LEARNING IMBALANCED DATA WITH VISION TRANSFORMERS

ZHENGZHUO XU, RUIKANG LIU, SHUO YANG, ZENGHAO CHAI, CHUN YUAN.

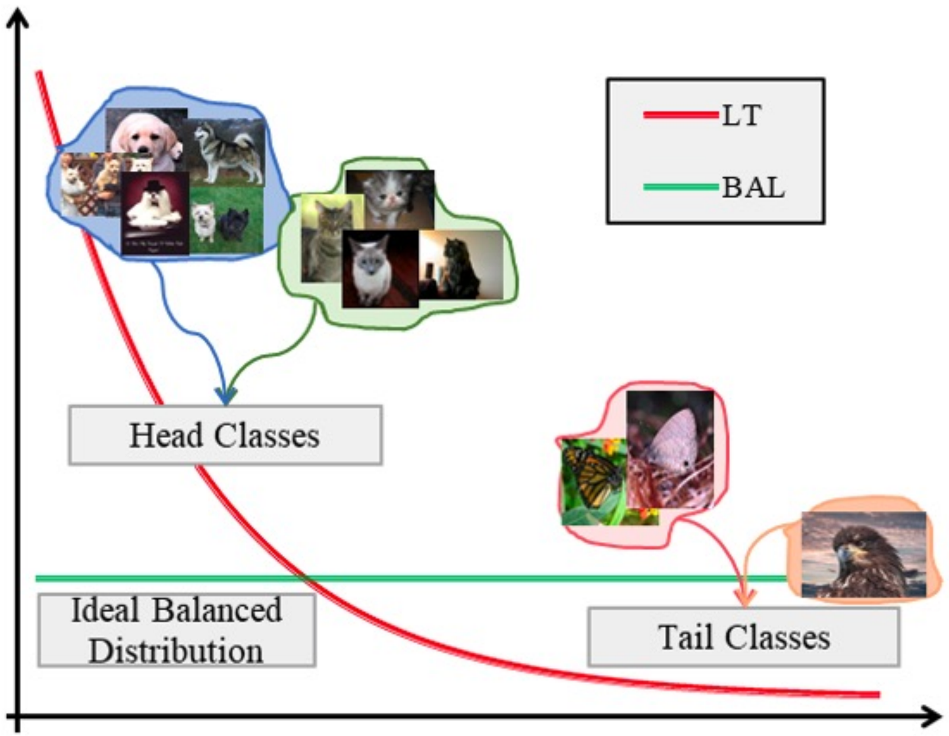
TSINGHUA UNIVERSITY

2023 / 5 / 15

LONG-TAILED RECOGNITION

What is long-tailed recognition?

- Training samples exhibit a long-tailed class distribution, where a small portion of classes have a massive number of sample points but the others are associated with only a few samples.
- The trained model can be easily biased towards head classes with massive training data, leading to poor model performance on tail classes that have limited data.
- Existing LTR methods seldom train Vision Transformers (ViTs) with Long-Tailed (LT) data, while the off the-shelf pretrain weight of ViTs leads to unfair comparisons.



ABSTRACT

We train the vision transformers from scratch with Long-Tailed data.

In summary, our main contributions:

- **To our best knowledge, we are the first to investigate training ViTs from scratch with LT data systematically.**
- **We pinpoint that the masked generative pretraining is robust to LT data, which avoids the toxic influence of imbalanced labels on feature learning.**
- **With a solid theoretical grounding, we propose the balanced version of BCE loss (Bal-BCE), which improves the vanilla BCE by a large margin in LTR.**
- **We propose LiVT recipe to train ViTs from scratch, and the performance of LiVT achieves state-of-the-art across various benchmarks for long-tailed recognition.**

PREVIOUS RECIPES TO TRAIN VIT-B

- *Previous recipes are difficult to train vision transformers.*
- *Self-supervised training is more robust than label-supervised methods.*
- *We select masked generative pretraining to learn feature from LT data.*

Dataset	ViT	Δ	DeiT III	Δ	MAE	Δ
ImageNet-BAL	38.7	-	67.2	-	69.2	-
ImageNet-LT	31.6	-7.0	48.4	-18.8	54.5	-14.7

Top-1 accuracy (%) of different recipes to train ViT-B-16 from scratch on ImageNet-LT/BAL. All perform much worse on LT than BAL.

START FROM BALANCED CROSS-ENTROPY

- **Balanced Cross-Entropy (BalCE) is proposed by Ren et al in NeurIPS 2020, which reweight the *softmax* logits with training instance numbers.**
- **If we implement it via logit adjustment, we have the following theorem 1.**
- **The bias item of logits will be the negative log of the number of training samples.**

Theorem 1. Logit Bias of Balanced CE. Let $\pi_{\mathbf{y}_i} = n_{\mathbf{y}_i}/N$ be the training label $\mathbf{y}_i$ distribution. If we implement the balanced cross-entropy loss via logit adjustment, the bias item of logit $\mathbf{z}_{\mathbf{y}_i}$ will be $\mathcal{B}^{\text{ce}}_{\mathbf{y}_i} = \log \pi_{\mathbf{y}_i}$, i.e.,

$$\begin{aligned}
\mathcal{L}_{\text{Bal-CE}} &= \log[1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\log n_{\mathbf{y}_j} - \log n_{\mathbf{y}_i}} \cdot e^{\mathbf{z}_{\mathbf{y}_j} - \mathbf{z}_{\mathbf{y}_i}}] \\
&= \log[1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{(\mathbf{z}_{\mathbf{y}_j} + \log n_{\mathbf{y}_j}) - (\mathbf{z}_{\mathbf{y}_i} + \log n_{\mathbf{y}_i})}] \\
&= \log[1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{(\mathbf{z}_{\mathbf{y}_j} + \log \pi_{\mathbf{y}_j}) - (\mathbf{z}_{\mathbf{y}_i} + \log \pi_{\mathbf{y}_i})}].
\end{aligned} \tag{3}$$

BALANCED BINARY CROSS ENTROPY

- **Why binary cross-cross-entropy?**
  - **Generally, Binary Cross-Entropy loss performs better than Cross-Entropy loss when collaborating with ViTs.**
  - **It fails to catch up with widely adopted Balanced Cross-Entropy loss and shows severe training instability in LTR.**
  - **Following Ren et al, we add the training instance numbers to the sigmoid logits.**

Theorem 2. Logit Bias of Balanced BCE. Let $\pi_{\mathbf{y}_i} = n_{\mathbf{y}_i}/N$ be the class $\mathbf{y}_i$ distribution. If we implement the balanced binary cross-entropy loss via logit adjustment, the bias item of logit $\mathbf{z}_{\mathbf{y}_i}$ will be $\mathcal{B}^{\text{bce}}_{\mathbf{y}_i} = \log \pi_{\mathbf{y}_i} - \boldsymbol{\log(1 - \pi_{\mathbf{y}_i})}$,

$$\mathcal{L}_{\text{Bal-BCE}} = -\sum_{\mathbf{y}_i \in \mathcal{C}} w_i[\mathbb{1}(\mathbf{y}_i) \cdot \log \frac{1}{1 + e^{-[\mathbf{z}_{\mathbf{y}_i} + \log \pi_{\mathbf{y}_i} - \log(1-\pi_{\mathbf{y}_i})]}} + (1 - \mathbb{1}(\mathbf{y}_i)) \cdot \log(1 - \frac{1}{1 + e^{-[\mathbf{z}_{\mathbf{y}_i} + \log \pi_{\mathbf{y}_i} - \log(1-\pi_{\mathbf{y}_i})]}})] \tag{6}$$

A SIMPLE PROOF

- **We start by revising the sigmoid activation function:**

$$\sigma(\mathbf{z}_{\mathbf{y}_i}) = \frac{1}{1+e^{-\mathbf{z}_{\mathbf{y}_i}}} = \frac{e^0}{e^0+e^{-\mathbf{z}_{\mathbf{y}_i}}} = \frac{e^{\mathbf{z}_{\mathbf{y}_i}}}{e^{\mathbf{z}_{\mathbf{y}_i}}+e^0}$$

- **If we view it as the binary version of softmax, e^{x} (e^{0}) will be the normalized probability to indicate *yes* (*no*).**

$$\begin{aligned}
\hat{\sigma}(\mathbf{z}_{\mathbf{y}_i}) &= \frac{n_{\mathbf{y}_i} \cdot e^{\mathbf{z}_{\mathbf{y}_i}}}{n_{\mathbf{y}_i} \cdot e^{\mathbf{z}_{\mathbf{y}_i}} + (N - n_{\mathbf{y}_i}) \cdot e^0} \\
&= \frac{\pi_{\mathbf{y}_i} \cdot e^{\mathbf{z}_{\mathbf{y}_i}}}{\pi_{\mathbf{y}_i} \cdot e^{\mathbf{z}_{\mathbf{y}_i}} + (1 - \pi_{\mathbf{y}_i}) \cdot e^0} \\
&= \frac{1}{1 + \frac{1-\pi_{\mathbf{y}_i}}{\pi_{\mathbf{y}_i}} \cdot e^{-\mathbf{z}_{\mathbf{y}_i}}}
\end{aligned}$$

A SIMPLE PROOF

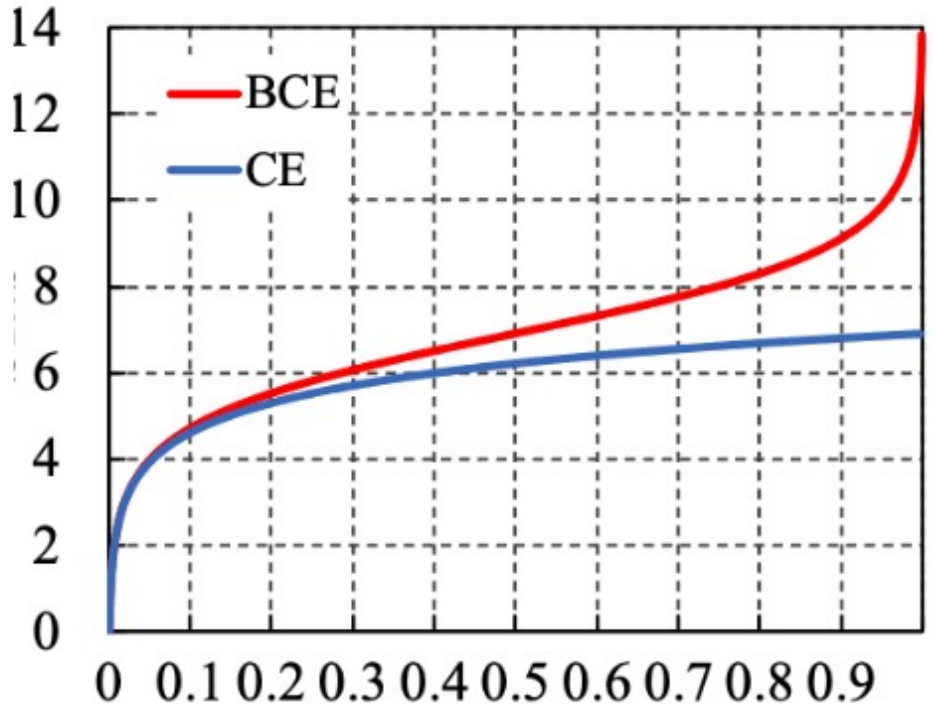
- **Considering the *log-sum-exp* for numerical stability:**

$$
\begin{aligned}
\hat{\sigma}(\mathbf{z}_{\mathbf{y}_i}) &= \frac{1}{1+\frac{1-\pi_{\mathbf{y}_i}}{\pi_{\mathbf{y}_i}}\cdot e^{-\mathbf{z}_{\mathbf{y}_i}}} = \frac{1}{1+e^{-\mathbf{z}_{\mathbf{y}_i}+\log\frac{1-\pi_{\mathbf{y}_i}}{\pi_{\mathbf{y}_i}}}} \\
&= \frac{1}{1+e^{-\mathbf{z}_{\mathbf{y}_i}+\log(1-\pi_{\mathbf{y}_i})-\log\pi_{\mathbf{y}_i}}} \\
&= \frac{1}{1+e^{-[\mathbf{z}_{\mathbf{y}_i}+\log\pi_{\mathbf{y}_i}-\log(1-\pi_{\mathbf{y}_i})]}}
\end{aligned}
$$

- **Please refer to Supp for another derivation from the Bayesian Theorem perspective.**

INTERPRETATION

- **Similar to Bal-CE , it enlarges the margins to increase the difficulty of the tail (smaller π_{yi}).**
- **Bal-BCE further reduces the head (larger π_{yi}) inter-class distances with larger positive values.**
- **BCE is not class-wise mutually exclusive, and the smaller head inter-class distance helps the networks focus more on the tail's contributions.**



PIPELINE

- **MGP for feature learning.**
  - **This stage adopts the masked auto encoder.**

Algorithm 1 LiVT Training Pipeline.

Input: $\mathcal{D}, \mathscr{F}, \mathscr{W}, \mathscr{D}, T_{pt}, T_{ft}, \mathcal{A}_{pt}, \mathcal{A}_{ft}, \pi_{\mathbf{y}_i}, \tau$
Output: Optimized θ_f, θ_w.

Initialize θ_f, θ_d randomly. ▷ MGP Stage
for $t = 1$ **to** T_{pt} **do**
  for $\{\mathbf{x}, \mathbf{y}\}$ sampled from $\mathcal{D}$ **do**
    $\mathbf{x} := \mathcal{A}_{pt}(\mathbf{x})$
    $\hat{\mathbf{x}} = \mathscr{D}(\mathscr{F}(\mathbf{M} \odot \mathbf{x} \mid \theta_f) \mid \theta_d)$
    $\mathcal{L}_{MSE}(\hat{\mathbf{x}}, \mathbf{x}) = ||\hat{\mathbf{x}} - \mathbf{x}||_2$
    $\{\theta_f, \theta_d\} \leftarrow \{\theta_f, \theta_d\} - \alpha \nabla_{\{\theta_f, \theta_d\}} \cdot \mathcal{L}_{MSE}(\hat{\mathbf{x}}, \mathbf{x})$
  end for
end for

- **BFT for unbiased classifier learning.**
  - **This stage adopts the Bal-BCE loss.**

Initialize θ_w randomly. ▷ BFT Stage
Calculate logit bias $\mathcal{B}^{\text{bce}}_{\mathbf{y}_i}$ via Eq. 10.
for $t = 1$ **to** T_{ft} **do**
  for $\{\mathbf{x}, \mathbf{y}\}$ sampled from $\mathcal{D}$ **do**
    $\mathbf{x} := \mathcal{A}_{ft}(\mathbf{x})$
    $\mathbf{v} = \mathscr{F}(\mathbf{x} \mid \theta_f)$
    $\mathbf{z} = \mathscr{W}(\mathbf{v} \mid \theta_w) + \tau \cdot \mathcal{B}^{\text{bce}}$
    Calculate $\mathcal{L}_{BCE}$ via Eq. 5 with calibrated $\mathbf{z}$.
    $\{\theta_f, \theta_w\} \leftarrow \{\theta_f, \theta_w\} - \alpha \nabla_{\{\theta_f, \theta_w\}} \cdot \mathcal{L}_{BCE}$
  end for
end for

EXPERIMENT

Table 2. Top-1 accuracy (%) of ResNet50 on ImageNet-LT. † indicates results with ResNeXt50. ∗: training with 384 resolution.

Method	Ref.	Many	Med.	Few	Acc
CE [13]	CVPR 19	64.0	33.8	5.8	41.6
LDAM [4]	NeurIPS 19	60.4	46.9	30.7	49.8
c-RT [29]	ICLR 20	61.8	46.2	27.3	49.6
τ-Norm [29]	ICLR 20	59.1	46.9	30.7	49.4
Causal [54]	NeurIPS 20	62.7	48.8	31.6	51.8
Logit Adj. [47]	ICLR 21	61.1	47.5	27.6	50.1
RIDE(4E)† [61]	ICLR 21	68.3	53.5	35.9	56.8
MiSLAS [80]	CVPR 21	62.9	50.7	34.3	52.7
DisAlign [75]	CVPR 21	61.3	52.2	31.4	52.9
ACE† [3]	ICCV 21	71.7	54.6	23.5	56.6
PaCo† [12]	ICCV 21	68.0	56.4	37.2	58.2
TADE† [77]	ICCV 21	66.5	57.0	**43.5**	58.8
TSC [36]	CVPR 22	63.5	49.7	30.4	52.4
GCL [35]	CVPR 22	63.0	52.7	37.1	54.5
TLC [33]	CVPR 22	68.9	55.7	40.8	55.1
BCL† [83]	CVPR 22	67.6	54.6	36.6	57.2
NCL [34]	CVPR 22	67.3	55.4	39.0	57.7
SAFA [23]	ECCV 22	63.8	49.9	33.4	53.1
DOC [58]	ECCV 22	65.1	52.8	34.2	55.0
DLSA [69]	ECCV 22	67.8	54.5	38.8	57.5
ViT-B training from scratch					
ViT [15]	ICLR 21	50.5	23.5	6.9	31.6
MAE [18]	CVPR 22	74.7	48.2	19.4	54.5
DeiT [55]	ECCV 22	70.4	40.9	12.8	48.4
LiVT	-	73.6	56.4	41.0	60.9
LiVT *	-	**76.4**	**59.7**	42.7	**63.8**

Table 3. Top-1 accuracy (%) of ResNet50 on iNaturalist 2018. ∗: training with 384 resolution.

Method	Ref.	Many	Med.	Few	Acc
CE [13]	CVPR 19	72.2	63.0	57.2	61.7
OLTR [44]	CVPR 19	59.0	64.1	64.9	63.9
c-RT [29]	ICLR 20	69.0	66.0	63.2	65.2
τ-Norm [29]	ICLR 20	65.6	65.3	65.9	65.6
LWS [29]	ICLR 20	65.0	66.3	65.5	65.9
BBN [81]	CVPR 20	61.8	73.6	66.9	69.6
BS [51]	ICLR 21	70.0	70.2	69.9	70.0
RIDE(4E) [61]	ICLR 21	70.9	72.5	73.1	72.6
DisAlign [75]	CVPR 21	69.0	71.1	70.2	70.6
MiSLAS [80]	CVPR 21	73.2	72.4	70.4	71.6
DiVE [21]	ICCV 21	70.6	70.0	67.6	69.1
ACE(4E) [3]	ICCV 21	-	-	-	72.9
TADE [77]	ICCV 21	74.4	72.5	73.1	72.9
PaCo [12]	ICCV 21	70.4	72.8	73.6	73.2
ALA [79]	AAAI 22	71.3	70.8	70.4	70.7
TSC [36]	CVPR 22	72.6	70.6	67.8	69.7
LTR-WD [1]	CVPR 22	71.2	70.4	69.7	70.2
GCL [35]	CVPR 22	67.5	71.3	71.5	71.0
BCL [83]	CVPR 22	66.7	71.0	70.7	70.4
NCL [34]	CVPR 22	72.0	74.9	73.8	74.2
DOC [58]	ECCV 22	72.8	71.7	70.0	71.0
DLSA [69]	ECCV 22	-	-	-	72.8
ViT-B training from scratch					
ViT [15]	ICLR 21	65.4	55.3	50.9	54.6
MAE [18]	CVPR 22	79.6	70.8	65.0	69.4
DeiT [55]	ECCV 22	72.9	62.8	55.8	61.0
LiVT	-	78.9	76.5	74.8	76.1
LiVT *	-	**83.2**	**81.5**	**79.7**	**81.0**

Table 4. Top-1 accuracy (%) of ResNet152 (with ImageNet-1K pretrained weight) on Places-LT. ∗: training with 384 resolution.

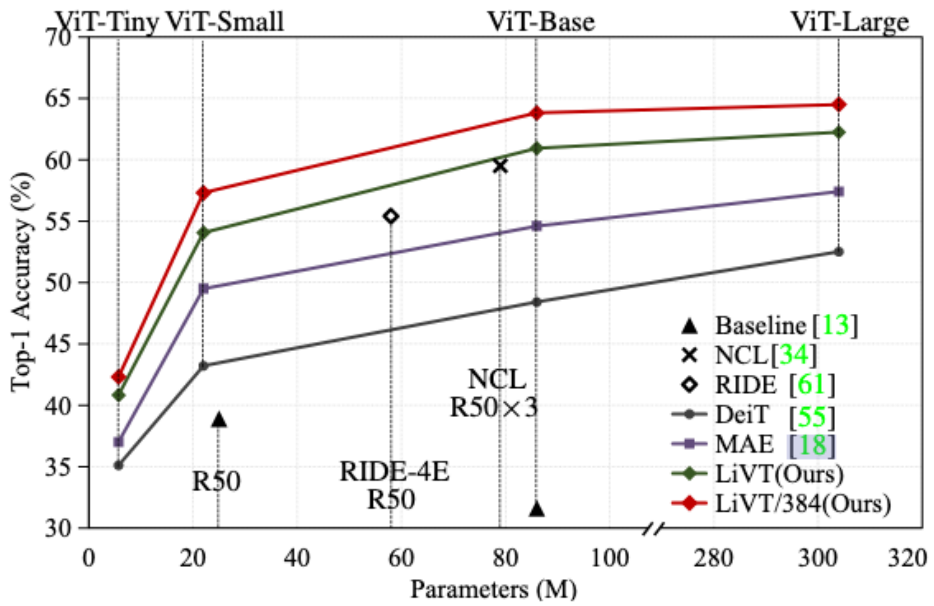
Method	Ref.	Many	Med.	Few	Acc
CE [13]	CVPR 19	45.7	27.3	8.2	30.2
Focal [38]	ICCV 17	41.1	34.8	22.4	34.6
Range [76]	CVPR 17	41.1	35.4	23.2	35.1
OLTR [44]	CVPR 19	44.7	37.0	25.3	35.9
FSA [10]	ECCV 20	42.8	37.5	22.7	36.4
LWS [29]	ICLR 20	40.6	39.1	28.6	37.6
Causal [54]	NeurIPS 20	23.8	35.8	**40.4**	32.4
BS [51]	NeurIPS 20	42.0	39.3	30.5	38.6
DisAlign [75]	CVPR 21	40.4	42.4	30.1	39.3
LADE [22]	CVPR 21	42.8	39.0	31.2	38.8
RSG [59]	CVPR 21	41.9	41.4	32.0	39.3
TADE [77]	ICCV 21	43.1	42.4	33.2	40.9
PaCo [12]	ICCV 21	36.1	**47.9**	35.3	41.2
ALA [79]	AAAI 22	43.9	40.1	32.9	40.1
NCL [34]	CVPR 22	-	-	-	41.8
BF [24]	CVPR 22	44.0	43.1	33.7	41.6
CKT [48]	CVPR 22	41.6	41.4	35.1	40.2
GCL [35]	CVPR 22	-	-	-	40.6
Bread [40]	ECCV 22	40.6	41.0	33.4	39.3
ViT-B training from scratch					
MAE [18]	CVPR 22	48.9	24.6	8.7	30.3
DeiT [55]	ECCV 22	**51.6**	31.0	9.4	34.2
LiVT	-	48.1	40.6	27.5	40.8
LiVT *	-	50.7	42.4	27.9	**42.6**

EXPERIMENT

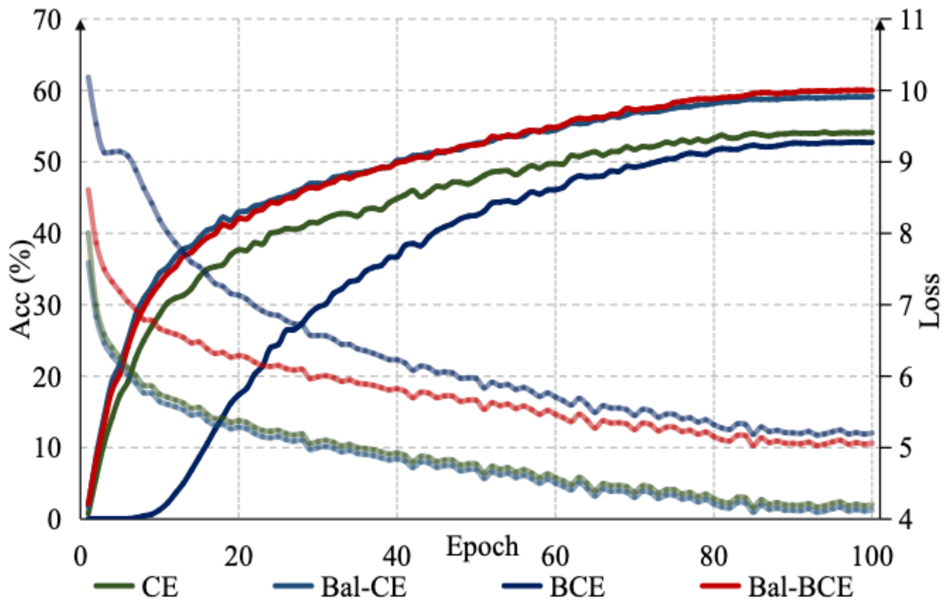
Model	Size	Loss	Many ↑	Med. ↑	Few ↑	Acc ↑	ECE ↓	MCE ↓
ViT-Tiny [55]	5.7M	CE	56.1	29.2	10.5	37.0	3.7	6.1
		Bal-CE	48.8 (-7.3)	39.2 (+10.0)	28.1 (+17.6)	**41.4** (+4.4)	2.6 (-1.1)	4.6 (-1.6)
		BCE	42.1	11.1	0.9	21.6	2.9	8.6
		Bal-BCE	50.6 (+8.4)	37.2 (+26.1)	26.1 (+25.2)	40.8 (+19.2)	3.1 (+0.1)	6.8 (-1.8)
ViT-Small [55]	22M	CE	68.9	43.1	17.3	49.5	4.7	9.2
		Bal-CE	62.7 (-6.2)	52.0 (+8.9)	36.3 (+19.0)	54.0 (+4.5)	0.9 (-3.8)	2.4 (-6.8)
		BCE	62.4	30.6	8.4	39.8	5.7	11.1
		Bal-BCE	65.8 (+3.4)	50.6 (+20.0)	32.9 (+24.6)	**54.1** (+14.2)	4.8 (-0.9)	9.0 (-2.2)
ViT-Base [15]	86M	CE	74.7	48.2	19.4	54.5	5.1	6.8
		Bal-CE	70.5 (-4.3)	56.8 (+8.6)	43.7 (+24.3)	60.1 (+5.6)	3.7 (-1.4)	4.9 (-1.9)
		BCE	73.7	46.5	15.6	52.4	5.6	7.9
		Bal-BCE	73.6 (-0.1)	55.8 (+9.3)	41.0 (+25.4)	**60.9** (+8.6)	2.4 (-3.1)	3.2 (-4.7)
ViT-Large [15]	304M	CE	77.3	51.5	21.7	57.4	3.6	7.4
		Bal-CE	72.7 (-4.5)	60.1 (+8.6)	41.9 (+20.3)	62.1 (+4.8)	2.1 (-1.5)	4.2 (-3.2)
		BCE	74.7	46.7	17.0	53.4	8.4	15.9
		Bal-BCE	75.3 (+0.6)	58.8 (+12.1)	37.5 (+20.5)	**62.6** (+9.2)	6.6 (-1.8)	14.8 (-1.1)

Performance on ImageNet-LT with different LT loss.

EXPERIMENT



Top-1 Acc v.s. Model size on ImageNet-LT.



Convergence

清華大學
Tsinghua University

THANKS FOR LISTENING！