studies on Hyper-parameters

In this section, we provide some ablation studies on the hyper-parameters of GRBS and BET. All the experiments are constructed based on the proposed Whitening-Net on the large scaled iNaturalist-LT dataset.

Table 1 presents all the hyper-parameters of GRBS and BET. In our experiments, S_0 and α are fixed across all the imbalanced datasets. We make batches sampled from GRBS always have a fixed number of categories ($F = 10$), which makes G easy to compute.

Hyper-parameters of GRBS. The hyperpameters of GRBS include group number G, minimum sampling ratio r_0. As shown in the Table 2, the performances under different hyper-parameter G of GRBS are similar, i.e., the proposed Whitening-Net is not sensitive to the choice of hyper-parameters. At the same time, we can control the classification accuracy of different shots by selecting different group number G, e.g., small group number G means that the class imbalance in

Table 1: The hyper-parameters of GRBS and BET on different datasets. G is the group number, r_0 is the basic sampling probability, T is the iteration interval for BET.

Hyper-parameter	G	r_0	S_0	α	T
CIFAT-10-LT	1	0.05	10	2	60
CIFAT-100-LT	10	0.01	10	2	30
ImageNet-LT	100	0.01	10	2	60
iNaturalist-LT	815	0.01	10	2	200

Table 2: Top 1 accuracy by varying group number G on iNaturalist-LT dataset.

Group number #G	Many	Medium	Few	All
200	47.2	53.3	55.8	53.0
400	48.4	53.1	55.0	53.1
800	49.3	53.4	53.8	**53.2**
1000	48.6	52.8	54.2	53.0

Table 3: Top 1 accuracy by varying minimum sampling ratio r_0 on iNaturalist-LT dataset.

Sampling ratio #r_0	Many	Medium	Few	All
0.002	48.8	53.3	53.4	52.8
0.01	49.3	53.4	53.8	**53.2**
0.02	48.9	53.3	54.9	53.1
0.03	48.3	53.4	55.6	53.1
0.04	46.9	53.2	56.7	53.0

Table 4: Top 1 accuracy by varying iteration interval T ($G = 800, r_0 = 0.01$) on iNaturalist-LT dataset.

Iteration interval #T	Many	Medium	Few	All
100	49.1	53.4	53.9	53.0
200	49.3	53.4	53.8	**53.2**
300	49.4	53.4	53.6	53.1

each group is more serious, and more tail classes will be sampled to alleviate the imbalance in each group, because we fix the minimum sampling ratio r_0 of the tail classes. The experimental results in Table 3 also demonstrate that the minimum sampling ratio r_0 can control the classification accuracy of the samples in each shot, and a larger r_0 will enhance the model's ability to recognize tail classes.

Hyper-parameters of BET. The parameter of iteration interval T denotes that in each epoch, the samples in the proposed GRBS participate in training after T iterations of random sampler, i.e., samll T means that the samples in GRBS participate in more training to augment the representation learning of tail class. The experimental results illustrated in Table 4 demonstrate that the proposed BET training strategy is robust to hyper-parameter T.

References

Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 103–112, 2019.

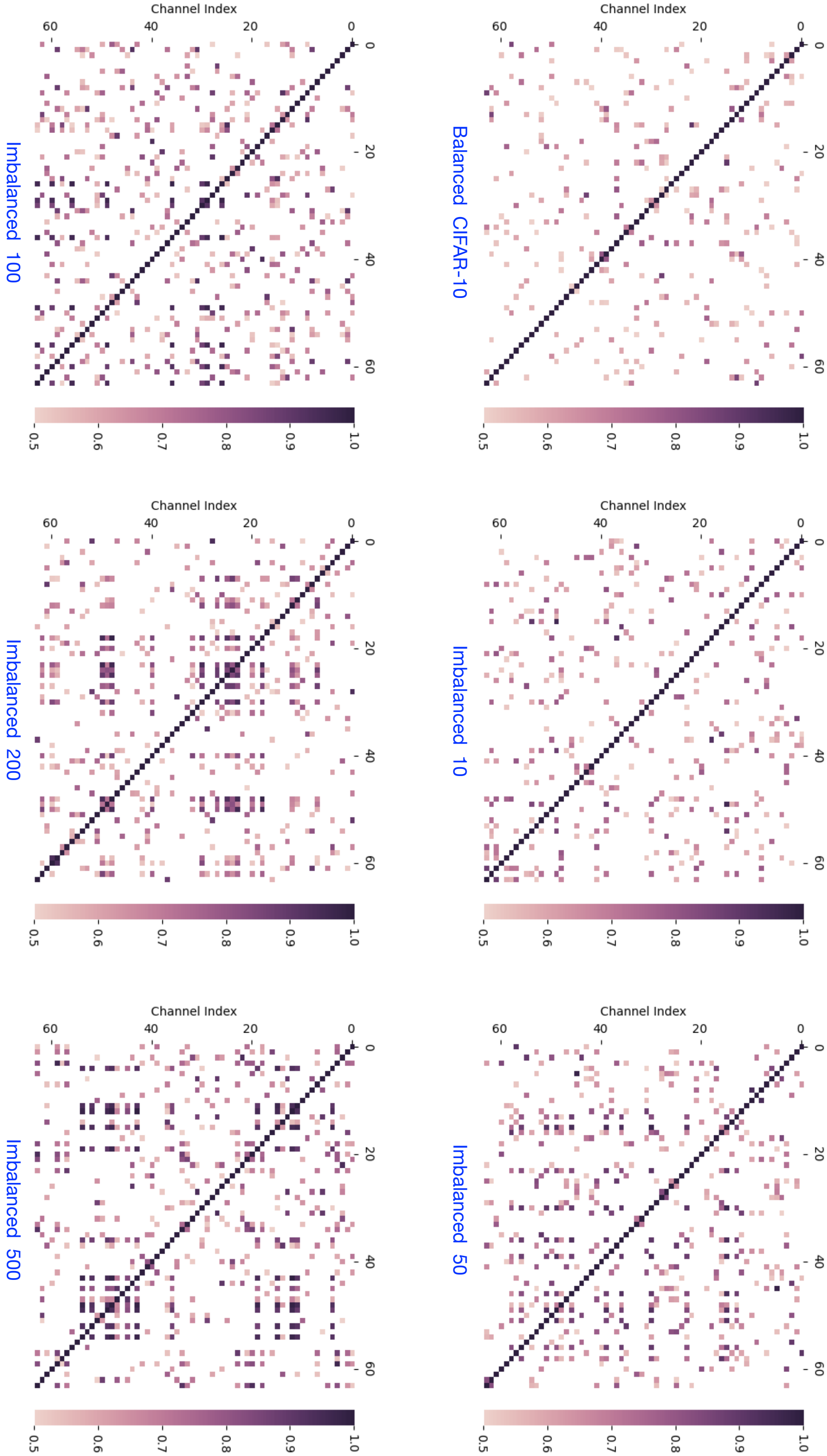


Figure 3: The correlation coefficients between channel-wised features fed into the classifier at the last epochs. The results are obtained on CIFAR-10-LT dataset using ResNet-110.

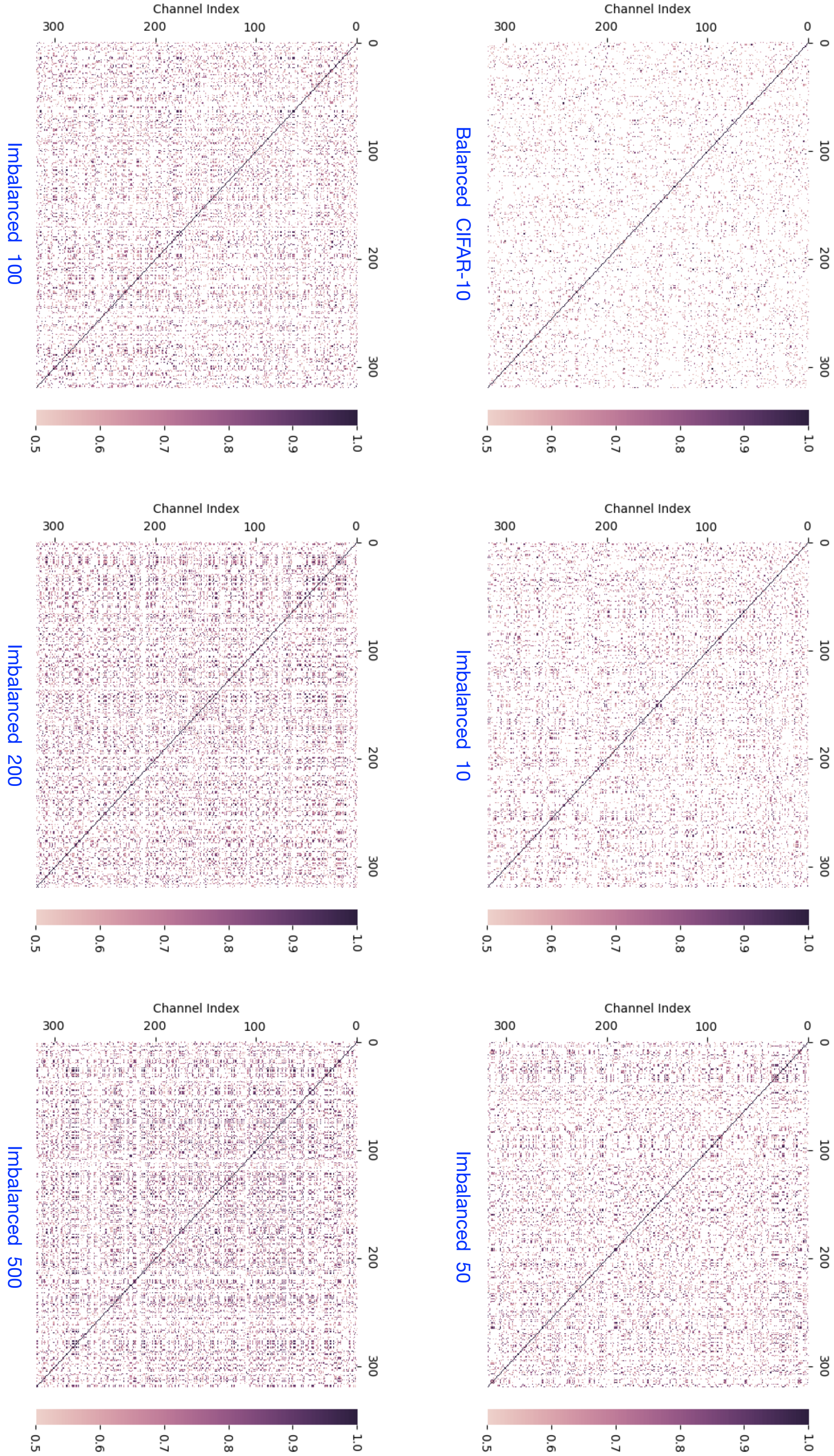


Figure 4: The correlation coefficients between channel-wised features fed into the classifier at the last epochs. The results are obtained on CIFAR-10-LT dataset using EfficientNet-B0.

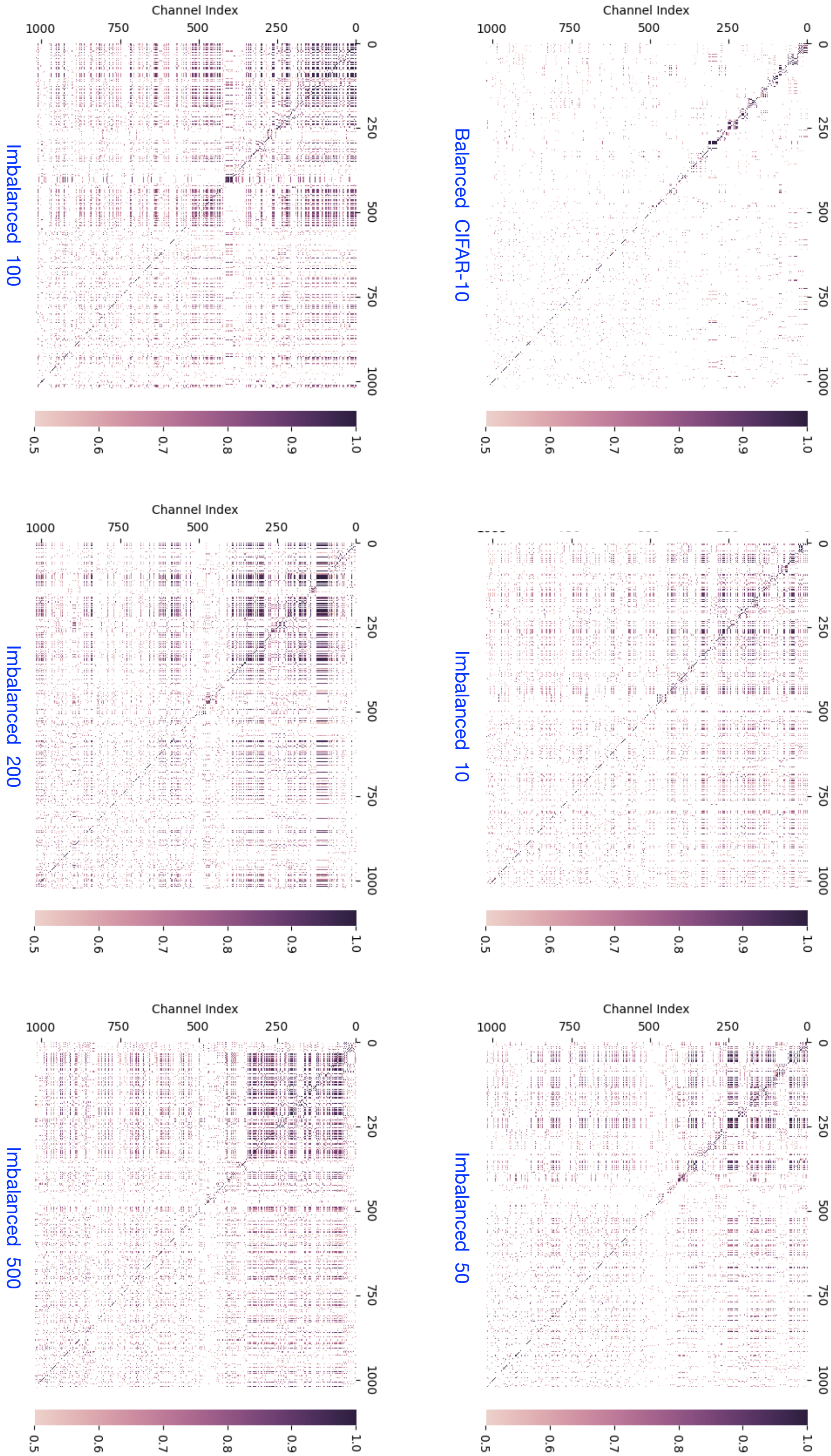
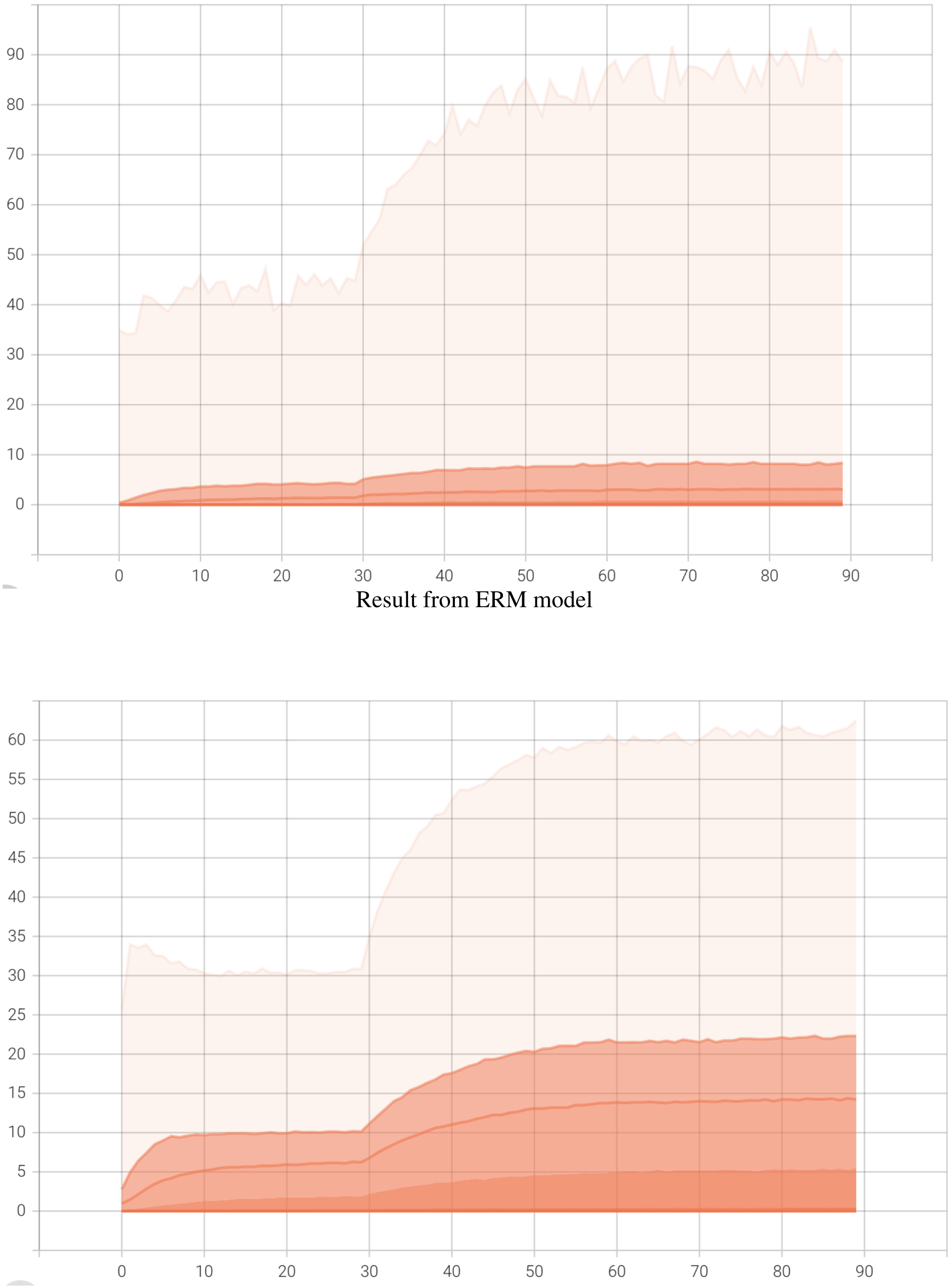


Figure 5: The correlation coefficients between channel-wised features fed into the classifier at the last epochs. The results are obtained on CIFAR-10-LT dataset using DenseNet-121.



Result from ERM model

Result from Whitening–Net

Figure 6: Singular value distributions of features fed into the classifier. The experiments are constructed on iNaturalist-LT dataset using ResNet-10. x-axis stands for number of epoch, y-axis is the singular value. The curves from the top to the bottom are maximum-value, 99.7% quantile, 95% quantile, 68% quantile and the minimum-value, respectively. The result in top figure shows that when trained with ERM, more than 95% of the singular values are smaller than 10. **In contrast, we can see that when trained with Weighting-Net, the learned features have more large-valued singular values, implying that the features are effectively decorrelated.**