Exploiting Synthetic Data for Data Imbalance Problems: Baselines from a Data Perspective

Moon Ye-Bin[1*] Nam Hyeon-Woo[1*]
Wonseok Choi[2] Nayeong Kim[3] Suha Kwak[2,3] Tae-Hyun Oh[1,2,4†]

[1]Dept. of Electrical Engineering and [2]Grad. School of Artificial Intelligence, POSTECH
[3]Dept. of Computer Science and Engineering, POSTECH
[4]Institute for Convergence Research and Education in Advanced Technology, Yonsei University

Abstract

We live in a vast ocean of data, and deep neural networks are no exception to this. However, this data exhibits an inherent phenomenon of imbalance. This imbalance poses a risk of deep neural networks producing biased predictions, leading to potentially severe ethical and social consequences. Given the remarkable advancements of recent diffusion models generating high-quality images, we believe that using generative models is a promising approach to address these challenges. In this work, we propose a strong baseline, SYNAuG, that utilizes synthetic data as a preliminary step before employing task-specific algorithms to address data imbalance problems. This simple approach yields impressive performance improvement on the data imbalance task such as CIFAR100-LT and ImageNet100-LT. While we do not claim that our approach serves as a complete solution to the problem of data imbalance, we argue that supplementing the existing data with synthetic data proves to be a crucial preliminary step in addressing data imbalance concerns. Note that this is a work in progress.

1. Introduction

Deep neural networks (DNNs) have achieved strong performance on visual tasks. The outstanding performance has been demonstrated by training a model with abundant and diverse labeled data, *e.g.*, [7, 10], suggesting a strong dependency on the quality and scale of datasets. However, state-of-the-art DNNs suffer from much lower performance in some classes where rich and well-organized data are not available.

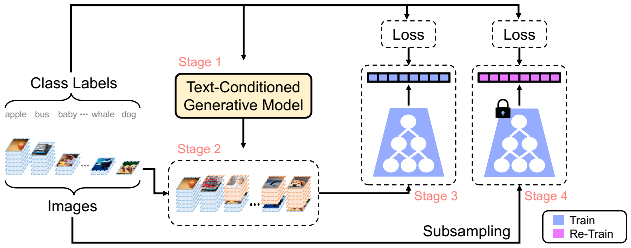


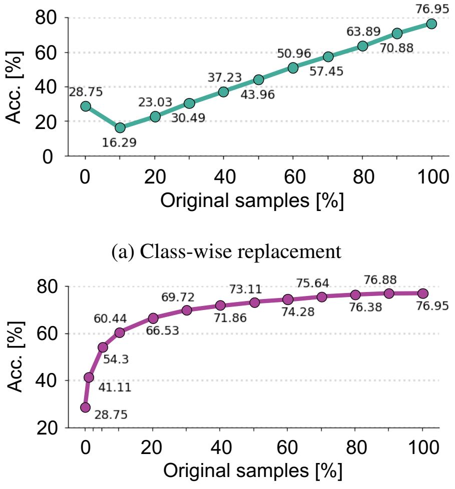
Figure 1. **Overview of SYNAuG process.** We first generate the synthetic samples from the text-conditioned generative model and then fill the imbalanced distribution. We train the model with the uniformized training data and re-train the classifier with the uniformly sub-sampled real data.

The data imbalance is often accompanied by the collection of massive data, especially when data is collected from the internet [7, 17] due to the natural image statistics that exhibit power law cluster size distributions [4, 18]. These natural biases lead the dataset to be imbalanced in terms of classes [6, 21]. With such a natural dataset, the classifier is typically trained by the standard supervised learning algorithms based on the empirical risk minimization (ERM) principle [19]. ERM deals with all the data equally and does not address imbalance; thus, training with ERM has been known to cause a biased classifier toward the majority of training data [8]. Researchers have independently developed various algorithms since these problems lead to substantial performance degradation.

In this work, we suggest using generative models to fill up and uniformize the number of samples in all the classes. Our approach differs from the prior arts limited to using a fixed and bounded dataset, mainly focusing on algorithmic strategies. Beyond the given dataset, we focus on exploiting recent generative models. Recently, generative diffusion models [16, 11] have shown potential as synthetic

[*] Authors contributed equally to this work.

[†] This work was supported by Institute of Information communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities; No.2021-0-02068, Artificial Intelligence Innovation Hub)



(a) Class-wise replacement

(b) Instance-wise replacement

Figure 2. **Replacement test.** We replace the original data with synthetic ones with (a) class-wise and (b) the same ratio of instances of all classes. We use CIFAR100 for the test dataset.

training data generation [14, 15]. In this regard, we propose a simple unified preceding step SYNAuG to deal with two data imbalance problems by leveraging a pre-trained generative diffusion model, as shown in Fig. 1. In the generative model era, we argue that restricting training data to the given dataset is rather not practical. Furthermore, data imbalance problems should be tackled from the data level before deploying algorithmic approaches as in the prior arts, so that the practitioners take the controllability of data to mitigate and stabilize the base conditions of datasets.

To mitigate the domain gap, we introduce Mixup, which interpolates the real and synthetic samples, and Re-train, which trains the last layer with real samples after pretraining on the real and synthetic samples; we call it SYNAug. This outperforms the other baselines on the long-tailed recognition benchmark. Our contributions are as follows:

- We propose SYNAug that balances the number of class instances, not restricted to the given datasets.
- Mixup and Re-train are introduced to deal with the domain gap between real and synthetic data.
- Balancing class distribution with synthetic data can significantly improve the long-tailed recognition accuracy.

2. Method

Motivation. The distribution and data amount of the original dataset affects the performance. To investigate the characteristics, we devise two experiments by changing the original and synthetic data ratio. The first (Fig. 2a) is to replace the original with the synthetic for each class; the second (Fig. 2b) does for entire classes gradually.

We observe that the second shows the performance degradation alleviate rather than the first. Interestingly, the performance of 1% of original data in the second is similar to the performance of 50% of original data in the first. These results imply that (1) synthetic data is still insufficient to replace original data fully, (2) we need a few original samples when adding the synthetic data, and (3) there might be additional room for improvement due to the domain gap.

Method	**IF=100**				**50**	**10**
	Many	**Medium**	**Few**	**All**		
CE [6]	68.31	36.88	4.87	37.96	43.54	59.50
SSD [9]	-	-	-	46.0	50.5	62.3
PaCo [5]	-	-	-	52.0	56.0	64.2
RISDA [3]	-	-	-	50.16	53.84	62.38
CE + CMO [13]	70.4	42.5	14.4	43.9	48.3	59.5
LDAM + CMO [13]	61.5	48.6	28.8	47.2	51.7	58.4
RIDE (3 experts) + CMO [13]	-	-	-	50.0	53.0	60.2
Weight Balancing [1]	72.60	51.86	32.63	53.35	57.71	**68.67**
WebAug	72.71	51.21	36.13	54.06	56.40	63.86
Intra-class Image Translation	71.86	45.88	22.97	47.87	53.33	64.95
Inter-class Image Translation	73.49	45.77	19.00	47.17	51.33	64.11
Class Distribution Fitting	**74.83**	50.79	26.03	51.53	55.60	65.60
SYNAuG	73.97	**56.26**	**39.10**	**57.31**	**60.34**	67.90

Table 1. **Long-tailed recognition performance on CIFAR100-LT.** We report the Top-1 accuracy (%) with different imbalance factors, *i.e.*, IF={100, 50, 10}. **Bold** stands for the highest accuracy in each IF or class.

SYNAuG. Based on our findings, we propose SYNAuG to exploit the synthetic data to alleviate the imbalance in the data perspective. To mitigate data imbalance, SYNAuG generates synthetic data, populates the imbalanced data to become a uniform distribution, and trains the model with uniformized data. We use the recent powerful generative models, *e.g.*, Stable Diffusion [16]. We generate the samples with the diversified prompt, "a photo of {`modifier`} {`class`}". We use the large language model, ChatGPT [12], to augment {`modifier`}. Then, we utilize two simple methods to reduce the domain gap: (1) Mixup [20] that interpolates samples between real and synthetic samples, and (2) Re-training the classifier with re-initialization on the uniformly sampled real samples.

Other Baselines. We construct three baselines using the generative model: (a) Intra-class Image Translation is that real samples with the same class are used as the guidance image; (b) Inter-class Image Translation is that random real samples are exploited as the guidance image, regardless of the class; and for (c) Class Distribution Fitting, we fine-tune the generative models with real samples.

3. Experiments

In this section, we demonstrate the effectiveness of our SYNAuG for the long-tailed recognition task.

Experimental setting. We employ two long-tailed recognition datasets: CIFAR100-LT and ImageNet100-LT. These datasets are curated by making the class distribution imbalanced from the original datasets, CIFAR100 and Ima-

	Modifier	Mixup	Re-train	IF		
				100	50	10
(a)				52.41	56.99	66.34
(b)	✓			53.54	57.09	66.66
(c)	✓	✓		55.45	58.69	66.84
(d)	✓	✓	✓	**57.31**	**60.34**	**67.90**

Table 2. **Ablation study of the components of SYNAuG on CIFAR100-LT.** (d) stands for SYNAuG.

geNet100. The classes in the long-tailed datasets are divided into three groups: Many-shot (more than 100 samples), Medium-shot (20-100 samples), and Few-shot (less than 20 samples). For CIFAR100-LT, the imbalance factor (IF) can be controlled by computing the ratio of samples in the head to tail class, N_1/N_K, where $N_k = |\mathcal{D}_k|$, and $\mathcal{D}_k$ is the set of samples belonging to the class $k \in \{1, \cdots, K\}$. When the IF value is large, the skewness of the training data is more severe, which has fewer samples and is more challenging. We evaluate under the different IFs of 100, 50, and 10. ResNet32 and ResNet50 are used for CIFAR100-LT and ImageNet100-LT, respectively.

Experiments on CIFAR100-LT. As shown in Table 1, compared to the vanilla method [6] using Cross Entropy (CE) loss only, we achieve a large improvement when exploiting the generated samples regardless of the skewness of the training data. Our method also shows outperformed performance compared to most of the previous works. The results demonstrate that relieving the imbalance from a data point of view is simple but more effective than conventional complex methods. Compared to the case that uses real-world web data[1], it shows that the generated images are of sufficient quality to mitigate the class imbalance problem.

Three baselines perform better than the other training only with the original long-tailed data but lower than our method. This result implies that the domain gap between the original and synthetic data is hard to narrow during the generation process.

Ablation study. Table 2 shows the ablation study of our method. Comparing (a) and (b), the diversified prompt by ChatGPT [12] improves performance. The current text-to-image generative models provide the controllability of synthetic data, which is beneficial. When we utilize Mixup (c) for interpolating between original and synthetic data, we can achieve further improvement. The mixed-up with original data serves as a bridge between the two groups, affecting largely performance improvement. Finally, we can get an additional gain by re-training the classifier (d), which stands for our SYNAuG.

Performance according to the number of synthetic data.

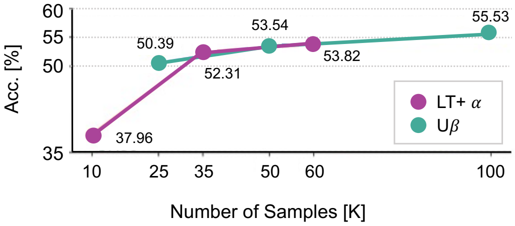


Figure 3. **Accuracy [%] (y-axis) vs. number of samples [K] (x-axis).** As we expected, performance improves as more synthetic samples are added. Also, it is improved significantly when the Few class disappears as the number of samples per class increases.

We explore the performance by varying the number of synthetic data. We use CIFAR100-LT with IF $=$ 100, so the total number of the original data is 10,847. For LT+α, we add the synthetic data to all classes equally without care of Many, Medium, and Few classes where α is the number of synthetic samples. In this case, the absolute difference in sample amount between classes remains. For Uβ, we add synthetic data to be the total number of samples in each class same where β is the upper bound of the data amount. We cut the original samples if it overflows than β. In Uβ case, the data distribution becomes uniform.

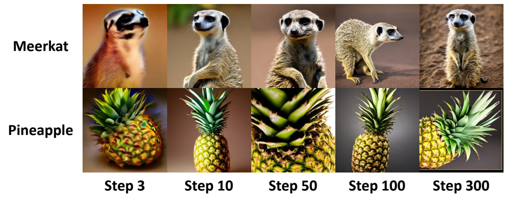
As shown in Fig. 3, the performance is increased as the synthetic data is added. For LT case, the performance is increased significantly at first because the Few class becomes not Few anymore. LT+α and Uβ are similar when we add much synthetic data. We hypothesize that LT+α becomes far from the long-tailed distribution although the difference in data amount across classes remains in LT+α.

Performance according to the quality of synthetic data. We evaluate SYNAuG on ImageNet100-LT considering the large resolution. Since it is known that the step value is related to generated image quality in Stable Diffusion, we also perform an ablation study according to the step value. As shown in Fig. 4 (Top), the generation quality is low when the number of steps is very small, but there is no big difference to the naked eye when it goes up to a certain number. These results lead to quantitative results as well in Fig. 4 (Bottom). Compared to the CE method trained on the original long-tailed data, we achieve large improvements in all the classes regardless of the synthetic image quality. There is a considerable margin between using synthetic samples having low quality and having a certain level of quality, but the difference is marginal when the image quality, according to the step value, exceeds the threshold.

4. Conclusion

We propose SYNAuG that improves the performance consistently on long-tailed recognition. While we focus on the data perspective, we believe that improving the model in multiple views is necessary for effective solutions to data imbalance. We notice a gradual decline in performance

[1] We collected image data for WebAug from Google image search. Google image search returns images very favorable to DNNs, because Google use CNN-based image search since March 2013 [2]. Thus, WebAug is analogous to the distillation of a Google internal model, *i.e.*, very strong baseline.



Method	Many	Medium	Few	All
CE	61.85	15.83	0.29	32.06
SYNAuG (step 3)	67.95	31.74	9.71	43.14
SYNAuG (step 10)	68.60	37.00	11.14	46.02
SYNAuG (step 50)	68.20	37.65	14.00	46.56
SYNAuG (step 100)	68.70	37.17	11.00	46.12
SYNAuG (step 300)	69.30	37.52	12.29	46.70

Figure 4. **(Top) quality of the generated samples according to the number of step, (Bottom) long-tailed recognition performance on ImageNet100-LT with ResNet50.**

when substituting real samples with synthetic data, suggesting the potential need for domain adaptation. There could be future research direction that is more sophisticated data augmentation, automatic data curation, transfer learning, the usage of differentiability of the generative models, and understanding taxonomies across classes. Thus, we emphasize that this work suggests a promising way to redraw the direction to overcome the long-standing data imbalance problems in the data perspective, and more interesting future work will come with integrating multiple levels.

References

[1] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[2] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[3] Xiaohua Chen, Yucan Zhou, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2

[4] National Research Council et al. *Frontiers in massive data analysis*. National Academies Press, 2013. 1

[5] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1

[8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 1

[9] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[10] 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 (ECCV)*, 2014. 1

[11] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *International Conference on Machine Learning (ICML)*, 2022. 1

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[13] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[14] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[15] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. 2023. 2

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[17] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations (ICLR)*, 2020. 1

[18] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001. 1

[19] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 1

[20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[21] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. 2023. 1