

SOLVING THE LONG-TAILED PROBLEM VIA INTRA- AND INTER-CATEGORY BALANCE

Renhui Zhang, Tiancheng Lin, Rui Zhang, Yi Xu

Shanghai Jiao Tong University

ABSTRACT

Benchmark datasets for visual recognition assume that data is uniformly distributed, while real-world datasets obey long-tailed distribution. Current approaches handle the long-tailed problem to transform the long-tailed dataset to uniform distribution by re-sampling or re-weighting strategies. These approaches emphasize the tail classes but ignore the hard examples in head classes, which result in performance degradation. In this paper, we propose a novel gradient harmonized mechanism with category-wise adaptive precision to decouple the difficulty and sample size imbalance in the long-tailed problem, which are correspondingly solved via intra- and inter-category balance strategies. Specifically, intra-category balance focuses on the hard examples in each category to optimize the decision boundary, while inter-category balance aims to correct the shift of decision boundary by taking each category as a unit. Extensive experiments demonstrate that the proposed method consistently outperforms other approaches on all the datasets.

Index Terms— long-tailed distribution, image classification, adaptive precision, gradients harmonizing

1. INTRODUCTION

Long-tailed distribution is a common phenomenon in the real world [1]. In a deep learning task, most examples in the dataset belong to a few dominant categories (head classes), while a large number of other categories (tail classes) claim very few examples [2, 3]. In fact, the most popular benchmark datasets for deep learning tasks of visual recognition are nearly uniformly distributed, such as CIFAR [4, 5] and ImageNet ILSVRC 2012 [6, 7]. As a result, standard methods designed for benchmarks perform poorly on long-tailed problems [8] in that the extreme imbalance between head and tail classes leads to the decision boundary bias towards head classes, as shown in Fig. 1(a). In order to remove the bias, current approaches deal with the long-tailed problem with an attempt to transform the long-tailed dataset to uniform distribution by re-sampling [9, 10, 11, 12, 13] or re-weighting strategies [14, 15, 16, 17]. In particular, they assign different weights (based on effectiveness [18] or inverting data presenting frequency [19]), margins (symmetric [17] or asymmetric [20]) and sampling strategies (over-sampling [10] or

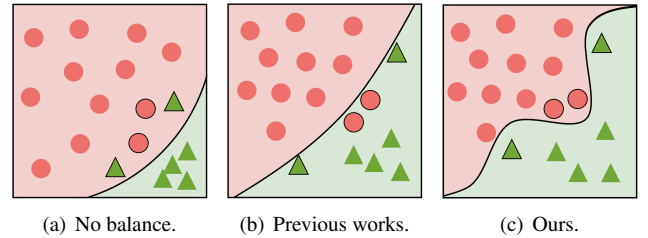


Fig. 1. Red circles and green triangles denote examples from head class and tail class respectively, and hard examples are annotated with black boarder. (a) Decision boundary (the black line) bias towards head classes in long-tailed dataset. (b) Previous works attempt to correct the bias by inter-category balance but ignores the hard examples (red circles or green triangles with black border). (c) Our method corrects the bias problem through emphasizing both hard examples and tail-class by intra- and inter-category balance.

under-sampling [9]) to the examples from different categories, which is called inter-category balance in this paper. However, most of the previous works increase the accuracy of tail classes at the cost of the head classes. It is mainly caused by the under-fitting of hard examples in head classes. As shown in Fig. 1(b), when the decision boundary of the tail class is expanded, the hard examples near the decision boundaries are more likely to be misclassified than the easy counterparts in the head class.

In this work, we point out that the imbalance in long-tailed problems can be summarized as the imbalance in difficulty and sample size. The former one implies that the contributions of easy and hard examples should be different to the model learning process while the latter leads to the bias of the decision boundary. Accordingly, we propose a novel gradient harmonized mechanism with category-wise adaptive precision (GHM-CWAP) to decouple the difficulty and sample size. Correspondingly, intra- and inter-category balance strategies are introduced to calculate the weight of each example, which tackle the imbalance of both difficulty and sample size respectively. Specifically, intra-category balance seeks to rectify the decision boundary by focusing on the hard examples while inter-category balance aims to correct the shift of decision boundary by taking all examples of one category as a unit. By rationally integrating two balance strategies, we can

obtain a better decision boundary as illustrated in Fig. 1(c).

Extensive experiments show that the proposed method achieves the comprehensive balance for the long-tailed problem and consistently outperforms other approaches on all datasets without sacrificing the head classes.

2. RELATED WORK

Intra-category balance aims to harmonize the contributions of hard and easy examples. Previous works, which study example difficulty in terms of loss, assign higher weights to hard examples [21, 22, 23, 24]. Since hard examples are much less than easy examples, these methods couple the sample size and difficulty, which result in that examples from head classes tend to be regarded as easy. Therefore, directly applying these methods to long-tailed problem is unsuitable, as the hard example mining process may be dominated by head classes.

Inter-category balance strategy is commonly used in imbalance problem, and mainly contains two ways, re-sampling [9, 10, 11, 12, 13] and re-weighting [14, 15, 16, 17]. Re-sampling aims to balance the sample size of each category used for training by data or feature augmentation. In contrast, re-weighting aims to harmonize the contribution of examples during training, assigning weight to examples or adding margin to the model output logits adaptively. Above methods attempt to convert the long-tailed dataset to uniform distribution by emphasizing the tail classes. However, they ignore the imbalance of difficulty among the examples during learning process, which result in the performance degradation on head classes.

3. METHOD

In this section, we first briefly review the Gradient harmonized mechanism (GHM) and explain its deficiency in long-tailed problems. Then, we introduce the proposed category-wise GHM to decouple difficulty and sample size in imbalance issues for intra- and inter-category balance strategies respectively. For practical implementation, we further introduce the adaptive unit region approximation to attain a more precise gradient distribution for GHM without increasing the complexity, which results in performance improvement.

3.1. Preliminaries

Gradient or loss are widely used as the criterion for choosing the easy and hard examples [23, 24]. As shown in gradient norm distribution (the black line) in Fig. 2(a), the contributions of the hard examples can be overwhelmed by the huge amount of easy examples. To solve the problem, GHM [24] is proposed to harmonize easy and hard examples by assigning weights proportional to the reciprocal of gradient norm distribution, which is implemented by Unit Region Approximation

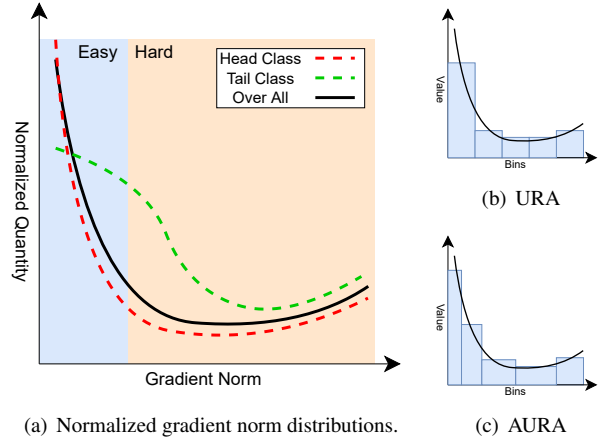


Fig. 2. (a) Head class dominates the gradient norm distribution. (b) and (c) are the approximation to the real gradient norm distribution by URA and AURA respectively. The black line denotes the real distribution and the blue bins denotes the histogram.

(URA) as a trade-off between the computational complexity and performance. However, GHM is not designed for the long-tailed problem. As shown in Fig. 2(a), the overall gradient norm distribution (the black line) would be dominated by head class (the red dotted line), while the distribution of tail class (the green dotted line) would be submerged. This can be easily understood because the sample sizes of tail classes are much smaller than that of head classes.

3.2. Category-Wise GHM

Instead of calculating the overall gradient norm distribution, we propose the category-wise gradient harmonized mechanism (CWGHM) to separate the gradient norm distributions of head classes from tail classes. Essentially, CWGHM decouples the difficulty and sample size, which correspond to intra-category balance and inter-category balance respectively. The former means that CWGHM harmonizes the easy and hard of examples in each category separately, while the latter indicates that CWGHM encodes the difference in sample size among the categories, thus could be further used for inter-category balance.

Formally, we denote the gradient norm distribution of category c as \mathcal{H}_c and suppose it contains N unit regions. In the i th unit region, we count the number of examples and denote it as $v_{c,i}$. Actually, new examples are more likely to be represented by previous examples as the sample size increasing [18]. Therefore, we also introduce the effectiveness factor α to prevent model learning from repeated examples. Then the effective sample size of category c is defined as $S_c = \sum_{i=1}^N (v_{c,i})^\alpha$. To achieve the harmony of easy and hard examples via intra-category balance, the weight $W_{c,i}$ for the

Table 1. Top-1 accuracy (%) of ResNet-32 with various loss function on long-tailed CIFAR-10/100 and TinyImageNet. Imbalance facotr means the ratio of sample size of head classes to tail classes.

Dataset	Long-Tailed CIFAR-10				Long-Tailed CIFAR-100				Long-Tailed TinyImagenet			
Imbalance factor	500	100	10	1	500	100	10	1	500	100	10	1
Softmax	59.76	71.74	86.69	93.00	30.90	38.77	57.21	71.00	31.42	39.06	54.51	63.32
Class Balanced	58.50	73.82	87.18	92.71	31.23	39.12	56.45	70.75	30.90	39.48	54.17	63.06
Focal [23]	57.72	72.43	86.70	92.66	30.31	38.12	56.12	70.13	30.93	39.23	54.00	63.22
GHM [24]	41.50	51.66	67.87	79.35	12.85	14.02	24.49	34.04	8.00	8.03	9.92	10.13
Effective Number [18]	57.97	71.47	87.13	92.98	30.40	39.32	57.37	70.75	30.90	39.48	54.17	63.06
Equalization [19]	58.29	72.61	87.08	93.04	30.82	39.84	58.14	71.66	29.63	37.39	52.75	62.84
Seesaw [20]	64.48	75.18	87.76	93.03	33.63	40.87	57.83	71.35	34.07	40.98	54.60	63.05
ours	66.13	75.39	87.29	92.75	33.85	41.59	57.81	71.41	34.79	42.66	55.35	63.21

examples in the i th unit region of category c is formulated as:

$$W_{c,i} = \frac{1}{v_{c,i}^\alpha / S_c} = \frac{S_c}{v_{c,i}^\alpha} \quad (1)$$

In addition, we also deploy an asymmetric margin strategy based on the effective sample size of each category to correct the decision boundary shift, which is the inter-category balance. Suppose an example from category m , $M_{m,n}$ denotes the margin added to the n th output logit z_n of model, which is the probability that the example belongs to category n .

$$M_{m,n} = \gamma \log[\min(1, S_n/S_m)], \quad (2)$$

where γ is a hyperparameter to adjust the value of margin, For the influence of head class on tail class, it will be reduced as $S_n/S_m < 1$ and $M_{m,n} < 0$. In contrast, the influence of tail classes on head classes stays the same since $S_n/S_m > 1$ and $M_{m,n} = 0$. Therefore, the head classes are protected from being influenced while the tail classes are emphasized.

Overall, the loss function can be formulated as (3), in which C denotes all categories and y_m denotes the ground truth,

$$\mathcal{L} = - \sum_{m \in C} \left[W_{m,i} \cdot y_m \log \left(\frac{e^{z_m}}{\sum_{n \in C} e^{z_n + M_{m,n}}} \right) \right] \quad (3)$$

3.3. Adaptive Unit Region Approximation

Since the performance of GHM is dependent on the approximation precision, the adaptive unit region approximation (AURA) is proposed to improve the approximation precision. As shown in figure 2(b), a number of unit regions (blue bins) with equal width are employed to approach the real distribution (black line) in URA. In fact, the real gradient norm distribution should be fitted with narrower bins at the parts of higher steepness to obtain a rather higher quantization accuracy, while wider bins are used to fit the parts of larger smoothness to reduce the computational complexity, as indicated in Fig. 2(c).

Algorithm 1 Adaptive Unit Region Approximation

Initialize \mathcal{H}_c with $d_i = 1/N$ (the width of unit region i) and $v_i = 0$ (the value of unit region i) for $i = 1, 2, 3, \dots, N$;
for each epoch do
 for each example do
 Calculate the gradient norm g ;
 Find which unit region that g lies in;
 Update the value of the unit region, $v_i \leftarrow v_i + 1/(N * d_i)$;
 end for
 Reassign the width of each unit region, $d_i \leftarrow 1/\log(v_i)$;
 Normalize the width of all unit region to 1;
 Reset $v_{c,i}$ to 0;
end for

Given gradient norm distribution \mathcal{H}_c , we denote the width and value of the i th unit region in \mathcal{H}_c as $d_{c,i}$ and $v_{c,i}$ respectively. Then \mathcal{H}_c is updated following Algorithm 1. All unit regions are initialized with the same width $d_{c,i} = 1/N$ and $v_{c,i} = 0$. Then \mathcal{H}_c will count the gradient norm of every example during previous epoch. After each epoch, the width of each unit region is reassigned according to the gradient norm distribution counted in the last epoch.

With AURA, we can approximate the gradient norm distribution with a higher precision, which result in the performance improvement of the proposed method.

4. EXPERIMENTS

4.1. Experimental Settings

Our experiments are performed on CIFAR [4, 5], TinyImageNet¹, ImageNet ILSVRC 2012 [6, 7] and iNaturalist dataset. Following [18], long-tailed versions of these datasets with varied imbalance factors, which means the ratio of the

¹The dataset can be downloaded from <https://www.kaggle.com/c/thu-deep-learning/data>

Table 2. Top-1 accuracy (%) of ResNet-50 on long-tailed ImageNet and iNaturalist2017 dataset. Categories are divided into four groups and sample size gradually decreases from “Many” to “Rare”.

	Long-Tailed ImageNet(imbalance factor=100)					iNaturalist2017(imbalance factor=435)				
	Many	Medium	Few	Rare	Over All	Many	Medium	Few	Rare	Over All
Softmax	63.91	48.07	22.15	12.55	36.67	84.92	69.88	50.21	40.63	54.64
Class Balance	63.30	47.71	22.26	12.63	36.48	85.29	70.66	50.88	41.31	55.31
Focal[23]	62.70	47.70	23.42	12.86	36.67	84.41	70.01	51.96	41.13	55.15
EffectiveNumber[18]	63.86	47.91	23.42	13.70	37.22	87.77	72.96	54.45	44.49	58.25
Seesaw[20]	63.51	48.69	24.37	15.52	38.02	87.48	73.60	55.64	46.78	59.15
ours	63.61	48.73	25.78	16.95	38.77	87.84	73.59	56.75	48.02	59.85

Table 3. Top-1 accuracy (%) of ResNet-50 on long-tailed ImageNet. “R” represents intra-category balance, “A^{URA}” and “A^{AURA}” represent inter-category balance implemented by URA and AURA respectively.

	ImageNet-LT(imbalance factor=100)				
	Many	Medium	Few	Rare	Over All
Softmax	63.91	48.07	22.15	12.55	36.67
R	63.51	48.69	24.37	15.52	38.02
A ^{URA}	63.93	48.32	22.60	13.44	37.07
R+A ^{URA}	63.55	48.67	25.16	16.30	38.42
R+A ^{AURA}	63.61	48.73	25.78	16.95	38.77

sample size of head classes to tail classes, are constructed by randomly removing examples. Besides, iNaturalist dataset is a long-tailed large-scale real-world dataset, whose imbalance factor is 435. With the above datasets, extensive studies are conducted to demonstrate the effectiveness of the proposed method. For our method, the best hyper-parameter is chosen via cross-validation, $\alpha = 0.9$. Besides, $N = 30$ and $\gamma = 0.8$, which are following [24] and [20]. For other methods, we use the hyper-parameters provided by the authors.

4.2. Experimental Results

Extensive studies on long-tailed CIFAR-10/100 and TinyImageNet datasets with varied imbalance factors are conducted with ResNet-32. Table 1 shows the performance of ResNet-32 using various methods in terms of classification accuracy. We present results of using softmax cross-entropy loss, class balance loss, which assigns weights to each classes according to the inverse of sample size, focal loss [23], GHMc loss [24], effective number loss [18], Equalization loss [19], Seesaw loss [20], and our proposed method.

From the results in Table 1, we have the following observations: (1) Our method achieves the best performance on datasets with large imbalance factors. (2) For balanced datasets, whose imbalance factor is 1, our method has almost the same performance as softmax cross-entropy loss. (3) The original form of GHM is not suitable for the multi-class clas-

sification task.

We also present results on ImageNet-LT and iNaturalist2017 dataset to demonstrate the performance on large-scale real-world datasets. Table 2 summarizes the top-1 accuracy on ImageNet-LT and iNaturalist2017 datasets. Our method achieves the best performance on large-scale real-world datasets, and the gain mainly comes from tail classes while keeping the performance on the head class without sacrificing head classes. Still, our method obtains a more significant gain on iNaturalist2017 than ImageNet-LT, even with a larger imbalance factor.

4.3. Ablation Studies

Table 3 shows the ablation study of each module of the proposed method on Long-Tailed ImageNet. It can be seen that each module can improve the performance separately. Intra-category balance enhances the classification performance on both head classes and tail classes. As a complement, inter-category balance imposes a significant improvement on the tail classes. Furthermore, the proposed AURA achieves superior performance to URA as AURA provides a better approximation to the real distribution of gradient norm distribution. Based on all these modules, our method achieves the overall best performance.

5. CONCLUSION

In this paper, we propose a novel gradient harmonized mechanism with category-wise adaptive precision to decouple the difficulty and sample size in long-tailed problems, corresponding to the intra- and inter-category balance strategies. Intra-category balance emphasizes hard examples in each category to optimize the decision boundary, while inter-category balance aims to correct the shift of decision boundary by regarding all examples of one category as a unit. Besides, adaptive unit region approximation is proposed to improve the performance further. It demonstrates that the proposed method achieves the comprehensive balance for the long-tailed problem and outperforms other approaches on all datasets.

6. REFERENCES

- [1] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie, “The inaturalist species classification and detection dataset,” in *CVPR*, 2018, pp. 8769–8778.
- [2] Grant Van Horn and Pietro Perona, “The devil is in the tails: Fine-grained classification in the wild,” *arXiv preprint arXiv:1709.01450*, 2017.
- [3] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu, “Large-scale long-tailed recognition in an open world,” in *CVPR*, 2019, pp. 2537–2546.
- [4] Alex Krizhevsky, Geoffrey Hinton, et al., “Learning multiple layers of features from tiny images,” 2009.
- [5] Antonio Torralba, Rob Fergus, and William T Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR. Ieee*, 2009, pp. 248–255.
- [7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [8] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, pp. 249–259, 2018.
- [9] Miroslav Kubat, Stan Matwin, et al., “Addressing the curse of imbalanced training sets: one-sided selection,” in *Icml. Citeseer*, 1997, vol. 97, pp. 179–186.
- [10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [11] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling, “Feature space augmentation for long-tailed data,” in *ECCV*. 2020, vol. 12374, pp. 694–710, Springer International Publishing.
- [12] Ren Jiawei, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al., “Balanced meta-softmax for long-tailed visual recognition,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [13] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker, “Feature transfer learning for face recognition with under-represented data,” in *CVPR*, 2019.
- [14] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” in *NIPS2019*. pp. 1567–1578, Curran Associates, Inc.
- [15] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong, “Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective,” in *CVPR*, pp. 7610–7619.
- [16] Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu, “Learning with average top-k loss,” in *Advances in Neural Information Processing Systems*. 2017, vol. 30, Curran Associates, Inc.
- [17] Arya Iranmehr, Hamed Masnadi-Shirazi, and Nuno Vasconcelos, “Cost-sensitive support vector machines,” *Neurocomputing*, vol. 343, pp. 50–64, 2019.
- [18] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, “Class-balanced loss based on effective number of samples,” in *CVPR*, 2019, pp. 9268–9277.
- [19] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan, “Equalization loss for long-tailed object recognition,” in *CVPR*, 2020, pp. 11662–11671.
- [20] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin, “Seesaw loss for long-tailed instance segmentation,” in *CVPR*, 2021, pp. 9695–9704.
- [21] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros, “Ensemble of exemplar-svms for object detection and beyond,” in *ICCV. IEEE*, 2011, pp. 89–96.
- [22] Qi Dong, Shaogang Gong, and Xiatian Zhu, “Class rectification hard mining for imbalanced deep learning,” in *ICCV*, 2017, pp. 1851–1860.
- [23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in *ICCV2017*, 2017, pp. 2980–2988.
- [24] Buyu Li, Yu Liu, and Xiaogang Wang, “Gradient harmonized single-stage detector,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8577–8584.