Learning Imbalanced Datasets with Maximum Margin Loss

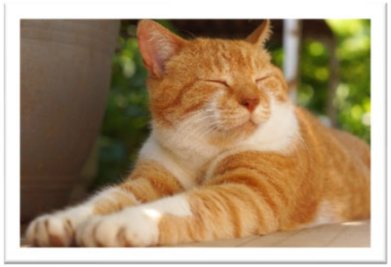
ICIP2021

KAIST

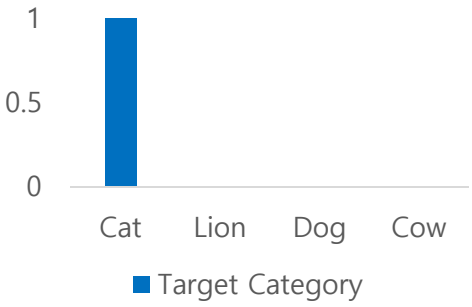
Haeyong Kang

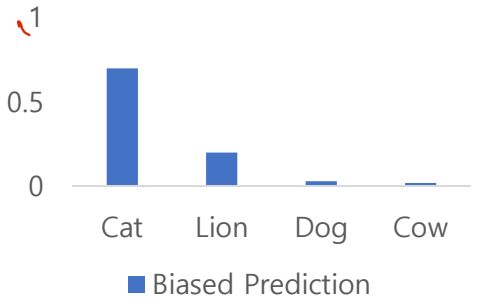
The problem of Imbalanced Dataset Classification

The Problem : Biased Prediction

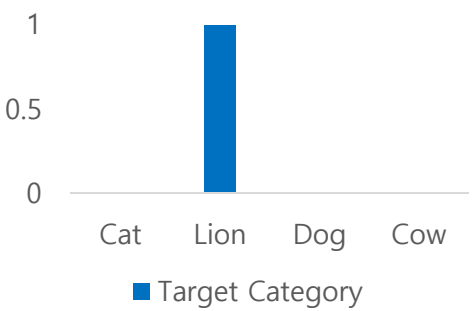


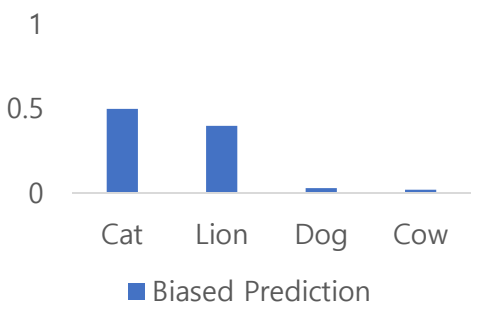
Head Class (1K)



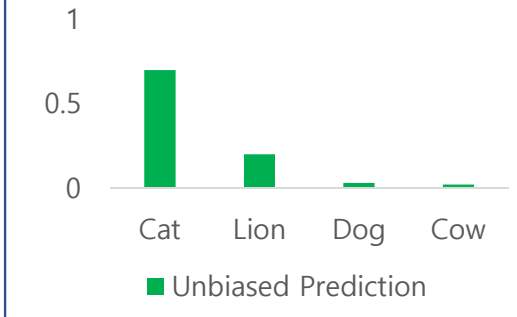


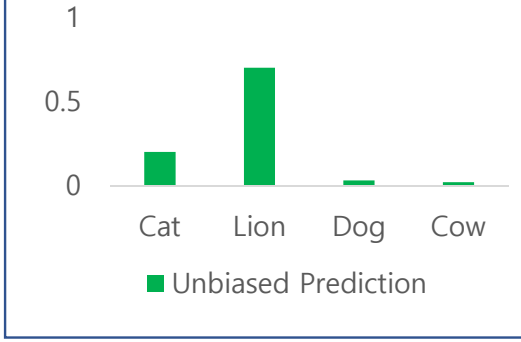
Tail Class (10)





The Target : Unbiased Prediction





Re-sampling

- Over-sampling the minority classes [1]

- Over-sampling is effective in a lot of cases but can lead to over-fitting of the minority classes
- Stronger data augmentation for minority classes can help alleviate the over-fitting

- Under-sampling the frequent classes [2]

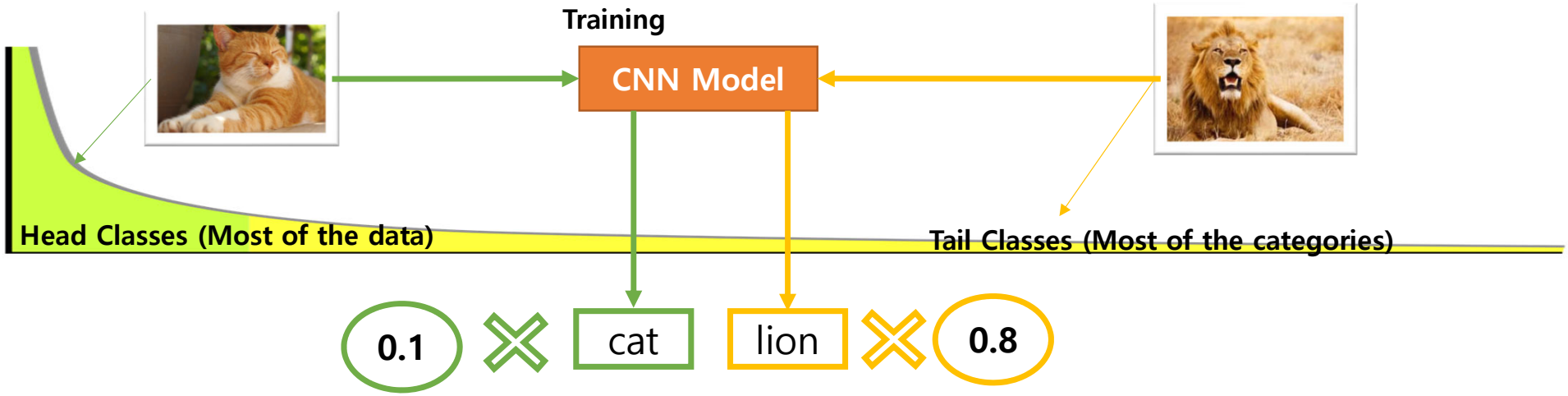
- it discards a large portion of the data and thus is not feasible when data imbalance is extreme.

- Balancing-sampling (decoupling representation) [3, 4]

- At first stage, pre-train model on uniform-samples, at the final stage, fine-tuned model on balancing-samples

[1] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In ICML2019.
[2] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 2018
[3] Kang, Bingyi, et al. Decoupling representation and classifier for long-tailed recognition. In ICLR 2020.
[4] Ren, Jiawei, et al. Balanced meta-softmax for long-tailed visual recognition." NeuralPS2020.

Re-weighting



- **The vanilla scheme re-weights [5]** classes proportionally to the inverse of their frequency.

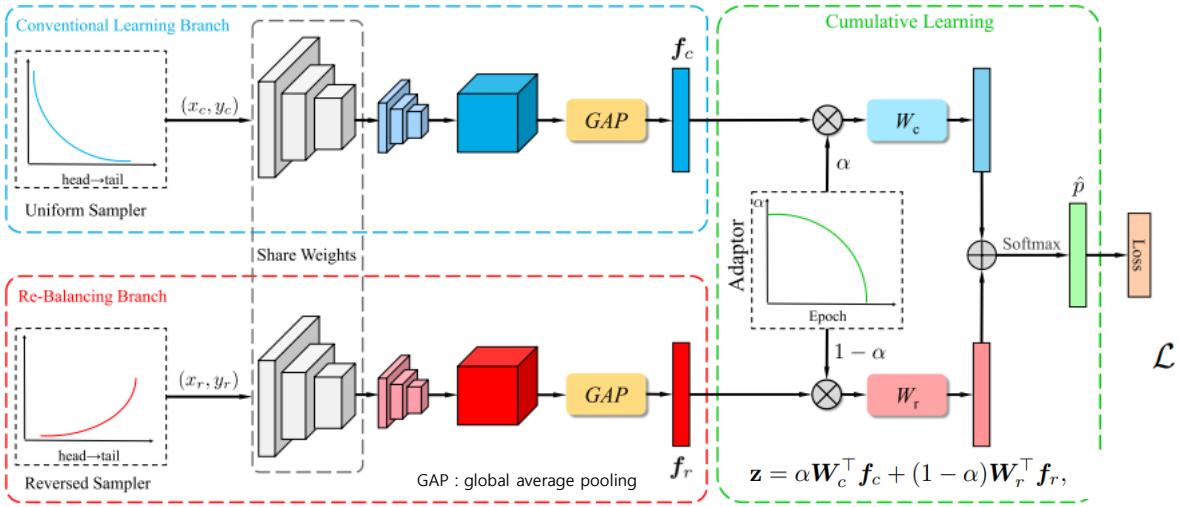
- **Focal loss [6]** down-weights the well-classified examples

- **Class Balance (CB) [7]** is that re-weighting by inverse class frequency yields poor performance on frequent classes, and thus propose re-weighting by the inverse effective number of samples.

[5] Chen Huang, Yining Li, Change Loy Chen, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. In PAMI2019.
[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In ICCV2017.
[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In CVPR2019

Two-Stage Re-balancing

- Smoothly adapted bilateral-branch training [9]
- Decoupled two-stage training [10, 11]



$$\mathcal{L} = \alpha E(\hat{\boldsymbol{p}}, y_c) + (1-\alpha)E(\hat{\boldsymbol{p}}, y_r),$$

Bilateral-Branch Network (BBN) [9]

Meta-learning

- **Meta-learning [12]** is also used in improving the performance on imbalanced datasets or the few shot learning settings.

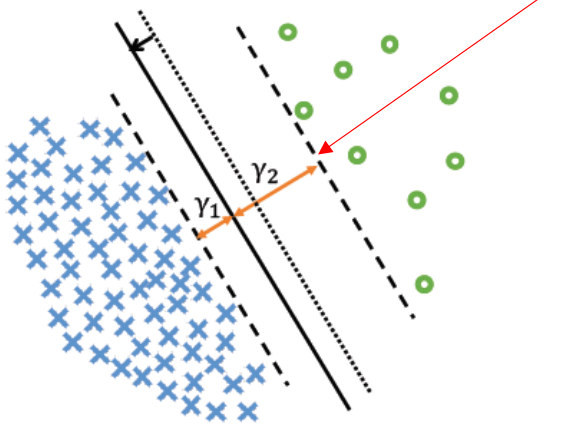
[9] BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition, In CVPR 2020
[10] Decoupling Representation and Classifier for Long-Tailed Recognition, In ICLR 2020
[11] Ren, Jiawei, et al. Balanced meta-softmax for long-tailed visual recognition." NeuralPS2020.
[12] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In NeuralPS2019.

Margin loss [12,13,14]

- Hinge Loss [12]

$$\ell(y) = \max(0, 1 - t \cdot y) \qquad \text{where } y = w \cdot x + b$$

- Label-Distribution-Aware-Margin (LDAM) loss [13]



-This paper encourages rare classes to have higher margin.

$$\frac{1}{\gamma_1 \sqrt{n_1}} + \frac{1}{\gamma_2 \sqrt{n_2}}$$

few sample class leads to higher margin.

➔ However, they don't converge to a max margin solution

➔ In this paper, we deal with a max-margin solution for imbalanced dataset learnings.

[12] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. In Neural processing letters1999.

[13] Liu, Weiyang, et al. Large-margin softmax loss for convolutional neural networks. In *ICML*2016.

[14] Cao, Kaidi, et al. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In NeualPS2019.

Maximum-Margin

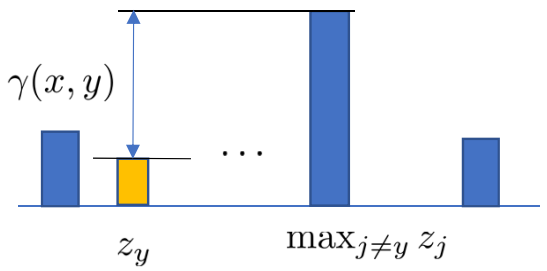
- Input and label space

-the input space : $x \in \mathbb{R}^d$

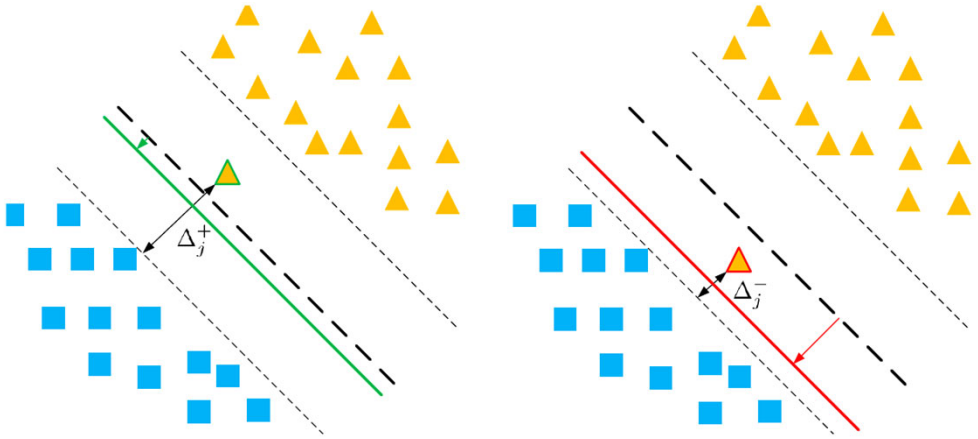
-the label space : $\{1, \cdots, j, \cdots, k\}$

- the maximum margin of an example (x, y)

$$\gamma(x, y) = z_y - \max_{\ell \neq y} z_\ell$$



Maximum-Margin Loss Function (1)



- An assumption that the decision boundaries are shifted by two types of the hard maximum margin of samples: hard positive margin and hard negative margin.

- our loss function device to occupy more margin, such that the red decision boundary shifts more than the green one.

Maximum Margin (MM) Loss

$$\Delta_y^{MM} = \begin{cases} \Delta_y^+ & \text{if } \arg\max_j f_j(x) = y; \\ \Delta_y^- & \text{otherwise.} \end{cases}$$

$$\Delta_y^+ = \exp\left(-\max(z_y - \max_{j \neq y} z_j, 0) - \delta^+\right),$$

and

$$\Delta_y^- = \exp\left(-\max(\max_{j \neq y} z_j - z_y, 0) - \delta^-\right).$$

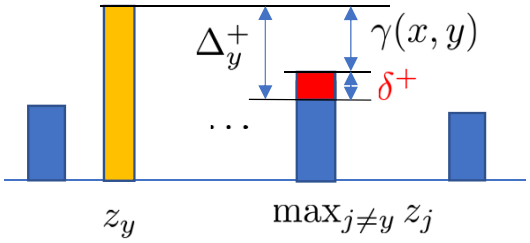
Cross Entropy Loss with MM

$$\mathcal{L}_{MM}((x,y);f) = -\log \frac{e^{z_y - \Delta_y^{MM}}}{e^{z_y - \Delta_y^{MM}} + \sum_{j \neq y} e^{z_j}}$$

Maximum-Margin Loss Function (2)

Maximum Margin (MM) Loss

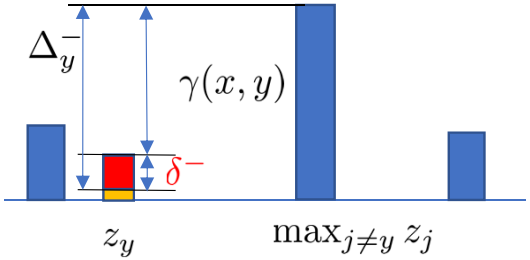
$$\Delta_y^{MM} = \begin{cases} \Delta_y^+ & \text{if } \arg\max_j f_j(x) = y; \\ \Delta_y^- & \text{otherwise.} \end{cases}$$



$$\Delta_y^+ = \exp\left(-\max(z_y - \max_{j\neq y} z_j, 0) - \delta^+\right),$$

Cross Entropy Loss with MM

$$\mathcal{L}_{MM}((x,y);f) = -\log \frac{e^{z_y - \Delta_y^{MM}}}{e^{z_y - \Delta_y^{MM}} + \sum_{j\neq y} e^{z_j}}$$



$$\Delta_y^- = \exp\left(-\max(\max_{j\neq y} z_j - z_y, 0) - \delta^-\right).$$

Maximum-Margin Loss Function - DRW

Algorithm 1 Imbalanced Learning with MM Loss

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, A model f_θ

Initialize the model parameters θ randomly
for $t = T_0, T_1, , \ldots, T_S$ **do**
    $\mathcal{B} \leftarrow$ SampleMinibatch($\mathcal{D}$, m)
    $\mathcal{L}(f_\theta) \leftarrow \frac{1}{m} \sum_{(x,y)\in\mathcal{B}} \cdot \mathcal{L}_{MM}((x, y); f_\theta)$
    $f_\theta \leftarrow f_\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta)$ ▷ one SGD step
end for
for $t = T_S, , \ldots, T_E$ **do**
    $\mathcal{B} \leftarrow$ SampleMinibatch($\mathcal{D}$, m)
    $\mathcal{L}(f_\theta) \leftarrow \frac{1}{m} \sum_{(x,y)\in\mathcal{B}} n_y^{-1} \cdot \mathcal{L}_{MM}((x, y); f_\theta)$
    $f_\theta \leftarrow f_\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta)$ ▷ one SGD step
end for

Experiments – image classification

Table 1: Top-1 validation errors of ResNet-32 on imbalanced CIFAR-10 and CIFAR-100. The MM-LDAM-DRW, achieves better performances, and each of them individually is beneficial when combined with LDAM loss or DRW schedules.

Dataset	Imbalanced CIFAR-10				Imbalanced CIFAR-100			
Imbalance Type	long-tailed		step		long-tailed		step	
Imbalance Ratio	100	10	100	10	100	10	100	10
ERM [5]	29.64	13.61	36.70	17.50	61.68	44.30	61.45	45.37
Focal [20]	29.62	13.34	36.09	16.36	61.59	44.22	61.43	46.54
LDAM [5]	26.65	13.04	33.42	15.00	60.40	43.09	60.42	43.73
MM (ours)	**26.56**	**12.34**	**33.19**	**13.99**	**60.29**	**42.63**	**60.25**	**43.55**
CB-RS [5]	29.45	13.21	38.14	15.41	66.56	44.94	66.23	46.92
CB-RW [6]	27.63	13.46	38.06	16.20	66.01	42.88	78.69	47.52
CB-Focal [6]	25.43	12.90	39.73	16.54	63.98	42.01	80.24	49.98
HG-DRS [5]	27.16	14.03	29.93	14.85	-	-	-	-
LDAM-HG-DRS [5]	24.42	12.72	24.53	12.82	-	-	-	-
M-DRW [5]	24.94	13.57	27.67	13.17	59.49	43.78	58.91	44.72
LDAM-DRW [5]	22.97	11.84	23.08	12.19	57.96	41.29	54.64	40.54
LDAM-DRW + SSP [11]	22.17	11.47	22.95	11.83	56.57	41.09	54.28	40.33
MM-DRW (ours)	**21.98**	**11.44**	**22.83**	**11.48**	**57.14**	**40.63**	**54.57**	**40.28**
MM-LDAM-DRW (ours)	**21.37**	**11.26**	**21.82**	**11.33**	**56.53**	**40.54**	**53.70**	**40.07**

Maximum-Margin Loss Function - DRW

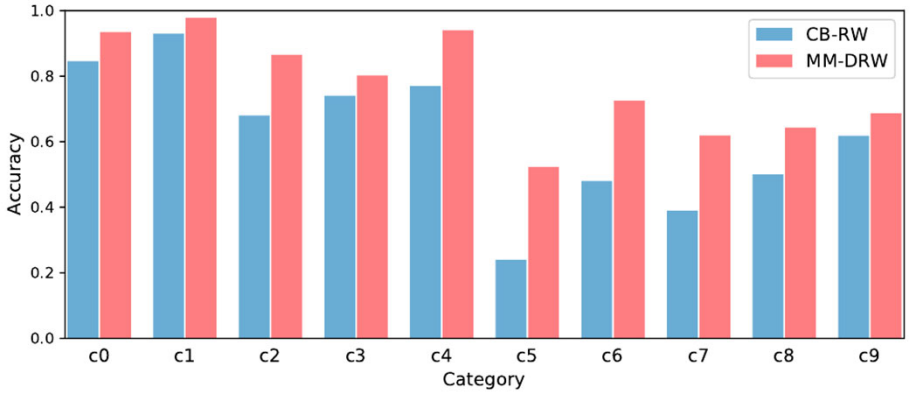


Fig. 3: Per-class top-1 error on CIFAR-10 with step imbalance ($\rho = 100$). Classes $C1$ to $C4$ are majority classes, and the rest are minority classes. The performances of MM-DRW w.r.t. all classes are better than the CB-RW [6].

Table 2: Ablation Study : top-1 validation errors of hyper-parameters $\delta^+ = \delta^- * \beta$ (Eq. 2 and 3) on CIFAR-10.

Dataset	Imbalanced CIFAR-10							
Type	long-tailed		step		long-tailed		step	
Ratio	100	β/δ^-	100	β/δ^-	10	β/δ^-	10	β/δ^-
MM-DRW	22.24	1.4 / 0.6	22.92	1.2 / 0.6	11.66	1.1 / 0.7	**11.48**	1.0 / 2.1
	21.98	1.5 / 0.6	**22.83**	1.3 / 0.6	**11.44**	1.2 / 0.7	11.64	1.1 / 2.1
	22.43	1.6 / 0.6	23.29	1.4 / 0.6	11.86	1.3 / 0.7	11.78	1.2 / 2.1

Table 3: Ablation Study : top-1 validation errors of hyper-parameters $\delta^+ = \delta^- * \beta$ (Eq. 2 and 3) on CIFAR-100.

Dataset	Imbalanced CIFAR-100							
Type	long-tailed		step		long-tailed		step	
Ratio	100	β/δ^-	100	β/δ^-	10	β/δ^-	10	β/δ^-
MM-DRW	58.02	1.2 / 1.2	54.65	1.7 / 1.8	40.97	1.3 / 1.5	40.42	1.0 / 2.4
	57.14	1.3 / 1.2	**54.57**	1.8 / 1.8	**40.63**	1.4 / 1.5	**40.28**	1.1 / 2.4
	57.24	1.4 / 1.2	54.76	1.9 / 1.8	40.95	1.5 / 1.5	40.48	1.2 / 2.4

Conclusions

- For better generalization on the minority classes, we designed the Maximum Margin (MM) loss function, motivated by minimizing a margin-based generalization bound through the shifting decision bound.

- To show the effectiveness, we conducted experiments on artificially imbalanced CIFAR-10/100: the MM outperformed the theoretically principled label-distribution-aware margin (LDAM); the per-class error of CB-RW was compared with that of MM.

- We concluded that the MM to enforce more margin non-linearly into minority class samples works better empirically.