A Unified Generalization Analysis of Re-Weighting and Logit-Adjustment for Imbalanced Learning

Zitai Wang, Qianqian Xu*, Zhiyong Yang,

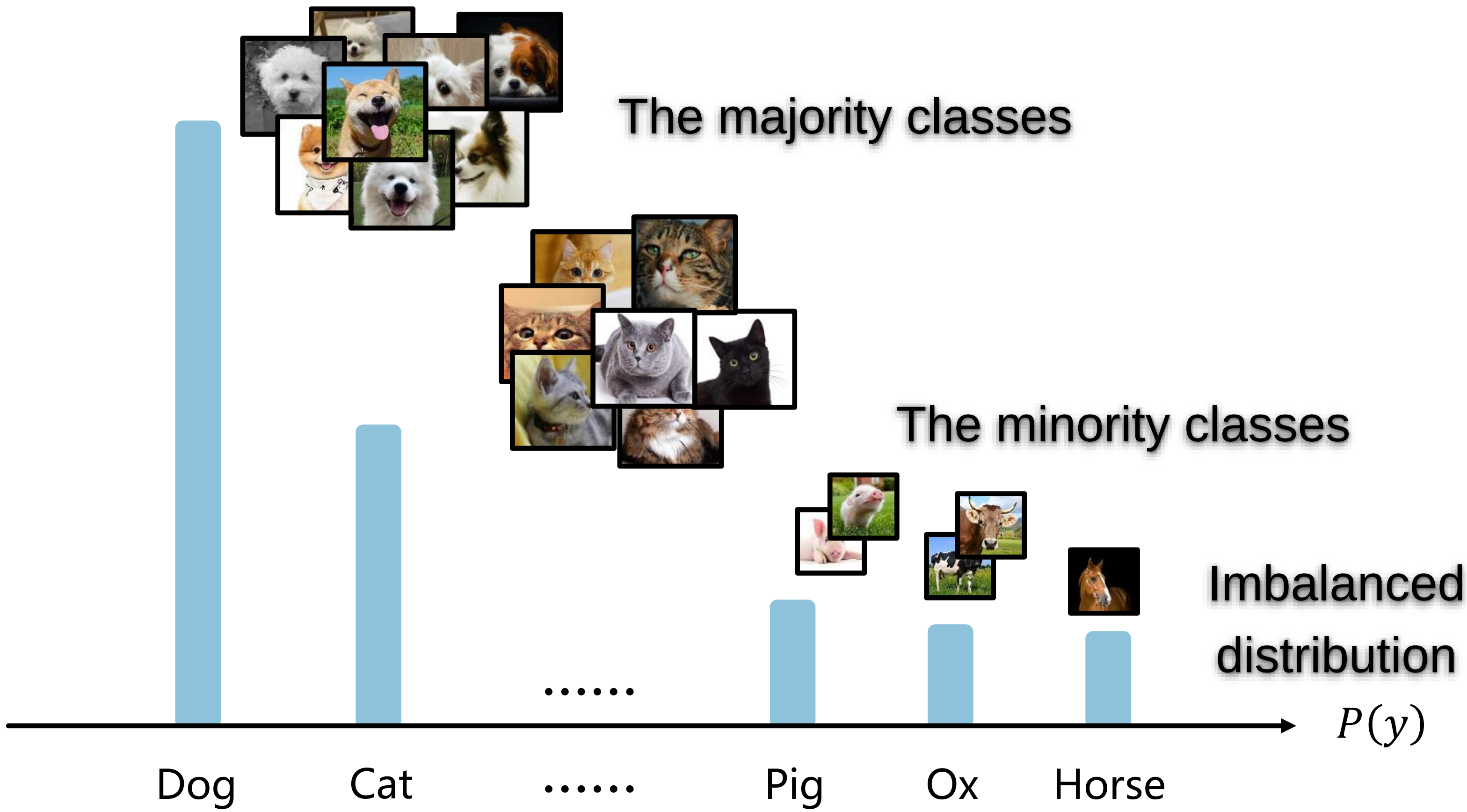
Yuan He, Xiaochun Cao, Qingming Huang*

Zitai Wang

2023.11

Background

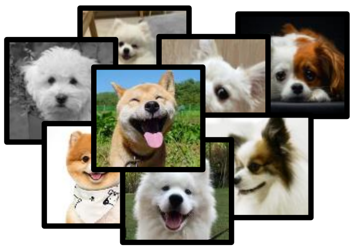
☐ Real-world datasets are generally imbalanced



A naïve ERM learning process will be biased!

Background

☐ Balanced Accuracy is a common metric in this case



99 Dogs

One Ox

A classifier only predicts dogs

Accuracy

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[f(x_i)=y_i]$$

× **99 / 100 = 99%, good model**

Balanced Accuracy

$$\frac{1}{C}\sum_{y=1}^{C}\frac{1}{N_y}\sum_{i=1}^{N_y}\mathbb{I}[f(x_i)=y_i]$$

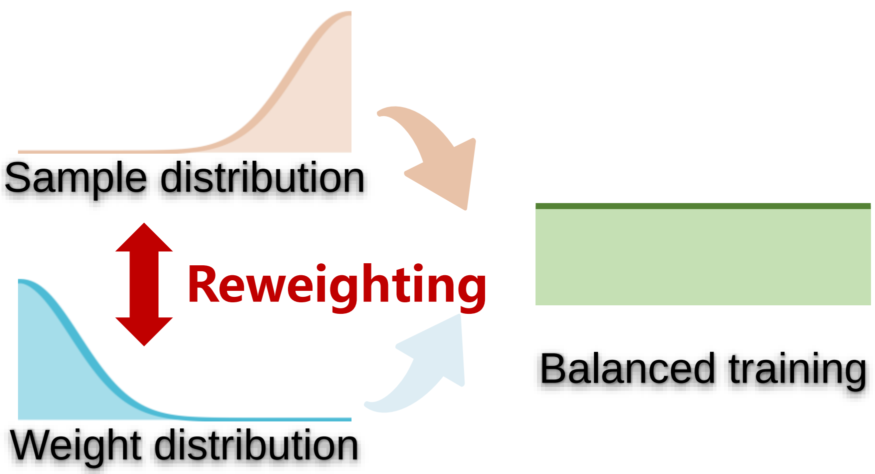
✓ **(0 + 1) / 2 = 50%, bad model**

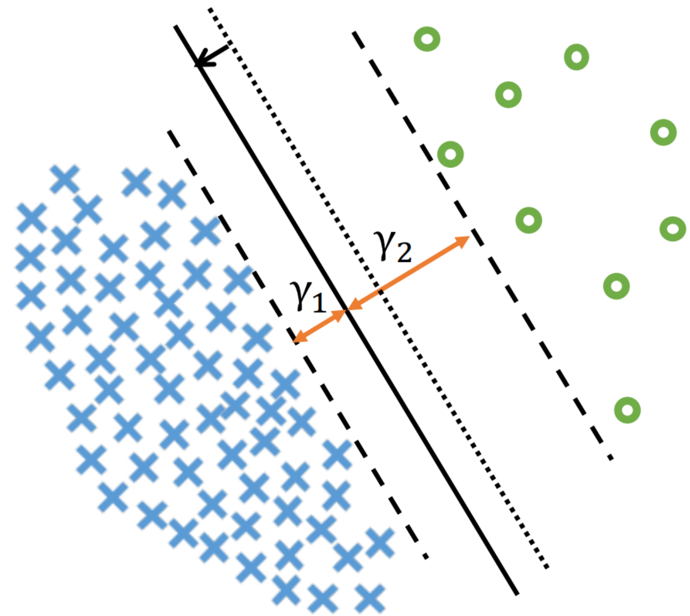
How to improve model performance on Balanced Acc.?

Prior arts

☐ Loss-modification approaches

- **Re-weighting [1, 2]**
- **Logit adjustment [3, 4, 5]**





[1] Combining statistical learning with a knowledge-based approach, ICML, 1999.
[2] Class-balanced loss based on effective number of samples, CVPR, 2019.
[3] Learning imbalanced datasets with label-distribution-aware margin loss, NeurIPS 2019.
[4] Long-tail learning via logit adjustment, ICLR, 2021
[5] Label-imbalanced and group-sensitive classification under overparameterization, NeurIPS 2021

Prior arts

☐ A unified formulation for RW and LA

- $\alpha_y = 1, \beta_y = 1, \Delta_y = 0 \rightarrow$ **Naïve CE loss**
- $\alpha_y = (1-p)/(1-p^{N_y}), \beta_y = 1, \Delta_y = 0 \rightarrow$ **CB loss [2]**
- $\alpha_y = 1, \beta_y = 1, \Delta_y = \tau \log \pi_y \rightarrow$ **LA loss [4]**
- $\alpha_y = 1, \beta_y = (N_y/N_1)^\gamma, \Delta_y = 0 \rightarrow$ **CDT loss [6]**

$$L_{\mathrm{VS}}(f(\boldsymbol{x}), y) = \underline{-\alpha_y} \log \left(\frac{e^{\beta_y f(\boldsymbol{x})_y + \Delta_y}}{\sum_{y'} e^{\underline{\beta_{y'}} f(\boldsymbol{x})_{y'} + \underline{\Delta_{y'}}}} \right)$$

Reweighting term

multiplicative adjustment term

additive adjustment term

[2] Class-balanced loss based on effective number of samples, CVPR, 2019.
[4] Long-tail learning via logit adjustment, ICLR, 2021
[6] Identifying and compensating for feature deviation in imbalanced deep learning, Arxiv, 2020

Limitation of prior arts

- **Theoretical insights are still fragmented and coarse-grained, failing to explain some empirical results**

Proposition (Union bound for Imbalanced Learning)

Given a function set $\mathcal{F}$ and a μ-Lipschitz continuous loss $L: \mathbb{R} \times C \to [0, M]$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the training set $\mathcal{S}$, the following generalization bound holds for all $g \in \mathcal{G}$:

$$\mathcal{R}_{bal}^{L}(f) = \frac{1}{C}\sum_{y=1}^{C} \mathcal{R}_{y}^{L}(f) \lesssim \frac{1}{C}\sum_{y=1}^{C}\left(\hat{\mathcal{R}}_{y}^{L}(f) + \mu\hat{\mathfrak{C}}_{S_y}(\mathcal{F}) + 3M\sqrt{\frac{\log 2C/\delta}{2N_y}}\right)$$

Balanced risk

Generation bound for each class, where **the Lipschitz Continuity is the only property of L utilized, which is global in nature**

[3] Learning imbalanced datasets with label-distribution-aware margin loss, NeurIPS 2019.

Main work

- **Theoretical insights:** propose **local Lipschitz continuity** and construct a **fine-grained generalization bound**

Theorem (Data-Dependent Bound for Imbalanced Learning)

Given a function set $\mathcal{F}$ and a loss function $L: \mathbb{R} \times C \to [0, M]$ with local Lipschitz constants $\{\mu_y\}_{y=1}^{C}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the training set $\mathcal{S}$, the following generalization bound holds for all $g \in \mathcal{G}$:

$$\mathcal{R}_{bal}^{L}(f) \lesssim \frac{1}{C\pi_C}\left[\hat{\mathcal{R}}^{L}(f) + 3M\sqrt{\frac{\log\frac{2C}{\delta}}{2N}}\right] + \frac{\hat{\mathfrak{C}}_{S}(\mathcal{F})}{C\pi_C}\sum_{y=1}^{C}\mu_y\sqrt{\pi_y}$$

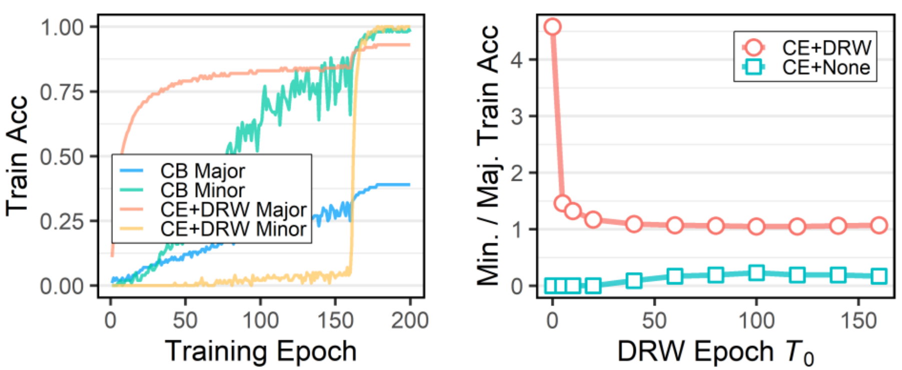
Balanced risk

Terms depending on training

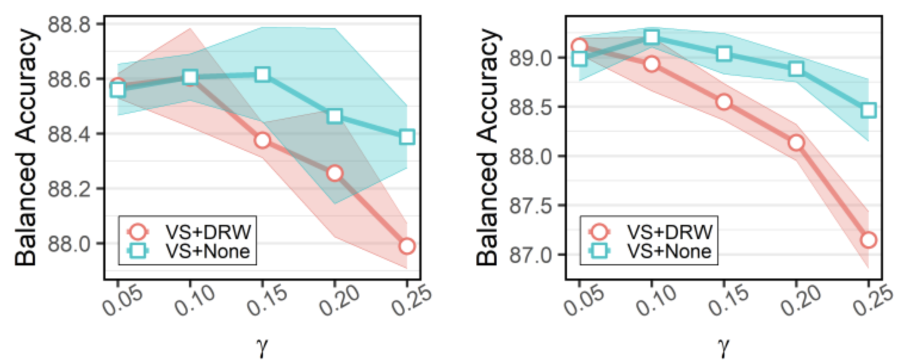
Loss dependent, where μ_y captures how the loss handles each class

Main work

- **Theoretical insights:** the fine-grained generalization bound is consistent with some empirical results
  - **Deferred scheme is necessary**
  - **Reweighting term and multiplicative adjustment term might be incompatible**



(a) CIFAR-100 LT ($\rho = 100$) (b) CIFAR-100 LT ($\rho = 100$)



(a) CIFAR-10 LT (b) CIFAR-10 Step

Main work

- **Improved algorithm**: **a principled learning algorithm induced by the theoretical insights**

Algorithm 1: Principled Learning Algorithm induced by the Theoretical Insights

---

Require: Training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N}$ and a model f parameterized by Θ.

```
Initialize the model parameters Θ randomly.
for t = 1, 2, ··· , T do
    B ← SampleMiniBatch(S, m)                          ▷ A mini-batch of m samples
    if t < T_0 then
        Set α = 1, β_y, Δ_y                            ▷ Adjust logits during the initial phase
    else
        Set α_y ∝ π_y^{-ν}, β_y = 1, Δ_y, ν > 0
    end if
    L(f, B) ← (1/m) Σ_{(x,y)∈B} L_VS(f(x), y)          ▷ Calculate the loss
    Θ ← Θ − η ∇_Θ L(f, B)                              ▷ One SGD step
    Optional: anneal the learning rate η.              ▷ Required when t = T_0
end for
```

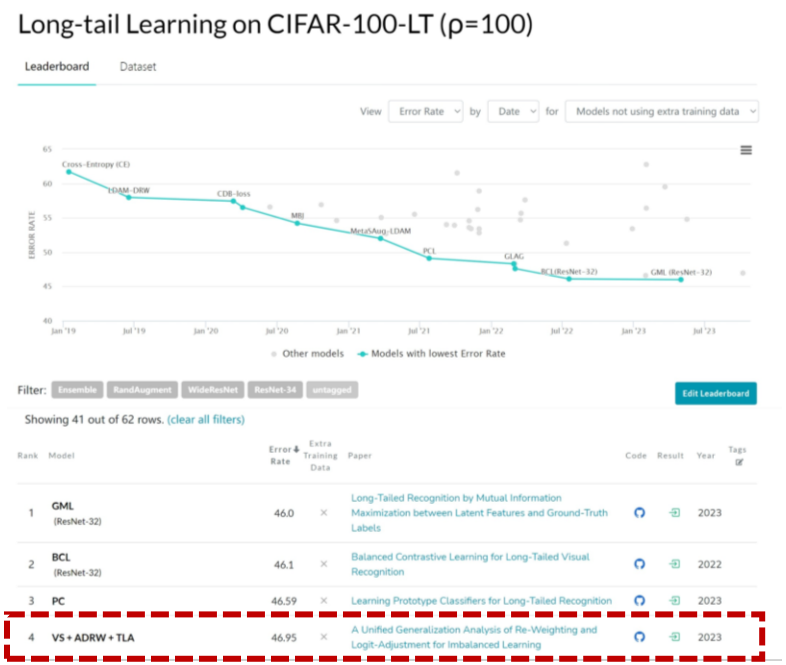
- ✓ **Reweighting is deferred and aligns with the bound**
- ✓ $\beta_y = 1$ **when reweighting is used**

Main work

- **Improved algorithm**: a principled learning algorithm induced by the theoretical insights

Dataset	CIFAR-10		CIFAR-100	
Imbalance Type	LT	Step	LT	Step
w/o SAM				
CE	71.5$_{\pm 0.4}$	64.8$_{\pm 0.9}$	38.3$_{\pm 0.4}$	38.6$_{\pm 0.2}$
LDAM	73.8$_{\pm 0.4}$	65.8$_{\pm 0.6}$	39.9$_{\pm 0.7}$	39.2$_{\pm 0.0}$
VS	78.8$_{\pm 0.2}$	76.1$_{\pm 0.7}$	41.8$_{\pm 0.7}$	**46.2**$_{\pm 0.3}$
CE+DRW	75.8$_{\pm 0.3}$	72.2$_{\pm 0.8}$	40.8$_{\pm 0.6}$	45.4$_{\pm 0.4}$
LDAM+DRW	77.7$_{\pm 0.4}$	77.8$_{\pm 0.5}$	**42.7**$_{\pm 0.5}$	45.3$_{\pm 0.6}$
VS+DRW	**80.1**$_{\pm 0.1}$	**78.2**$_{\pm 0.2}$	41.3$_{\pm 0.4}$	44.0$_{\pm 0.3}$
CE+ADRW	78.6$_{\pm 0.5}$	75.5$_{\pm 0.6}$	41.8$_{\pm 0.6}$	46.5$_{\pm 0.3}$
LDAM+ADRW	79.1$_{\pm 0.2}$	78.5$_{\pm 0.4}$	**43.0**$_{\pm 0.2}$	45.8$_{\pm 0.2}$
VS+TLA+DRW	**80.8**$_{\pm 0.2}$	**80.0**$_{\pm 0.1}$	43.0$_{\pm 0.4}$	**46.8**$_{\pm 0.1}$
VS+TLA+ADRW	**81.1**$_{\pm 0.2}$	**80.9**$_{\pm 0.2}$	**43.4**$_{\pm 0.6}$	**47.8**$_{\pm 0.1}$
w/ SAM				
CE+DRW	80.5$_{\pm 0.2}$	79.5$_{\pm 0.3}$	44.7$_{\pm 0.6}$	48.5$_{\pm 0.3}$
LDAM+DRW	81.6$_{\pm 0.2}$	81.2$_{\pm 0.7}$	45.2$_{\pm 0.3}$	**49.1**$_{\pm 0.2}$
VS	**82.6**$_{\pm 0.2}$	**83.2**$_{\pm 0.4}$	**45.9**$_{\pm 0.3}$	47.4$_{\pm 0.3}$
CE+ADRW	82.6$_{\pm 0.2}$	**82.8**$_{\pm 0.9}$	44.9$_{\pm 0.6}$	48.9$_{\pm 0.2}$
LDAM+ADRW	**83.0**$_{\pm 0.1}$	82.4$_{\pm 0.3}$	**46.3**$_{\pm 0.4}$	**49.3**$_{\pm 0.4}$
VS+TLA+ADRW	**83.6**$_{\pm 0.2}$	**83.8**$_{\pm 0.1}$	**46.4**$_{\pm 0.6}$	**49.1**$_{\pm 0.2}$

✓ **Rank 4th if using more techniques (models with extra training data or ensemble are filtered)**



Thanks for your listening!