VIRTUAL JUNE 19-25

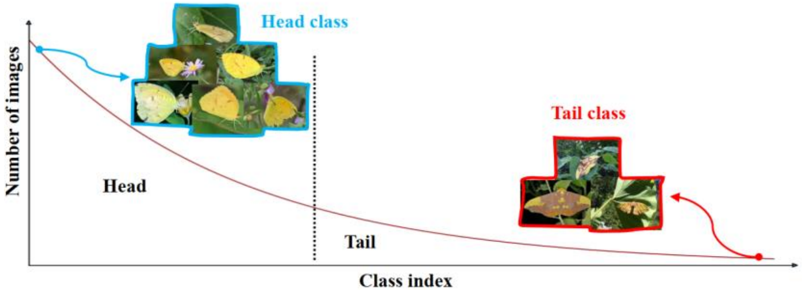
Improving Calibration for Long-Tailed Recognition

Zhisheng Zhong **Jiequan Cui** **Shu Liu** **Jiaya Jia**

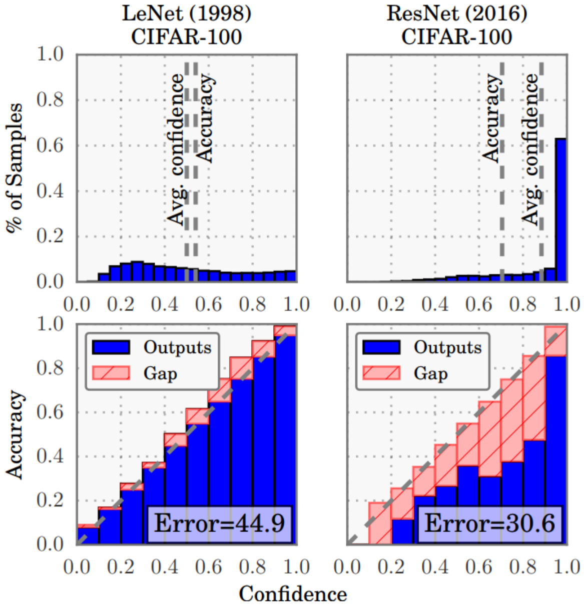
Chinese University of Hong Kong SmartMore

Background

Long-tailed Recognition



Confidence Calibration

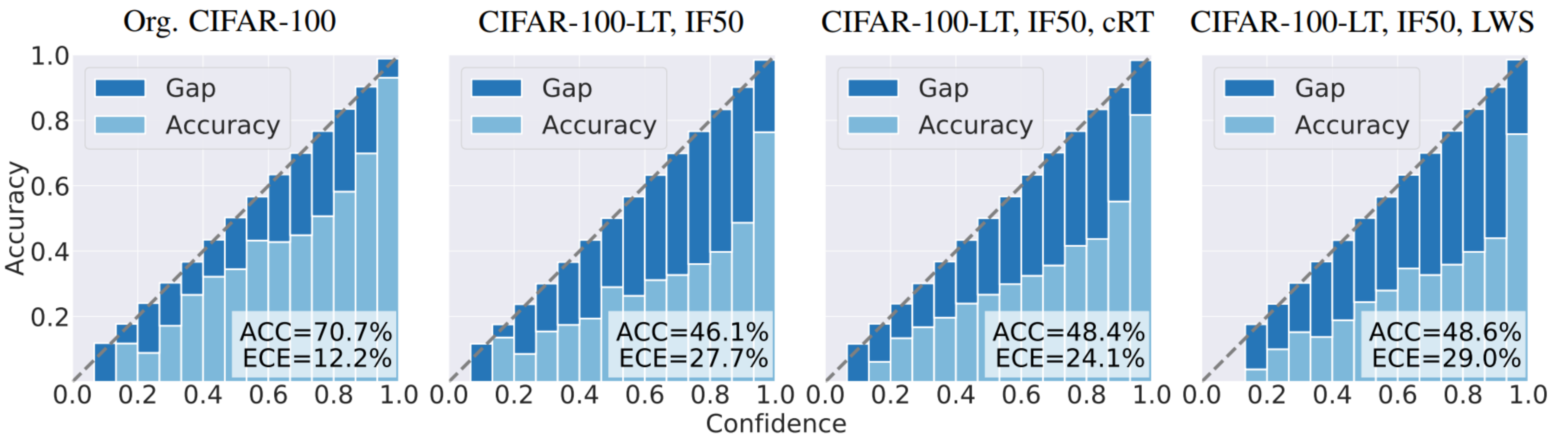


[1] Zhou, et.al, BBN Bilateral Branch Network with Cumulative Learning for Long-Tailed Visual Recognition, CVPR 2020
[2] Guo, et.al, On Calibration of Modern Neural Networks, ICML2017

Background

Plain CE model | **Plain CE model** | **SOTA two-stage decoupling models**

Balanced data | **Imbalanced data** | **ICLR 2020**

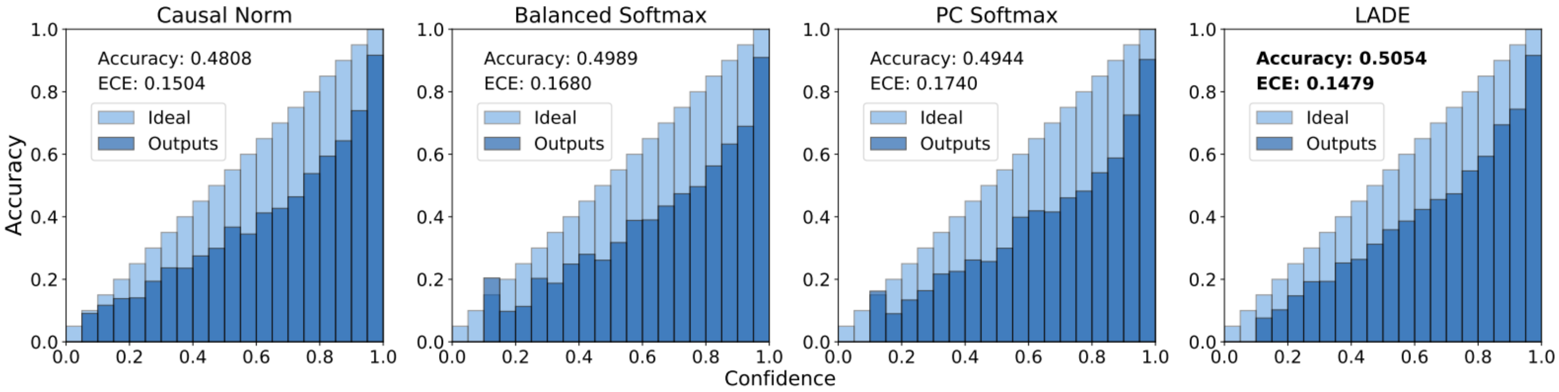


*** Backbone: ResNet-32**

Background

SOTA one-stage models

NeurIPS 2020 **NeurIPS 2020** **CVPR 2021**



*** Backbone: ResNet-32**

Best Top-1 Acc.: 50.6% & Best ECE: 14.7%

Conclusion: because of the **imbalanced composition ratio of each class**, networks trained on long-tailed datasets are more **mis-calibrated** and **over-confident**. The SOTA one-stage models, and two-stage models suffer terrible over-confidence as well.

[3] Hong, et.al, Disentangling Label Distribution for Long-tailed Visual Recognition, CVPR 2021

MiSLAS

In this paper, we focus on the **two-stage decoupling models (cRT & LWS)**:

To relieve the over-confidence issue in long-tailed recognition,

We explore **soft label** methods for long-tailed recognition

mixup

Label smoothing

Knowledge distillation

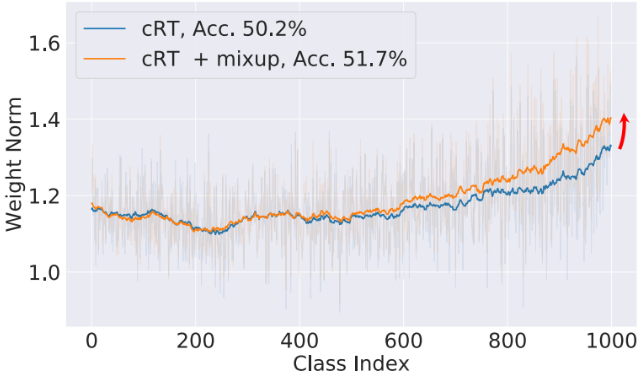
MiSLAS

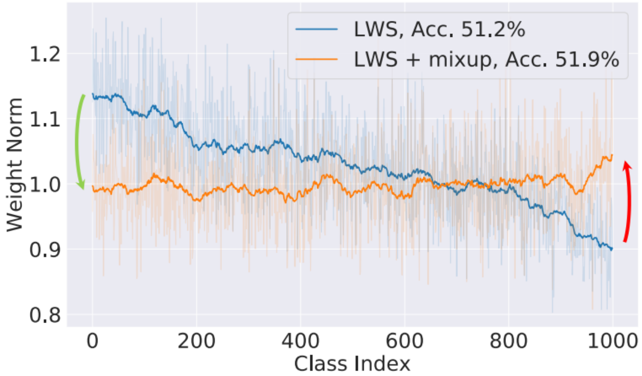
Study of mixup Strategy

Mark	Stg.-1	Stg.-2	ResNet-50	ResNet-101	ResNet-152
CE	☒		45.7 / 13.7	47.3 / 13.7	48.7 / 14.5
CE	☑		45.5 / 7.98	47.7 / 10.1	48.3 / 10.2
cRT	☒	☒	50.3 / 8.97	51.3 / 9.34	52.7 / 9.05
cRT	☒	☑	50.2 / 3.32	51.3 / 3.38	52.8 / 3.60
cRT	☑	☒	**51.7** / 5.62	**53.1** / 6.86	**54.2** / 6.02
cRT	☑	☑	51.6 / **3.13**	53.0 / **2.93**	54.1 / **3.37**

Mark	Stg.-1	Stg.-2	ResNet-50	ResNet-101	ResNet-152
CE	☒		45.7 / 13.7	47.3 / 13.7	48.7 / 14.5
CE	☑		45.5 / 7.98	47.7 / 10.1	48.3 / 10.2
LWS	☒	☒	51.2 / 4.89	52.3 / 5.10	53.8 / 4.48
LWS	☒	☑	51.0 / 5.01	52.2 / 5.38	53.6 / 5.50
LWS	☑	☒	52.0 / **2.23**	**53.5** / **2.73**	**54.6** / **2.46**
LWS	☑	☑	52.0 / 8.04	53.3 / 6.97	54.4 / 7.74

Top-1 Acc. (↑, the higher the better) / ECE (↓, the lower the better)





MiSLAS

Cross-entropy

Formulation: $$l(y, \boldsymbol{p}) = -\log(\boldsymbol{p}_y) = -\boldsymbol{w}_y^\top \boldsymbol{x} + \log(\sum \exp(\boldsymbol{w}_i^\top \boldsymbol{x}))$$

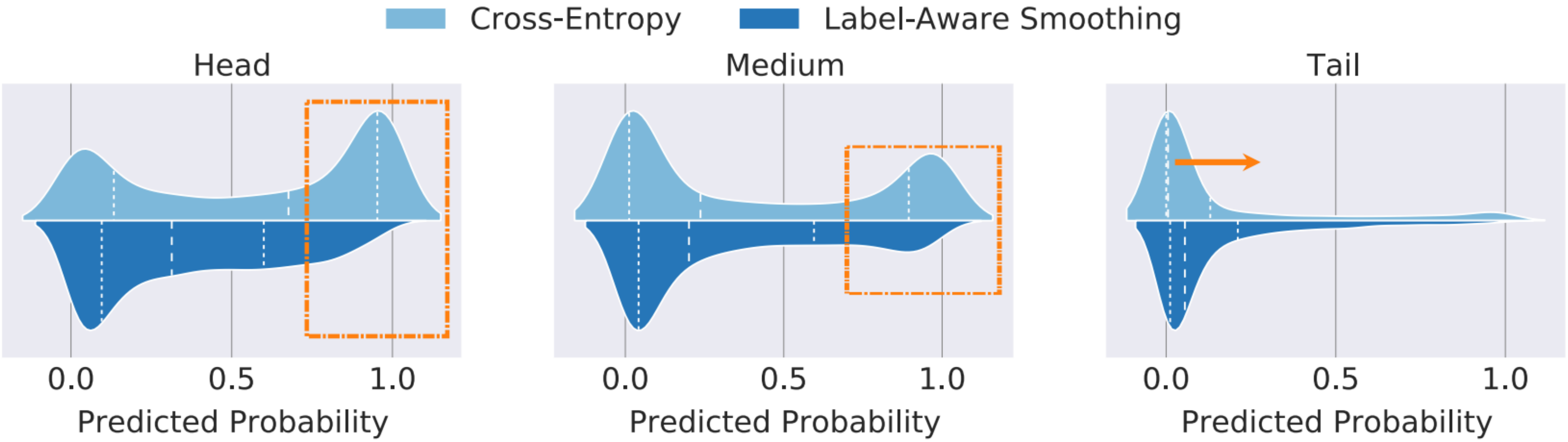
Optimal solution: ${\boldsymbol{w}_y^*}^\top \boldsymbol{x} = \inf$ while keeping others $\boldsymbol{w}_i^\top \boldsymbol{x}$, $i \neq y$, small enough.

Label-aware Smoothing

Formulation: $$l(\boldsymbol{q}, \boldsymbol{p}) = -\sum_{i=1}^{K} \boldsymbol{q}_i \log \boldsymbol{p}_i, \quad \boldsymbol{q}_i = \begin{cases} 1 - \epsilon_y = 1 - f(N_y), & i = y, \\ \frac{\epsilon_y}{K-1} = \frac{f(N_y)}{K-1}, & \text{otherwise}, \end{cases}$$

Optimal solution: $${\boldsymbol{w}_i^*}^\top \boldsymbol{x} = \begin{cases} \log\left(\frac{(K-1)(1-\epsilon_y)}{\epsilon_y}\right) + c, & i = y, \\ c, & \text{otherwise}, \end{cases}$$

Label-aware Smoothing



Label-aware Smoothing

Three types of related function $\boldsymbol{f}(\cdot)$:

$$\epsilon_y = f(N_y) = \begin{cases} \text{(Concave)} & \epsilon_K + (\epsilon_1 - \epsilon_K)\sin\left[\frac{\pi(N_y - N_K)}{2(N_1 - N_K)}\right], & y = 1, 2, ..., K, \\ \text{(Linear)} & \epsilon_K + (\epsilon_1 - \epsilon_K)\frac{N_y - N_K}{N_1 - N_K}, & y = 1, 2, ..., K, \\ \text{(Convex)} & \epsilon_1 + (\epsilon_1 - \epsilon_K)\sin\left[\frac{3\pi}{2} + \frac{\pi(N_y - N_K)}{2(N_1 - N_K)}\right], & y = 1, 2, ..., K, \end{cases}$$

A more powerful classifier framework for Stage-2:

$$\boldsymbol{z} = \mathrm{diag}(\boldsymbol{s})\left(r\boldsymbol{W} + \Delta\boldsymbol{W}\right)^{\top}\boldsymbol{x}$$

Shift Learning on Batch Normalization

The SOTA two-stage decoupling methods ignore the dataset bias or domain shift between these two stages (the distributions of the dataset with **different sampling manners** are inconsistent for two stages). We focus on BN to relieve the dataset bias problem.

$$\boldsymbol{\mu}_{\mathrm{I}}^{(j)} = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{x}_i)^{(j)}, \quad \boldsymbol{\sigma}_{\mathrm{I}}^{2\,(j)} = \frac{1}{m}\sum_{i=1}^{m}\left[g(\boldsymbol{x}_i)^{(j)} - \boldsymbol{\mu}_{\mathrm{I}}^{(j)}\right]^2, \quad \boldsymbol{x}_i \sim P_{\mathcal{D}_{\mathrm{I}}}(\boldsymbol{x}, y),$$

$$\boldsymbol{\mu}_{\mathrm{C}}^{(j)} = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{x}_i)^{(j)}, \quad \boldsymbol{\sigma}_{\mathrm{C}}^{2\,(j)} = \frac{1}{m}\sum_{i=1}^{m}\left[g(\boldsymbol{x}_i)^{(j)} - \boldsymbol{\mu}_{\mathrm{C}}^{(j)}\right]^2, \quad \boldsymbol{x}_i \sim P_{\mathcal{D}_{\mathrm{C}}}(\boldsymbol{x}, y).$$

we unfreeze the update procedures of the running mean $\boldsymbol{\mu}$ and running variance $\boldsymbol{\sigma}$ but fix the learnable linear transformation parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for a better normalization in Stage-2.

Experiment

Recognition Accuracy

Method	CIFAR-10-LT			CIFAR-100-LT		
	100	50	10	100	50	10
CE	70.4	74.8	86.4	38.4	43.9	55.8
mixup [37]	73.1	77.8	87.1	39.6	45.0	58.2
LDAM+DRW [4]	77.1	81.1	88.4	42.1	46.7	58.8
BBN(include mixup) [39]	79.9	82.2	88.4	42.6	47.1	59.2
Remix+DRW(300 epochs) [5]	79.8	-	89.1	46.8	-	61.3
cRT+mixup	79.1 / 10.6	84.2 / 6.89	89.8 / 3.92	45.1 / 13.8	50.9 / 10.8	62.1 / 6.83
LWS+mixup	76.3 / 15.6	82.6 / 11.0	89.6 / 5.41	44.2 / 22.5	50.7 / 19.2	62.3 / 13.4
MiSLAS	**82.1 / 3.70**	**85.7 / 2.17**	**90.0 / 1.20**	**47.0 / 4.83**	$\boxed{\textbf{52.3 / 2.25}}$	**63.2 / 1.73**

Best Top-1 accuracy and the lowest ECE than all previous SOTA methods

Experiment

Recognition Accuracy

Method	ResNet-50
CE	44.6
CE+DRW [4]	48.5
Focal+DRW [18]	47.9
LDAM+DRW [4]	48.8
CRT+mixup	51.7 / 5.62
LWS+mixup	52.0 / 2.23
MiSLAS	**52.7 / 1.80**

(a) ImageNet-LT

Method	ResNet-50
CB-Focal [7]	61.1
LDAM+DRW [4]	68.0
BBN(include mixup) [39]	69.6
Remix+DRW [5]	70.5
cRT+mixup	70.2 / **1.79**
LWS+mixup(under-conf.)	70.9 / 9.41
MiSLAS(under-conf.)	**71.6** / 7.67

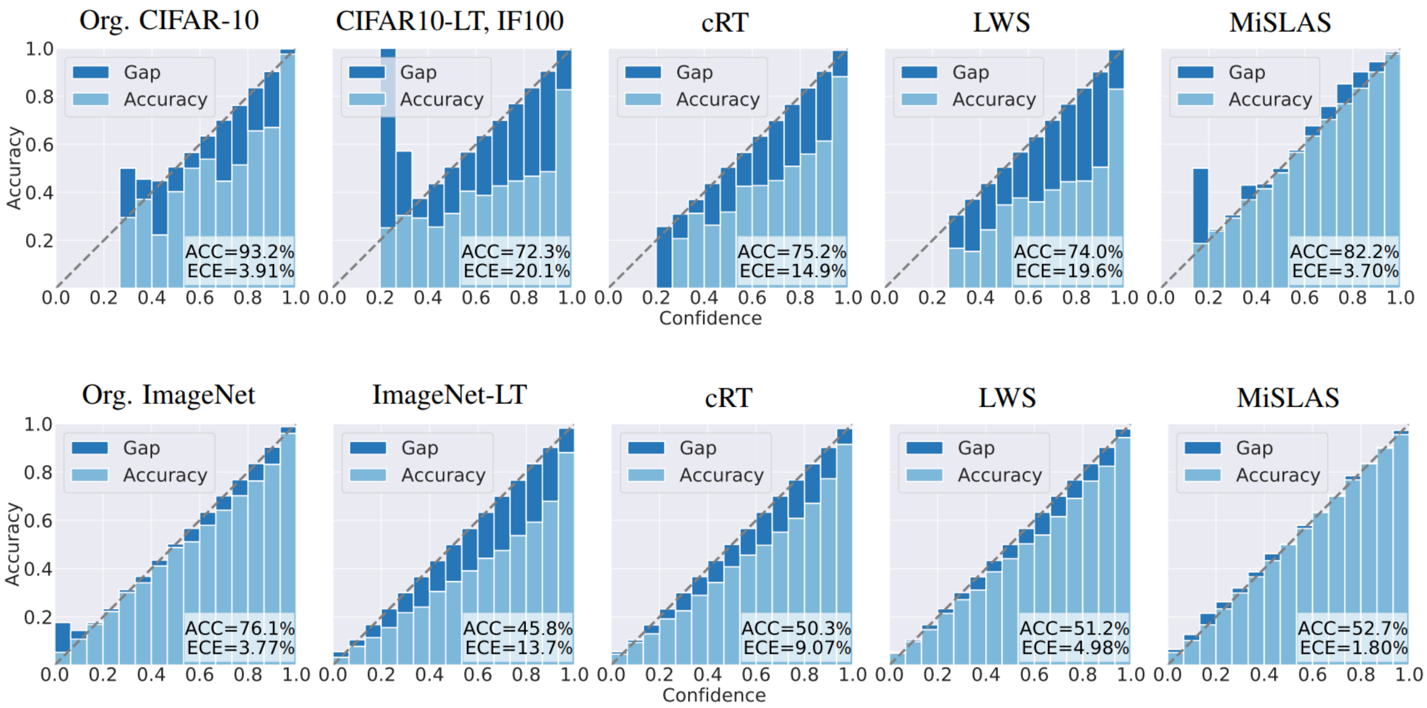
(b) iNaturalist 2018

Method	ResNet-152
Range Loss [38]	35.1
FSLwF [8]	34.9
OLTR [20]	35.9
OLTR+LFME [35]	36.2
cRT+mixup	38.3 / 12.4
LWS+mixup	39.7 / 11.7
MiSLAS	**40.4 / 3.60**

(c) Places-LT

Experiment

Confidence Calibration



Improving Calibration for Long-Tailed Recognition

Zhisheng Zhong **Jiequan Cui** **Shu Liu** **Jiaya Jia**

Chinese University of Hong Kong SmartMore

Thanks for listening!

Our paper

For more details

Our code