南京航空航天大學

Nanjing University of Aeronautics and Astronautics

ParNeC 模式识别与神经计算研究组

PAttern Recognition and NEural Computing

Two Papers for Class-Imbalanced Learning

Feng Sun

2021-03-24

Class Imbalanced Learning

- Why we never heard about class imbalance problem in ML before?

ImageNet (Image Classification)

Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE CVPR. 2009.

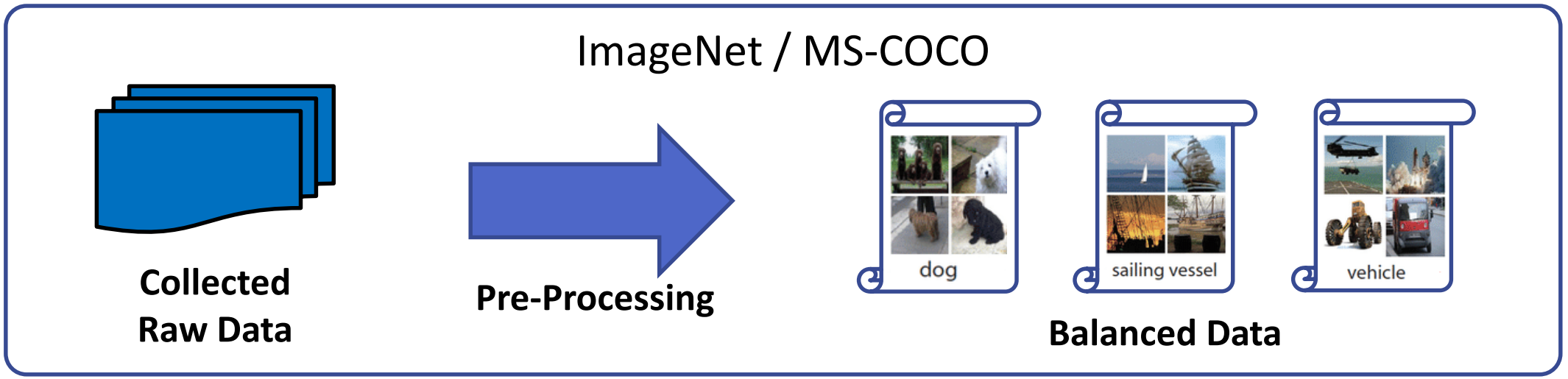
MS-COCO (Object Detection & Instance Segmentation)

Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." ECCV. Springer, Cham, 2014.

Class Imbalanced Learning

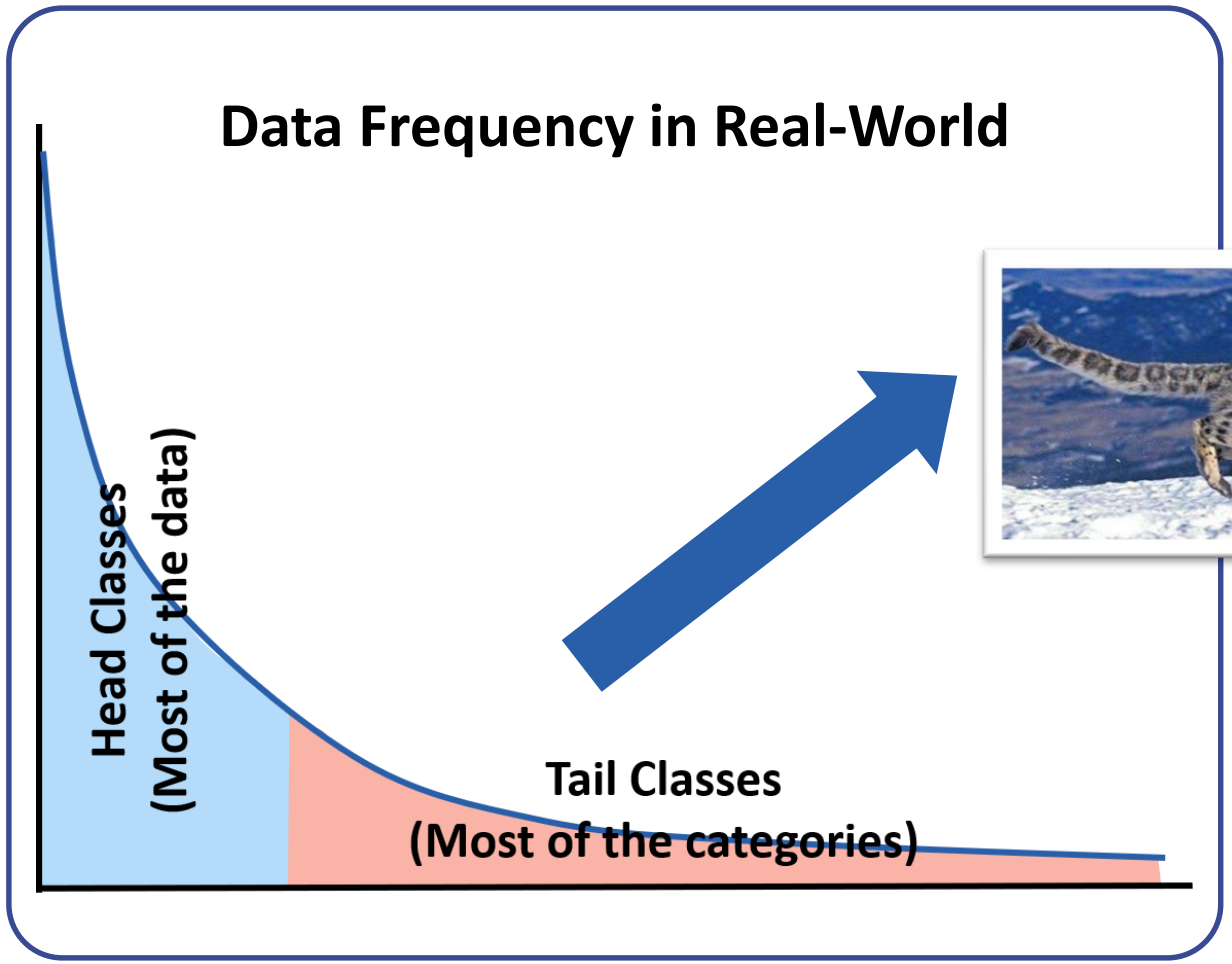
- Why we never heard about class imbalance problem in ML before?

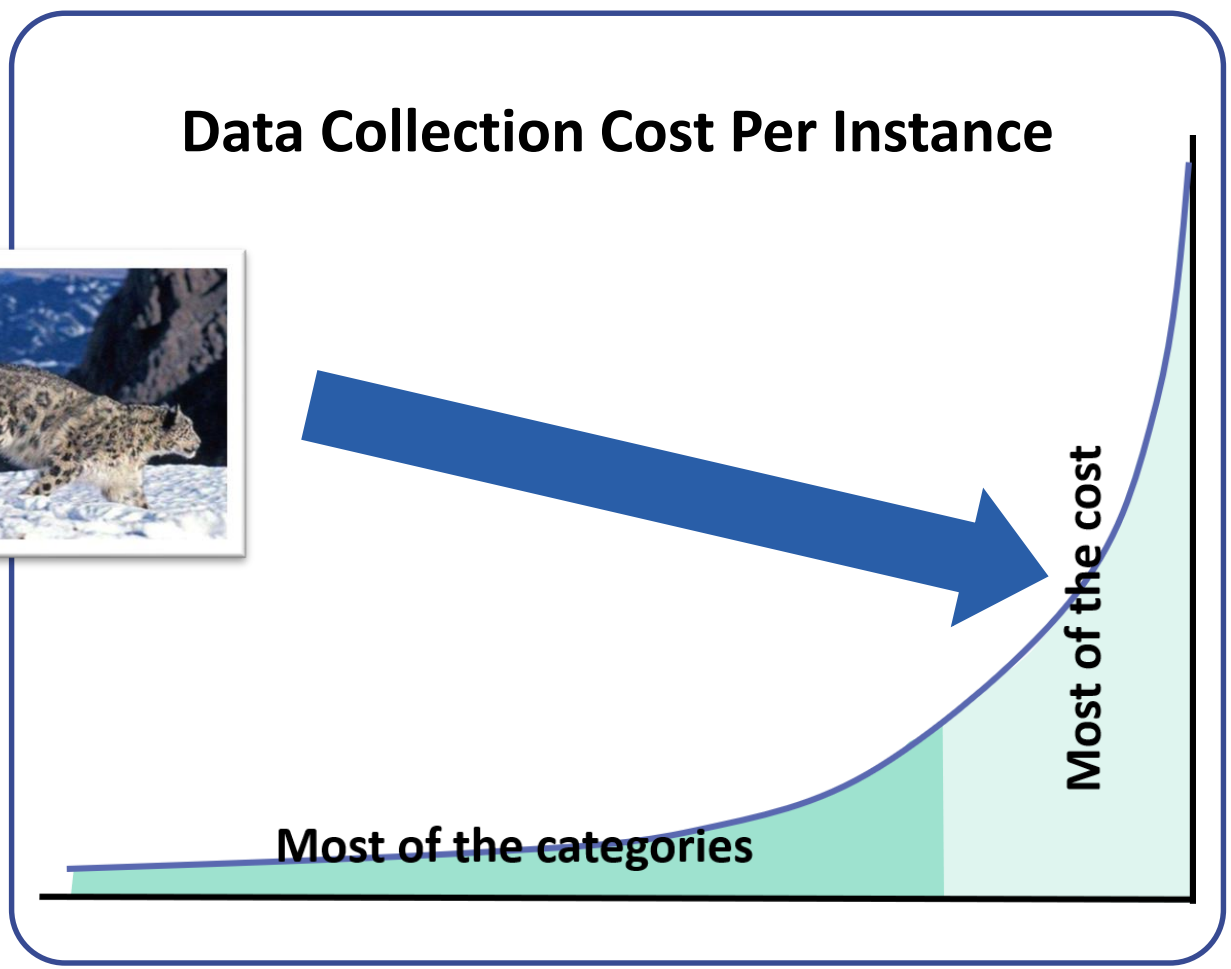
It's because the dataset we saw has already been balanced by the pre-processing in the data collection stage



Limitation of Balanced Datasets

- What's the problem of balancing all the dataset?

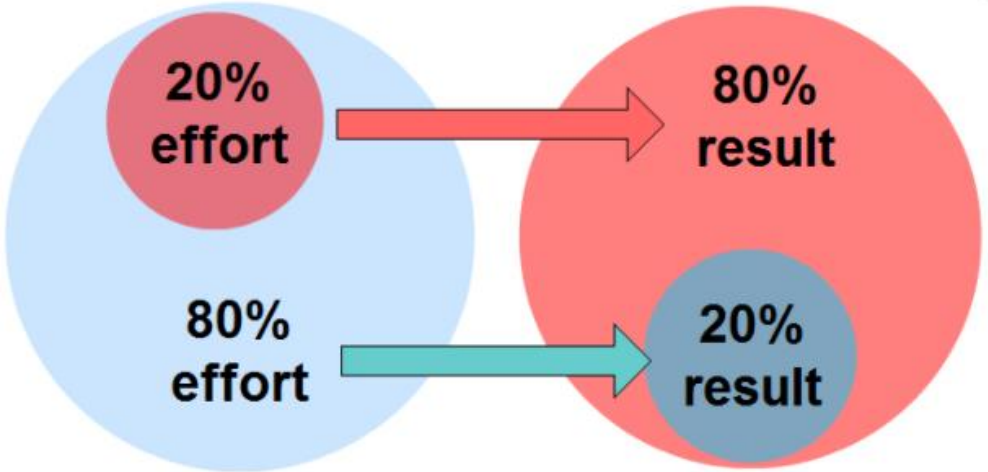


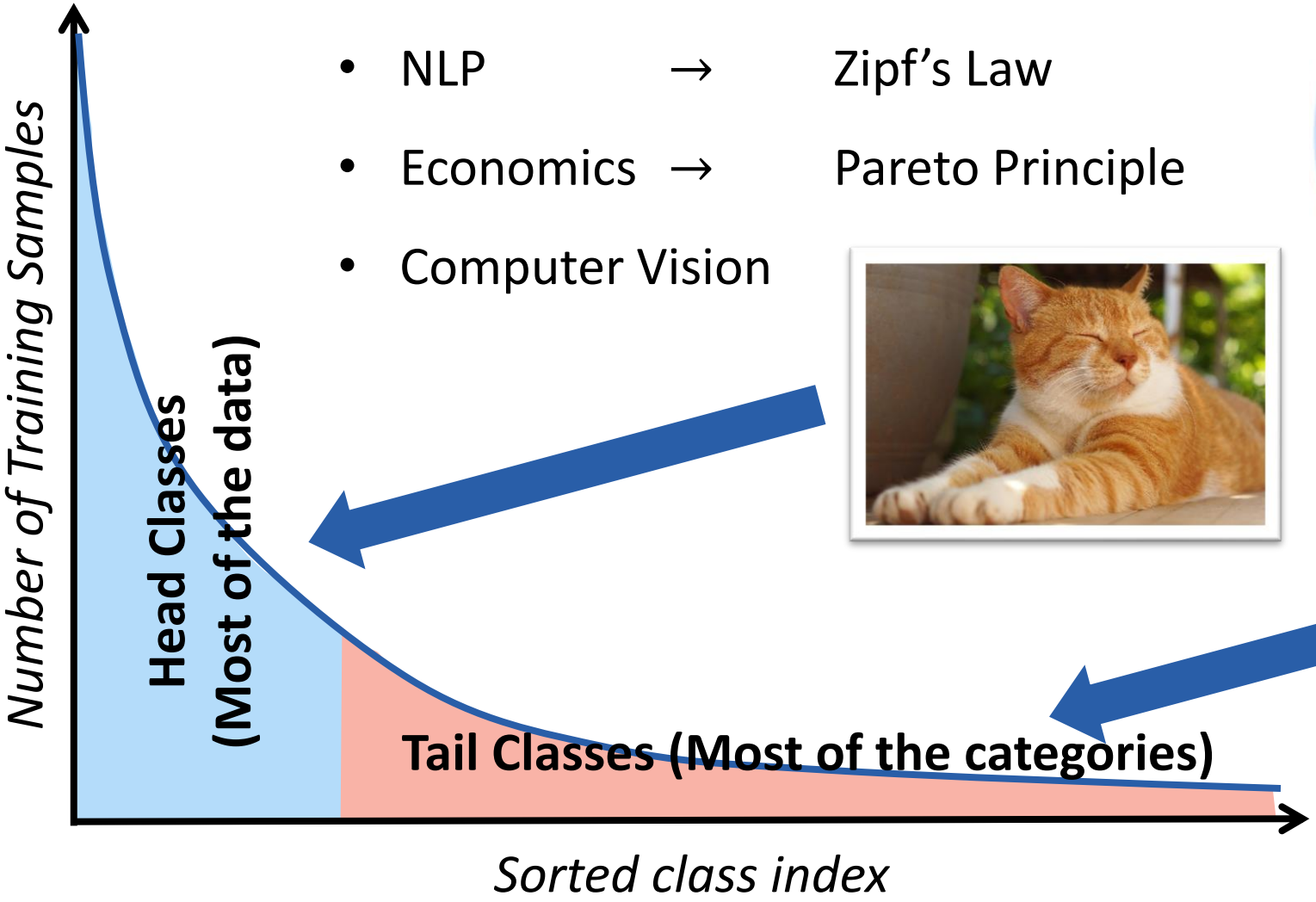


Long-Tailed Distribution

What is the long-tailed distribution?

- NLP → Zipf's Law
- Economics → Pareto Principle
- Computer Vision

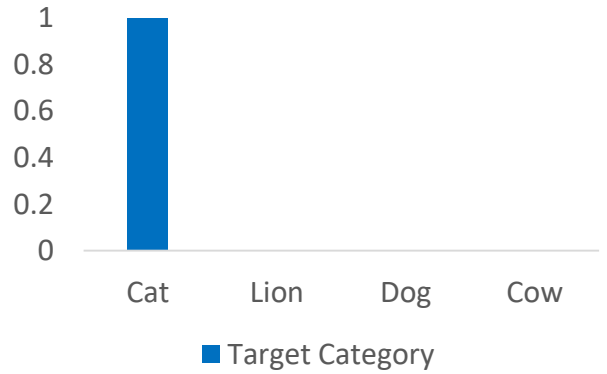


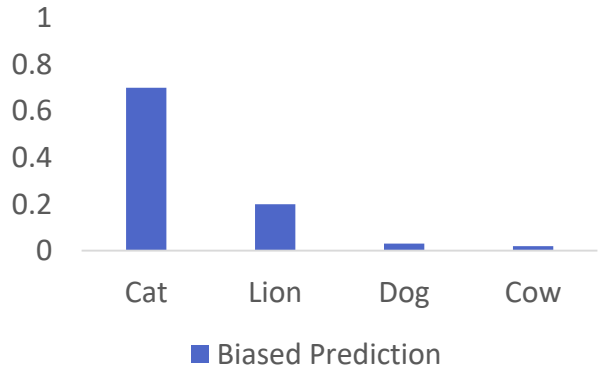


Long-Tailed Classification

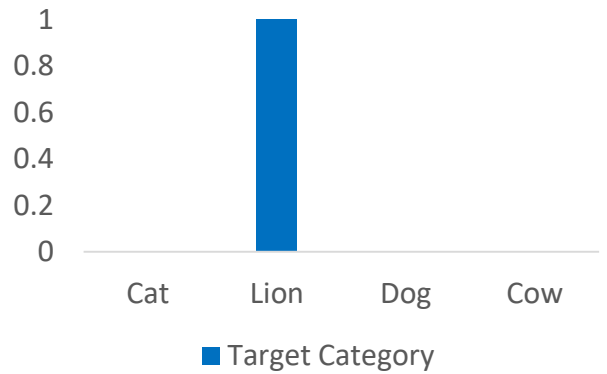
- The problem of long-tail datasets.

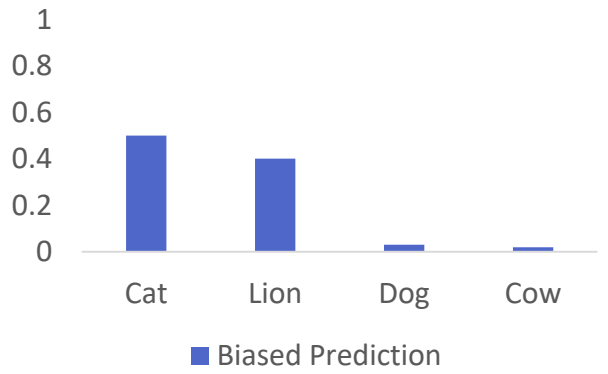
Head Class (1K)





Tail Class (10)

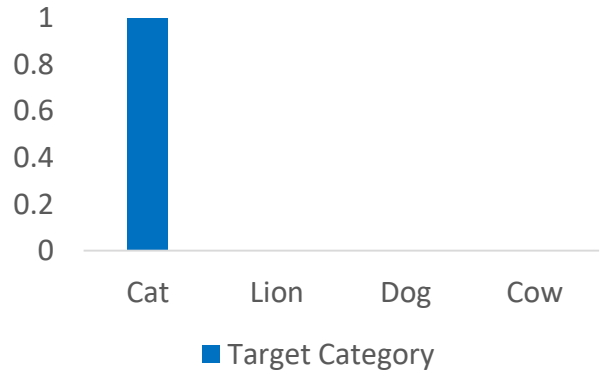


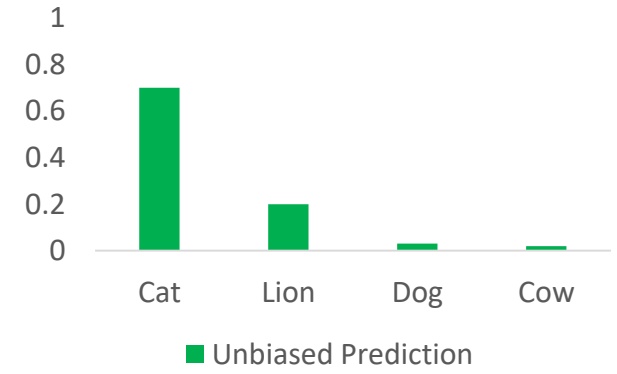


Long-Tailed Classification

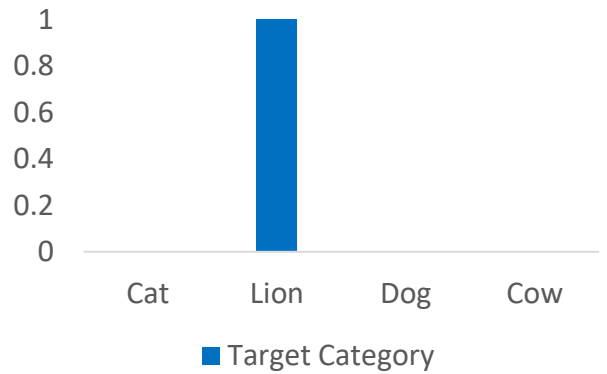
- The target of long-tail classification.

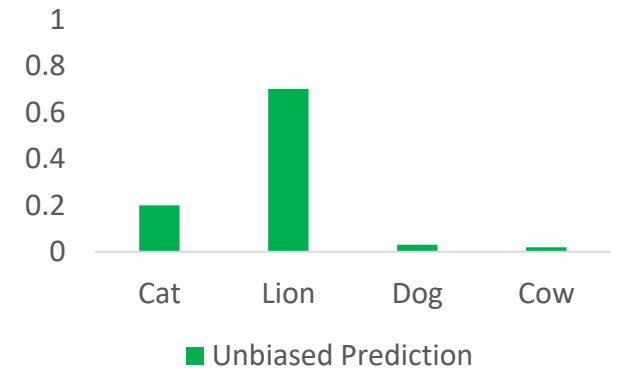
Head Class (1K)





Tail Class (10)

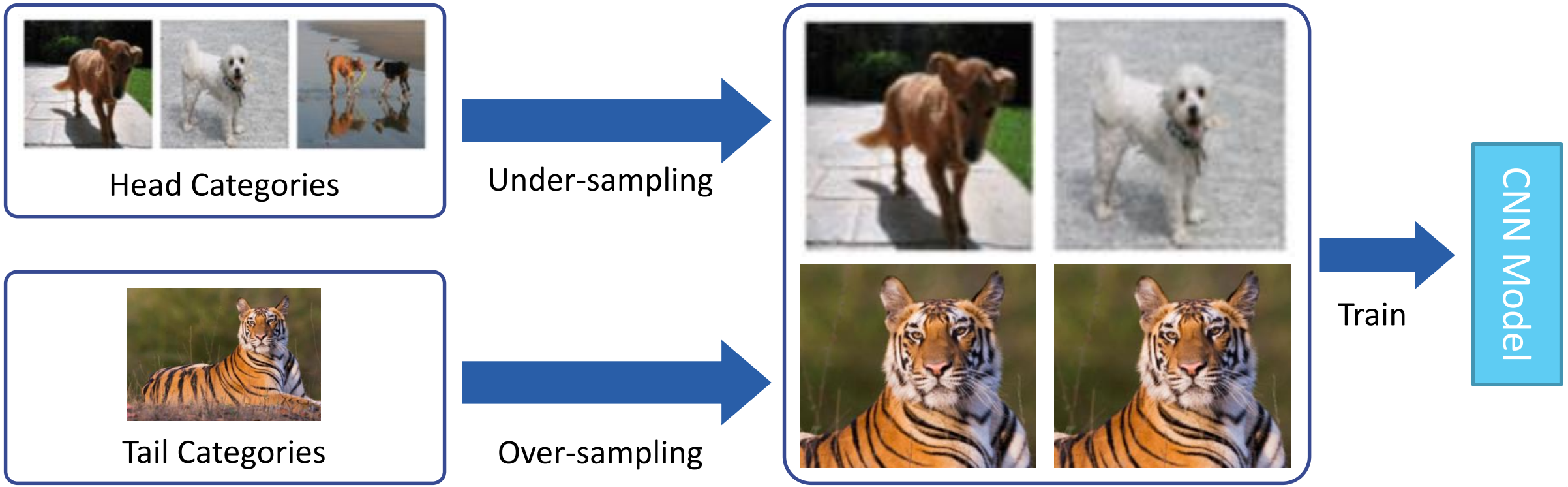




Re-balancing(Re-Sampling/Re-Weighting)

- Re-Sampling
  - Over-sampling for the tail categories.
  - Under-sampling for the head categories.

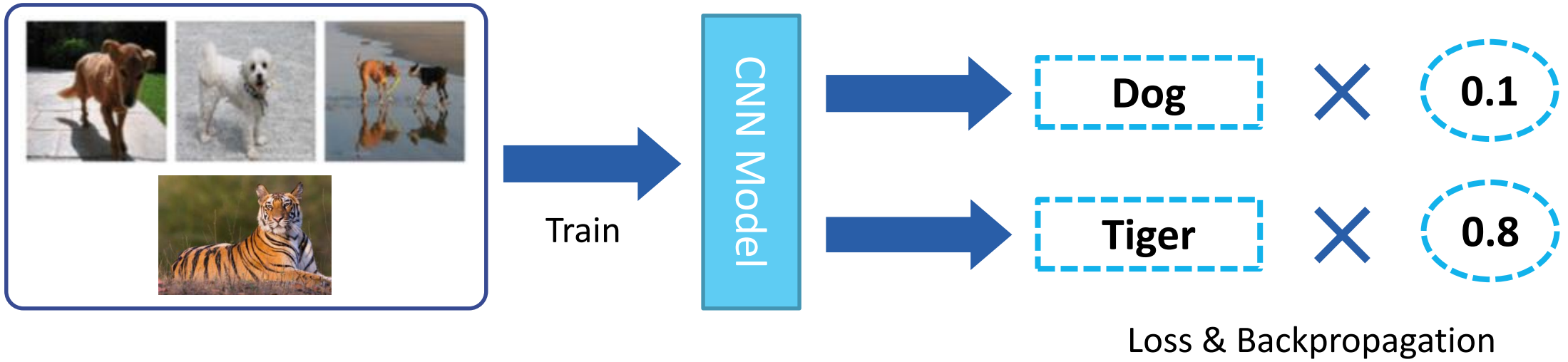
- Drawbacks
  - Over-fitting to the tail.
  - Under-fitting to the head.



Re-balancing(Re-Sampling/Re-Weighting)

- Re-Weighting
  - Weighting by inverse class frequency.
  - Weighting by inverse square root of class frequency.

- Drawbacks
  - Over-fitting to the tail.
  - Under-fitting to the head.



How can we design a better class-balanced loss that is applicable to a diverse array of datasets with drastically different scale and imbalance ?

Class-Balanced Loss Based on Effective Number of Samples

Yin Cui[1,2*] Menglin Jia[1] Tsung-Yi Lin[3] Yang Song[4] Serge Belongie[1,2]
[1]Cornell University [2]Cornell Tech [3]Google Brain [4]Alphabet Inc.

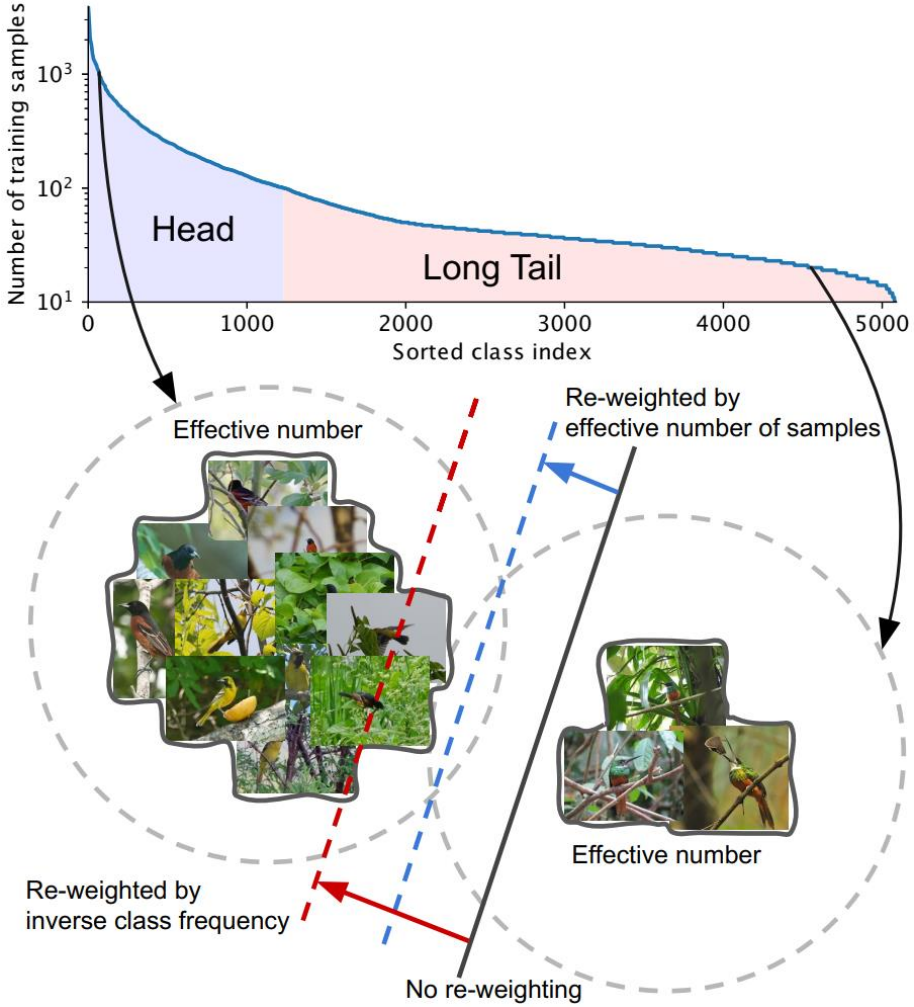
Motivations & Contributions

Motivations

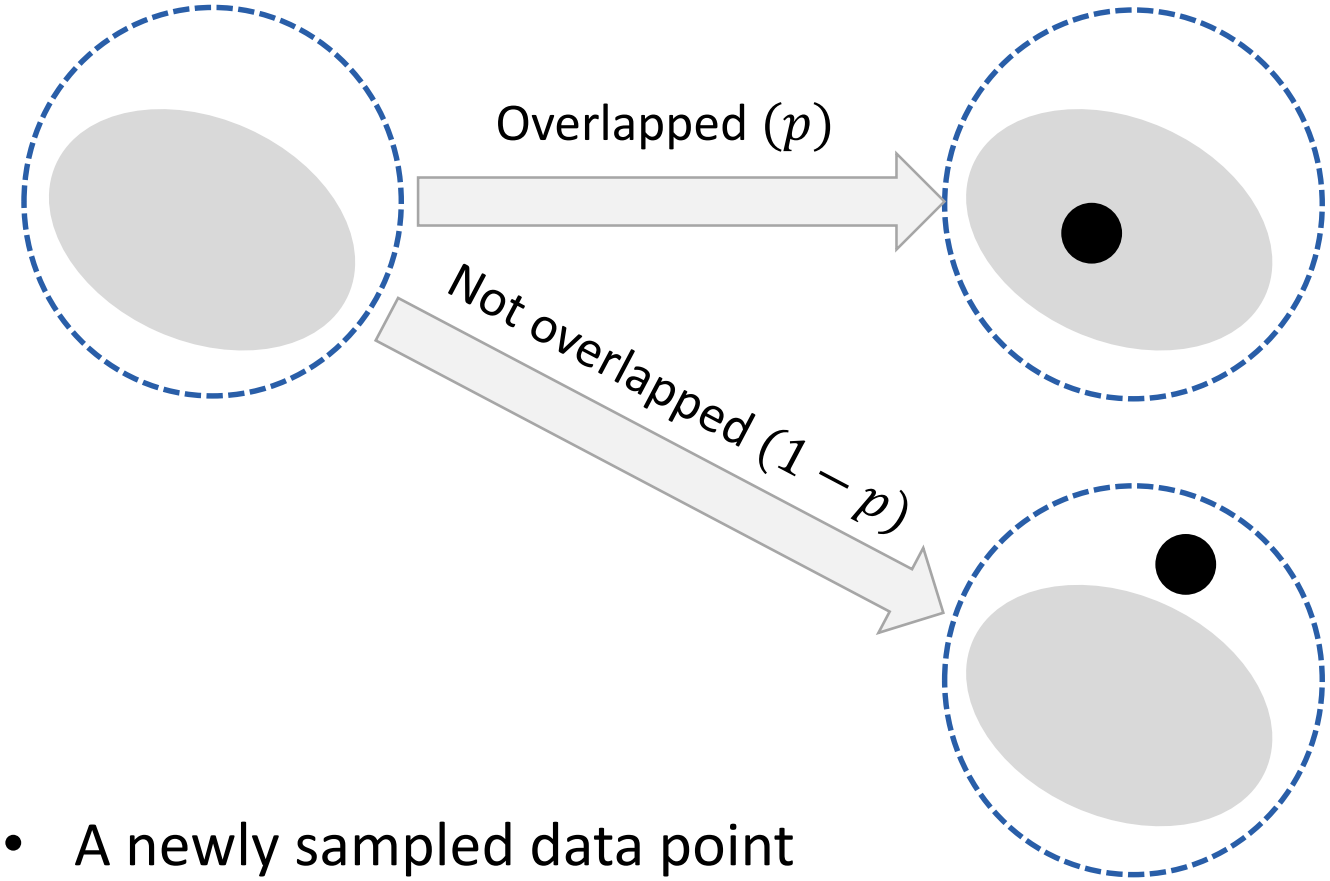
- Re-weighting loss by inverse class frequency cannot perform well.
- There is information overlap among data.
- The benefit from data will diminish gradually.

Contributions

- A novel theoretical framework:
  - To characterize data overlap.
  - To calculate the effective number of samples.
- A class-balanced re-weighting term:
  - Inversely proportional to the effective number of samples.



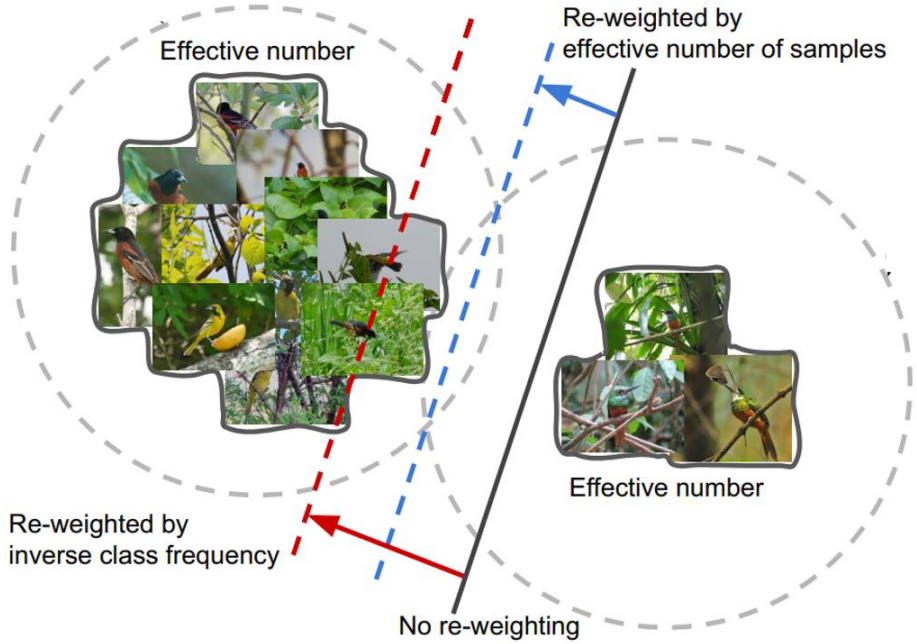
Sampling Process



Feature space $\mathcal{S}$ with volume of N

Previously sampled data

Newly sampled data with volume of 1



- A newly sampled data point
  - Inside the set of previously sampled data with the probability of p.
  - Entirely outside with the probability of $1-p$.

Effective Number

Definition 1 (Effective Number). The *effective number* of samples is the expected volume of samples.

n is the number of samples

Proposition 1 (Effective Number). $E_n = (1-\beta^n)/(1-\beta)$, where $\beta = (N-1)/N$.

The hyperparameter $\beta \in [0, 1)$ controls how fast E_n grows as n increase.

Proof. We prove the proposition by induction.

Step 1. If $n = 1$, there is no overlapping. Therefore, $E_1 = (1-\beta^1)/(1-\beta) = 1$ holds.

Step 2. If $n-1$, assume $E_{n-1} = (1-\beta^{n-1})/(1-\beta)$ holds.

$$p = \frac{E_{n-1}}{N}$$

Then.

$$E_n = pE_{n-1} + (1-p)(E_{n-1}+1) = 1 + \frac{N-1}{N}E_{n-1}$$

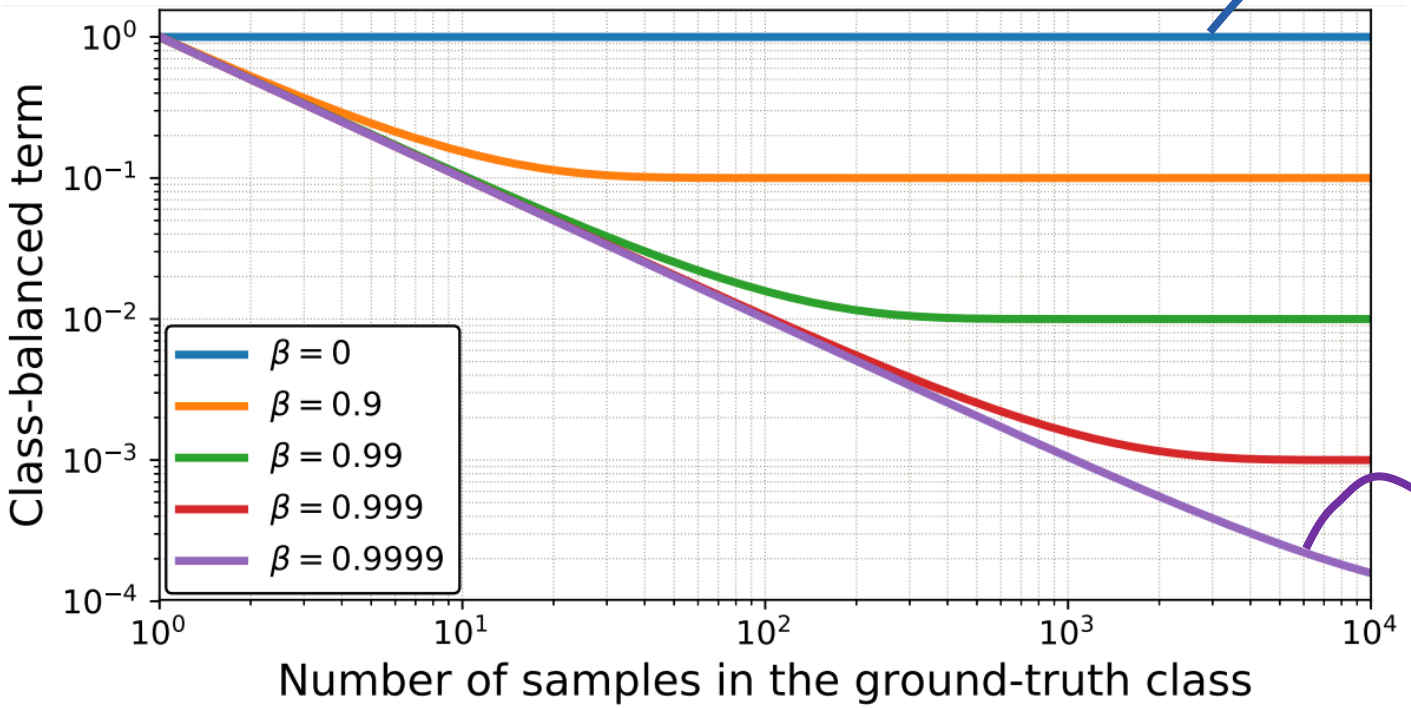
$$= 1 + \beta\frac{1-\beta^{n-1}}{1-\beta} = \frac{1-\beta+\beta-\beta^n}{1-\beta} = \frac{1-\beta^n}{1-\beta}$$

$$\begin{cases} E_n = 1 & \text{if } \beta = 0 (N = 1) \\ E_n \to n & \text{as } \beta \to 1 (N \to \infty) \end{cases}$$

Class-Balanced Loss

The ***Class-Balanced Loss*** is introduced a weighting factor that is inversely proportional to the effective number of samples.

$$\mathrm{CB}(\mathbf{p}, y) = \frac{1}{E_{n_y}} = \frac{1-\beta}{1-\beta^{n_y}}\mathcal{L}(\mathbf{p}, y) \qquad E_n = \frac{1-\beta^n}{1-\beta}$$



No re-weighting

By using hyperparameter β to smoothly adjust the class-balanced term between no re-weighting and re-weighting by inverse class frequency.

Re-weighting by inverse class frequency

Class-Balanced Loss

- Class-Balanced Softmax Cross-Entropy Loss:

$$\mathrm{CB}_{\mathrm{softmax}}(\mathbf{z}, y) = -\frac{1-\beta}{1-\beta^{n_y}} \log\left(\frac{\exp(z_y)}{\sum_{j=1}^{C}\exp(z_j)}\right) \qquad \text{where } \mathbf{z} = [z_1, z_2, \cdots, z_c]^T.$$

- Class-Balanced Sigmoid Cross-Entropy Loss:

$$\mathrm{CB}_{\mathrm{sigmoid}}(\mathbf{z}, y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} \log\left(\frac{1}{1+\exp(-z_i^t)}\right) \qquad \text{where } z_i^t = \begin{cases} z_i, & \text{if } i = y. \\ -z_i, & \text{otherwise.} \end{cases}$$

- Class-Balanced Focal Loss:

$$\mathrm{CB}_{\mathrm{focal}}(\mathbf{z}, y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} (1-p_i^t)^{\gamma} \log(p_i^t) \qquad \text{where } p_i^t = \mathrm{sigmoid}(z_i^t) = \frac{1}{1+\exp(-z_i^t)}$$

Experimental Results

Visual Recognition on Long-Tailed CIFAR

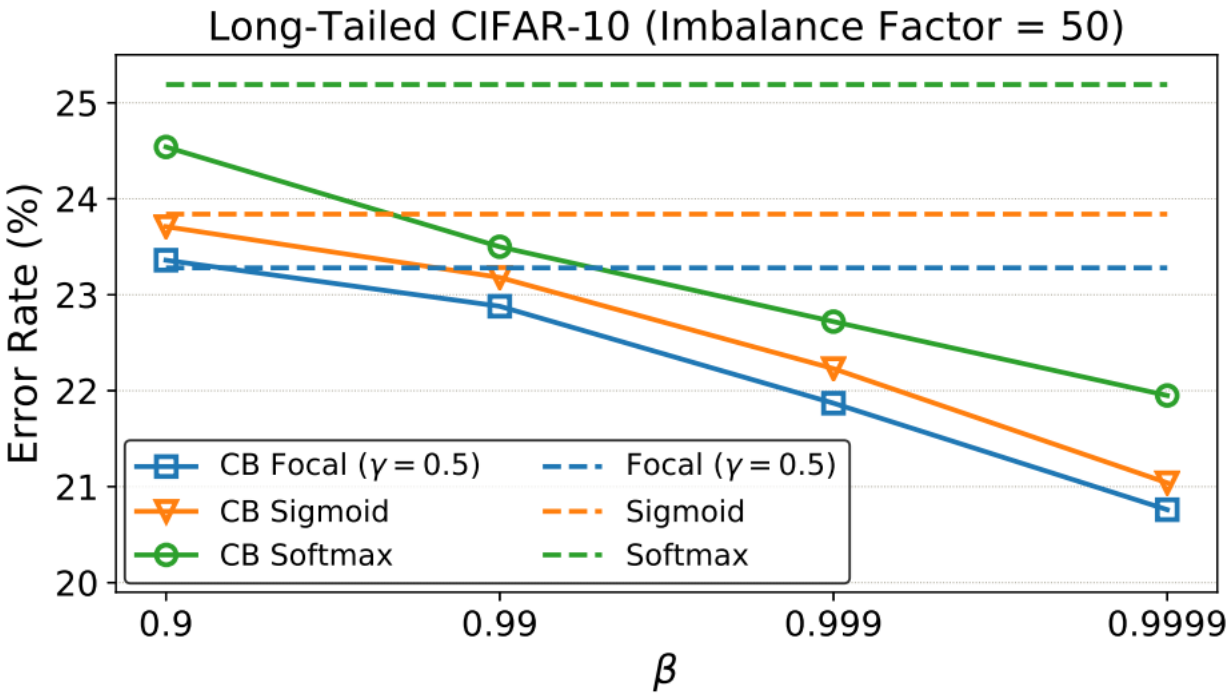
Dataset Name	Long-Tailed CIFAR-10						Long-Tailed CIFAR-100					
Imbalance	200	100	50	20	10	1	200	100	50	20	10	1
Softmax	**34.32**	29.64	25.19	17.77	13.61	6.61	65.16	61.68	56.15	48.86	44.29	29.07
Sigmoid	34.51	**29.55**	23.84	**16.40**	**12.97**	**6.36**	64.39	**61.22**	55.85	48.57	44.73	**28.39**
Focal ($\gamma = 0.5$)	36.00	29.77	**23.28**	17.11	13.19	6.75	65.00	61.31	55.88	48.90	44.30	28.55
Focal ($\gamma = 1.0$)	34.71	29.62	23.29	17.24	13.34	6.60	**64.38**	61.59	**55.68**	**48.05**	**44.22**	28.85
Focal ($\gamma = 2.0$)	35.12	30.41	23.48	16.77	13.68	6.61	65.25	61.61	56.30	48.98	45.00	28.52
Class-Balanced	**31.11**	**25.43**	**20.73**	**15.64**	**12.51**	**6.36***	**63.77**	**60.40**	**54.68**	**47.41**	**42.01**	**28.39***
Loss Type	SM	Focal	Focal	SM	SGM	SGM	Focal	Focal	SGM	Focal	Focal	SGM
β	0.9999	0.9999	0.9999	0.9999	0.9999	-	0.9	0.9	0.99	0.99	0.999	-
γ	-	1.0	2.0	-	-	-	1.0	1.0	-	0.5	0.5	-

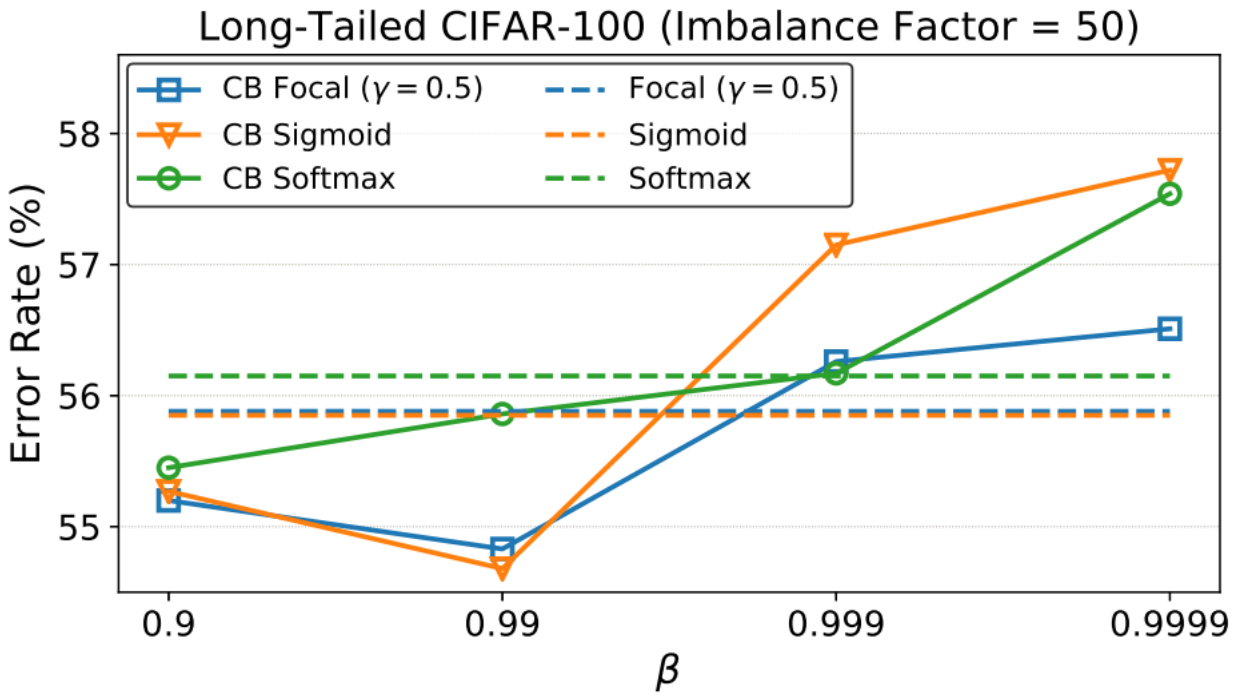
$$\text{loss type} \in \{\text{softmax}, \text{sigmoid}, \text{focal}\} \qquad \beta \in \{0.9, 0.99, 0.999, 0.9999\} \qquad \gamma \in \{0.5, 1.0, 2.0\}$$

- Class-Balanced loss significantly improve the performance of commonly used loss functions on long-tailed datasets.
- Sigmoid cross-entropy and focal loss are outperform softmax cross-entropy in most cases.

Experimental Results

Visual Recognition on Long-Tailed CIFAR





- On CIFAR-10, class-balanced loss yields consistent improvement across different β and the larger the β is, the larger the improvement is.
- On CIFAR-100, $\beta = 0.9$ or $\beta = 0.99$ improves the original loss, whereas a larger β hurts the performance.

Experimental Results

Visual Recognition on Large-Scale Datasets

					iNaturalist 2017		iNaturalist 2018		ILSVRC 2012	
Network	Loss	β	γ	Input Size	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ResNet-50	Softmax	-	-	224 × 224	45.38	22.67	42.86	21.31	23.92	7.03
ResNet-101	Softmax	-	-	224 × 224	42.57	20.42	39.47	18.86	22.65	6.47
ResNet-152	Softmax	-	-	224 × 224	41.42	19.47	38.61	18.07	21.68	5.92
ResNet-50	CB Focal	0.999	0.5	224 × 224	41.92	20.92	38.88	18.97	22.71	6.72
ResNet-101	CB Focal	0.999	0.5	224 × 224	39.06	18.96	36.12	17.18	21.57	5.91
ResNet-152	CB Focal	0.999	0.5	224 × 224	38.06	18.42	35.21	16.34	20.87	5.61
ResNet-50	CB Focal	0.999	0.5	320 × 320	38.16	18.28	35.84	16.85	21.99	6.27
ResNet-101	CB Focal	0.999	0.5	320 × 320	34.96	15.90	32.02	14.27	20.25	5.34
ResNet-152	CB Focal	0.999	0.5	320 × 320	33.73	14.96	30.95	13.54	19.72	4.97

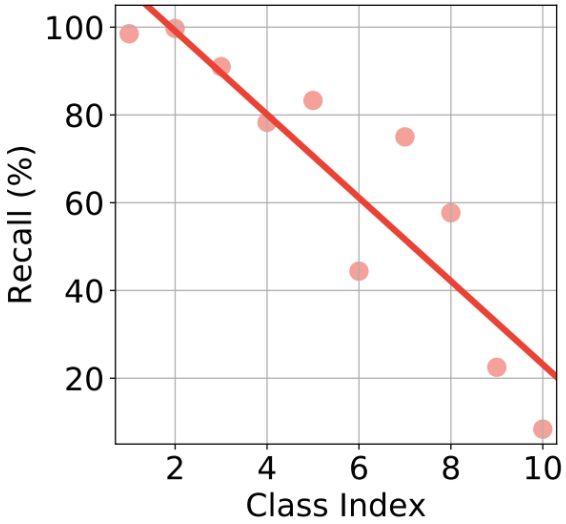
Table 3. Classification error rate on large-scale datasets trained with different loss functions. The proposed class-balanced term combine with focal loss (CB Focal) is able to outperform softmax cross-entropy by a large margin.

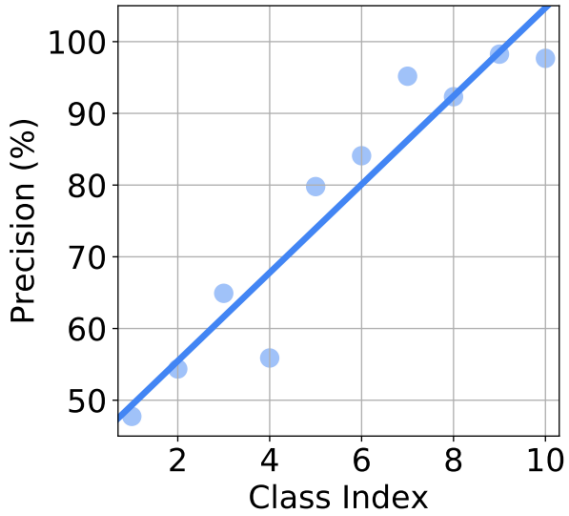
CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning

Chen Wei[1]* Kihyuk Sohn[2] Clayton Mellina[2] Alan Yuille[1] Fan Yang[2]
[1]Johns Hopkins University [2]Google Cloud AI

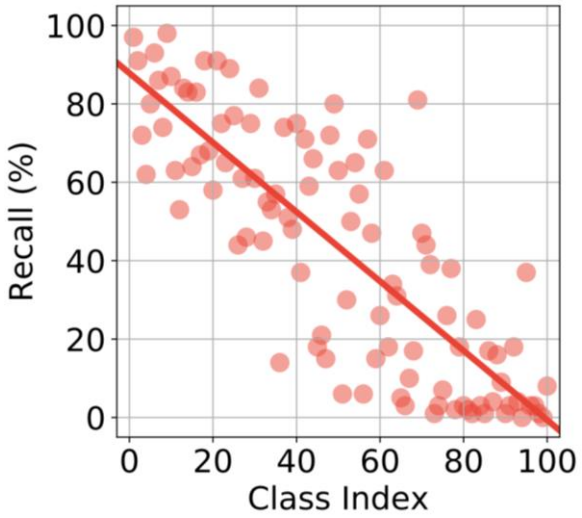
Motivation

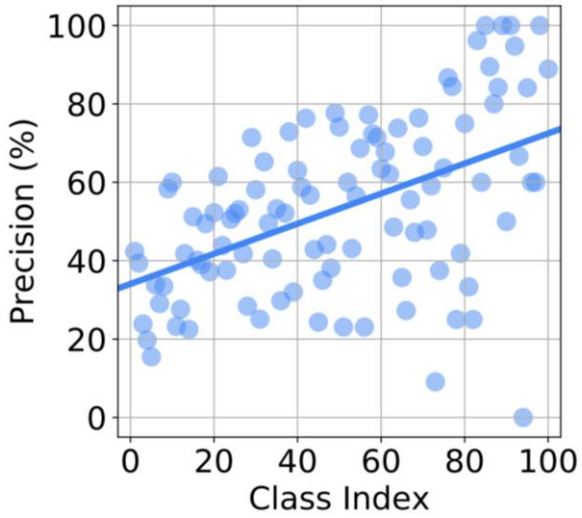
Bias of a FixMatch model on class-imbalanced data





Per-class recall and precision on CIFAR10-LT





Per-class recall and precision on CIFAR100-LT

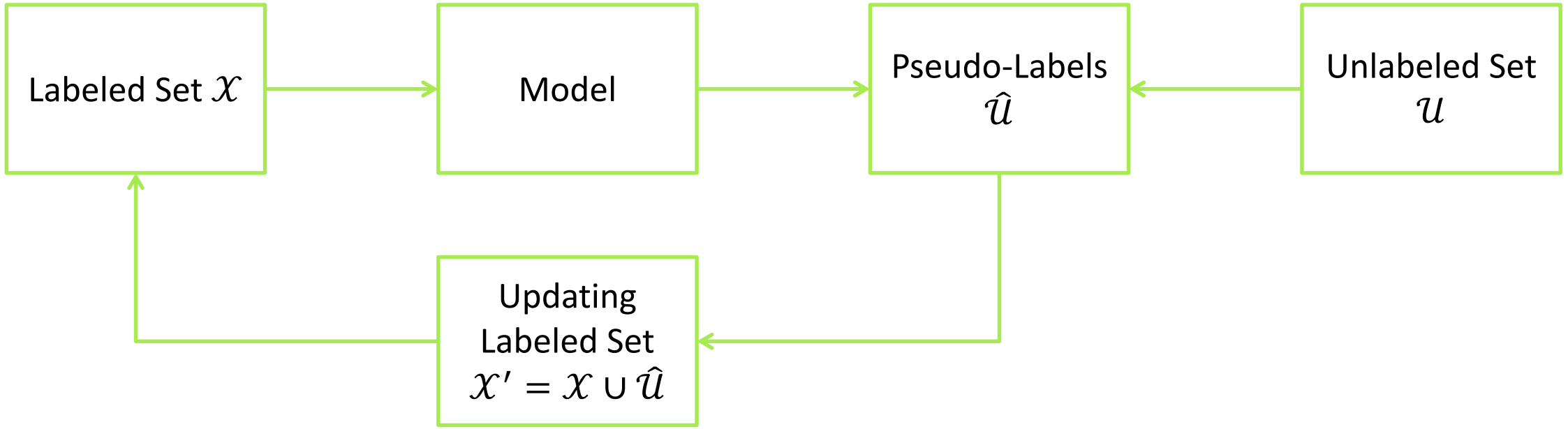
- The assumption that the performance of the head classes is better than that of the tail classes may be partially true.

- The model obtains high recall but low precision on head classes, while obtaining low recall but high precision on tail classes.

Problem setup and notations

- Labeled set: $\mathcal{X} = \{(x_n, y_n) : n \in (1, \cdots, N)\}$
- Training examples: $x_n \in \mathbb{R}^d$
- Class labels: $y_n \in \{1, \cdots, L\}$
- The training examples in $\mathcal{X}$ of class l: N_l, i.e., $\sum_{l=1}^{L} N_l = N$
- Assuming that the classes are sorted by index in descending order: $N_1 \geq N_2 \geq \cdots \geq N_L$
- Unlabeled set: $\mathcal{U} = \{u_m \in \mathbb{R} : m \in (1, \cdots, M)\}$
- Imbalance ratio: $\gamma = \frac{N_1}{N_L}$
- Label fraction: $\beta = \frac{N}{N+M}$
- Given class-imbalanced sets $\mathcal{X}$ and $\mathcal{U}$, to learn a classifier: $f : \mathbb{R}^d \to \{1, \cdots, L\}$

Self-training

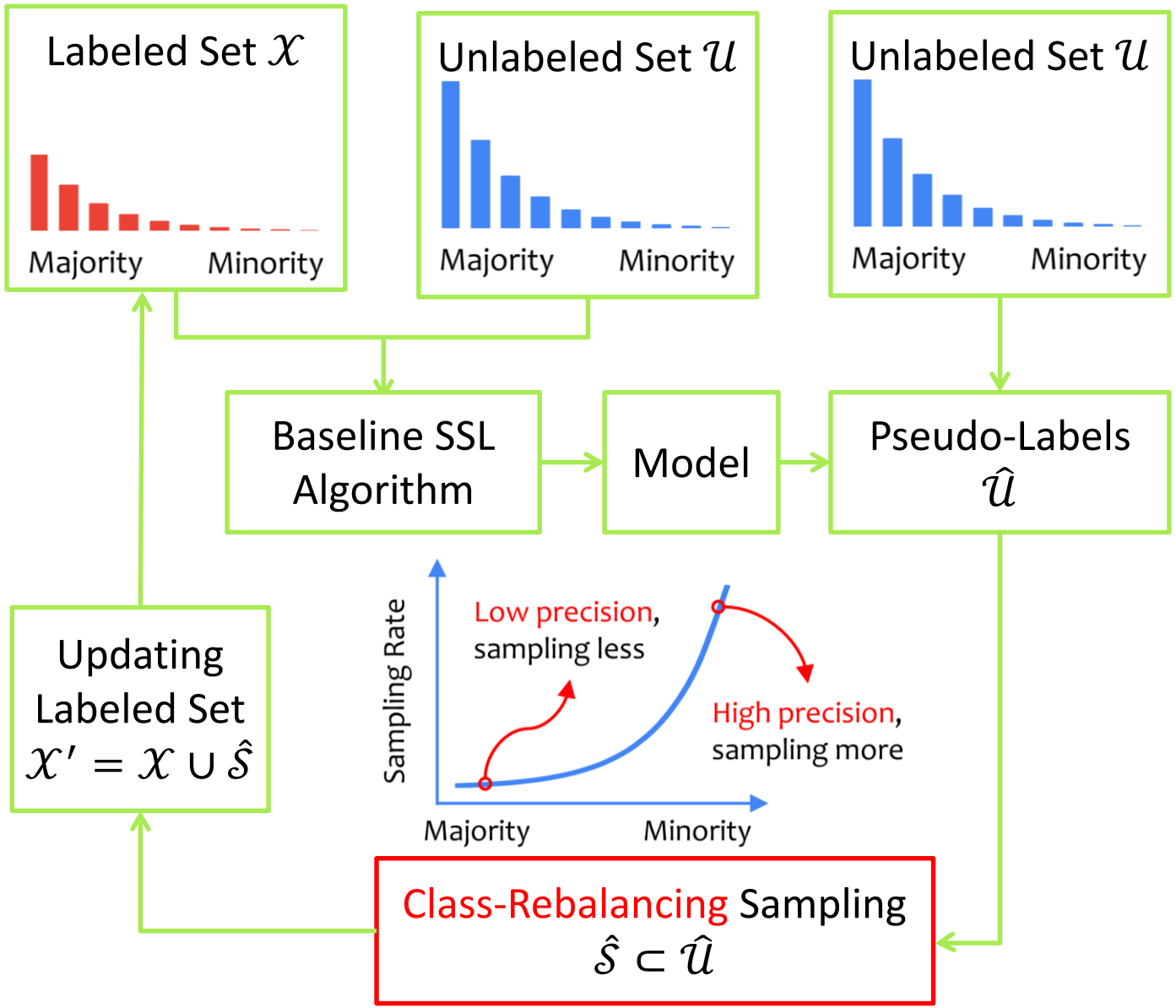


Step 1. The model is trained on labeled set to obtain a teacher model.

Step 2. The teacher model's predictions are used to generate pseudo-labels $\hat{y}_m$ for unlabeled data $\mathcal{U}_m$.

Step 3. Add the pseudo-labeled set $\hat{\mathcal{U}} = \{(u_m, \hat{y}_m)\}_{m=1}^{M}$ to label set, for the next generation.

Class-rebalancing self-training



Step 1. Use SSL algorithms to exploit both labeled and unlabeled data to get a better teacher model.

Step 3. Select a subset $\hat{\mathcal{S}} \subset \hat{\mathcal{U}}$ add to labeled set, for the next generation.

Class-Rebalancing Rule. The less frequent a class l is, the more unlabeled samples that are predicted as class l are include into the pseudo-labeled set $\hat{\mathcal{S}}$.

Class-rebalancing self-training

- Unlabeled samples that are predicted as class l are added to $\hat{S}$ as the rate of:

$$\mu_l = \left(\frac{N_{L+1-l}}{N_1}\right)^{\alpha}$$

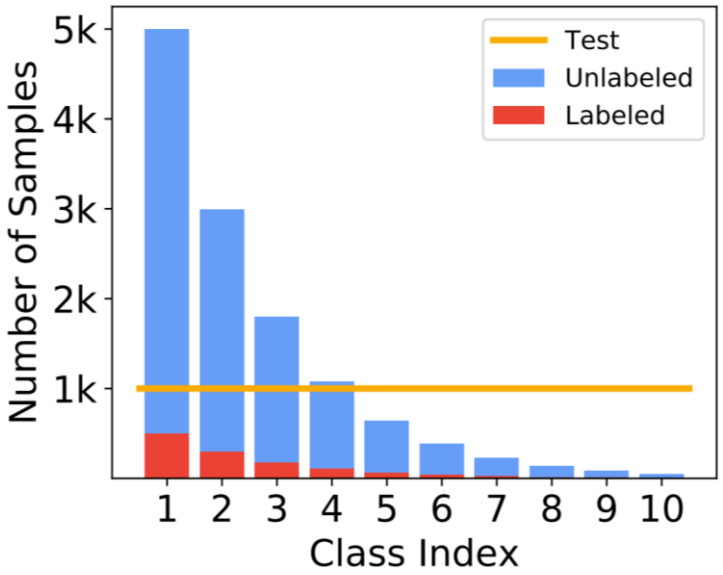
Where $\alpha \geq 0$ tunes the sampling rate and thus the size of $\hat{S}$.

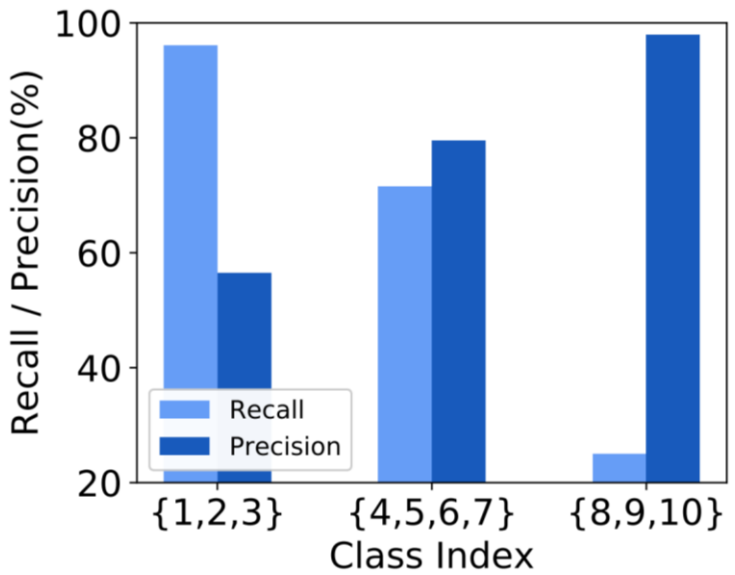
- For instance, assume that $L = 10$ and $\gamma = \frac{N_1}{N_{10}} = 100$, then:

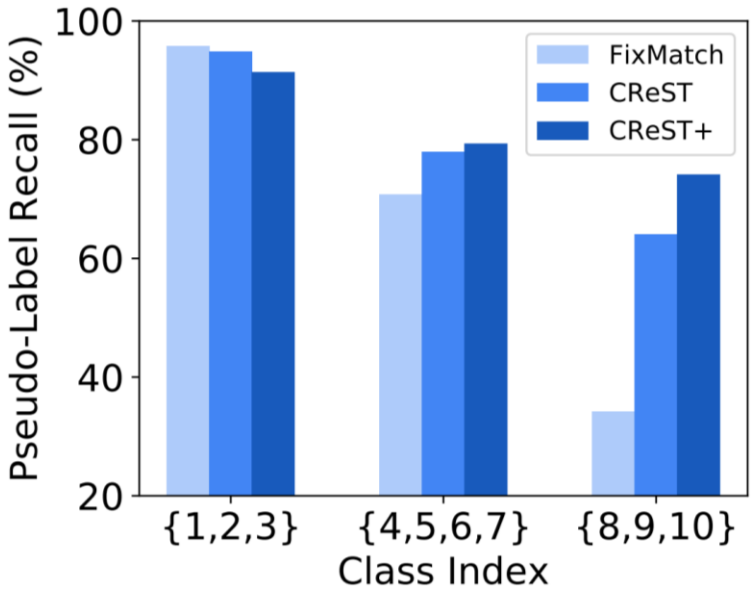
$\mu_{10} = (\frac{N_{10+1-10}}{N_1}) = 1$ Keep all samples predicted as the most minority class.

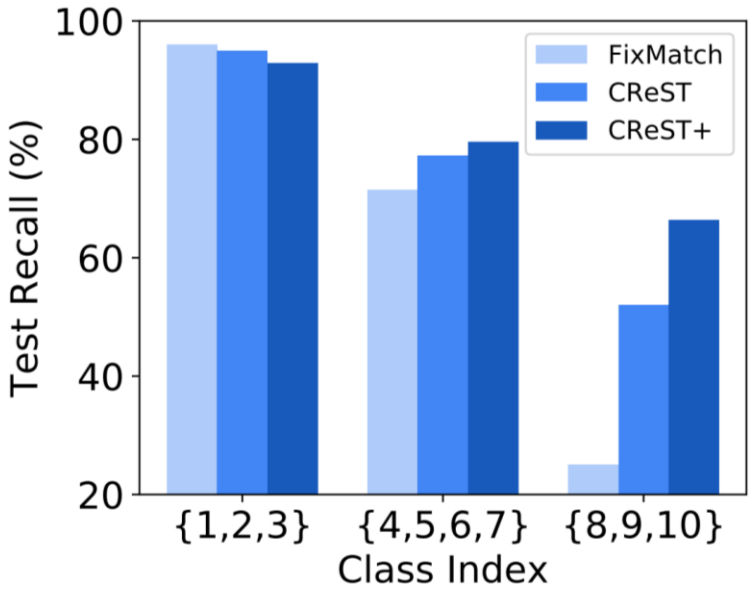
$\mu_1 = (\frac{N_{10+1-1}}{N_1}) = 0.01^{\alpha}$ Keep only 0.01^{α} of samples predicted as the most majority class.

Experimental Results on CIFAR10-LT









- Both labeled and unlabeled sets are class-imbalanced.
- Precision and Recall of a FixMatch model.
- The Proposed CReST and CReST+ improve the quality of pseudo-labels.
- The Proposed CReST and CReST+ improve the recall on balanced test set.

Experimental Results

Method	CIFAR10-LT						CIFAR100-LT			
	$\beta = 10\%$			$\beta = 30\%$			$\beta = 10\%$		$\beta = 30\%$	
	$\gamma = 50$	$\gamma = 100$	$\gamma = 200$	$\gamma = 50$	$\gamma = 100$	$\gamma = 200$	$\gamma = 50$	$\gamma = 100$	$\gamma = 50$	$\gamma = 100$
FixMatch [38]	$79.4_{\pm 0.65}$	$66.3_{\pm 1.74}$	$59.7_{\pm 0.74}$	$81.9_{\pm 0.30}$	$73.1_{\pm 0.58}$	$64.7_{\pm 0.69}$	$33.7_{\pm 0.94}$	$28.3_{\pm 0.66}$	$43.1_{\pm 0.24}$	$38.6_{\pm 0.45}$
w/ CReST	$83.8_{\pm 0.45}$	$75.9_{\pm 0.62}$	$64.1_{\pm 0.23}$	$84.2_{\pm 0.13}$	$77.6_{\pm 0.86}$	$67.7_{\pm 0.82}$	$37.4_{\pm 0.29}$	$32.1_{\pm 1.52}$	$45.6_{\pm 0.19}$	$40.2_{\pm 0.53}$
w/ CReST+	$\mathbf{84.2}_{\pm 0.39}$	$\mathbf{78.1}_{\pm 0.84}$	$\mathbf{67.7}_{\pm 1.39}$	$\mathbf{84.9}_{\pm 0.27}$	$\mathbf{79.2}_{\pm 0.20}$	$\mathbf{70.5}_{\pm 0.56}$	$\mathbf{38.8}_{\pm 1.03}$	$\mathbf{34.6}_{\pm 0.74}$	$\mathbf{46.7}_{\pm 0.34}$	$\mathbf{42.0}_{\pm 0.44}$

Table 1. Classification accuracy (%) on CIFAR10-LT and CIFAR100-LT under various label fraction β and imbalance ratio γ. The numbers are averaged over 5 different folds.

THANKS