# LEARNING IMBALANCED DATASETS WITH MAXIMUM MARGIN LOSS

*Haeyong Kang*, Thang Vu, and Chang D. Yoo*

Korea Advanced Institute of Science and Technology
School of Electrical Engineering
291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506389

## ABSTRACT

A learning algorithm referred to as Maximum Margin (MM) is proposed for considering the class-imbalance data learning issue: the deep model tends to predict the majority classes rather than the minority ones. For better generalization on the minority classes, the proposed Maximum Margin (MM) loss function is newly designed by minimizing a margin-based generalization bound through the shifting decision bound. As a prior study, the theoretically principled label-distribution-aware margin (LDAM) loss had been successfully applied with classical strategies such as re-weighting or re-sampling. However, the maximum margin loss function has not been investigated so far. In this study, we evaluate the two types of hard maximum margin-based decision boundary shift with training schedule on artificially imbalanced CIFAR-$10/100$ and show the effectiveness.

***Index Terms***— Maximum Margin (MM) Loss, Hard Positive/Negative Margin, Label-Distribution-Aware Margin (LDAM)

## 1. INTRODUCTION

With the advancement of deep neural networks, large-scale datasets have appeared. In general, these real-world large data sets often have shown long-tailed label distributions [1, 2, 3, 4] as depicted in Fig.1. On these datasets, the models [3] have a tendency to perform poorly on the minority classes due to the over-fitting. This tendency has to do with biased predictions, i.e., the trained deep model tends to predict the majority classes rather than the minor ones. Thus, overfitting to minority classes seems to be one of the challenges of generalization.

For robustness to the over-fitting to minority classes, it needs to design a training loss that is in expectation closer to the test distribution or to regularize the parameters to achieve better trade-offs between the accuracies of the majority classes and the minority classes. Instead of depending on the sampling size-dependent margins [5], we design a novel maximum margin loss function that encourages the model to have the optimal trade-off between per-class margins.

This work was partly supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-01396) and partly supported by the IITP grant (2017-0-01780, The technology development for event recognition/relational reasoning and learning knowledge based system for video understanding)

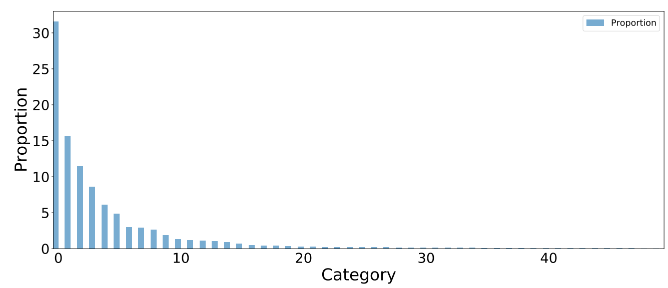


**Fig. 1**: Long-tail distribution of the real-world large datasets: the foreground predicate proportion of the Visual Genome [3], containing $150$ object classes and $50$ predicate classes.

To achieve an optimal trade-off between the margins of the classes, we design a loss function to maximize the per-class margins with the following assumption. See figure 2 for an illustration in the binary classification case. We assume that the decision boundary is shifted by the hard samples that are defined by the maximum margin. The hard samples compose of two types of margins: hard positive margin $\Delta_j^+$ and hard negative margin $\Delta_j^-$. The hard positive margins $\Delta_j^+$ w.r.t $j$-th class are defined by the maximum margin with correctly classified samples; the hard negative margins $\Delta_j^-$ w.r.t $j$-th class are defined by the maximum margin with miss-classified samples. In the training, the hard negative margins shift the model's decision boundary more than the positive margins.

In summary, our main contributions are (i) we design a maximum margin loss function to encourage larger sample margins for hard negative sample classes such that the smaller the maximum margins are the greater the shifting margins are. (2) we applied the maximum loss to the deferred re-balancing optimization procedure [5] for more generalization, and (3) our practical implementation shows significant improvements on two benchmark vision tasks, such as artificially imbalanced CIFAR-$10/100$ for fair comparisons.

This paper is organized as follows. Section 2 provides discussions of related works on imbalanced learning and maximum margin loss function. In Section 3, the proposed maxi-

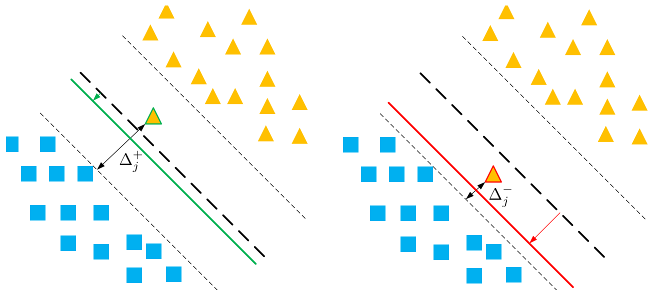


**Fig. 2**: For classification with a linearly separable classifier, the maximum margin of the $j$-th class $\Delta_j$ is defined to be the minimum distance of the data in the $j$-th class to the decision boundary (dotted line). As illustrated here, we assume that the decision boundaries are shifted by two types of the hard maximum margin of samples: hard positive margin $\Delta_j^+$ and hard negative margin $\Delta_j^-$ respectively; our loss function device $\Delta_j^-$ to occupy more margin than $\Delta_j^+$ such that the red decision boundary shifts more than the green one.

mum margin loss function is discussed. In Section 4, experimental results on the artificially imbalanced CIFAR-$10/100$ are discussed along with ablation tests. Finally, Section 5 concludes.

## 2. RELATED WORKS

The two classical approaches for learning long-tailed data are re-weighting [6] the losses of the examples and re-sampling (over-sampling the minority classes [7] and under-sampling the majority classes [8]) the examples in the SGD mini-batch. They both devise a training loss that is in expectation closer to the test distribution to achieve better trade-offs between the accuracies of the majority classes and the minority classes. Recently, other learning paradigms have also been explored such as transfer learning [4], metric learning [9], meta-learning [10], semi-supervised and self-supervised learning [11], and decoupled representation and classifier [12, 13].

A maximum-margin classifier is typically obtained by using the hinge loss function in SVMs [14, 18]. The maximum-margin classifier benefits from margins to minimize intra-class variation in predictions and to maximize the inter-class margin. With the benefits of the maximum-margin, Large-Margin Softmax [15], Angular Softmax [16], and Additive Margin Softmax [17] have been proposed recently. In contrast to these class-independent margins, Label-Distribution-Aware Margin (LDAM) encourages bigger margins for minority classes, providing a concrete formula for the desired margins of the classes. Uneven margins for imbalanced datasets are also proposed and studied in [19]. However, they didn't investigate the maximum margin loss led by training samples yet. In this study, we investigate the performances of two types of hard maximum margin-based decision boundary shift, comparing the results with current state-of-the-art methods.

## 3. MAXIMUM MARGIN (MM) LEARNING

### 3.1. Maximum Margin (MM) Loss

Inspired by the trade-off between the class margins for two classes, we define two types of maximum margins for multiple classes of the following form as follows :

$$\Delta_y^{MM} = \begin{cases} \Delta_y^+ & \text{if } \arg\max_j f_j(x) = y; \\ \Delta_y^- & \text{otherwise.} \end{cases} \tag{1}$$

where an example $(x, y)$, a deep model $f$ with logits $\boldsymbol{z}$, hyper-parameters $\delta^+/\delta^-$ to acquire empirical sample maximum margins,

$$\Delta_y^+ = \exp\left(-\max(z_y - \max_{j\neq y} z_j, 0) - \delta^+\right), \tag{2}$$

and

$$\Delta_y^- = \exp\left(-\max(\max_{j\neq y} z_j - z_y, 0) - \delta^-\right). \tag{3}$$

To achieve an optimal trade-off between the margins of the classes, we design a Maximum Margin (MM) loss function to maximize the per-class margins with the following assumption. See figure 2 for an illustration in the binary classification case. We assume that the decision boundary is shifted by the hard samples that are defined by the maximum margin. The hard samples compose of two types of margins: hard positive margin $\Delta_j^+$ and hard negative margin $\Delta_j^-$ with hyper-parameter $\delta$ for smoothing effects. The hard positive margins $\Delta_j^+$ w.r.t $j$-th class are defined by the maximum margin with correctly classified samples ($z_y > \max_{j\neq y} z_j$). The hard negative margins $\Delta_j^-$ w.r.t $j$-th class are defined by the maximum margin with miss-classified samples ($\max_{j\neq y} z_j > z_y$). We take exponential function to get more non-linearity such that the smaller the maximum margins are the greater the shifting margins are.

We design a maximum margin loss function to encourage the network to have the margins above. Let $(x, y)$ be an example and $f$ be a model. For simplicity, we use $z_j = f(x)_j$ to denote the $j$th-output of the model for the $j$-th class. Following the previous work [5], in order to tune the margin more easily, we effectively normalize the logits (the input to the loss function) by normalizing the last hidden activation to $\ell_2$ norm $1$ and normalizing the weight vectors of the last fully-connected layer to $\ell_2$ norm $1$. Notice that we then scale the logits by a constant $s = 10$. Empirically, the non-smoothness of hinge loss may pose difficulties for optimization. The smooth relaxation of the hinge loss is the following cross-entropy loss with enforced margins:

$$\mathcal{L}_{MM}((x,y); f) = -\log \frac{e^{z_y - \Delta_y^{MM}}}{e^{z_y - \Delta_y^{MM}} + \sum_{j\neq y} e^{z_j}} \tag{4}$$

where $\Delta_j^{MM}$ in Eq. 1 for $j \in \{1, \cdots, k\}$.

### 3.2. MM's Hyper-parameters $\delta^+/\delta^-$

To achieve the best performances of MM, we set the relationship between the hyper-parameters $\delta^+, \delta^-$ in Eq. 2 and 3 as follows:

$$\delta^+ = \delta^- * \beta \quad (5)$$

where the $\beta$ is a scaler. In the experiments, the best performances were acquired by setting both $\delta^-$ and $\beta > 1.0$ empirically, meaning that the hard negative margins $\Delta_j^-$ shift more than $\Delta_j^+$ in the process of training. To further enforce a class-dependent margin for multiple classes, we add the class-distribution-aware margin $\gamma_j = \frac{C}{n_j^{1/4}}$ [5] for some constant $C$ to Eq. 5 as follows:

$$\delta_j^+ = (\delta^- - \gamma_j) * \beta, \quad (6)$$

and

$$\delta_j^- = \delta^- - \gamma_j. \quad (7)$$

### 3.3. Deferred Re-balancing Optimization Schedule [5]

For a fair comparison, the proposed MM loss is also applied to the deferred re-balancing training procedure [5] as shown in Algorithm 1, which first trains using vanilla ERM with the MM loss before annealing the learning rate, and then deploys a re-weighted MM loss with a smaller learning rate. In the following experiments with the MM loss function, the first stage of training leads to better initialization for the second stage of training with re-weighted losses. With the non-linear MM loss of the hyper-parameter $\delta^+, \delta^-$ and deferred re-balancing training, this re-weighting scheme works stable more.

**Algorithm 1** Imbalanced Learning with MM Loss

**Require:** Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, A model $f_\theta$
Initialize the model parameters $\theta$ 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**

## 4. EXPERIMENTS

**Datasets.** We evaluate our proposed MM loss function on artificially created versions of CIFAR-10 and CIFAR-100 with controllable degrees of data imbalance.

**Baselines.** We compare our MM with the standard training and other state-of-the-art algorithms. For a fair compararision, we follow the prior experiment setting [5]: (1) Empirical risk minimization (ERM) loss: all the examples have the same weights; by default, all model use standard cross-entropy loss; (2) Re-Weighting (RW): the model re-weights each sample by the inverse of the sample size of its class, and then re-normalize to make the weights 1 on average in the mini-batch; (3) Re-Sampling (RS): each example is sampled with probability proportional to the inverse sample size of its class; (4) CB : the examples are re-weighted or re-sampled according to the inverse of the effective number of samples in each class, defined as (1-$\beta^{n_i}$) = (1-$\beta$), instead of inverse class frequencies; (5) Focal: we use the recently proposed focal loss; (6) SGD schedule: by SGD, we also refer to the standard schedule where the learning rates are decayed a constant factor at certain steps; we follow the same standard learning rate decay schedule.

**Our proposed algorithms** We evaluate the following algorithms: (1) MM: the proposed Maximum Margin losses; (2) MM-DRW: following the training Algorithm 1, the MM with DRW Eq.5 is evaluated and (3) MM-LDAM-DRW with Eq. 6 and 7 is also performed with the parameter settings in Table. 2 and 3.

**Implementation details for CIFAR.** For CIFAR-10 and CIFAR-100, we follow the simple data augmentation in [21] for training: 4 pixels are padded on each side, and a $32 \times 32$ crop is randomly sampled from the padded image or its horizontal flip. We use ResNet-32 [21] as our base network, and use stochastic gradient descend with the momentum of 0.9, weight decay of $2 \times 10^{-4}$ for training. The model is trained with a batch size of 128 for 200 epochs. For a fair comparison, we use an initial learning rate of 0.1, then decay by 0.01 at the 160th epoch and again at the 180th epoch. We also use linear warm-up learning rate schedule [22] for the first 5 epochs.

### 4.1. Experimental results on CIFAR

**Imbalanced CIFAR-10 and CIFAR-100.** The original version of CIFAR-10 and CIFAR-100 contains 50,000 training images and 10,000 validation images of size $32 \times 32$ with 10 and 100 classes, respectively. We evaluate the MM loss function on their imbalanced version that reduces the number of training examples per class and keeps the validation set unchanged. To ensure that our methods are compared with a variety of settings, we consider two types of imbalance: long-tailed imbalance [6] and step imbalance [8]. We use imbalance ratio $\rho$ to denote the ratio between sample sizes of the most frequent and least frequent class, i.e., $\rho = \max_i\{n_i\}/\min_i\{n_i\}$. A long-tailed imbalance follows an exponential decay in sample sizes across different classes. For step imbalance setting, all minority classes have the same sample size, as do all frequent classes. This gives a clear distinction between minority classes and majority classes.

**Performances.** We report the top-1 validation error of various

**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** |

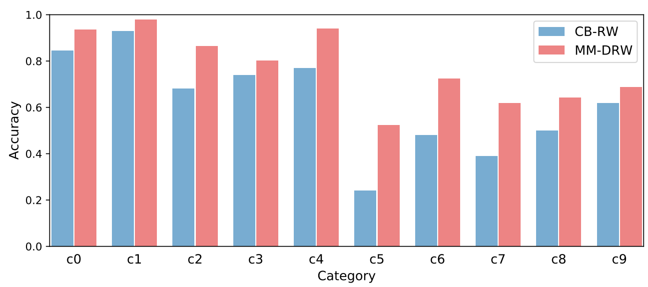


**Fig. 3**: Per-class top-1 error on CIFAR-10 with step imbalance ($\rho = 100$). Classes $C1$ to $C4$ are majority classes, and the rest are minority classes. The performances of MM-DRW w.r.t. all classes are better than the CB-RW [6].

methods for imbalanced versions of CIFAR-10 and CIFAR-100 in Table 1. We evaluate the performances of MM as well as MM with the DRW training schedule. The overall performance of MM is better than LDAM; in those of MM, the MM method performs better in the imbalance ratio of 10 than 100. With the re-weighting loss in Algorithm 1, the MM-DRW also shows the effectiveness, compared with LDAM-DRW [5]; this explains that the MM method works well under the re-weighting scheme. In particular, MM-LDAM-DRW, our key method to enforce more margin non-linearly into minority class samples in Eq. 6 and 7, works better in the imbalance ratio of 100 than 10.

### 4.2. Ablation study

**Generalization.** To show the effectiveness of the MM loss function, we compare the per-class error of CB-RW with that of MM as shown in Fig. 3 on the imbalanced CIFAR-10.

**Hyper-parameters.** To show how to achieve the performances of MM-DRW, we prepare the ablation study on the hyper-parameters $\delta^+, \delta^-$ in Eq. 2 and 3, as shown in Table. 2 and 3, respectively. The results show that the ratio of $\delta^+/\delta^- > 1.0$ depends on label distributions (long-tailed or step) and dataset size.

**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 | $\beta/\delta^-$ | 100 | $\beta/\delta^-$ | 10 | $\beta/\delta^-$ | 10 | $\beta/\delta^-$ |
| 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 | $\beta/\delta^-$ | 100 | $\beta/\delta^-$ | 10 | $\beta/\delta^-$ | 10 | $\beta/\delta^-$ |
| 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 |

## 5. CONCLUSION

The Maximum Margin (MM) was proposed for considering the class-imbalance data learning issue: the trained model tends to predict the majority classes rather than the minority ones. 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.

## 6. REFERENCES


[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[2] Grant Van Horn and Pietro Perona, "The devil is in the tails: Fine-grained classification in the wild," *arXiv preprint arXiv:1709.01450*, 2017.

[3] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

[4] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2537–2546.

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems*, 2019, pp. 1567–1578.

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[7] Jonathon Byrd and Zachary Lipton, "What is the effect of importance weighting in deep learning?," in *International Conference on Machine Learning*. PMLR, 2019, pp. 872–881.

[8] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[9] Chong You, Chi Li, Daniel P Robinson, and René Vidal, "Scalable exemplar-based subspace clustering on class-imbalanced data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 67–83.

[10] 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," *arXiv preprint arXiv:1902.07379*, 2019.

[11] Yuzhe Yang and Zhi Xu, "Rethinking the value of labels for improving class-imbalanced learning," *arXiv preprint arXiv:2006.07529*, 2020.

[12] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9719–9728.

[13] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2019.

[14] Johan AK Suykens and Joos Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[15] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang, "Large-margin softmax loss for convolutional neural networks.," in *ICML*, 2016, vol. 2, p. 7.

[16] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[17] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[18] Haeyong Kang, Chang D Yoo, and Yongcheon Na, "Maximum margin learning of t-spns for cell classification with filtered input," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 1, pp. 130–139, 2015.

[19] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao, "Striking the right balance with uncertainty," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 103–112.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.