Adaptive Logit Adjustment Loss for Long-Tailed Visual Recognition

Yan Zhao [1], Weicong Chen [2], Xu Tan [3], Kai Huang [2], Jihong Zhu [1]*

[1] Tsinghua University, [2] Bytedance, [3] Microsoft
zhao-y18@mails.tsinghua.edu.cn, {chenweicong.do, huangkai.honka}@bytedance.com,
xuta@microsoft.com, jhzhu@tsinghua.edu.cn

Abstract

Data in the real world tends to exhibit a long-tailed label distribution, which poses great challenges for the training of neural networks in visual recognition. Existing methods tackle this problem mainly from the perspective of data quantity, i.e., the number of samples in each class. To be specific, they pay more attention to tail classes, like applying larger adjustments to the logit. However, in the training process, the quantity and difficulty of data are two intertwined and equally crucial problems. For some tail classes, the features of their instances are distinct and discriminative, which can also bring satisfactory accuracy; for some head classes, although with sufficient samples, the high semantic similarity with other classes and lack of discriminative features will bring bad accuracy. Based on these observations, we propose Adaptive Logit Adjustment Loss (ALA Loss) to apply an adaptive adjusting term to the logit. The adaptive adjusting term is composed of two complementary factors: 1) *quantity factor*, which pays more attention to tail classes, and 2) *difficulty factor*, which adaptively pays more attention to hard instances in the training process. The difficulty factor can alleviate the over-optimization on *tail yet easy* instances and under-optimization on *head yet hard* instances. The synergy of the two factors can not only advance the performance on tail classes even further, but also promote the accuracy on head classes. Unlike previous logit adjusting methods that only concerned about data quantity, ALA Loss tackles the long-tailed problem from a more comprehensive, fine-grained and adaptive perspective. Extensive experimental results show that our method achieves the state-of-the-art performance on challenging recognition benchmarks, including ImageNet-LT, iNaturalist 2018, and Places-LT.

1 Introduction

With the development of deep learning, the computer vision community has witnessed the immense breakthrough of visual recognition on the classic benchmarks, such as ImageNet (Russakovsky et al. 2015), COCO (Lin et al. 2014) and Places (Zhou et al. 2017). In contrast to these artificially balanced datasets, real-world scenarios usually subject to a long-tailed label distribution. A few classes (head classes) contain most of the data, while most classes (tail classes) occupy relatively few samples (Liu et al. 2019; Gupta, Dollar, and Girshick 2019). Unfortunately, confronted with such imbalanced distribution, the performance of these neural networks is found to degrade notably, especially on tail classes (Cao et al. 2019; Kang et al. 2019; Liu et al. 2019).

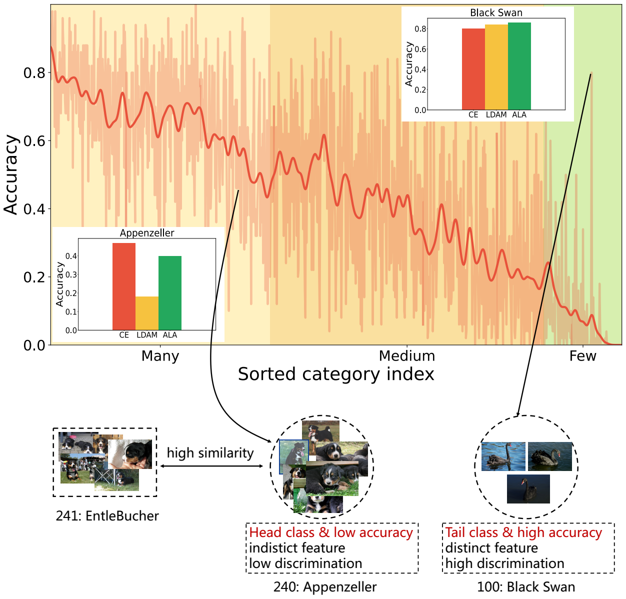


Figure 1: Per-class accuracy of Cross Entropy (CE) method on ImageNet-LT dataset. The x-axis represents the class index sorted by the sample number. The y-axis shows the per-class accuracy. Best view in color and zoom in.

Most existing long-tailed visual recognition methods address the problem by emphasizing the optimization on tail classes. These works can be roughly divided into three paradigms: re-sampling the training data (Buda, Maki, and Mazurowski 2018; Chawla et al. 2002; Wallace et al. 2011), re-weighting the coefficients of loss formulations (Menon et al. 2013; Cui et al. 2019; Ren et al. 2018) and adjusting the logit (Cao et al. 2019; Tan et al. 2020; Menon et al. 2020). Data re-sampling increases the sampling rate for tail classes and decreases it for head classes. Loss re-weighting guides the network to pay more attention to tail samples by up-weighting the tail classes and down-weighting the head

[*] Jihong Zhu is the corresponding author.

classes. Logit adjusting methods subtract a positive adjusting term from logit. This term is usually in reverse proportion to the frequency of each class, which encourages more optimization on tail classes. They all tackle the long-tailed problem from the perspective of data quantity, sharing the same design philosophy: emphasizing more on tail classes, less on head classes. However, according to our observations, *data quantity is a necessary but insufficient condition.*

As shown in Figure 1, we plot the accuracy of each class on ImageNet-LT (Liu et al. 2019), which is split by the number of training instances into few (1-20), medium (20-100) and many ($>$100) classes. It is noticeable that although there is a certain correlation between accuracy and data quantity in general, it is not absolute from the perspective of each class. For instance, class "Appenzeller" and "Black Swan" belong to the head and tail classes respectively. For "Black Swan", despite comprising relatively few samples, it has high accuracy. After searching all the bird group, we find that the characteristic of black swan is so distinct and discriminative, such as the black feather, the slender neck and the red beak, so that it can be easily distinguished. However, for "Appenzeller", even with sufficient samples, it still leads to a poor accuracy. The indistinct property and high semantic similarity with other classes (such as "EntleBucher") reduce its differentiability in the feature space, which greatly increases the risk of misclassification. The above observations indicate that larger regularization is not needed for tail yet easy classes (like "Black Swan"), but is urgent for head yet hard classes (like "Appenzeller").

To further shed light on the drawback of only focusing on data quantity, we also present specific comparison of accuracy between Cross Entropy (CE) and LDAM (Cao et al. 2019). LDAM is a prominent and effective method for long-tailed classification. It modifies the initial loss of CE by logit adjustment, but only from the perspective of data quantity. As shown in the accuracy histogram of Figure 1, for the tail yet easy class "Black Swan" (in the upper right corner), the accuracy of CE is good enough (0.8). Although LDAM indeed has slight promotion (0.8 to 0.84), the gain is just marginal. In contrast, for the head yet hard class "Appenzeller" (in the lower left corner), CE achieves bad performance (0.47). Under this situation, LDAM deteriorates the performance severely (0.47 to 0.18).

Based on these observations, we propose a novel Adaptive Logit Adjustment Loss (ALA Loss), which encourages more regularization on not only tail classes (including all the instances in tail classes), but also hard instances (in both head and tail classes). The adjusting term of ALA Loss is composed of two complementary factors: 1) quantity factor, which pays more attention to tail classes; 2) difficulty factor, which adaptively regularizes more on hard instances in the training process by binding with the value of logit. The synergy of the two factors can advance the performance on tail classes even further. More importantly, it mitigates the under-optimization on the hard samples of head classes, which promotes the accuracy on head classes at the same time. As we intended, compared to logit adjusting method LDAM, ALA Loss achieves better results on both the tail class "Black Swan" and the head class "Appenzeller" in Figure 1.

The contributions of our work can be summarized as follows:

1. We develop a novel Adaptive Logit Adjustment Loss (ALA Loss), which contains a quantity factor and a difficulty factor. ALA Loss works in a more comprehensive and fine-grained way. To be specific, previous methods only regularize more on tail classes. In comparison, ALA Loss takes both the quantity and difficulty of data into consideration, adjusting from the perspective of both class level and instance level. Supplemented with our difficulty factor, the over-optimization on tail yet easy and under-optimization on head yet hard instances can be efficaciously alleviated.
2. We propose to adaptively apply regularization in the training process. Specifically, previous methods employ prior data quantity related to class frequency for logit adjustment. Our adjusting term takes a step further by adaptively choose which instances to regularize based on the value of the predicted logit. It can make the learning process more efficient, and boost the performance of both tail classes and hard samples.
3. We conduct extensive and comprehensive experiments. ALA Loss shows consistent and significant improvements on three challenging large-scale long-tailed datasets, including ImageNet-LT, iNaturalist 2018 and Places-LT.

2 Related Works

2.1 Long-Tailed Classification

Existing techniques for long-tailed classification mainly involves data re-sampling (Chawla et al. 2002; Han, Wang, and Mao 2005; Drumnond 2003), loss modifying (Cui et al. 2019; Khan et al. 2017; Cao et al. 2019; Ren et al. 2018; Lin et al. 2017), knowledge transferring (Yin et al. 2019; Liu et al. 2019) and network structure designing (Kang et al. 2019; Zhou et al. 2020; Wang et al. 2020).

As for data re-sampling, two common techniques are: over-sampling (Chawla et al. 2002; Han, Wang, and Mao 2005) for tail classes and under-sampling (Drumnond 2003) for head classes. As for loss modifying, it can be roughly classified into re-weighting based (Cui et al. 2019; Ren et al. 2018; Shu et al. 2019; Lin et al. 2017) and logit adjusting based (Cao et al. 2019; Tan et al. 2020; Menon et al. 2020) methods. Apart from the aforementioned strategies, knowledge transferring usually occurs from head to tail classes and the knowledge can be intra-class variance (Yin et al. 2019) or semantic feature (Liu et al. 2019). Recently, the methods of network structure designing also show promising success. Kang et al. (2019) proposes a commonly used two-stage training strategy. Xiang, Ding, and Han (2020); Zhou et al. (2020); Wang et al. (2020) introduce multi-expert structure into long-tailed problem, sharing the same principle of divide-and-conquer. In this paper, we mainly focus on the simple but efficient logit adjusting losses, which can be easily integrated into other methods.

2.2 Logit Adjustment

Logit adjusting based loss is first proposed in face recognition (Liu et al. 2016, 2017; Wang et al. 2018a,b; Deng et al. 2019), which encourages larger inter-class margin and enforces extra intra-class compactness. LDAM (Cao et al. 2019) introduces this idea to long-tailed recognition for the first time, and proposes a class-dependent adjusting term to enlarge margins for tail classes. Equalization Loss (Tan et al. 2020) applies logit adjustment to alleviate the overwhelmed discouraging gradients from head to tail classes in the field of object detection. Logit Adjust (Menon et al. 2020) analyzes from Fisher consistency and proposes a general form for logit adjustment. Among them, LDAM is a prominent and effective method for long-tailed classification, so we take it as the baseline of logit adjusting losses.

3 Method

As shown in Figure 1, for long-tailed recognition, the quantity and difficulty of data are two intertwined and equally crucial part. Based on this observation, we propose Adaptive Logit Adjustment Loss (ALA Loss), which consists of two complementary factors, i.e., quantity factor and difficulty factor.

3.1 Preliminary

We first revisit the widely used softmax Cross-Entropy (CE) loss:

$$\begin{aligned}\mathcal{L}_{CE}(y, f_\theta(x)) &= -\log(\sigma_{CE}(y, f_\theta(x))),\\ \sigma_{CE}(y, f_\theta(x)) &= \frac{e^{f_\theta(x)[y]}}{\sum_{j=1}^{C} e^{f_\theta(x)[j]}},\end{aligned} \tag{1}$$

where x is the input instance, $y \in \{1, 2, \cdots, C\}$ is the corresponding target class. C is the total number of classes. $f_\theta(x)[j]$ is the predicted logit of the j-th class. $\sigma_{CE}(y, f_\theta(x))$ is the predicted probability of the classifier.

Next, we review logit adjusting losses in the field of long-tailed recognition. The common methods (Cao et al. 2019; Menon et al. 2020; Tan et al. 2020) tackle the long-tailed classification by subtracting a positive adjusting term $\mathcal{A} \in \mathbb{R}^C$ from logit $f_\theta(x)$. Therefore, the formulation of logit adjusting loss can be written as:

$$\begin{aligned}\mathcal{L}_{LA}(y, f_\theta(x), \mathcal{A}) &= -\log(\sigma_{LA}(y, f_\theta(x), \mathcal{A})),\\ \sigma_{LA}(y, f_\theta(x), \mathcal{A}) &= \frac{e^{f_\theta(x)[y] - \mathcal{A}[y]}}{\sum_{j=1}^{C} e^{f_\theta(x)[j] - \mathcal{A}[j]}},\end{aligned} \tag{2}$$

Furthermore, the gradients of the loss $\mathcal{L}_{LA}$ on the logit $f_\theta(x)$ can be formulated as:

$$\frac{\partial \mathcal{L}_{LA}}{\partial f_\theta(x)} = \begin{cases}\sigma_{LA}(y, f_\theta(x)[y], \mathcal{A}[y]) - 1, & \text{for } j = y,\\ \sigma_{LA}(y, f_\theta(x)[j], \mathcal{A}[j]), & \text{for } j \neq y,\end{cases} \tag{3}$$

In previous works, $\mathcal{A}$ is only related to the data quantity. Specifically, it is a class-dependent term and negatively related to the number of samples in each class (Cao et al. 2019; Tan et al. 2020; Menon et al. 2020). To further shed light on the effect of logit adjusting losses, we analyze it from the perspective of gradient. According to Equation (3), for the target class y, the gradient on logit is $\sigma_{LA}(y, f_\theta(x)[y], \mathcal{A}[y]) - 1$. For two samples with the same logit, denoting $f_\theta(x)[y_h]$ as the logit for the sample from head classes and $f_\theta(x)[y_t]$ for the one from tail classes, $f_\theta(x)[y_h] = f_\theta(x)[y_t]$. However, $\mathcal{A}[y_h] < \mathcal{A}[y_t]$, thus $\sigma_{LA}(y_h, f_\theta(x), \mathcal{A}) > \sigma_{LA}(y_t, f_\theta(x), \mathcal{A})$. Because of $\sigma_{LA}(y, f_\theta(x)[y], \mathcal{A}) - 1 < 0$, thus $\left|\frac{\partial \mathcal{L}_{LA}}{\partial f_\theta(x)}\right|_{y_h} < \left|\frac{\partial \mathcal{L}_{LA}}{\partial f_\theta(x)}\right|_{y_t}$. It means that, for two samples with the same logit, the one from tail classes will get a larger scale of gradient than that from head classes, making the model focus more on tail classes.

3.2 ALA Loss

Previous logit adjusting methods can effectively ameliorate the long-tailed situation from the perspective of data quantity. However, there are still some limitations. As shown in Figure 1, the over-optimization on tail yet easy and under-optimization on head yet hard samples are urgent issues to be settled. Therefore, we propose ALA Loss, whose form is the same as common logit adjusting losses shown in Equation (2), but the adjusting term $\mathcal{A}$ is not only related to the data quantity but also correlated with the instance difficulty. ALA Loss designs $\mathcal{A}$ as the combination of a difficulty factor ($\mathcal{DF}$) and a quantity factor ($\mathcal{QF}$), as

$$\mathcal{A}^{ALA} = \mathcal{DF} \cdot \mathcal{QF}, \tag{4}$$

We will discuss the difficulty factor and quantity factor in detail, respectively.

Difficulty Factor ($\mathcal{DF}$). $\mathcal{DF}$ is an instance-specific term, which aims to make the model pay more attention to hard instances. *Since hard instances are those with worse predicted results, the design principle is that $\mathcal{DF}$ should be negatively related to the target prediction.* Predicted logit and probability can both be utilized as the signal to measure the difficulty. We empirically find logit works better. The reason behind it is two folds: 1) Due to softmax, the predicted probability is sharper compared with the corresponding logit, which will lead to a over-large or over-small adjusting term. 2) The predicted logit have the same form with the original logit to be adjusted, which is more consistent and coherent.

However, as the value range of logit is unknown, it is hard to design the specific formulation of $\mathcal{DF}$. Therefore, we restrict $f_\theta(x)$ to $[-1, 1]$ by weight normalization and feature normalization following LDAM. Specifically, we make the following transformations to $f_\theta(x_i)$. x_i is the i-th sample and it belongs to the j-th class:

$$\begin{aligned}f_\theta(x_i) &:= W_j^T x_i + b_j\\ &:= \|W_j\| \, \|x_i\| \cos\theta_{ij} \quad \text{// by setting } b_j = 0\\ &:= \|x_i\| \cos\theta_{ij} \quad \text{// by weight normalization}\\ &:= \cos\theta_{ij} \quad \text{// by feature normalization}\end{aligned} \tag{5}$$

Taking above transformations and design principle into consideration together, $\mathcal{DF}$ is designed negatively related to the value of $\cos\theta_{ij}$. At the same time, the value range of $\mathcal{DF}$ is restricted to $[0, 1]$. Then the formulation of $\mathcal{DF}$ is designed as:

$$\mathcal{DF} = \frac{1 - \cos\theta_{iy}}{2}, \tag{6}$$

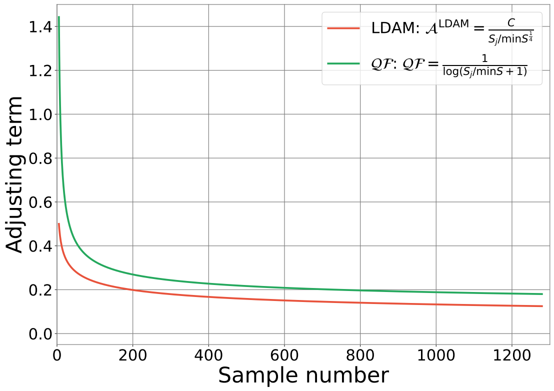


Figure 2: Comparison between LDAM and the $\mathcal{QF}$ term in ALA loss.

Note that we detach the gradient of $\cos\theta_{iy}$.

To further understand $\mathcal{DF}$, we give a more intuitive interpreting way. θ_{ij} denotes the angle between x_i and W_j. Since W_j is generally considered as the target center of the j-th class (Deng et al. 2019), θ_{iy} represents the angle between x_i and its target class center, which is preferred to be as small as possible. For easy samples, θ_{iy} tends to be small, and $\cos\theta_{iy}$ tends to be large, which is vice versa for hard samples. Following Equation (6), hard samples get more regularization than easy samples by our $\mathcal{DF}$, which brings improvements for discriminative learning.

Quantity Factor ($\mathcal{QF}$). $\mathcal{DF}$ effectively makes the model focus more on hard instances, including those in tail classes. However, it is not enough. A further consideration is that instances in the same class jointly contribute to learn the representation and determine the classifier boundary. Hard samples in head classes can benefit from other samples of the same class due to the large number of samples, while hard samples in tail classes benefit less due to the lack of data samples in tail classes.

Therefore, we design a class-dependent term $\mathcal{QF}$ to better combine with $\mathcal{DF}$, making the network focus more on tail classes. *The design principle is similar with previous logit adjusting losses: $\mathcal{QF}$ should be negatively related to the data quantity.* However, different from the common power function ($1/x^n$) used in other methods (Cao et al. 2019; Menon et al. 2020), we empirically find log function ($1/\log(x+1)$) is a better selection, which applies stronger regularization for tail classes.

Denoting $S = \{S_1, S_2, ..., S_C\}$ as a set of sample number for each class, $\mathcal{QF}$ is formulated as following:

$$\mathcal{QF} = \frac{1}{\log(\frac{S_j}{\min S} + 1)}, \tag{7}$$

where S_j is normalized by the minimum number of samples $\min S$ following LDAM. Obviously, $\mathcal{QF}$ is class-dependent, whose value is larger for tail classes with smaller S_j.

To further demonstrate the new function form of $\mathcal{QF}$ is essential, we intuitively show the comparison of our $\mathcal{QF}$ with LDAM in Figure 2. It can be clearly seen that $\mathcal{QF}$ assigns larger adjusting term to tail classes, which can boost the performance of them. The experimental results in Section 4.4 verify that $\mathcal{QF}$ is a more appropriate complementary term for $\mathcal{DF}$.

Final Formation. Following LDAM (Cao et al. 2019), we only adjust the target logit, which means $\mathcal{A}[j] = 0$ when $j \neq y$, and we also re-scale the logit by the same constant s. Thus the final fomulation of our ALA Loss is

$$\mathcal{L}_{ALA} = -\log \frac{e^{s(f_\theta(x)[y] - \mathcal{A}^{ALA}[y])}}{e^{s(f_\theta(x)[y] - \mathcal{A}^{ALA}[y])} + \sum_{j \neq y}^{C} e^{s f_\theta[j]}}. \tag{8}$$

Note that the design principle for $\mathcal{DF}$ is to be negatively related to the instance difficulty, and for $\mathcal{QF}$ is to be negatively related to the data quantity. Our formulations are designed following the motivation of being simple and practical. Other forms that conform to the principles are also suitable.

3.3 Advantages over Previous Methods

In summary, two appealing properties of ALA Loss make it stand out among previous logit adjusting losses.

1) *ALA Loss is comprehensive and fine-grained.* The $\mathcal{DF}$ term in ALA Loss considers from the perspective of instance difficulty, which alleviates the over-optimization on tail yet easy and under-optimization on head yet hard instances. For the $\mathcal{QF}$ term in ALA Loss, its design principle is the same as previous logit adjusting methods. However, we redesign the function to assign larger adjustments on tail classes, which makes it more suitable to integrate with $\mathcal{DF}$.

2) *ALA Loss is adaptive.* Similar to previous methods (Cao et al. 2019; Menon et al. 2020), $\mathcal{QF}$ tackles the long-tailed recognition only considering the prior data quantity. In contrast, $\mathcal{DF}$ is related with the predicted logit, taking the dynamic status of training process into account. Consequently, our ALA Loss can adaptively focus on those poorly performing instances at present by giving them larger regularization, which is more rational and effective.

4 Experiments

4.1 Dataset

To evaluate the effectiveness and generality of our method, we conduct a series of experiments on three widely used large-scale long-tailed datasets: ImageNet-LT, iNaturalist 2018 and Places-LT.

ImageNet-LT. The ImageNet-LT (Liu et al. 2019) dataset is an artificially sampled subset of ImageNet-2012 (Deng et al. 2009), with 115.8K images. In this dataset, the overall number of classes is 1,000, while the maximum and minimum number of samples per class are 1,280 and 5, respectively.

iNaturalist 2018. The iNaturalist 2018 (Van Horn et al. 2018) dataset is a real-world imbalanced dataset, with 437.5K images. The overall number of classes is 8,142, with

Method	Top-1 Accuracy @ R-50				Top-1 Accuracy @ X-50			
	Many	Medium	Few	All	Many	Medium	Few	All
Cross Entropy †	64.0	33.8	5.8	41.6	65.9	37.5	7.7	44.4
Focal Loss ‡	-	-	-	-	64.3	37.1	8.2	43.7
OLTR ‡	-	-	-	-	51.0	40.8	20.8	41.9
Decouple-LWS †	57.1	45.2	29.3	47.7	60.2	47.2	30.3	49.9
DisAlign †	59.9	49.9	31.8	51.3	61.5	50.7	33.1	52.6
Casual Norm ‡	-	-	-	-	62.7	48.8	31.6	51.8
Balanced softmax ‡	-	-	-	-	62.2	48.8	29.8	51.4
PC softmax ‡	-	-	-	-	60.4	46.7	23.8	48.9
LADE ‡	-	-	-	-	62.3	49.3	31.2	51.9
Logit Adjust (loss)	-	-	-	51.0	-	-	-	-
LDAM + DRW $\ast$	61.8	47.2	31.4	50.7	62.9	47.5	31.9	51.3
ALA Loss	62.4	49.1	35.7	**52.4**	64.1	49.9	34.7	**53.3**

Table 1: **Top-1 accuracy on the test set of ImageNet-LT equipped with ResNet50 (R-50) and ResNeXt50 (X-50).** The superscript † denotes that the results are from MisAlign (Zhang et al. 2021), ‡ are from LADE (Hong et al. 2021) and $\ast$ means our reproduced results. The lower part from Causal Norm to LDAM + DRW are the logit adjusting methods.

Method	Many	Medium	Few	All
Cross Entropy †	72.2	63.0	57.2	61.7
CB-Focal ‡	-	-	-	61.1
Decouple-LWS †	65.0	66.3	65.5	65.9
BBN †	49.4	70.8	65.3	66.3
DisAlign †	-	-	-	67.8
Casual Norm ‡	-	-	-	63.9
Balanced Softmax ‡	-	-	-	69.8
PC Softmax ‡	-	-	-	69.3
LADE ‡	-	-	-	70.0
LDAM + DRW ‡	-	-	-	68.0
ALA Loss	71.3	70.8	70.4	**70.7**

Table 2: **Top-1 accuracy on the validation set of iNaturalist 2018 with ResNet-50.** † indicates that the results are from MisAlign. ‡ indicates that the results are from LADE.

Method	Many	Medium	Few	All
Cross Entropy †	45.7	27.3	8.2	30.2
Focal Loss †	41.1	34.8	22.4	34.6
Range Loss †	41.1	35.4	23.2	35.1
OLTR †	44.7	37.0	25.3	35.9
Feature Aug †	42.8	37.5	22.7	36.4
Decouple-LWS †	40.6	39.1	28.6	37.6
DisAlign †	40.4	42.4	30.1	39.3
Causal Norm ‡	23.8	35.8	40.4	32.4
Balanced Softmax ‡	42.0	39.3	30.5	38.6
PC Softmax ‡	43.0	39.1	29.6	38.7
LADE ‡	42.8	39.0	31.2	38.8
ALA Loss	43.9	40.1	32.9	**40.1**

Table 3: **Top-1 accuracy on the test set of Places-LT with ResNet-152.** † indicates that the results are from DisAlign. ‡ indicates that the results are from LADE.

the maximum and minimum number of samples per class as 1,000 and 2, respectively.

Places-LT. The Places-LT (Liu et al. 2019) dataset is a long-tailed subset of the dataset Places (Zhou et al. 2017), with 62.5K images. It consists of 365 categories, the samples of each class ranging from 5 to 4,980.

4.2 Experimental Setting

Implementation Details. For ImageNet-LT, ResNet-50 and ResNeXt-50 (32x4d) (He et al. 2016) are adopted as backbones. And we mainly use ResNeXt-50 for ablation studies. The batch size is set as 256 with an initial learning rate of 0.1 and a weight decay of 0.0005. For iNaturalist 2018, ResNet-50 is used as the backbone. And the batch size is set as 512 with an initial learning rate of 0.2 and a weight decay of 0.0002. For Places-LT, we utilize ResNet-152 as the backbone and pretrain it on the full ImageNet dataset following (Hong et al. 2021) for fair comparison.

All networks are trained on 2 Tesla V100 GPUs, 90 epochs for ImageNet-LT and iNaturalist 2018, while 30 epochs for Places-LT. The scale factor s in Equation (8) is set to 30 by default. And we use the same training strategies as LDAM (Cao et al. 2019) .

Evaluation Protocol. All networks are trained on the long-tailed training datasets, and then evaluated on the corresponding balanced validation or test datasets. Top-1 accuracy is used as the evaluation metric, in the form of percentages. In order to better analyze the performance on classes of different data frequency, we report the accuracy on four class subsets according to the number of training instances in each class: *Many-shot* ($>$100), *Medium-shot* (20-100), *Few-shot* (1-20) and *All* as in (Liu et al. 2019).

	LDAM	$\mathcal{DF}$	$\mathcal{QF}$	Many	Medium	Few	All
Cross Entropy (CE)	✗	✗	✗	65.9	37.5	7.7	44.4
LDAM	✓	✗	✗	62.9	47.5	31.9	51.3
$\mathcal{DF}$	✗	✓	✗	64.7	47.3	28.6	51.5
$\mathcal{QF}$	✗	✗	✓	61.4	47.7	33.0	51.0
$\mathcal{DF} \cdot \mathcal{A}^{LDAM}$	✓	✓	✗	63.5	48.4	31.9	52.0
$\mathcal{DF} \cdot \mathcal{QF}$ (ALA Loss)	✗	✓	✓	64.1	49.9	34.7	**53.3**

Table 4: **Ablation studies for ALA Loss.** Results on the test set of ImageNet-LT with ResNeXt-50. The first line is the results of Cross Entropy (CE); the last line is our ALA Loss. $\mathcal{DF}$ denotes the difficulty factor in Equation (6). $\mathcal{QF}$ denotes the quantity factor in Equation (7).

4.3 Main Results

In this section, we present the performance of our method and compare with previous state-of-the-art works. The results on ImageNet-LT, iNaturalist 2018 and Places-LT are shown in Table 1, Table 2 and Table 3 respectively.

Experimental Results on ImageNet-LT. As shown in Table 1, we have the following observations: 1) ALA Loss achieves superior performance than existing methods on both networks. Comparing with the state-of-the-art method DisAlign, ALA Loss gets 1.1% accuracy gain on ResNet50 and 0.7% gain on ResNeXt50. 2) Comparing with other logit adjusting losses: Logit Adjust (Menon et al. 2020) and LDAM, ALA Loss gets better results on all the three subsets, showing the full range of advantages of our method. 3) Unlike other methods that improve tail classes at the sacrifice of head classes, our method not only achieves better results on tail classes, but also improves significantly on head classes, and achieves comparable results to Cross Entropy (CE), owing to the proposed Difficulty Factor.

Experimental Results on iNaturalist 2018. As shown in Table 2, ALA Loss achieves better results than other methods on this real-world long-tailed dataset, showing the effectiveness of our method. Remarkably, ALA Loss obtains nearly equal result on all three subsets, which is a rather ideal result for long-tailed visual recognition (Wang et al. 2020). Most of the existing methods improve the results of tail classes at the expense of head classes, while our method takes both into consideration.

Experimental Results on Places-LT. As shown in Table 3, ALA Loss again outperforms other methods, especially on many- and few-shot subsets. Compared with other logit adjusting methods: from Casual Norm to LADE, ALA Loss achieves promising results, with 0.9% gain on the many- and 0.8% gain on the medium-shot. For the few-shot, ALA Loss achieves the best result other than Casual Norm. Casual Norm applies the post-hoc logit adjustment in the phase of inference. It can boost the accuracy of tail, however at the expense of head classes. Quantity and Difficulty Factor both contribute to the improvements of the few-shot subset, while the results of many-shot subset are mainly attributed to the Difficulty Factor.

4.4 Ablation Study

Quantitative analysis. In this section, a series of experiments are conducted to examine the effect of each component in ALA Loss. According to the results shown in Table 4, we have the following observations:

1) $\mathcal{DF}$ aims to optimize hard instances in all subsets, which can boost the performance of both head and tail classes, improving the long-tailed classification to a certain extent. Compared with CE, $\mathcal{DF}$ achieves considerable better results on both medium- and few-shot subsets, with only a slight decline on many-shot subset. Moreover, it even outperforms data quantity based adjusting method LDAM and $\mathcal{QF}$, indicating the advantage of tackling the long-tailed problem from the fine-grained perspective of instance difficulty.

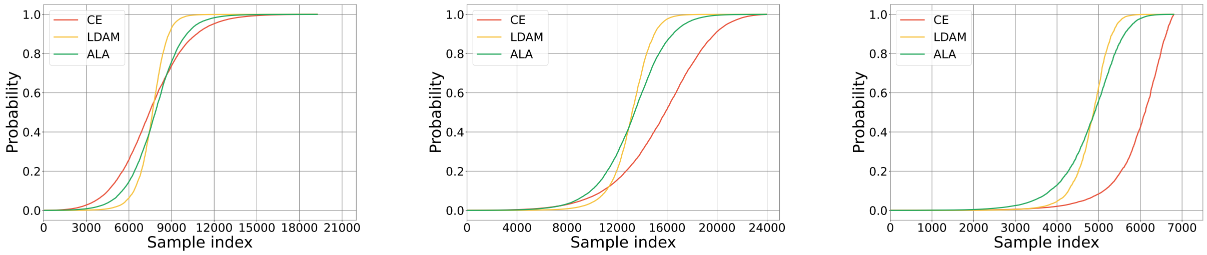
2) $\mathcal{QF}$ performs better on tail classes than LDAM. The comparison between LDAM and $\mathcal{QF}$ reveals that our proposed quantity factor term is able to achieve comparable overall performances as LDAM. What's more, on the few-shot subset, $\mathcal{QF}$ brings significant gains, which is consistent with the analysis in Figure 2.

3) $\mathcal{DF}$ performs better on the many-shot subset, while LDAM and $\mathcal{QF}$ get higher accuracy on the few-shot subset. It is consistent with our design principle. That is: $\mathcal{DF}$ focus more on hard samples, since it tackles the long-tailed problem from the perspective of instance difficulty. $\mathcal{QF}$ and LDAM pay more attention to tail classes, since they are only related with the data quantity.

4) Both $\mathcal{QF}$ and LDAM can be further boosted when combined with our $\mathcal{DF}$. According to the results shown in Table 4, the combination setting $\mathcal{DF} \cdot \mathcal{QF}$ achieves the best result, and other settings like $\mathcal{DF} \cdot \mathcal{A}^{LDAM}$ also performs better than using either only.

Qualitative analysis In this section, we conduct two qualitative analysis to characterize our ALA Loss intuitively and comprehensively. Concretely, we further visualize and analyze the advantage of ALA Loss from the perspective of probability and adaptability.

Probabily analysis. To intuitively reflect the advantages of our ALA Loss over data quantity based logit adjusting methods, we conduct an experiment to compare ALA Loss with CE and LDAM. According to the results shown in Figure 3, we roughly consider those samples with predicted probabili-

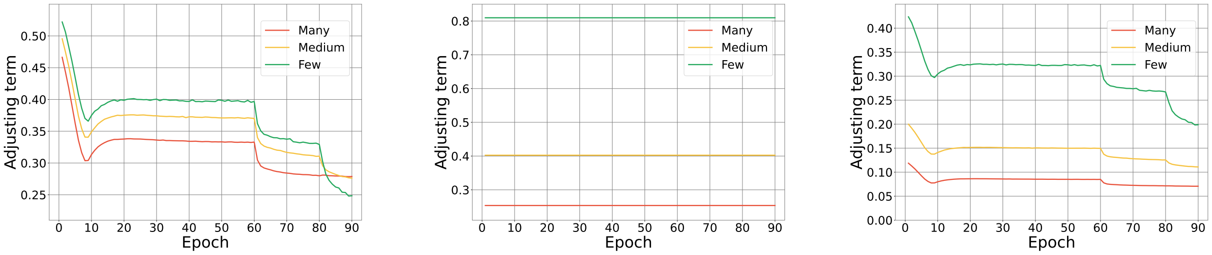


(a) Many-shot

(b) Medium-shot

(c) Few-shot

Figure 3: The predicted probability distributions for each class subset among CE, LDAM and ALA Loss. The x-axis represents the sample index, which is sorted by the predicted probability. There are more hard samples when the curve goes to the right.



(a) $\mathcal{DF}$

(b) $\mathcal{QF}$

(c) $\mathcal{A}^{ALA} = \mathcal{DF} * \mathcal{QF}$

Figure 4: Adjusting curves of $\mathcal{DF}$, $\mathcal{QF}$ and $\mathcal{A}^{ALA}$ in the training process. In order to compare the difference on Many / Medium / Few, the value denotes the average adjusting terms in the same subset.

ties < 0.2 as hard samples, and those > 0.8 as easy samples. It is worth noticing that: 1) For the medium- and few-shot subsets shown in Figure 3(b) and Figure 3(c), both LDAM and ALA Loss can significantly improve the predicted probabilities compared with CE; 2) For the many-shot subset shown in Figure 3(a), LDAM has much more hard samples than CE, which brings significant drop to the accuracy. However, ALA Loss alleviates the excessive hard samples caused by LDAM, with only a slight accuracy decline on many-shot subset compared with CE. Again, from the perspective of probability, it verifies that our method does not need to excessively sacrifice the performance of head classes in exchange for the promotion of tail classes.

Adaptability analysis. We also visualize the trend of adjusting term in the training process in Figure 4. We discuss it from three aspects:

Firstly, as shown in Figure 4(a), we obtain the following observations about $\mathcal{DF}$:

1) From head to tail classes, the adjusting term increases gradually, which indicates that the proportion of hard samples in tail is larger than head classes. It can be verified by the ablation experiments in Table 4: $\mathcal{DF}$ bring significant performance gains for tail classes compared with CE.

2) $\mathcal{DF}$ adaptively changes as the performance of network fluctuates in the training process, and gets smaller with the optimization of the network.

Secondly, as shown in Figure 4(b), $\mathcal{QF}$ keeps unchanged in the whole training process, encouraging larger regularization for tail classes.

Lastly, as shown in Figure 4(c), $\mathcal{A}^{ALA}$ adjusts more on the few-shot subset. But the relative difference between the adjusting term of head and tail classes shrinks compared with $\mathcal{QF}$ in Figure 4(b), which can alleviate the under-optimization for head yet hard and over-optimization for tail yet easy instances.

5 Conclusion

In this work, we analyze the issues behind the existing methods for long-tailed classification and propose to revisit this problem from the perspective of not only data quantity but also instance difficulty. Our analysis shows that: it is unreasonable for previous logit adjusting methods to simply regularize more on tail classes, which leads to the over-optimization on tail yet easy instances and under-optimization on head yet hard instances. Therefore, we propose an Adaptive Logit Adjustment (ALA) loss that contains a difficulty factor to focus the model more on hard instances and a quantity factor to make the model pay more attention to tail classes. Extensive and comprehensive experimental results show that our method outperforms the existing SOTA methods on three widely used large-scale long-tailed benchmarks including ImageNet-LT, iNaturalist 2018 and Places-LT. We will also apply the proposed ALA Loss to the long-tailed object detection and instance segmentation datasets in the future work.

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.

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*.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

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.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Drumnond, C. 2003. Class Imbalance and Cost Sensitivity: Why Undersampling beats Oversampling. In *ICML-KDD 2003 Workshop: Learning from Imbalanced Datasets*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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, 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.

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling Label Distribution for Long-tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6626–6636.

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*.

Khan, S. H.; Hayat, M.; Bennamoun, M.; Sohel, F. A.; and Togneri, R. 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8): 3573–3587.

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, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 212–220.

Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, 7.

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.

Menon, A.; Narasimhan, H.; Agarwal, S.; and Chawla, S. 2013. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, 603–611. PMLR.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

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.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *arXiv preprint arXiv:1902.07379*.

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.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2011. Class imbalance, redux. In *2011 IEEE 11th international conference on data mining*, 754–763. IEEE.

Wang, F.; Cheng, J.; Liu, W.; and Liu, H. 2018a. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7): 926–930.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018b. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. *arXiv preprint arXiv:2010.01809*.

Xiang, L.; Ding, G.; and Han, J. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, 247–263. Springer.

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2019. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5704–5713.

Zhang, S.; Li, Z.; Yan, S.; He, X.; and Sun, J. 2021. Distribution Alignment: A Unified Framework for Long-tail Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2361–2370.

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, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.