Distilling Long-tailed Datasets

Supplementary Material

6. Training Details

Experiment Setup. We compare our method with various coreset selection methods and state-of-the-art dataset distillation methods. Consistent with these works, we use ConvNets [23] for the training and evaluation. For trajectory matching methods such as MTT, DATM, and our method, we trained and saved 100 experts with 100 epochs. During evaluation, the models are trained for 1000 epochs on the synthetic dataset. For the experts of our method, the experts are trained in a decoupled manner. For representation training, the experts are trained for 100 epochs with weight decay. For classifier fine-tuning, the experts are fine-tuned for 10 epochs with MaxNorm constraint [16, 37]. For TinyImageNet-LT and ImageNet-LT, we use a depth-4 ConvNet in our experiments. For experiments on ImageNet-LT, we adopt the Tesla [6] code base to reduce memory usage.
Hyper-parameters. We report the hyper-parameters of our method under different settings in Table 10. For expert epochs, image learning rate, label learning rate, and other hyper-parameters, we follow the previous works [4, 13].

7. More Experiments

7.1. Long-tailed Test Set

We perform experiments on CIFAR10-LT so that the train/test set follows the same long-tailed distribution, and the number of the samples for each class in the test set is (1000, 555, 308, 170, 94, 52, 29, 16, 9, 5). The results are shown in Table 6. We further compare the precision, recall and F1-score of the prediction results to obtain better insights. "-" indicates that the method cannot converge during training. Our analysis reveals the following:

Metrics	Acc.(%)		Precision		Recall		F1	
IPC	10	50	10	50	10	50	10	50
Random	27.3±1.4	54.2±1.1	0.21±0.01	0.35±0.01	0.28±0.01	0.53±0.01	0.17±0.01	0.34±0.01
MTT [4]	44.7±0.0	12.4±1.0	0.04±0.00	0.20±0.01	0.10±0.00	0.25±0.01	0.06±0.00	0.14±0.01
DATM [13]	-	72.7±0.5	-	0.42±0.01	-	0.47±0.01	-	0.44±0.01
Ours	**51.6±1.4**	**73.2±1.2**	**0.37±0.01**	**0.46±0.01**	**0.53±0.02**	**0.62±0.01**	**0.35±0.01**	**0.48±0.01**

Table 6. **Quantitative comparisons on long-tailed test set.**

1. The accuracy of a long-tailed testset cannot reflect the actual model performance. Take MTT as an example; the accuracy of IPC 10 is much higher than that of IPC 50. We find that this is because the model trained in IPC 10 predicts all samples to the first class so that its accuracy reaches 44.7% (the percentage of the first class sample numbers). However, this definitely does not indicate the model obtained by MTT IPC 10 performs well.
2. The effectiveness of our method is consistent with the results of the balanced test set. Instead of only focusing on accuracy, for an imbalanced test set, we should also involve precision, recall, and F1-score, as provided in Table 6. From the table, we can observe that our methods have a small leading on precision, outperform the baselines with a large margin on recall, and have the best performance on the f1 score. The results indicate that we can preserve the head class accuracy and improve the tail class accuracy. To support this conclusion, we show the class-wise accuracy comparison under IPC 50 in Table 7.

Method	cls0	cls1	cls2	cls3	cls4	cls5	cls6	cls7	cls8	cls9
DATM [13]	**78.1**	86.3	**58.8**	**60.0**	57.4	38.5	51.7	31.3	22.2	0.0
Ours	76.8	**91.4**	58.1	53.5	**59.6**	**53.8**	**75.9**	**68.8**	**55.6**	**20.0**

Table 7. **Class-wise accuracy on long-tailed test set.**

7.2. DD-Ranking Evaluation

We further evaluate our method with DD-Ranking [25] to provide a fair evaluation for LTDD, reducing the impacts from knowledge distillation and data augmentation to reflect the real informativeness of the distilled data. We compared our method with the hard-label-based method MTT and the soft-label-based method DATM in Table 8. These results indicate the effectiveness of our proposed method.

Metrics	Hard Label Recovery (HLR) ↓	Improvement Over Random ↑
MTT [4]	31.3%	-
DATM [13]	13.6%	-0.3%
Ours	**6.6%**	**21.6%**

Table 8. **DD-Ranking Evaluation.**

8. Compute Resources

The computational cost comparisons of our method with the other trajectory matching methods [4, 13] are listed in Table 9. The comparisons are done under the same hardware (NVIDIA A6000) and software environments. We can see that our computational cost is in a reasonable range.

Dataset	CIFAR-10-LT	CIFAR-100-LT	TinyImageNet-LT
MTT [4]	10.0	40.8	50.2
DATM [13]	22.7	98.2	124.0
Ours	15.0	56.8	85.9

Table 9. **Computation cost comparison.**

9. Dataset Images

The distilled dataset is visualized in Figure 8. We visualized the synthetic dataset images of DATM and our method under three different imbalance factors, $\beta = 50$, $\beta = 100$, $\beta = 200$. We observe in the figure that: As the imbalance factor increases, the DATAM distilled image quality degrades and contains more noise and distortions. Instead, our distilled dataset can still maintain good quality.

Dataset	β	IPC	N_{rep}	N_{cls}	T_{rep}^{-}	T_{rep}	T_{rep}^{+}	T_{cls}^{-}	T_{cls}	T_{cls}^{+}	λ_{rep}	λ_{cls}
CIFAR-10-LT	10	10	80	80	0	10	20	0	1	1	1.0	0.2
		20	80	80	0	10	20	0	1	1	1.0	0.2
		50	80	80	0	20	40	0	1	1	1.0	0.2
	50	10	80	80	0	10	20	0	1	1	0.5	0.5
		20	80	80	0	10	20	0	1	1	0.5	0.5
		50	80	80	0	20	40	0	1	1	0.5	0.5
	100	10	80	80	0	10	20	0	1	1	1.0	0.5
		20	80	80	0	10	20	0	1	1	1.0	0.5
		50	80	80	0	20	40	0	1	1	1.0	0.5
	200	10	80	80	0	10	20	0	1	1	0.1	1.0
		20	80	80	0	10	20	0	1	1	0.1	1.0
		50	80	80	0	20	40	0	1	1	0.1	1.0
CIFAR-100-LT	10	10	40	20	0	30	50	0	1	1	1.0	0.1
		20	40	20	20	70	70	0	1	1	1.0	0.1
		50	40	20	20	70	70	0	1	1	1.0	0.1
	20	10	40	20	0	30	50	0	1	1	1.0	0.1
		20	40	20	20	70	70	0	1	1	1.0	0.1
		50	40	20	20	70	70	0	1	1	1.0	0.1

Table 10. **Hyper-parameters for different settings.**

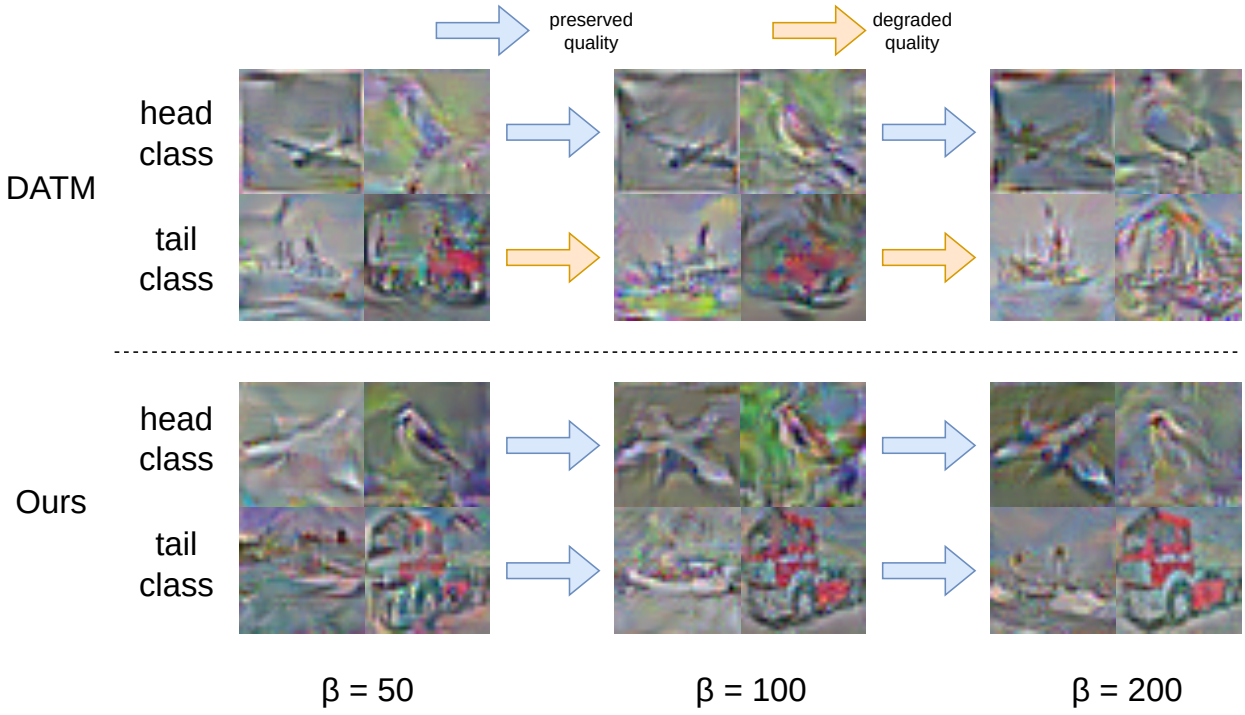

Figure 8. **Visualization of distilled datasets.** We visualize the images from the distilled dataset for DATM and our method. We can observe that as the imbalance factor increases, images from DATM preserve the quality on head classes, but degrade on tail classes. On the contrary, our method is able to preserve good quality in all classes.