Probability Guided Loss for Long-Tailed Multi-Label Image Classification

Dekun Lin

Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, China
University of Chinese Academy of Sciences, Beijing 100049, China
kunonkey@163.com

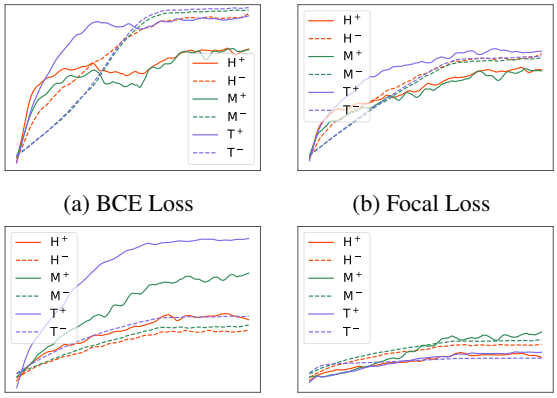
Abstract

Long-tailed learning has attracted increasing attention in very recent years. Long-tailed multi-label image classification is one subtask and remains challenging and poorly researched. In this paper, we provide a fresh perspective from probability to tackle this problem. More specifically, we find that existing cost-sensitive learning methods for long-tailed multi-label classification will affect the predicted probability of positive and negative labels in varying degrees during training, and different processes of probability will affect the final performance in turn. We thus propose a probability guided loss which contains two components to control this process. One is the probability re-balancing which can flexibly adjust the process of training probability. And the other is the adaptive probability-aware focal which can further reduce the probability gap between positive and negative labels. We conduct extensive experiments on two long-tailed multi-label image classification datasets: VOC-LT and COCO-LT. The results demonstrate the rationality and superiority of our strategy.

Introduction

Deep convolutional neural networks (CNNs) have achieved extraordinary success on computer vision so far (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; He et al. 2016), which is largely owing to the artificial balanced datasets (e.g., CIFAR (Krizhevsky 2009), ImageNet ILSVRC (Deng et al. 2009) and MS COCO (Lin et al. 2014)). However, real-world data often exhibits long-tailed distribution. Namely, a few dominant categories have the majority of samples while most of the classes are instance-scarce. Such a universal phenomenon poses severe challenges to a number of computer vision tasks, e.g., image classification (Cui et al. 2019; Zhou et al. 2020), object detection (Tan et al. 2020; Li et al. 2020) and instance segmentation (Hsieh et al. 2021; Wang et al. 2021). Models are prone to overfit on minority classes and suffer from performance decline dramatically.

There are multiple objects in a daily captured image typically. Multi-label classification (MLC) is to identify these objects concurrently. Compared to classic single label classification (SLC) such as ImageNet ILSVRC 2012, the MLC is undoubtedly more practical and challenging. Long-tailed multi-label classification (LTMLC) which combines the characteristics of long-tailed learning and MLC is underestimated to date.



(a) BCE Loss (b) Focal Loss

(c) DB Loss (d) Our PG Loss

Figure 1: Predicted mean probability curves by different cost-sensitive methods on long-tailed multi-label classification dataset VOC-LT during training. H denotes the head classes, M and T indicate the medium and tail classes, respectively. Both the superscript + and the solid line denote the probability of positive labels, while subscript − and the dotted line represent the probability of negative labels.

Cost-sensitive learning methods (Elkan 2001; Zhou and Liu 2005) adjust loss values for various classes during training for re-balancing. In contrast with complicated model design, such approaches are more simple, interpretable and may be even more powerful. CB loss (Cui et al. 2019) calculates class-wise effective samples to re-weight loss and achieve nice performance on SLC. Ridnik et al. who are aware of the imbalance between positive and negative labels in a multi-label setting, i.e., the number of negative labels is much larger than positive labels, propose the ASL loss (Ridnik et al. 2021) to slow down the learning of massive easy negative labels on MLC. Similarly, Wu et al. put forward the DB loss (Wu et al. 2020) which focuses more on the learning of negative labels during training and make remarkable advance in LTMLC.

Long-tailed datasets can be split into head classes with many samples, tail classes with few samples and medium classes between them. We visualize the mean probability of both positive and negative labels for these three splitting classes on LTMLC by several representative cost-sensitive methods during training as is shown in Fig. 1. For brevity, we denote the probability of positive labels as positive probability and term the probability of negative labels as negative probability. By the naive binary-cross-entropy (BCE) loss, both the positive and negative probabilities of each category increase to a big value rapidly. Besides, the probability gap between positive and negative labels is somewhat large during training. While with the focal loss, the probability gap is significantly reduced and the probability rises at a decreasing rate. The negative probability gets a tremendous suppression by DB loss who has considered the positive-negative imbalance between labels.

The above mentioned approaches do have made progress, we hence can have following hypotheses naturally: 1) slowing down the growth of probability who increases fast helps to improve performance, scilicet, probability who increases too fast may suffer from overfitting; 2) it may be beneficial to narrow the probability gap between positive and negative labels. Based on the above two points, we can guess that the remaining problem of focal loss may be that the probability still increases at a relatively fast rate in early training stage. With regard to DB loss, the positive probability for tail classes grows too fast compared to others and the probability gap between positive and negative labels is large.

In this paper, we discover the key factors that affect the probability during training and thereby propose a probability re-balancing scheme which can control the probability change at will. Moreover, we propose the adaptive probability-aware focal to narrow the positive-negative probability gap dynamically. As is depicted in Fig. 1d, the predicted probability increase exceedingly slowly during training and the positive-negative probability gap is quite narrow by our proposed loss. The main contributions of this paper can be summarized as follows:

- We propose to handle the LTMLC from a novel perspective of probability. Our study reveals how the training probability influence the performance of models. Accordingly, we design a probability guided loss to adjust the probability along a right way during training.
- We conduct systematic experiments on two long-tailed multi-label classification datasets, the sound results demonstrate the rationality and validity of our strategy which outperforms previous state-of-the-art.

Related Works

Multi-Label Classification. Most existing methods aim at designing particular models to exploit high-level semantic information for MLC. Traditional approaches like dependency network (Guo and Gu 2011), co-occurrence adjacency matrices (Xue et al. 2011) and conditional graph (Li et al. 2016) explore high order label dependencies. With the popularity of CNNs all over the tasks in computer vision, methods resort to the Recurrent Neural Network (RNN) (Wang et al. 2016, 2017) and Graph Convolutional Network (GCN) (Kipf and Welling 2016) to similarly explore relations between labels. These methods usually require deliberate design and may be hard to optimize. Meanwhile, these approaches cannot have an excellent performance on LTMLC. Based on focal loss (Lin et al. 2017), Ridnik et al. who have been aware to the imbalance between positive and negative labels propose an asymmetric loss named ASL (Ridnik et al. 2021). Unfortunately, such a method cannot generalize well on LTMLC, either.

Long-Tailed Single-Label Classification. As has been common used in long-tailed classification, re-sampling (He and Garcia 2009; Shen, Lin, and Huang 2016; Byrd and Lipton 2019) can alleviate the imbalance problem to some extent. The basic idea is to over-sample the minority categories or to under-sample the frequent categories during training (Han, Wang, and Mao 2005; Buda, Maki, and Mazurowski 2018). Class-aware sampling (Shen, Lin, and Huang 2016) chooses samples of each category with equal probabilities. Re-weighting the loss is another straight strategy. The training loss can be controlled in class-wise level by re-weighting according to the frequency of each category (Huang et al. 2016; Wang, Ramanan, and Hebert 2017; Cui et al. 2019). A sample-level control of loss can also be achieved for different samples (Lin et al. 2017; Ren et al. 2018). Recent studies (Kang et al. 2019; Zhou et al. 2020) decouple the learning of representation and classifier head to improve the prediction. Despite their effectiveness, the performance on LTMLC is not so satisfactory.

Long-tailed Multi-Label Classification. Methods concerning with the MLC are numerous up to now. In contrast, researches on LTMLC are particularly rare. In common with the ASL loss (Ridnik et al. 2021) which is proposed for MLC, Wu et al. (Wu et al. 2020) propose to slow down the optimization rate of negative labels based on binary-cross-entropy. Such a design effectively improves the prediction accuracy on LTMLC. Nevertheless, there is still room for improvement of it.

Approach

We first briefly introduce the Focal Loss (Lin et al. 2017) and the negative-tolerance regularization (NTR) of DB Loss (Wu et al. 2020). Meanwhile, we give corresponding analysis on how they influence the predicted probability during training. Subsequently, we propose the probability re-balancing (PR) to adjust the growth rate of predicted probability for classes who increase too fast in early training stage. Afterwards, we present the adaptive probability-aware focal (APAF) which can adaptively alter the gradient scaling factors in each iteration and eventually effectively narrow the positive-negative probability gap. Finally, we formally define the Probability Guided (PG) loss which is composed of PR and APAF.

NTR and Focal Loss

The negative-tolerance regularization (NTR) is one component of DB loss (Wu et al. 2020). It is built upon the binary-

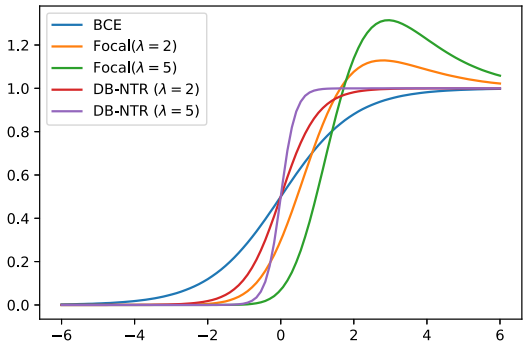


Figure 2: The gradient curves on negative labels by BCE, NTR and Focal Loss. X-axis denotes the logit of negative labels, and the y-axis is the corresponding gradients.

cross-entropy (BCE) loss. The total loss of BCE is given by:

$$L_{BCE}(X,Y) = \sum_{i=1}^{N}\sum_{c=1}^{K} L(p_c^i, y_c^i), \tag{1}$$

where (X,Y) is the training set and N is the number of training samples. The number of classes is K, then we have $y^i = [y_1^i, ..., y_K^i] \in \{0,1\}^K$. Let p_c^i denote the predicted probability of the cth label of sample x^i, it is obtained by sigmoid function. For brevity we denote the predicted probability as p and the ground-truth as y. The loss of each label is given by:

$$L = -y\log(p) - (1-y)\log(1-p). \tag{2}$$

Then the loss of NTR which is based on BCE is given by:

$$L_{NT-BCE} = \begin{cases} \frac{1}{\lambda}\log(1+e^{\lambda(z-v)}), & y=0; \\ \log(1+e^{-(z-v)}), & y=1, \end{cases} \tag{3}$$

where λ is the gradient scaling factor for negative labels. The z is the output logits of classifier and v is the class-specific bias. Such a design can slow down the learning rate of easy negative samples while accelerate the hard ones greatly, alleviate the problem of imbalance between positive and negative labels.

The loss of focal loss is given by:

$$L_{Focal} = \begin{cases} -p^{\lambda}\log(1-p), & y=0; \\ -(1-p)^{\lambda}\log(p), & y=1, \end{cases} \tag{4}$$

where λ is the gradient scaling factor for both positive and negative labels. When $\lambda = 0$, it is the BCE loss; when $\lambda > 0$, it becomes the focal loss. The gradient curves of both NTR and focal loss are depicted in Fig. 2. As can be seen from it clearly, the common idea behind them is to down-weight the learning of easy negative samples while up-weight the hard ones. And this is exactly the key factor that affects the process of predicted probability during training. The bigger the scaling factor λ is, the stronger the probability will be suppressed. The difference lies that NTR which is tailored for LTMLC focuses only on negative labels, while focal is a symmetrical design for both positive and negative labels. Thus the positive-negative probability gap by focal loss is smaller than by NTR.

scaling factor	positive probability	negative probability
λ_1^+	↑	↓
λ_2^+	↓	↑
λ_1^-	↓	↓
λ_2^-	↑	↓

Table 1: The effects of different gradient scaling factors in PR for both positive and negative labels.

Probability Re-Balancing

Our proposed probability re-balancing (PR) is given by:

$$L_{PR-BCE} = \begin{cases} \frac{1}{\lambda_2^-}\log(1+e^{\lambda_1^-(z-v)}), & y=0; \\ \frac{1}{\lambda_2^+}\log(1+e^{-\lambda_1^+(z-v)}), & y=1. \end{cases} \tag{5}$$

It has four gradient scaling factors: λ_1^-, λ_2^- for negative labels and λ_1^+ with λ_2^+ for positive labels respectively. To have a closer look at how these four different factors affect the predicted probability during training, we conduct experiments on two long-tailed datasets: VOC-LT and COCO-LT. Limited by space, we put only results of VOC-LT here as is shown in Fig. 5. We can actually draw the same conclusion by COCO-LT just as VOC-LT without losing generality.

We can see from Fig. 3a and Fig. 3b that when λ_1^+ increases, the predicted positive probability increases while the negative probability decreases. On the contrary, the predicted positive probability decreases and the negative probability increases when λ_2^+ increases as shown in Fig. 3e and Fig. 3f. As for the negative scaling factors, both the positive and negative probability will decrease when increasing λ_1^- as illustrated in Fig. 3c and Fig. 3d. When increasing λ_2^-, the predicted positive probability increases while the negative probability decreases. Such a rule is also shown in Tab. 1 to keep things straight.

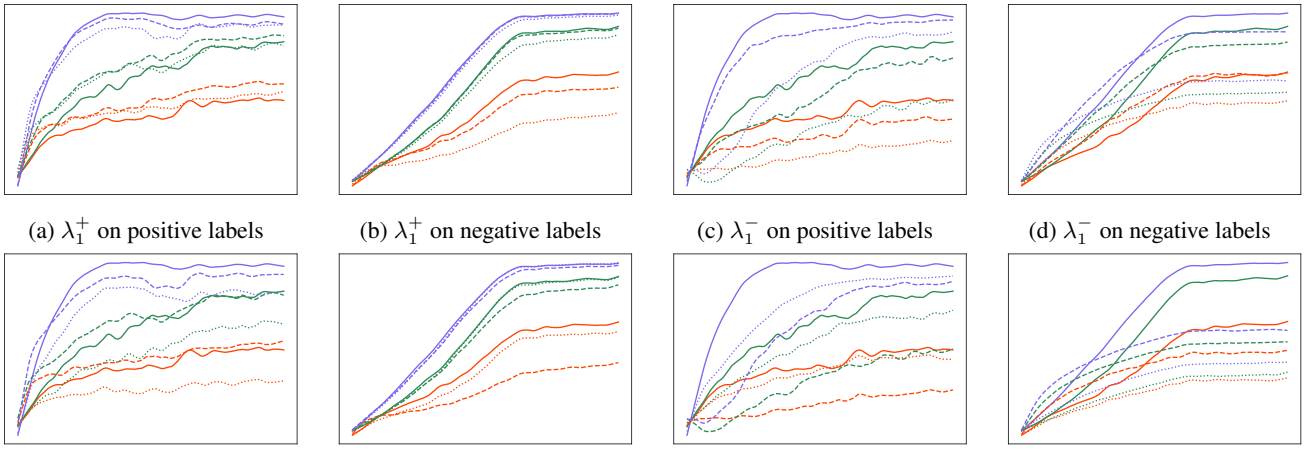
Actually, PR is a generalized version of NTR, or named NTR is a special case of PR. When $\lambda_1^- = \lambda_2^-, \lambda_1^+ = \lambda_2^+ = 1$, it becomes the NTR.

Adaptive Probability-Aware Focal

We propose the adaptive probability-aware focal (APAF) to reduce the positive-negative probability gap in a more effective manner, that is to utilize the class-wise predicted positive and negative probability to generate the gradient scaling factors. Specifically, we first compute the class-wise mean positive probability $\overline{p^+}$ and mean negative probability $\overline{p^-}$ in each iteration. They are given by:

$$\begin{cases} \overline{p^+} = \frac{1}{\sum_{j=1}^{m} y_j}\sum_{j=1}^{m} p_j^+ y_j; \\ \overline{p^-} = \frac{1}{\sum_{j=1}^{m}(1-y_j)}\sum_{j=1}^{m} p_j^-(1-y_j), \end{cases} \tag{6}$$

where y is the ground-truth, p is the predicted probability, the subscript j denotes the jth sample in a batch, the superscript $+$ denotes the positive labels and $-$ the negative labels, m is the number of samples in a batch. Note that the sum of the occurrences of some categories may be 0 in a batch, especially for the tail classes. We cannot



(a) λ_1^+ on positive labels (b) λ_1^+ on negative labels (c) λ_1^- on positive labels (d) λ_1^- on negative labels

(e) λ_2^+ on positive labels (f) λ_2^+ on negative labels (g) λ_2^- on positive labels (h) λ_2^- on negative labels

Figure 3: Predicted mean probability curves on VOC-LT by different gradient scaling factors during training. Red color indicates the head classes, green and blue colors denote the medium and tail classes, respectively. In all subfigures, the solid line denotes that both λ_1 and λ_2 equal to 1. For different subfigures with different scaling factors, (a)-(b) dashed line means $\lambda_1^+ = 2, \lambda_2^+ = 1$, dotted line denotes $\lambda_1^+ = 5, \lambda_2^+ = 1$; (c)-(d) dashed line means $\lambda_1^- = 2, \lambda_2^- = 1$, dotted line denotes $\lambda_1^- = 5, \lambda_2^- = 1$; (e)-(f) dashed line means $\lambda_1^+ = 5, \lambda_2^+ = 1$, dotted line denotes $\lambda_1^+ = 5, \lambda_2^+ = 5$; (g)-(h) dashed line means $\lambda_1^- = 5, \lambda_2^- = 1$, dotted line denotes $\lambda_1^- = 5, \lambda_2^- = 5$.

calculate the mean probability of ith class $\overline{p_i^+}$ at this moment. To tackle it, we split all classes $\{C_1, C_2, \cdots, C_K\}$ into $\{C_{head}, C_{medium}, C_{tail}\}$ according to the number of samples in each category, i.e. $\{C_1, C_2, \cdots, C_K\} = \{C_{head}, C_{medium}, C_{tail}\}$. We then replace this mean probability value by the mean of other classes. These classes represent classes which this class belongs to. For instance, if this class belongs to tail classes $\{C_{tail}\}$, then it is given by:

$$\overline{p_i^+} = \frac{1}{|\{C_{tail}\}|} \sum_j \overline{p_j^+}, i, j \in \{C_{tail}\}, \tag{7}$$

where the subscript i denotes the ith class, both ith and jth class belong to tail classes. $|\{C_{tail}\}|$ denotes the number of tail classes. If $\overline{p_i^+}$ still cannot be calculated after such an operation, that means all the tail classes in this iteration occur 0 times, we then set $\overline{p_i^+} = \overline{p_i^-}$ at this situation. The scaling factors for focal loss are then given by:

$$\begin{cases} \lambda^+ = \lambda_0 \frac{\overline{p^+}}{\overline{p^-}}; \\ \lambda^- = \lambda_0 \frac{\overline{p^-}}{\overline{p^+}}, \end{cases} \tag{8}$$

where λ_0 denotes the initial scaling factor. In general, this factor equals to 2. If the mean positive probability $\overline{p^+}$ is bigger than the negative probability $\overline{p^-}$, then we are to suppress the positive probability with increasing the positive gradient factor λ^+ and vice versa. Finally, we have adaptive probability-aware focal loss:

$$L_{APAF} = \begin{cases} p^{\lambda^-} \log(1-p), & y = 0; \\ (1-p)^{\lambda^+} \log(p), & y = 1, \end{cases} \tag{9}$$

In reality, PR can get scaling factors in an adaptive class-wise probability-aware way as well just as APAF. We term it as adaptive probability re-balancing (APR), which will be further discussed later on experiments part.

Probability Guided Loss

Thus far, we have PR-BCE to adjust the changing rate of probability freely. And the APAF can effectively reduce the positive-negative probability gap as well as reducing the growth rate of probability to some extent. These two strategies can be integrated as a unified probability guided loss for end-to-end training, which is finally given by:

$$L_{PG} = \begin{cases} \frac{1}{\lambda_2^-} p^{\lambda^-} \log(1 + e^{-\lambda_1^-(z-v)}), & y = 0; \\ \frac{1}{\lambda_2^+} (1-p)^{\lambda^+} \log(1 + e^{\lambda_1^+(z-v)}), & y = 1. \end{cases} \tag{10}$$

Experiments

Datasets

Just as Wu et al. (Wu et al. 2020), we conduct experiments on two long-tailed multi-label classification datasets: VOC-LT and COCO-LT. They are artificially constructed from Pascal Visual Object Classes Challenge (VOC) (Everingham et al. 2015) and MS-COCO (Lin et al. 2014), respectively.

VOC-LT: The VOC-LT is sampled from the 2012 train-val set of VOC (Everingham et al. 2015) based on the Pareto distribution which is the same as Liu et al. (Liu et al. 2019). There are 1142 images and 20 class in training set. The number of each class ranges from 4 to 775. These 20 classes are split into three groups according to the number of samples per class: a head class contains more than 100 samples, a medium class has 20 to 100 samples, and a tail class has less than 20 samples. The ratio of head, medium and tail classes

Datasets	VOC-LT				COCO-LT			
Methods	total	head	medium	tail	total	head	medium	tail
ERM	70.86	68.91	80.20	65.31	41.27	48.48	49.06	24.25
RM	74.70	67.58	82.81	73.96	42.27	48.62	45.80	32.02
Focal Loss (Lin et al. 2017)	73.88	69.41	81.43	71.56	49.46	49.80	54.77	42.14
RS (Shen, Lin, and Huang 2016)	75.38	70.95	82.94	73.05	46.97	47.58	50.55	41.70
ML-GCN (Chen et al. 2019)	68.92	70.14	76.41	62.39	44.24	44.04	48.36	38.96
LDAM (Cao et al. 2019)	70.73	68.73	80.38	69.09	40.53	48.77	48.38	22.92
CB Focal (Cui et al. 2019)	75.24	70.30	83.53	72.74	49.06	47.91	53.01	44.85
ASL Loss (Ridnik et al. 2021)	76.40	70.70	82.26	76.29	50.21	49.05	53.65	46.68
DB Focal (Wu et al. 2020)	78.29	72.67	83.17	78.75	53.45	50.91	56.58	51.52
Ours	**80.37**	**73.67**	**83.83**	**82.88**	**54.43**	**51.23**	**57.42**	**53.40**

Table 2: mAP performance of our proposed method and other comparison methods on LTMLC. The result of DB loss is re-trained by us, the other results are taken from Wu et al. (Wu et al. 2020).

is 6:6:8 after such splitting. The testing set contains 4952 images which is identical to the 2007 test set of VOC.

COCO-LT: The COCO-LT is sampled from the 2017 version of MS-COCO (Lin et al. 2014) in a similar way. There are 1909 images and 80 classes in the training set. The number of each class ranges from 6 to 1128. The splitting of classes is similar to VOC-LT. The ratio of head, medium and tail classes is 22:33:25. The testing set contains 5000 images which is identical to the 2017 testing set of MS-COCO.

Implementation Details

We use the mean average precision (mAP) metrics to evaluate the performance of methods for LTMLC just as DB loss (Wu et al. 2020). For fair comparison, we use configurations similar to DB loss. More specifically, We use the ResNet50 (He et al. 2016) which is pre-trained on ImageNet (Deng et al. 2009) as the backbone of model. The input images are all resized to a dimension of 224×224 and the batch size is 32. We adopt standard data augmentations the same as DB loss. The class-aware re-sampling (Shen, Lin, and Huang 2016) is applied the same as previous works (Shen, Lin, and Huang 2016; Cui et al. 2019). The optimizer we take is SGD whose momentum is 0.9 and weight-decay is 1e-4. The initial lr is 8e-3 for VOC-LT and 1e-2 for COCO-LT. We also use warm-up learning rate schedule (Goyal et al. 2017) for the first 500 iterations with a ratio of $\frac{1}{3}$. All the codes of DB loss are took out and retrained by us. We conduct all experiments on PyTorch1.8.0.

Comparisons

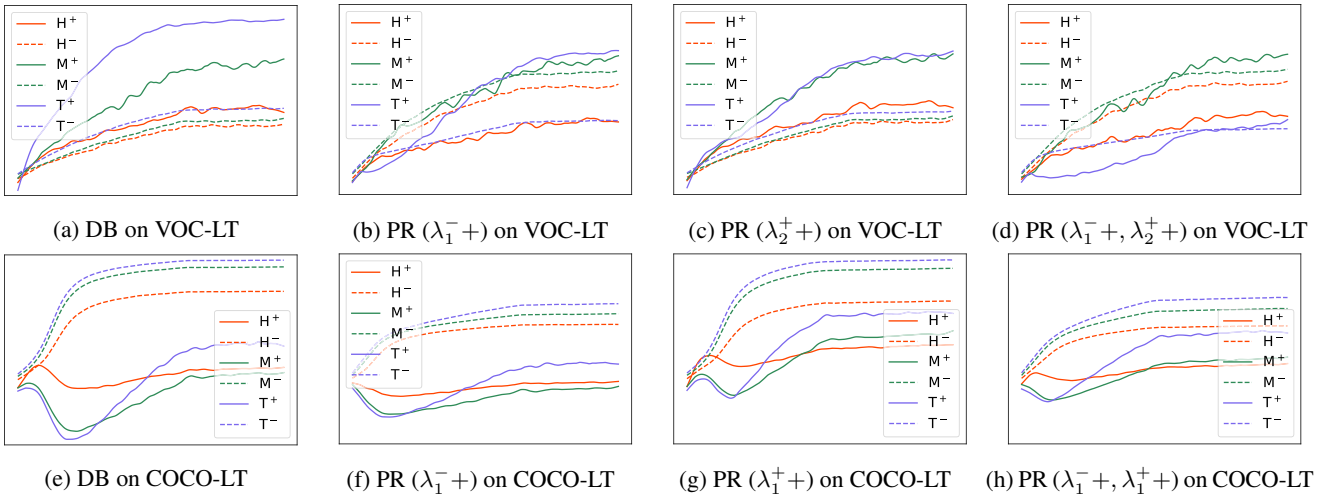
First of all, we compare the mAP performance between our method and previous methods on long-tailed datasets to verify the effectiveness of our proposed method. These comparison methods contain Empirical Risk Minimization (ERM), Re-Weighting (RW) which re-weights by the inverse proportion of the square root of class frequency, Re-Sampling (RS) (Shen, Lin, and Huang 2016), Focal Loss (Lin et al. 2017), ML-GCN (Chen et al. 2019), LDAM (Cao et al. 2019), CB Focal (Cui et al. 2019), DB Loss (Wu et al. 2020) and ASL Loss (Ridnik et al. 2021). The mAP performance of different methods are shown in Tab. 2. The experiment results show that our proposed method outperforms previous methods and has a significant advance. Specifically, our proposed PG loss achieves the best results of 80.37% and 54.43% total mAP scores on VOC-LT and COCO-LT, respectively. It can have about 2.1% and 1% total mAP performance gain on VOC-LT and COCO-LT compared with previous SOTA method DB Focal loss. Meanwhile, our results on three subset classes, i.e. head, medium and tail classes, are the best, too. The main performance gain arises in tail classes at an astonishing 4% mAP on VOC-LT. With regard to COCO-LT, performance advance mainly occurs in both medium and tail classes.

Ablation Study

To better observe how our proposed PR and APAF affect the probability and performance, we conduct ablation analyses on them, respectively.

Ablation Analysis on PR In this section, we elaborate on how we conduct our PR strategy according to the probability map. Our experimental results demonstrate our proposed PR can significantly have a positive effect on the predicted probability and therefore achieve great advance.

Fig. 4a shows the mean predicted probability of different splitting classes on VOC-LT by DB loss during training. The positive probability of tail classes grows rapidly compared with other classes, despite probability of other classes has been suppressed well. Hence, our goal is to re-balance the positive probability of tail classes, namely to suppress it. Under the pattern shown in Tab. 1, we here choose to increase λ_1^- and λ_2^+ for tail classes alone. When only increasing λ_1^-, the positive probability of tail classes gets a certain suppression as illustrated in Fig. 4b. Likewise, the positive probability of tail classes gets squeezed to some extent when merely increasing λ_2^+ as is shown in Fig. 4c. With increasing λ_1^- and λ_2^+ concurrently, this probability is hence greatly suppressed as is depicted in Fig. 4d. We can see from Fig.4e that completely different from results on VOC-LT, the negative probability of all three splitting classes remains grow fast and have a big gap with positive probability when adopting DB loss on COCO-LT. We thus aim at suppressing the negative probability while enhancing the positive probability. Refer-



(a) DB on VOC-LT (b) PR (λ_1^-+) on VOC-LT (c) PR ($\lambda_2^+ +$) on VOC-LT (d) PR ($\lambda_1^-+, \lambda_2^+ +$) on VOC-LT

(e) DB on COCO-LT (f) PR (λ_1^-+) on COCO-LT (g) PR ($\lambda_1^+ +$) on COCO-LT (h) PR ($\lambda_1^-+, \lambda_1^+ +$) on COCO-LT

Figure 4: Predicted mean probability curves by different scaling factors on both VOC-LT and COCO-LT during training. H indicates the head classes, M and T denote the medium and tail classes, respectively. Both the superscript $+$ and the solid line denote the probability of positive labels, while subscript $-$ and the dotted line represent the probability of negative labels. Subfigures (b) and (f) are results of only increasing λ_1^-, (c) means increasing λ_2^+ on VOC-LT and (g) shows raising λ_1^+ on COCO-LT. Probability maps of increasing two needful scaling factors are shown in (d) and (h).

VOC-LT	total	head	medium	tail
DB	77.29	72.17	**83.83**	76.37
PR (λ_1^-+)	77.85	72.17	83.50	78.00
PR ($\lambda_2^+ +$)	77.56	72.83	**83.83**	76.50
PR ($\lambda_1^-+, \lambda_2^+ +$)	**78.44**	**73.00**	82.50	**79.75**
COCO-LT	total	head	medium	tail
DB	51.91	49.50	54.88	50.12
PR (λ_1^-+)	52.24	50.00	55.70	49.64
PR ($\lambda_1^+ +$)	52.82	49.95	56.21	50.88
PR ($\lambda_1^-+, \lambda_1^+ +$)	**53.66**	**50.91**	**56.88**	**52.00**

Table 3: The mAP performance of DB loss and our proposed PR on long-tailed datasets.

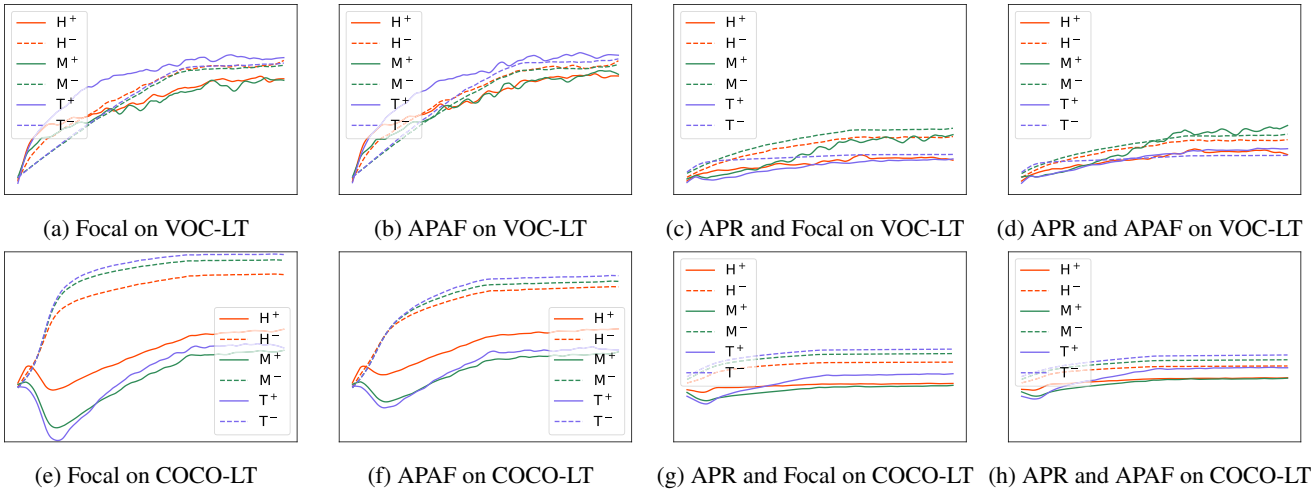
VOC-LT	total	head	medium	tail
PR ($\lambda_{1F}^-, \lambda_{2F}^+$)	79.46	73.67	**84.00**	80.50
PR ($\lambda_{1A}^-, \lambda_{2F}^+$)	**80.37**	73.67	83.83	**82.88**
PR ($\lambda_{1F}^-, \lambda_{2A}^+$)	79.70	**73.83**	83.33	81.38
PR ($\lambda_{1A}^-, \lambda_{2A}^+$)	79.95	73.33	83.67	82.00
COCO-LT	total	head	medium	tail
PR ($\lambda_{1F}^-, \lambda_{1F}^+$)	54.08	51.14	57.09	52.72
PR ($\lambda_{1A}^-, \lambda_{1F}^+$)	**54.43**	**51.22**	**57.42**	**53.40**
PR ($\lambda_{1F}^-, \lambda_{1A}^+$)	54.18	51.05	57.24	52.80
PR ($\lambda_{1A}^-, \lambda_{1A}^+$)	54.22	51.18	57.27	52.80

Table 4: The mAP performance of PR and APR on long-tailed datasets. The subscript A denotes the scaling factor is adaptive and F means the scaling factor is fixed.

ring to Tab. 1 similarly, we increase the scaling factors λ_1^- and λ_1^+ for all classes simultaneously. When only increasing λ_1^-, the negative probability of all classes is suppressed as is shown in Fig. 4f. Conversely, the positive probability is upgraded when merely increasing λ_1^+ as is illustrated in Fig. 4g. The probability map on COCO-LT after increasing both two scaling factors is depicted in Fig. 4h. Apparently, the growth rate of negative probability is declined. Meanwhile, the positive-negative probability gap is significantly narrowed by our PR.

Tab. 4 shows the mAP performance comparison between DB loss and our proposed PR. It can be clearly seen that our PR with multiple different gradient scaling factors guided by probability has a leap in performance compared to DB loss. Specifically, when only increasing λ_1^- for VOC-LT, we can have about 0.6% mAP gain which happens primarily in tail classes. And when merely increasing λ_2^+, the mAP gain is about 0.3%. The final performance gain is about 1.2% mAP on VOC-LT when concurrently increasing λ_1^- and λ_2^+. We mainly conduct PR on tail classes, thus the performance on tail class has a significant improvement up to about 3.4%. While there is a little performance loss on medium classes. On account for COCO-LT, the overall mAP gain is about 1.8%. When only increasing λ_1^-, we can have a mAP gain about 0.3%. While when increasing λ_1^+ merely, the mAP gain is about 0.9%. Since the PR is performed on all classes, the final performance gain is considerable on all three splitting classes consequently, that is about 1.5% on head classes, 2% on medium classes and 1.9% on tail classes.

Ablation Analysis on APR As discussed before, we increase the value of λ_1^- and λ_2^+ concurrently on VOC-LT, λ_1^- and λ_1^+ on COCO-LT. We here test whether the adaptive scaling factors for PR have influence on performance. As can be seen from Tab. 4, the best results are achieved when



(a) Focal on VOC-LT (b) APAF on VOC-LT (c) APR and Focal on VOC-LT (d) APR and APAF on VOC-LT

(e) Focal on COCO-LT (f) APAF on COCO-LT (g) APR and Focal on COCO-LT (h) APR and APAF on COCO-LT

Figure 5: Predicted mean probability curves by different components on both VOC-LT and COCO-LT during training. H indicates the head classes, M and T denote the medium and tail classes, respectively. Both the superscript $+$ and the solid line denote the probability of positive labels, while subscript $-$ and the dotted line represent the probability of negative labels.

fixing positive gradient factors while keeping negative scaling factors dynamic. We speculate that this may owing to the maximum importance of λ_1^- compared to other scaling factors. For the negative scaling factor λ_1^- contributes most to the suppression of negative probability. In addition, setting only one factor dynamic is sufficient. For if multiple factors are set dynamic, they may affect each other mutually.

Ablation Analysis on APAF Similar to analysis on PR, we observe how APAF affect the training by probability map. Fig. 5a shows the training probability map by focal loss on VOC-LT, we can see that the positive-negative probability gap of all classes has already been small. Thus, our APAF has little effect on the probability map compared to focal loss as shown in Fig. 5b. Yet for COCO-LT, the positive-negative probability gap remains large by focal loss as shown in Fig. 5e. Our APAF can effectively reduce such a gap as is shown in Fig. 5f. While Fig. 5c and Fig. 5g draw the probability maps on VOC-LT and COCO-LT by APR plus focal, respectively. And Fig. 5d and Fig. 5h show the probability maps by APR plus APAF on the two long-tailed datasets. We can see that the probability is more compact between positive and negative labels by replacing focal with APAF.

Tab. 5 presents the impact of our APAF on performance. Consistent with the above analysis on probability map, the mAP scores of APAF and focal are basically the same on VOC-LT which are both around 76.8%. For they have similar impact on the probability. Unlike on VOC-LT, APAF can achieve a certain performance gain for about 0.6% mAP on COCO-LT since it can effectively narrow the positive-negative probability gap compared to focal. This advance happens mainly on tail and head classes. After applying our APR, the APAF can eliminate the positive-negative probability gap more efficiently than focal. Accordingly, APAF can have about 1.2% performance advance on VOC-LT, which take place mainly on medium and tail classes. While for COCO-LT, the gain is about 0.2% which happens mainly in tail and head classes.

VOC-LT	total	head	medium	tail
Focal	76.86	72.17	83.50	75.50
APAF	76.77	72.17	83.33	75.13
APR + Focal	79.14	73.5	83.00	80.25
APR + APAF	**80.36**	**73.67**	**83.83**	**82.62**
COCO-LT	total	head	medium	tail
Focal	50.20	48.09	53.88	47.00
APAF	50.79	49.14	53.88	49.96
APR + Focal	54.24	51.00	57.39	53.12
APR + APAF	**54.43**	**51.23**	**57.42**	**53.40**

Table 5: The mAP performance of Focal Loss and our proposed APAF on long-tailed datasets.

Conclusion

In this paper, we propose to look upon long-tailed multi-label classification from the viewpoint of training probability. And further a simple yet powerful loss function is presented by us to tackle LTMLC. We find that existing cost-sensitive methods affect the training probability differently and thus affect final performance. We introduce a probability re-balancing scheme to reasonably adjust the probability. Besides, the adaptive probability re-balancing can further improve performance. Moreover, we propose the adaptive probability-aware focal to narrow the probability gap between positive and negative labels more effectively. Extensive experiments on two long-tailed datasets demonstrate the significance of proposed method.

Acknowledgments

This work was supported by the Major Science and Technology Project of Sichuan Province (No. 2019ZDZX0005)

and Sichuan Science and Technology Program (No. 2022ZHCG0007).

References

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259.

Byrd, J.; and Lipton, Z. 2019. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, 872–881. PMLR.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5177–5186.

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.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, 973–978. Lawrence Erlbaum Associates Ltd.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1): 98–136.

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Guo, Y.; and Gu, S. 2011. Multi-label classification using conditional dependency networks. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Han, H.; Wang, W.-Y.; and Mao, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, 878–887. Springer.

He, H.; and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9): 1263–1284.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hsieh, T.-I.; Robb, E.; Chen, H.-T.; and Huang, J.-B. 2021. Droploss for long-tail instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1549–1557.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2016. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5375–5384.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Li, Q.; Qiao, M.; Bian, W.; and Tao, D. 2016. Conditional graphical lasso for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2977–2986.

Li, Y.; Wang, T.; Kang, B.; Tang, S.; Wang, C.; Li, J.; and Feng, J. 2020. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10991–11000.

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.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2537–2546.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, 4334–4343. PMLR.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric Loss for Multi-Label Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 82–91.

Shen, L.; Lin, Z.; and Huang, Q. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, 467–482. Springer.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11662–11671.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2285–2294.

Wang, J.; Zhang, W.; Zang, Y.; Cao, Y.; Pang, J.; Gong, T.; Chen, K.; Liu, Z.; Loy, C. C.; and Lin, D. 2021. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9695–9704.

Wang, Y.-X.; Ramanan, D.; and Hebert, M. 2017. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 7032–7042.

Wang, Z.; Chen, T.; Li, G.; Xu, R.; and Lin, L. 2017. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, 464–472.

Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; and Lin, D. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, 162–178. Springer.

Xue, X.; Zhang, W.; Zhang, J.; Wu, B.; Fan, J.; and Lu, Y. 2011. Correlative multi-label multi-instance image annotation. In *2011 International Conference on Computer Vision*, 651–658. IEEE.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9719–9728.

Zhou, Z.-H.; and Liu, X.-Y. 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1): 63–77.