南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

No One Left Behind: Improving the Worst Categories in Long-Tailed Learning

CVPR 2023

Background

(1) No one makes great contribution on improving the lowest recall among all categories.

(2) Average accuracy of subset is not accurate.

Motivation

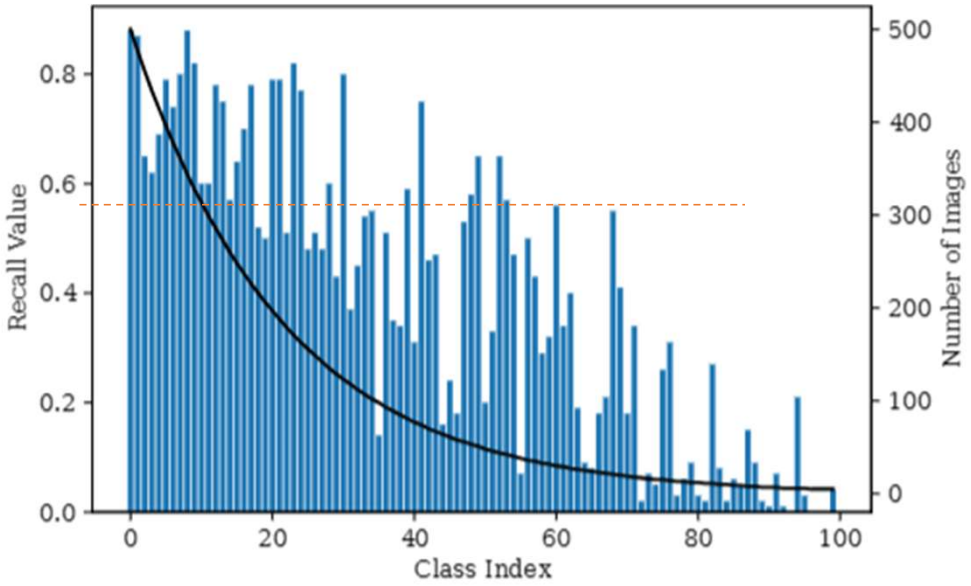


Figure 1. The per-class recall of models trained on the imbalanced CIFAR100 (with imbalance ratio 100). Per-class recall value varies a lot from category to category. Moreover, it is not necessarily true that all categories in the "Few" subset have lower accuracy than those in the "Many" or "Medium" subsets.

Few, Medium, Many

~~Accuracy within each subset often with overall accuracy~~

Motivation

make sure all classes are equally treated

use Harmonic Mean (sensitive to small numbers)

Motivation

Method	Mean Accuracy	Lowest Recall
BSCE [22]	42.24	3.00
DiVE [12]	45.11	2.00
MiSLAS [30]	47.05	5.00
RIDE [26]	48.64	2.00
PaCo [4]	51.24	5.00

Table 1. The lowest per-class recall of various state-of-the-art methods on the imbalanced CIFAR100 (with imbalance ratio 100). Although there are rapid improvements over the mean accuracy, the lowest per-class recall remains very low.

Therefore, in order to make full use of existing advances, we do not aim at building up a whole new training framework, but rather propose a simple plug-in method.

Related work - Re-sampling and Re-weighting

Re-sampling :

over-sampling the tail classes or

under-sampling the head classes.

Disadvantage :

either over-fitting of tail classes or under-fitting of head classes.

Re-weighting :

give each instance a weight based on its true label when computing the loss.

Disadvantage :

loss function hard to optimize.

Related work - Two-Stage Decoupling

decouple the learning of features and classifiers

Related work - Hybrid and Multiple Heads

decouple the learning of features and classifiers

Disadvantage :

these methods require a joint training of multiple heads, together with a complex routing module that dynamically determines the head to use during inference

Method

non-linear function : F

i-th training image (x^i, y^i) $x^i \in R^{H \times W \times D}$

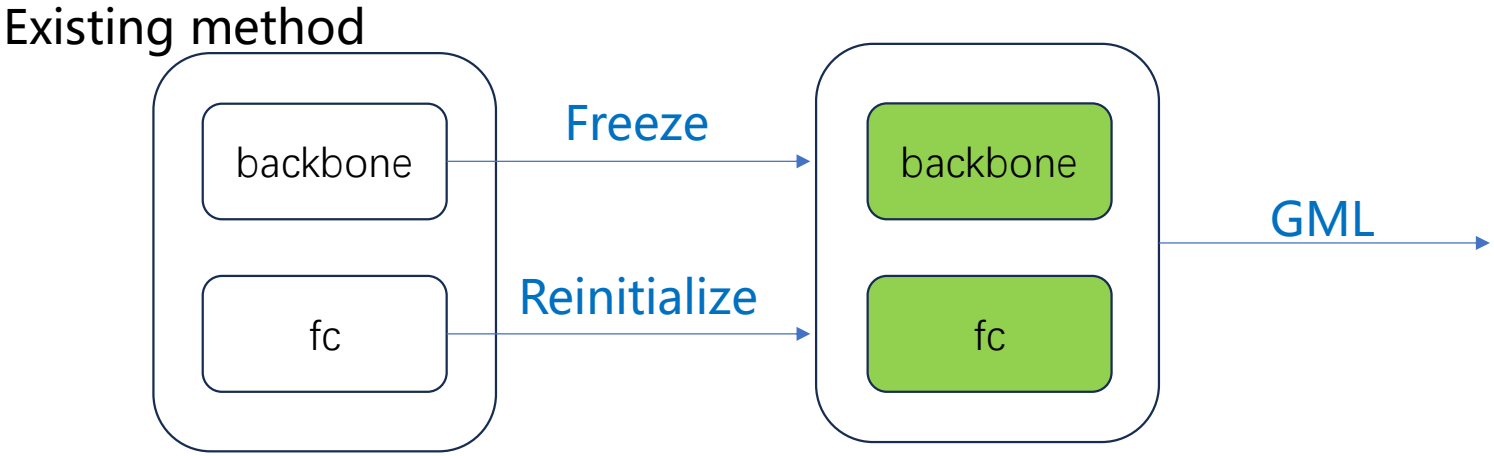
Imbalance dataset has N images

A forward propagation through the network yields logit $o^i = F(x^i)$

For a multi-class classification problem with C categories

$$\tilde{p}^i = \text{softmax}(o^i), \tilde{p}^i \in R^C$$

Method – GML (Geometric Mean Loss)



Not CE, but GML :

$$L_{GML} = -\frac{1}{C}\sum_{c=1}^{C}\log \bar{p}_c \qquad (1)$$

(borrow this idea from many two-stage decoupling methods)

Method – GML (Geometric Mean Loss)

Not CE, but GML :

$$L_{GML} = -\frac{1}{C}\sum_{c=1}^{C}\log \bar{p}_c \qquad (1)$$

average of across all training samples belonging to class c in this mini-batch

$$\bar{p}_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\tilde{p}^i_{y^i} \qquad (2)$$

Reweighting when computing :

$$\tilde{p}^i_j = \frac{N_j \exp(o^i_j)}{\sum_{c=1}^{C} N_c \exp(o^i_c)} \qquad (3)$$

Method – why GML work ?

Mean accuracy :

Low penalty on small recall

$$Acc = \frac{1}{N}\sum_{i=1}^{N} 1(y^i = \arg\max \tilde{p}^i) \qquad (4)$$

The same as arithmetic mean of per-class recall on balanced dataset

Harmonic mean :

High penalty on small recall

$$HM(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}} \qquad (5)$$

But using reciprocal, so it is hard and numeric unstable to be optimized

Method – why GML work ?

maximize the geometric mean of per-class recall :

$$GM(x_1,\ldots,x_n)=\sqrt[n]{\left|x_1\times\ldots\times x_n\right|} \qquad (6)$$

Heavily affected

$$\frac{0.01+0.99}{2}=0.5$$

Less sensitive

$$\mathrm{HM}(0.01,0.99)=0.02$$

$$\mathrm{GM}(0.01,0.99)=0.10$$

Method – why GML work ?

using a simple logarithm transformation :

$$\sqrt[C]{r_1 \dots r_C} = \exp\left(\log\left(r_1 \dots r_C\right)^{1/C}\right) \tag{7}$$

$$= \exp\left(\frac{1}{C}\log\left(r_1 \dots r_C\right)\right) \tag{8}$$

$$= \exp\left(\frac{1}{C}\sum_{i=1}^{C}\log r_i\right). \tag{9}$$

surrogate

$\bar{p}_c \longleftrightarrow r_c$

$$\sqrt[C]{r_1 \dots r_C} \propto \exp(-\mathcal{L}_{\mathrm{GML}}). \tag{10}$$

Method – combine

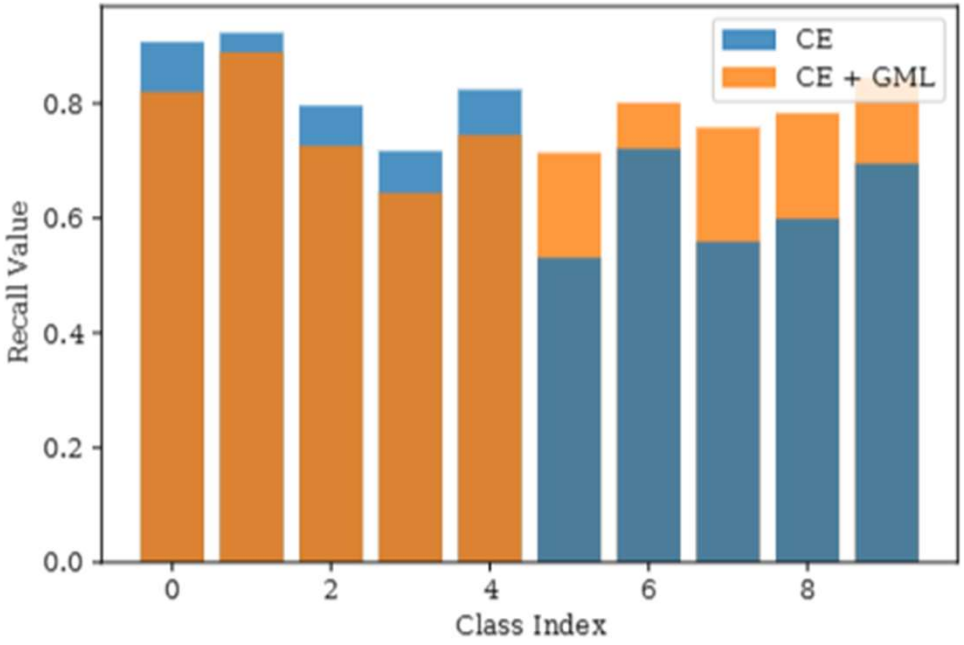


Figure 2. Bar plot of per-class recall on the imbalanced CIFAR10 (with imbalance ratio 100) before ('CE') and after ('CE+GML') the fine-tuning. For the fine-tuned model, the recall of the first 5 classes dropped while the recall of the latter 5 classes increased, motivating us to combine the prediction of both models.

Method – combine

$$\tilde{\boldsymbol{p}}_{\text{ensembled}} = \frac{\tilde{\boldsymbol{p}}_{\text{new}} + \tilde{\boldsymbol{p}}_{\text{old}}}{2} . \qquad (11)$$

$$\tilde{\boldsymbol{p}}_{\text{new}} = \text{softmax}\left(\frac{\boldsymbol{o}_{\text{new}}}{t_{\text{new}}}\right) , \qquad (12)$$

$$\tilde{\boldsymbol{p}}_{\text{old}} = \text{softmax}\left(\frac{\boldsymbol{o}_{\text{old}}}{t_{\text{old}}}\right) . \qquad (13)$$

output of old classifier

Method – training pipeline

1. Pre-training stage : Obtain the pre-trained model from scratch using any one of the existing methods.

2. **Fine-tuning stage** : Freeze the backbone and re-train the classifier using our proposed loss function.

3. **An optional ensemble stage** : Calibrate the prediction of two classifiers and combine them additively.

Method – training pipeline

Algorithm 1 The overall training procedure

Input: Training images $\boldsymbol{x}$ and their labels y.

1: Randomly initialize the network and train it from scratch to obtain the pre-trained model.
2: Use the pre-trained model to perform initialization.
3: Freeze the backbone and re-initialize the classifier.
4: Fine-tune the model using the loss function defined as Eq. (1) for a few epochs.
5: During inference, combine the results from both classifiers as defined in Eq. (11).

$$\mathcal{L}_{\mathrm{GML}} = -\frac{1}{C}\sum_{c=1}^{C}\log \bar{\boldsymbol{p}}_c\,, \tag{1}$$

$$\tilde{\boldsymbol{p}}_{\mathrm{ensembled}} = \frac{\tilde{\boldsymbol{p}}_{\mathrm{new}} + \tilde{\boldsymbol{p}}_{\mathrm{old}}}{2}\,. \tag{11}$$

Experiments - datasets

CIFAR100-LT (imbalance ratio = 100)

ImageNet-LT and Places-LT

Dataset	Number of Classes	# Training Images	# Test Images	Imbalance Ratio
CIFAR100-LT [16]	100	10,847	10,000	100
ImageNet-LT [6, 19]	1,000	115,846	50,000	256
Places-LT [32]	365	62,500	36,500	996

Table 2. Statistics of three imbalanced datasets used in our experiments.

Experiments - Evaluation Metrics

3 subsets

- Head : >= 100 images;
- Medium : others;
- Few : <= 20;

accuracy within each subset

overall accuracy

Experiments – on CIFAR100LT

Methods	Accuracy	Geometric Mean	Harmonic Mean	Lowest Recall
CE	38.74	21.03*	2.05*	0.00
BSCE [22]	42.24	35.16	26.24	3.00
LDAM [2]	43.51	33.61	21.49	3.00
LADE [13]	44.39	39.05	32.58	5.00
DiVE [12]	45.11	37.08	25.18	2.00
MiSLAS [30]	47.05	40.16	30.93	5.00
RIDE [26]	48.64	38.71	23.86	2.00
PaCo [4]	51.24	45.29	36.42	5.00
CE + GML	41.06	36.59	31.26	6.00
PaCo + GML	50.53	45.47	39.20	9.00
PaCo + GML (Ensemble)	49.82	**45.70**	**41.02**	**15.00**

Table 3. Results on the CIFAR100-LT dataset with imbalance ratio 100. Numbers with * are computed by substituting zero elements with a small number (10^{-3}) or else the geometric and harmonic mean will all be zero.

Experiments – on ImageNet-LT

Methods	Accuracy	Geometric Mean	Harmonic Mean	Lowest Recall
CE	43.90	23.25*	1.25*	0.00
BSCE [22]	50.48	42.32*	13.74*	0.00
cRT [14]	49.64	41.35*	13.82*	0.00
DiVE [12]	53.63	45.49*	12.76*	0.00
RIDE [26]	55.69	47.56*	17.32*	0.00
PaCo [4]	58.53	51.32*	21.79*	0.00
CE + GML	45.61	39.82	31.67	2.00
PaCo + GML	55.57	50.67	43.71	2.00
PaCo + GML (Ensemble)	57.22	**52.29**	**44.80**	**2.00**

Table 4. Results on the ImageNet-LT dataset. Numbers with * are computed by substituting zero elements with a small number (10^{-3}) or else the geometric and harmonic mean will all be zero.

Experiments – on ImageNet-LT

Methods	Accuracy	Geometric Mean	Harmonic Mean	Lowest Recall
CE	28.71	12.11*	0.73*	0.00
BSCE [22]	37.18	29.30*	5.64*	0.00
PaCo [4]	40.45	27.88*	2.53*	0.00
MiSLAS [30]	40.48	35.53	28.97	3.00
CE + GML	36.82	29.30*	8.00*	0.00
MiSLAS + GML	39.90	35.41	29.98	5.00
MiSLAS + GML (Ensemble)	40.37	**35.92**	**30.51**	**5.00**

Table 5. Results on the Places-LT dataset. Numbers with * are computed by substituting zero elements with a small number (10^{-3}) or else the geometric and harmonic mean will all be zero. Note that to plug GML into MiSLAS, since MiSLAS is also a two-stage method, we think it's fairer to use GML together with their proposed label aware smoothing loss.

Experiments – improve on some methods

Methods	G-Mean	H-Mean	Lowest Recall
CE	21.03*	2.05*	0.00
CE + GML	36.59 (15.56↑)	31.26 (29.21↑)	6.00 (6.00↑)
BSCE	35.16	26.24	3.00
BSCE + GML	36.52 (1.36↑)	30.88 (4.64↑)	7.00 (4.00↑)
MiSLAS	40.16	30.93	5.00
MiSLAS + GML	40.90 (0.74↑)	36.49 (5.56↑)	11.00 (6.00↑)
PaCo	45.29	36.42	5.00
PaCo + GML	45.47 (0.18↑)	39.20 (2.78↑)	9.00 (4.00↑)

Table 6. Results on the CIFAR100-LT dataset with imbalance ratio 100. Our proposed method is applicable to various methods. "G-Mean" is short for "Geometric Mean" and "H-Mean" is short for "Harmonic Mean".

Experiments – improve on some methods

Methods	G-Mean	H-Mean	Lowest Recall
CE	21.03*	2.05*	0.00
CE + GML	36.59 (15.56↑)	31.26 (29.21↑)	6.00 (6.00↑)
BSCE	35.16	26.24	3.00
BSCE + GML	36.52 (1.36↑)	30.88 (4.64↑)	7.00 (4.00↑)
MiSLAS	40.16	30.93	5.00
MiSLAS + GML	40.90 (0.74↑)	36.49 (5.56↑)	11.00 (6.00↑)
PaCo	45.29	36.42	5.00
PaCo + GML	45.47 (0.18↑)	39.20 (2.78↑)	9.00 (4.00↑)

Table 6. Results on the CIFAR100-LT dataset with imbalance ratio 100. Our proposed method is applicable to various methods. "G-Mean" is short for "Geometric Mean" and "H-Mean" is short for "Harmonic Mean".

Experiments – Ablation

Default : CIFAR100-LT

Methods	G-Mean	H-Mean	Lowest Recall
CE + GML	**36.59**	**31.26**	**6.00**
(w/o re-weighting)	25.45	5.11	0.00
(w/o re-weighting, re-sampling)	32.32	23.05	4.00

Table 8. Ablation on the re-weighting design in GML.

poor without re-weighting

Experiments – training from scratch

Methods	G-Mean	H-Mean	Lowest Recall
CE + GML	36.59	31.26	6.00
PaCo + GML	**45.47**	**39.20**	**9.00**
GML (train from scratch)	36.63	30.86	**9.00**

Table 9. Training from scratch using GML.

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

THANKS