Supplementary Material

M2m: Imbalanced Classification via Major-to-minor Translation

A. Details on the datasets

CIFAR-LT-10/100. CIFAR-10/100 datasets [7] consist of 60,000 RGB images of size 32×32, 50,000 for training and 10,000 for testing. Each image in the two datasets is corresponded to one of 10 and 100 classes, respectively. In our experiments, we construct "synthetically long-tailed" variants of CIFAR-10/100, namely CIFAR-LT-10/100, respectively [1]. We hold-out 10% of the test set to construct a validation set, and use the remaining for testing. We use ResNet-32 [5] with a mini-batch size 128, and set a weight decay of 2×10^{-4}. We train the network for 200 epochs with an initial learning rate of 0.1. We follow the learning rate schedule used by [2] for fair comparison: the initial learning rate is set to 0.1, and we decay it by a factor of 100 at 160-th and 180-th epoch. When the deferred scheduling [1] is used, *e.g.*, DRS, DRW and our method, it is applied after 160 epochs of standard training.

CelebA-5. CelebFaces Attributes (CelebA) dataset [9] is a multi-labeled face attributes dataset. It is originally composed of 202,599 number of RGB face images with 40 binary attributes annotations per image. We port this CelebA to a 5-way classification task by filtering only the samples with five non-overlapping labels about hair colors: namely, "blonde", "black", "bald", "brown", and "gray". This is in a similar manner as done in [11]. We denote the resulting dataset by CelebA-5. We pick out 50 and 100 samples per each class for validation and testing. We use ResNet-32 [5] with a mini-batch size 128, and set a weight decay of 2×10^{-4}. We train the network for 90 epochs with an initial learning rate of 0.1. We decay the learning rate by 0.1 at epoch 30 and 60. When the deferred scheduling is used, it is applied after 60 epochs of standard training.

SUN397. Scene UNderstanding (SUN) [13] is a dataset for a scene categorization. It originally consists of 108,754 RGB images which are labeled with 397 classes. For the inputs, center patches are first extracted and they are resized to 32×32. We hold-out 10 and 40 samples per each class for validation and testing, respectively, as the dataset itself does not provide any separated split for testing. We use pre-activation ResNet-18 [5] which roughly has $4\times$ more channels with a mini-batch size 128, and set a weight decay of 2×10^{-4}. We train the network for 90 epochs with an initial learning rate of 0.1. We decay the learning rate by 0.1 at epoch 30 and 60. When the deferred scheduling is used, it is applied after 60 epochs of standard training.

Twitter. Twitter [3] is a dataset for a part-of-speech (POS) tagging task in social media text with 25 classes. Each sample is a pair of a token and a tag, *e.g.*, "(books, common noun)" and "(#acl, hashtag)", where each token is embedded into a 50-dimensional vector via a pre-defined word-embedding [6]. We discarded two classes with zero test samples and obtained 14,614 training samples with 23 classes. We use 2-layer fully-connected network with a hidden layer size of 256 and a ReLU nonlinearity. We set a mini-batch size 64 and a weight decay of 5×10^{-5}. We train the network for 15 epochs with an initial learning rate 0.1 and decay the learning rate by 0.1 at epoch 10. When the deferred scheduling is used, it is applied after 10 epochs of standard training.

Reuters. Reuters [8] is a dataset for a text categorization task which predicts the subject of a given text. As an input, 1000-dimensional bag-of-words vectors are given, which are processed from a news story document. It is originally composed of 52 classes, but we discarded the classes that have less than 5 test samples for a reliable evaluation, obtaining a subset of the full dataset of 36 classes with 6436 training samples. We hold-out 10% of training samples to construct a validation set. We use 2-layer fully-connected network with a hidden layer size of 256 and a ReLU nonlinearity. We set a mini-batch size 64 and a weight decay of 5×10^{-5}. We train the network for 15 epochs with an initial learning rate 0.1 and decay the learning rate by 0.1 at epoch 10. When the deferred scheduling is used, it is applied after 10 epochs of standard training.

B. More results from ablation study

Generation from another classifier g. As mentioned, our method introduces another classifier g to generate synthetic minority x^* independently from the training classifier f. This is because using f itself instead of g in the optimization objective (2) would let the synthetic samples already confident in the target minority class to f, and this makes the overall training process redundant. To further validate the importance of using g, we consider an ablation called "M2m-Self": instead of using g, "M2m-Self" uses f for generating minority samples. As reported in Table 1, one could immediately see that M2m-Self only shows marginal improvement from DRS, which is much inferior than the original M2m.

Methods	bACC (Δ)	GM (Δ)
ERM-DRS	$75.2_{\pm 0.26}$ (-3.96%)	$73.9_{\pm 0.32}$ (-5.01%)
M2m-Self	$75.9_{\pm 0.27}$ (-3.07%)	$74.9_{\pm 0.32}$ (-3.73%)
M2m-No-Reject	$77.4_{\pm 0.33}$ (-1.15%)	$76.8_{\pm 0.40}$ (-1.29%)
M2m ($\gamma = 0$)	$76.9_{\pm 0.19}$ (-1.79%)	$76.4_{\pm 0.20}$ (-1.80%)
M2m	$78.3_{\pm 0.16}$ (-0.00%)	$77.8_{\pm 0.16}$ (-0.00%)
M2m-Ensemble	$78.5_{\pm 0.20}$ (+0.26%)	$78.0_{\pm 0.22}$ (+0.26%)

Table 1: Comparison of classification performance across various types of ablations. Δ indicates the relative gap from the original result presented in "M2m". All the values and error bars are mean and standard deviation across three random trials, respectively.

Using multiple classifiers for generation. Since our method is not restricted to use the only one pre-trained classifier g in the optimization (2), the multiple classifiers g_i for $i = 1, \ldots, m$ can be used to improve the quality of generation. To verify the additional gain from multiple classifiers, we consider an ablation called "M2m-Ensemble": use the ensemble of the classifiers ($m = 2$) for generation instead of the single classifier. Here, we use the same architecture ResNet-32 for g_1 and g_2 and use a higher γ due to the smoothed prediction from the ensemble. The results in Table 1 show that M2m-Ensemble slightly perform better than M2m. It indicates that our method can benefit from the stronger classifier.

Rejection criteria. We also propose a sample rejection criteria to alleviate the risk of unreliable generation, possibly due to a weak generalization of g. To verify the effect of this rejection criteria, we consider an ablation, namely "M2m-No-Reject", which does not use this rejection policy in training. In other words, all the generated samples are used to train f. The results in Table 1 show that M2m-No-Reject performs significantly worse than M2m. This indeed confirms the gain from using the proposed rejection criteria.

The effect of γ. As specified in Algorithm 1 in the main paper, we set a threshold γ to filter out the synthetic samples which the generation objective is not sufficiently minimized, mainly due to the limited budget. To evaluate the practical effectiveness of using γ, here we consider an ablation that this thresholding is not used, equivalently when $\gamma = \infty$. As reported in Table 1, we indeed observe a performance degradation by not using γ. This reveals that the confidence level in g affects the final quality of the generation.

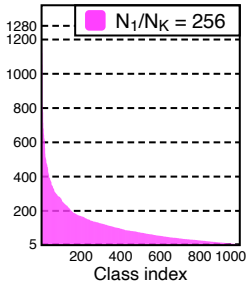

Figure 1: The class-wise distribution of ImageNet-LT.

Loss	Re-balancing	bACC (Δ)	GM (Δ)
ERM	-	$38.6_{\pm 0.75}$	$26.9_{\pm 0.78}$
ERM	DRS	$40.8_{\pm 0.67}$	$31.6_{\pm 1.05}$
ERM	**M2m (ours)**	$\mathbf{42.2}_{\pm 0.51}$	$\mathbf{33.1}_{\pm 0.64}$
LDAM	-	$41.0_{\pm 0.07}$	$28.5_{\pm 0.11}$
LDAM	DRW	$43.0_{\pm 0.17}$	$34.5_{\pm 0.17}$
LDAM	**M2m (ours)**	$\mathbf{43.7}_{\pm 0.26}$	$\mathbf{35.1}_{\pm 0.35}$

Table 2: Comparison of classification performance on ImageNet-LT. All the values and error bars are mean and standard deviation across three random trials, respectively.

C. Results on ImageNet-LT

We additionally evaluate our method on ImageNet-LT [10] dataset, a subset of ImageNet dataset [12] with a synthetic imbalance following the Pareto distribution of the power $\alpha = 6$. It is composed of 115,846 training samples with 1,000 categories, 1,280 images in the maximal class and 5 images in the minimal class. A more detailed distribution is presented in Figure 1. We use the randomly-resized cropping and the horizontal flipping as a data augmentation, and all the images are resized to 128×128. We hold-out 20 samples per class randomly from the original ImageNet training set to form a

validation set, and the original (roughly balanced) ImageNet validation set is used for testing. We use ResNet-50 [4] with a mini-batch size 256 and set a weight decay of 10^{-4}. We train the network for 200 epochs with an initial learning rate of 0.1 and it is decayed by 0.1 at epoch 160 and 180. When the deferred scheduling is used, *e.g.*, DRS, DRW and our method, it is applied after 160 epochs of standard training. We evaluate our method with followings which show the best performance among the baselines in the experiments in the main paper: (*a*) ERM-DRS and (*b*) LDAM-DRW [1]. We report the *balanced accuracy* (bACC) and the *geometric mean scores* (GM). As reported in Table 2, our method, M2m, significantly outperforms the baselines. In the case of ERM loss, compare to DRS, M2m shows 3.43 % and 4.75 % relative gains in bACC and GM, respectively. Furthermore, with a margin-based loss function LDAM, the improvement is much enlarged.

References

[1] 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 (NeurIPS)*, 2019. 1, 3

[2] 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 (CVPR)*, 2019. 1

[3] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2011. 1

[4] 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 (CVPR)*, 2016. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[6] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2016. 1

[7] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, 2009. 1

[8] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 2004. 1

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[10] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[11] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[13] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 1