Distribution-Balanced Loss for Multi-Label Classification in Long-Tailed Datasets

Tong Wu[1 [0000−0001−5557−0623]], Qingqiu Huang[2 [0000−0002−6467−1634]], Ziwei Liu[2 [0000−0002−4220−5958]], Yu Wang[1 [0000−0001−6108−5157]], and Dahua Lin[2 [0000−0002−8865−7896]]

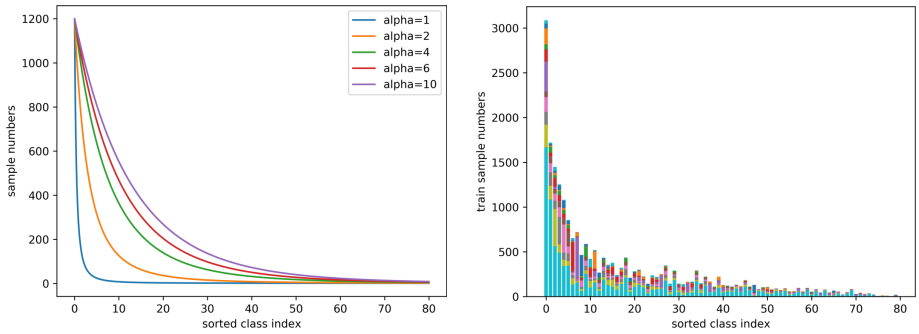
[1] Tsinghua University, Beijing, China
wutong16.thu@gmail.com, yu-wang@mail.tsinghua.edu.cn
[2] The Chinese University of Hong Kong, Hong Kong, China
zwliu.hust@gmail.com, {hq016,dhlin}@ie.cuhk.edu.hk

1 Dataset Construction

We create our multi-label long-tailed datasets by extracting subsets from two multi-label image recognition benchmarks, VOC and COCO, respectively. We adopt a *pareto distribution* $pdf(x) = \alpha \frac{x_{min}^{\alpha}}{x^{\alpha+1}}$ following Liu *et al* [4] with α controlling the shape of the distribution, as shown in Fig. 1a. Concretely, we cut off the probability distribution function(pdf) when the cumulative distribution function(CDF) achieves 0.99, and then we rescale the pdf with a maximum of N_{max}, which is the maximum of sample numbers per class. Finally, we evenly split the x-axis into the number of classes as the original dataset, and we get a reference distribution. We construct the datasets in a head-to-tail manner: we first rank all the classes by $\hat{p}_i$ in Eq.1 mentioned in the main paper calculated with original data, and the subset is empty. For each class i from head to tail, we compare the current sample number in the subset and the expected sample number by the reference distribution and then randomly *add* or *eliminate* certain instances accordingly. This way, we can constrain the tail classes to have a relatively small amount of data. As seen in Fig. 1b, the construction is processed incrementally. The distribution of the test set has a similar ranking order as the constructed train set, as shown in Fig. 2. Except for one class, "person", the rest part of the test set is only slightly imbalanced.

2 Implementation Details of Comparing Methods

Some of the comparing methods are designed mainly for single-label datasets, and we make slight adjustments so that they work the best with our datasets: For class-balanced(CB) loss [3], we set $\beta = 0.99$ and 0.9 for VOC-MLT and COCO-MLT, respectively, and we use an initial leanrning rate of 0.1, with an extra loss weight of 10 because we used an *average* manner in the loss reduction while [3] used *sum*; To calculate the effective numbers of a label set with multiple ground-truth, we adopt an average of E_{n_i} calculated from each positive class, $\bar{E_n} = \frac{1}{\sum y_i} \sum_{i,y_i>0} (1-\beta^{n_i})/(1-\beta)$. We find it works better than



(a) **(b)**

Fig. 1: (a). Pareto distribution with different α. (b). The construction of COCO-MLT, where the colors visualize the increment of sample number per-class after each sampling iteration

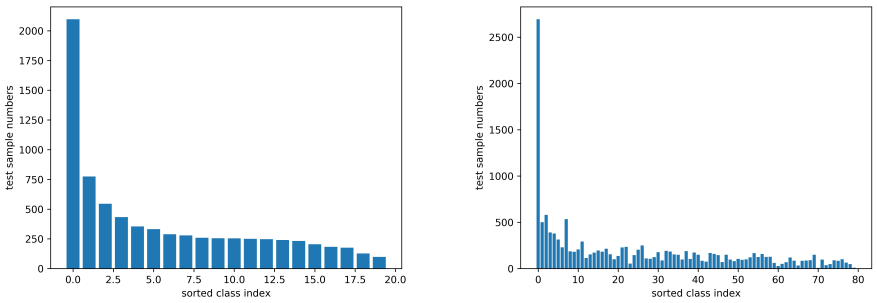
$E_{\bar{n}} = (1-\beta^{\bar{n}})/(1-\beta)$, where $\bar{n} = \frac{1}{\sum y_i}\sum_{i,y_i>0} n_i$. For ML-GCN [2], the dimension of the hidden layer in GCN is set as 256 and we use adjacent matrix generated from the long-tailed versions of datasets. We also experiment with adjacent matrix generated from the original datasets so that it better matches the distribution of the test set, while the results show little difference. For LDAM, we adopt a class-level margin, following [1] that we tune C in $\frac{C}{n_i^{1/4}}$ and normalize the largest margin to be 0.5.

3 The Effect of μ in Smoothing Function.

In our paper, we report how the results are affected by β of a smoothing function in Eq.5. And μ also controls the shape of the function in the actual range of variables. As shown in Fig. 3a, the influence is relatively small so we selected two insignificant peaks where $\mu = 0.2, 0.3$ for COCO-MLT and VOC-MLT, respectively, for the main experiments.

4 The Effect of ν in Negative-tolerant Regularization

To under stand how ν the netagive-tolerant regularization in Eq.12 in affect the results independently, we fix $\lambda = 2, 5$ for COCO-MLT and VOC-MLT, respectively, and change ν by changing κ, as shown in Fig. 3b. Setting $\kappa = 0$ has a relatively good result, which means that the thresholds in the regularization can be simply fixed as zero, and changing κ in a small range has little effect.

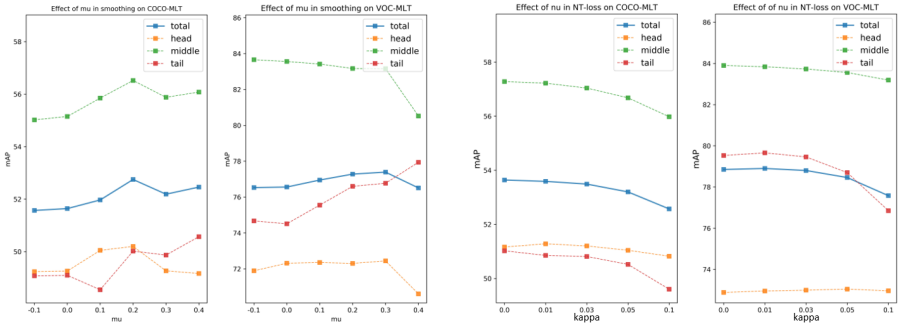


(a) **(b)**

Fig. 2: (a). The test set distribution of COCO2017, and we use the sorted sample number of the long-tailed training set as the x-axis index, which is relatively balanced except for one class(person). (b). The test set distribution of VOC2007

References

1. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems. pp. 1565–1576 (2019) 2
2. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2019) 2
3. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019) 1
4. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1



(a) **(b)**

Fig. 3: (a). The effect of μ in the smoothing fuction. (b). The effect of κ in netagive-tolerant regularization