Class-Balanced Loss Based on Effective Number of Samples

Paper Authors:

Yin Cui,   Menglin Jia, Tsung-Yi Lin

Yang Song, Serge Belongie

From CVPR 2019

Reporter: Cheng Kang

Outline

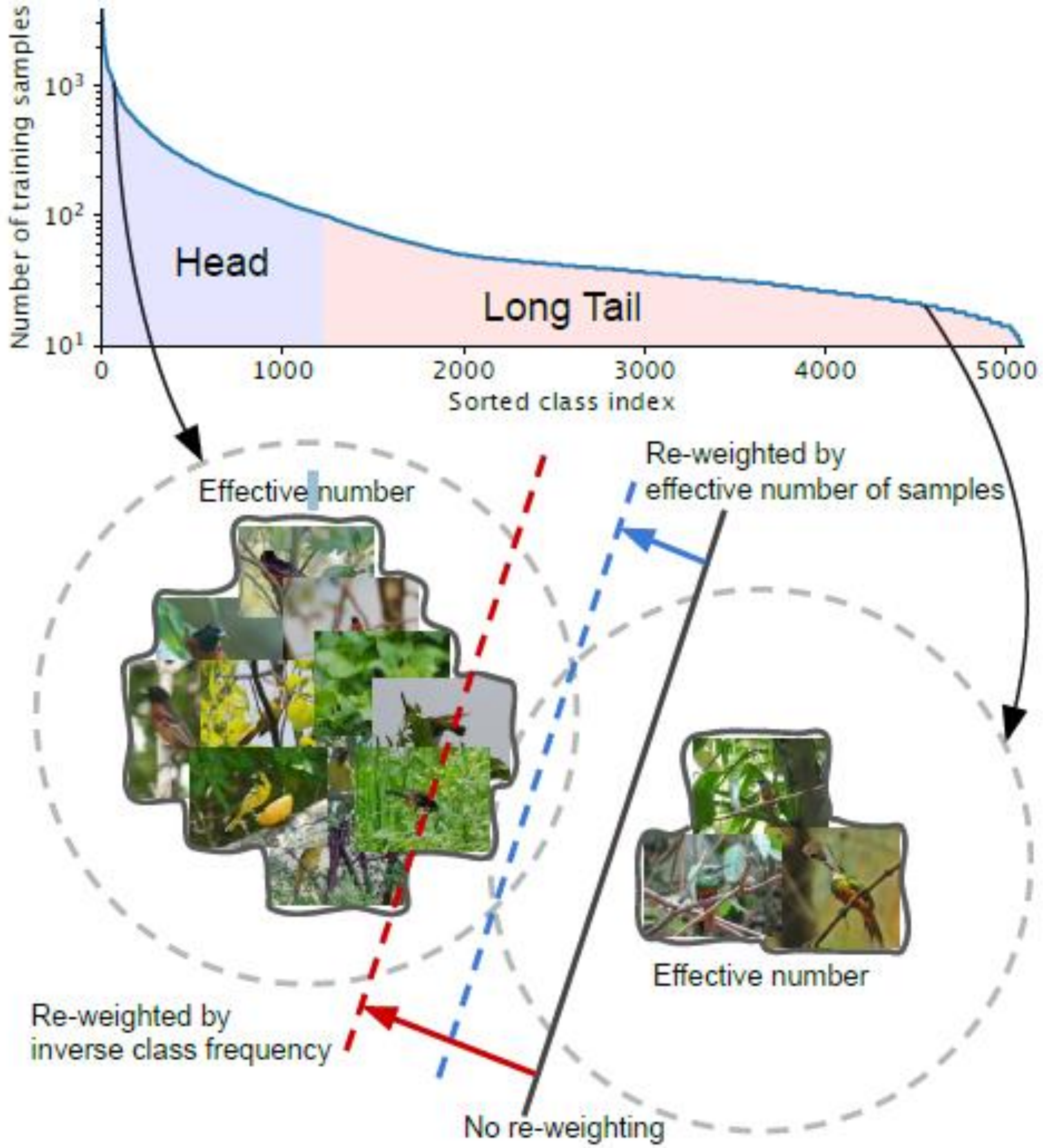
- 1. Introduction
- 2. Relative Work
- 3. Effective Number of Samples
- 4. Class-Balanced Loss
- 5. Experiments
- 6. Conclusion and Discussion

1. Introduction

Unbalanced problems and data, eg.

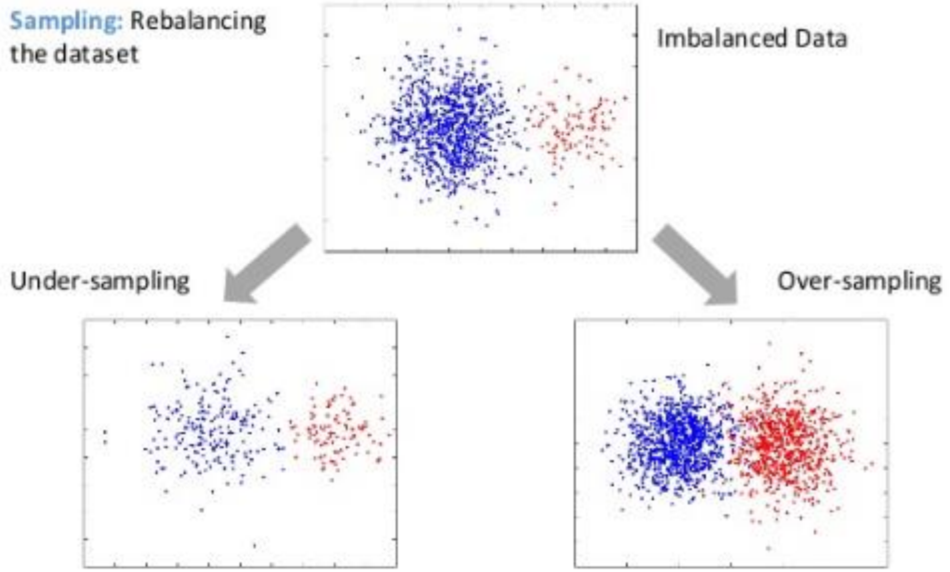
- Classes have often unequal frequency.
  - Medical diagnosis: 95 % healthy, 5% disease.
  - e-Commerce: 99 % do not buy, 1 % buy.
  - Security: 99.999 % of citizens are not terrorists.

Similar situation for multiclass classifiers. Majority class classifier can be 99 % correct but useless.



2. Related Work

A



- *There are mainly two strategies:*

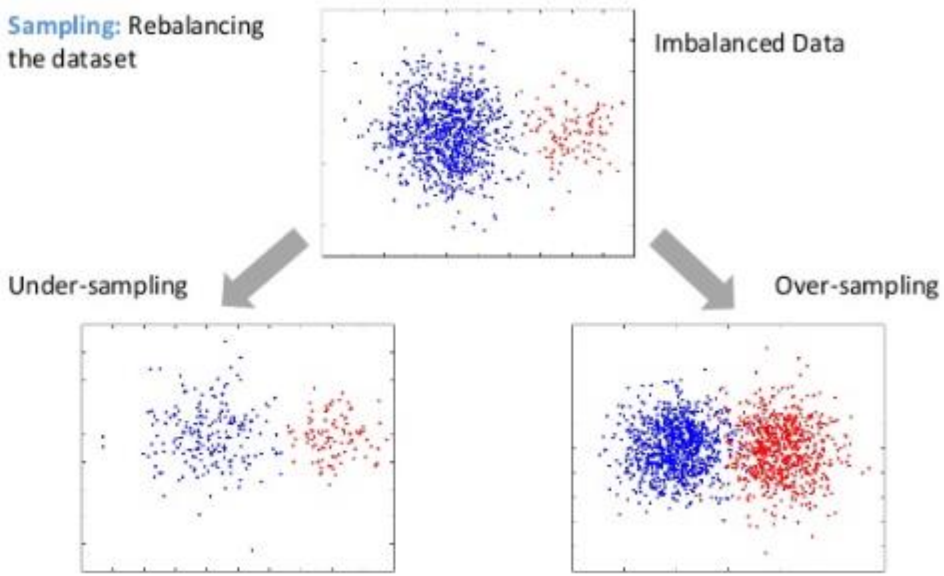
- **A. re-sampling or under-sampling**

  By over-sampling (adding repetitive data) for the minor class or under-sampling (removing data) for the major class, or both

  Drawbacks: cause the model to overfit.

2. Related Work

A



B

$$R(q^*) = \min_{q \in D} \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y)\, W(y, q(x))$$

$$R(q^*) = \min_{q(x) \in D} \sum_{x \in X} \sum_{y \in Y} p(x)\, p_{Y|X}(y|x)\, W(y, q(x))$$

- *There are mainly two strategies:*
- **A. re-sampling or under-sampling**

  By over-sampling (adding repetitive data) for the minor class or under-sampling (removing data) for the major class, or both

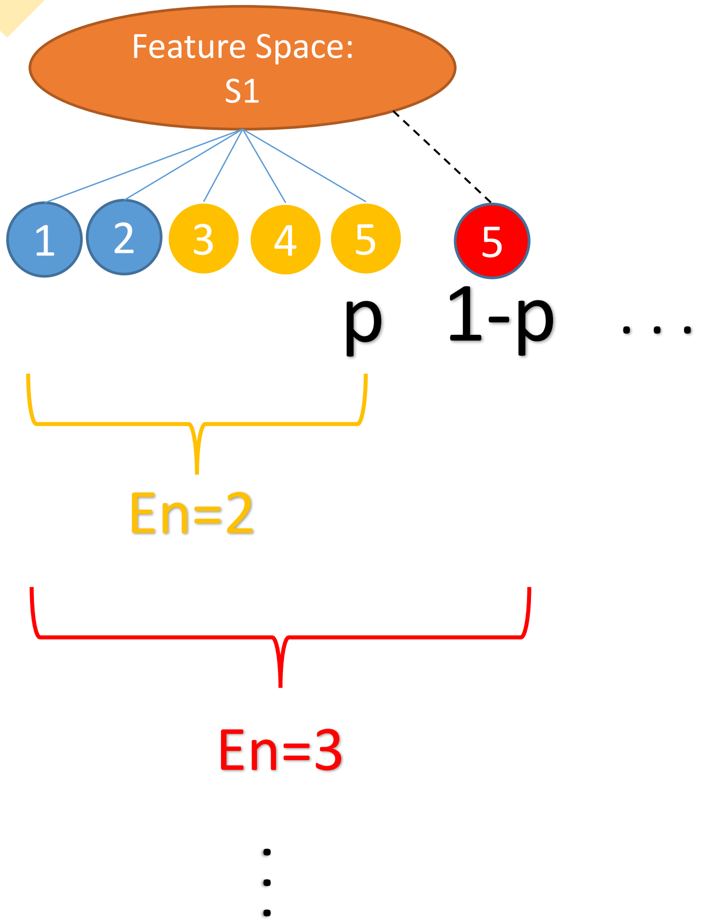
  Drawbacks: cause the model to overfit
- **B. cost-sensitive re-weighting**

  By influencing the loss function by assigning relatively higher costs to examples from minor classes

  Drawbacks: A side effect of assigning higher weights to hard examples is the focus on harmful samples.

(refers to Course XP33ROD in CVUT)

3. Effective Number of Samples



Definition:

S: is the feature space of a specific class

N: We assume the volume of S is N and N ≥ 1 (***the boundary of volume***)

Then, the expectation of Effective number (En) is that:

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

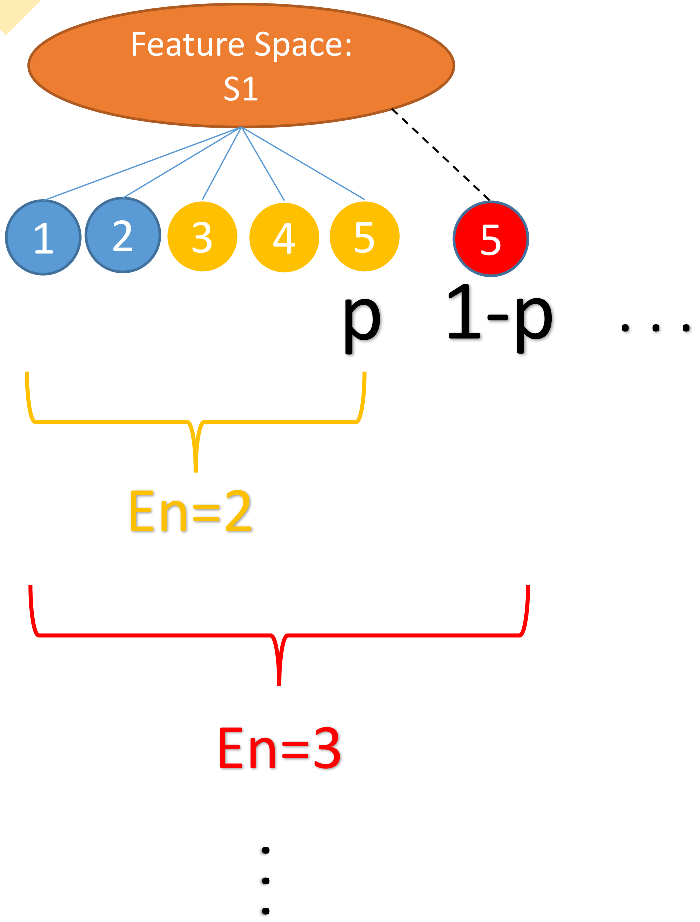
where p is the probability of whether the feature space of the n-th sample is inside the volume, if not, the probability is 1-p.

And $p = E_{n-1}/N$

Finally,

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

3. Effective Number of Samples



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

Then, set $\beta = (N-1)/N$.

E1=1;

E2=1+β*E1=1+β;

E3=1+β*E2=1+β+ β* β;

$\vdots$

En-1=1+β*En-2=1+β+ β*β+ β*β*β+…+ $\overbrace{\beta * \ldots * \beta}^{(n-1)-1}$***;***

En=1+β*En-1=1+β+ β*β+ β*β*β+…+ $\overbrace{\beta * \ldots * \beta}^{(n)-1}$***;***

Induction:

$$E_n = (1-\beta^n)/(1-\beta) = \sum_{j=1}^{n} \beta^{j-1}$$

Finally,

$$N = \lim_{n\to\infty} \sum_{j=1}^{n} \beta^{j-1} = 1/(1-\beta) \qquad \lim_{\beta\to 1} E_n = \lim_{\beta\to 1} \frac{f(\beta)}{g(\beta)} = \lim_{\beta\to 1} \frac{f'(\beta)}{g'(\beta)} = n$$

4. Class-Balanced Loss

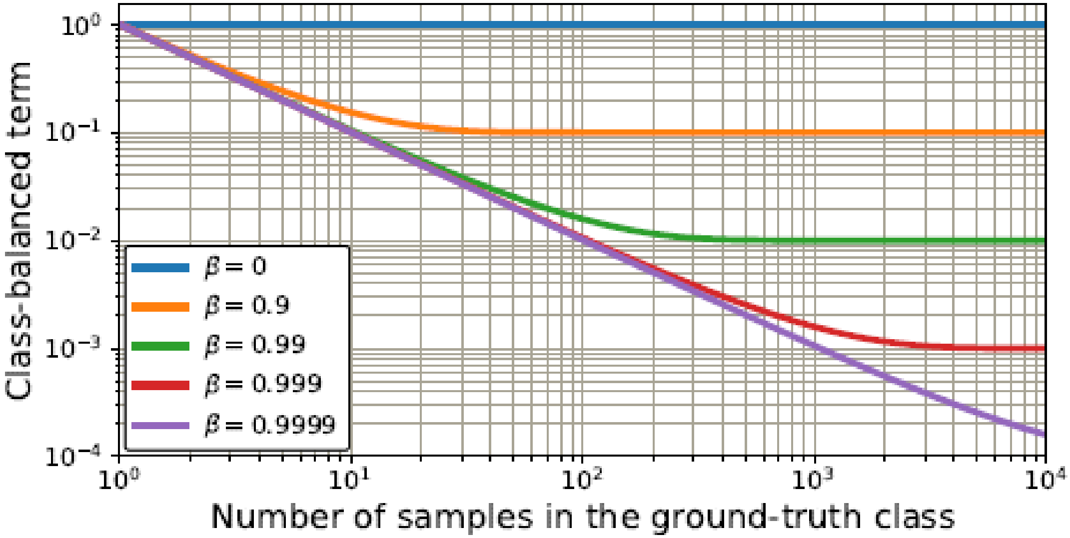


Figure 3. Visualization of the proposed class-balanced term $(1 - \beta)/(1 - \beta^{n_y})$, where n_y is the number of samples in the ground-truth class. Both axes are in log-scale. For a long-tailed dataset where major classes have significantly more samples than minor classes, setting β properly re-balances the relative loss across classes and reduces the drastic imbalance of re-weighing by inverse class frequency.

Suppose the number of samples for class *i* is n_i

$$E_{n_i} = (1 - \beta_i^{n_i})/(1 - \beta_i)$$

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

$$\mathrm{CB}_{\text{softmax}}(\mathbf{z}, y) = -\boxed{\frac{1 - \beta}{1 - \beta^{n_y}}} \log \left(\frac{\exp(z_y)}{\sum_{j=1}^{C} \exp(z_j)} \right)$$

$$\mathrm{CB}_{\text{sigmoid}}(\mathbf{z}, y) = -\boxed{\frac{1 - \beta}{1 - \beta^{n_y}}} \sum_{i=1}^{C} \log \left(\frac{1}{1 + \exp(-z_i^t)} \right)$$

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

5. Experiments

Dataset Name	# Classes	Imbalance
Long-Tailed CIFAR-10	10	10.00 - 200.00
Long-Tailed CIFAR-100	100	10.00 - 200.00
iNaturalist 2017	5,089	435.44
iNaturalist 2018	8,142	500.00
ILSVRC 2012	1,000	1.78

Table 1. Datasets that are used to evaluate the effectiveness of class-balanced loss. We created 5 long-tailed versions of both CIFAR-10 and CIFAR-100 with imbalance factors of 10, 20, 50, 100 and 200 respectively.

$$imbalance\ fators = \frac{N_{largest-class}}{N_{smallest-class}}$$

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	-

Table 2. Classification error rate of ResNet-32 trained with different loss functions on long-tailed CIFAR-10 and CIFAR-100. We show best results of class-balanced loss with best hyperparameters (SM represents Softmax and SGM represents Sigmoid) chosen via cross-validation. Class-balanced loss is able to achieve significant performance gains. $*$ denotes the case when each class has same number of samples, class-balanced term is always 1 therefore it reduces to the original loss function.

$$\mathrm{CB}_{\mathrm{focal}}(\mathbf{z}, y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} (1 - p_i^t)^{\boxed{\gamma}} \log(p_i^t)$$

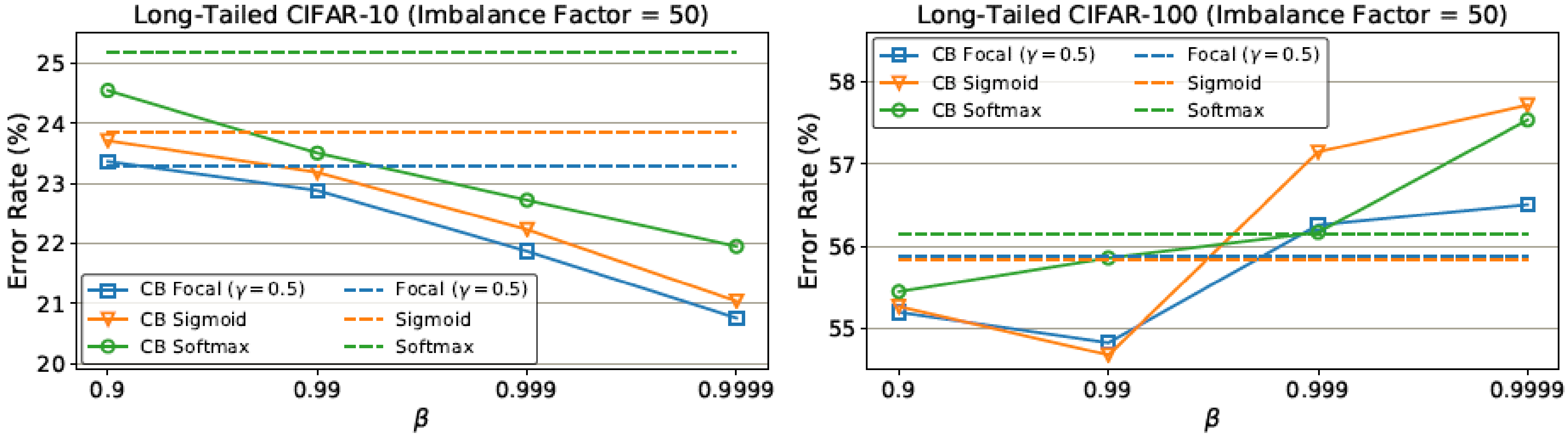


Figure 5. Classification error rate when trained with and without the class-balanced term. On CIFAR-10, class-balanced loss yields consistent improvement across different β and the larger the β is, the larger the improvement is. On CIFAR-100, $\beta = 0.99$ or $\beta = 0.999$ improves the original loss, whereas a larger β hurts the performance.

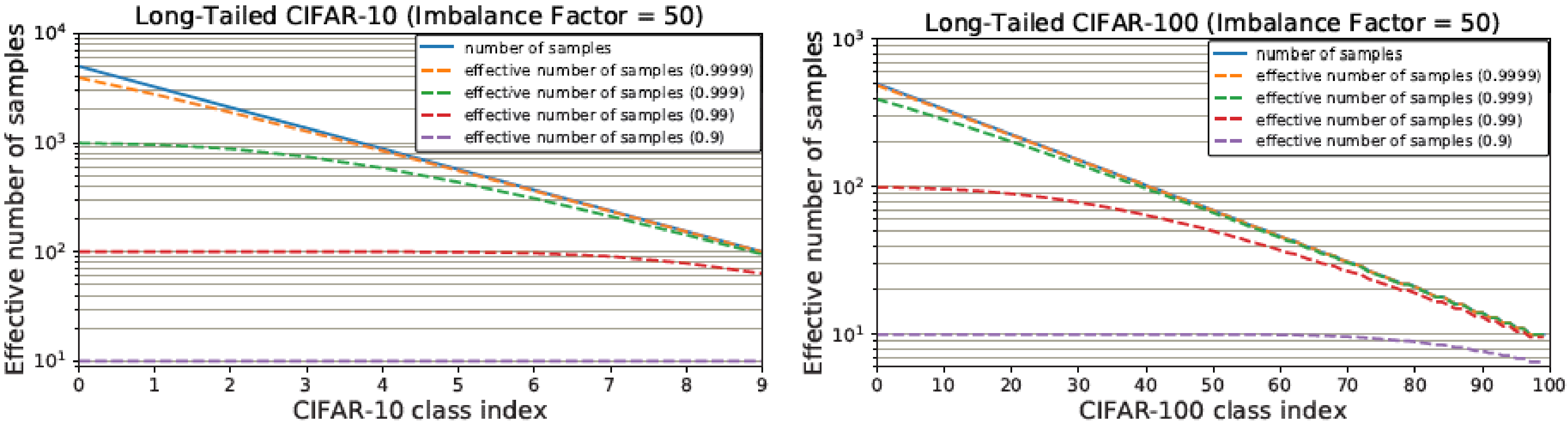


Figure 6. Effective number of samples with different β on long-tailed CIFAR-10 and CIFAR-100 with the imbalance of 50. This is a semi-log plot with vertical axis in log-scale. When $\beta \to 1$, effective number of samples is same as number of samples. When β is small, effective number of samples are similar across all classes.

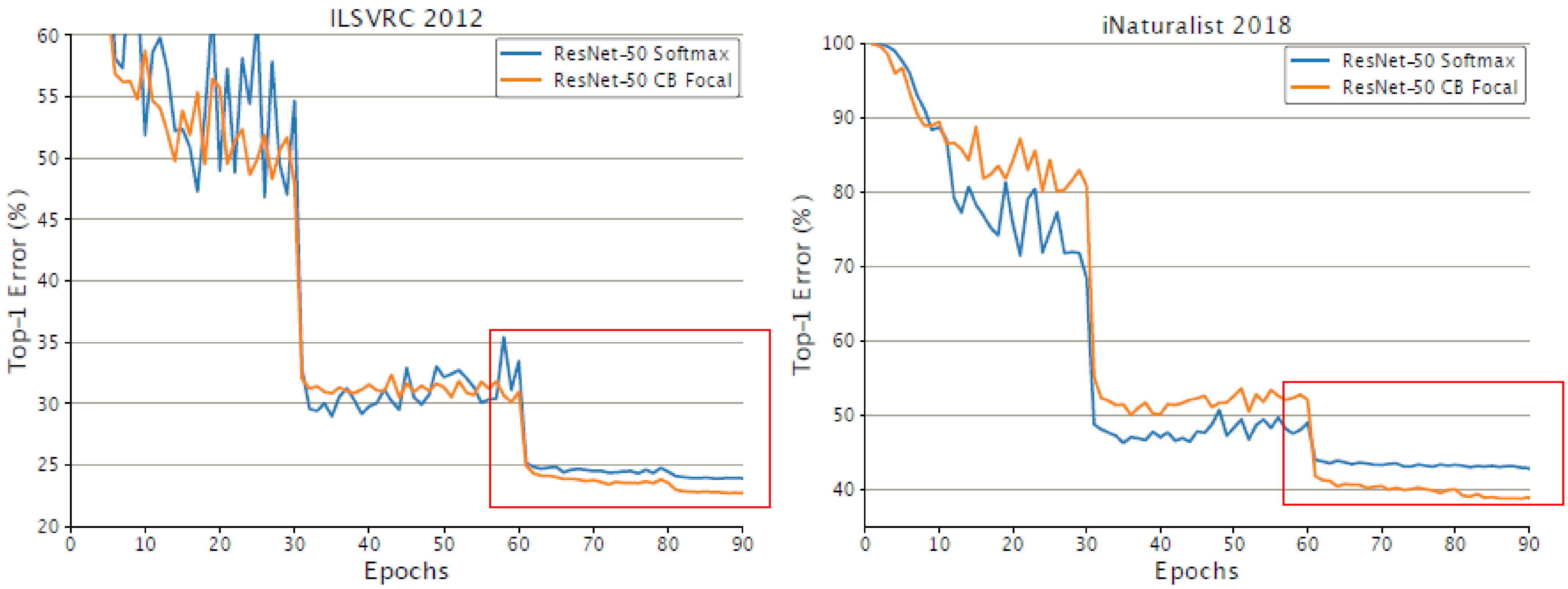


Figure 7. Training curves of ResNet-50 on ILSVRC 2012 (left) and iNaturalist 2018 (right). Class-balanced focal loss with $\beta = 0.999$ and $\gamma = 0.5$ outperforms softmax cross-entropy after 60 epochs.

6. Conclusion and Discussion

- The key idea is to take data overlap into consideration to help quantify the effective number of samples.
  - (a class-balanced loss to re-weight loss inversely with the effective number of samples per class)

- In the future, we plan to extend our framework by incorporating reasonable assumptions on the data distribution or designing learning-based, adaptive methods.