REPRODUCIBILITY CHALLENGE: A LABEL IS WORTH A THOUSAND IMAGES IN DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset distillation aims to compress large training datasets into smaller ones, with the aim of training models with less data without sacrificing performance. In this project we reproduce key experiments from "A Label is Worth a Thousand Images" by Qin et al. (2025) and perform additional extensions to investigate their claim that soft, or probabilistic, labels are the primary driver of performance in many data distillation methods. Our results confirm that soft labels are indeed important for dataset distillation and that structured semantic information encoded in them appears to be the determining factor of whether a soft label is beneficial. All code for this work can be at: https://github.com/COMP6258-Reproducibility-Challenge/no-distillation-reproduction.

1 BACKGROUND

The massive amount of data required for training deep learning models is associated with large computation costs and storage requirements. This has led to an interest in methods that can reduce dataset size without sacrificing model performance. One popular approach is dataset distillation, which aims to distil a large dataset into a smaller, synthetic one, such that a model trained on the distilled data achieves a comparable performance to one trained on the full dataset (Yu et al., 2024).

In the context of image classification, dataset distillation methods aim to generate a smaller set of high quality, synthetic images which retain a sufficient amount of information from the full dataset. Recent work by Qin et al. (2025) suggests that the driving factor behind the performance of many state-of-the-art (SOTA) data distillation techniques is the use of soft labels. These are labels which distribute the probability mass over all classes in contrast to hard labels where membership of a class is binary. Furthermore, they claim that good quality soft label softmax distributions should contain structured information from semantically similar classes. To support these claims, the authors devise a simple baseline model to compare with other SOTA distillation methods.

The simple baseline model proposed in Qin et al. (2025) is made up of two components. First, an expert model is trained on the entire original training dataset, $\mathcal{D}_{orig}$, with hard labels. For each image in the training data the trained expert model is able to output softmax probabilities over all the classes (the soft labels). The distilled dataset, $\mathcal{D}_{dist}$, is a randomly selected subset of the training dataset (such that $|\mathcal{D}_{dist}| \ll |\mathcal{D}_{orig}|$), with their expert assigned soft labels. Finally, a student model is trained on the distilled dataset. We base our reproduction and extension on CIFAR-100, for which the authors use a standard ConvNet architecture with three convolutional blocks.

2 EXPERIMENTS

In this section we reproduce the key experiments presented in Qin et al. (2025) and present extensions on some of those experiments.

2.1 BENCHMARKING DISTILLATION METHODS AGAINST THE SOFT LABEL BASELINE

The first results presented compare the authors' soft label baseline method against SOTA distillation methods for varying data budgets, measured by images per class (IPC). Table 1 displays our

reproduction of these results alongside the original results from the paper. We follow the experimental setup discussed in Appendix A.5 of Qin et al. (2025), using the recommended hyperparameter choices and expert model epoch. We observe similar test accuracies for the soft label baseline. The results suggest that they are correct in their assertion that the soft label baseline is competitive with SOTA distillation methods despite its simplicity, which supports their claim that soft labels may be the primary reason for the performance of data distillation methods.

Distill Method	Ra-BPTT	MTT	DM	SL Baseline	SL Baseline (Original)
IPC=1	35.3 (0.4)	24.3 (0.3)	11.4 (0.3)	15.0 (0.2)	16.0 (0.3)
10	47.5 (0.2)	40.1 (0.4)	29.7 (0.3)	34.9 (0.5)	34.3 (0.2)
50	50.6 (0.2)	47.7 (0.3)	43.6 (0.4)	46.9 (0.1)	47.1 (0.4)
100	–	–	–	48.7 (0.2)	51.3 (0.5)
Full	ConvNet(F) = 53.3%				
Full (Original)	ConvNet(F) = 56.4%				

Table 1: Reproduction of Table 2 (Qin et al., 2025) for the CIFAR-100 dataset. Student test accuracies averaged over 5 evaluations. Standard deviations reported in parentheses.

2.2 Label Entropy

Next, the authors present experimental results showing the relationship between student accuracy, expert accuracy, and soft label entropy. As expert accuracy increases, the entropy of the soft labels decreases due to the model becoming more certain in its predictions, resulting in the probability mass of the soft labels being concentrated in only a few classes. Because of this the authors argue that there is an optimal expert accuracy (and therefore epoch) with which to train the student model which balances confidence in predictions with the structural information conveyed by the soft labels.

Figure 1 shows the results from our version of the experiment. We observe a pattern similar to that reported by the authors in their experiments on ImageNet-1K. Across different data budgets, student accuracy peaks and then decreases as expert accuracy increases further and soft label entropy decreases. The divergence in test accuracy is particularly clear in the right figure, where the student was trained with IPC = 100. This supports the authors' claim that some amount of uncertainty in the soft labels is required for effective student training. However, it is yet unclear whether this is due to structural information in the soft labels or some regularising effect.

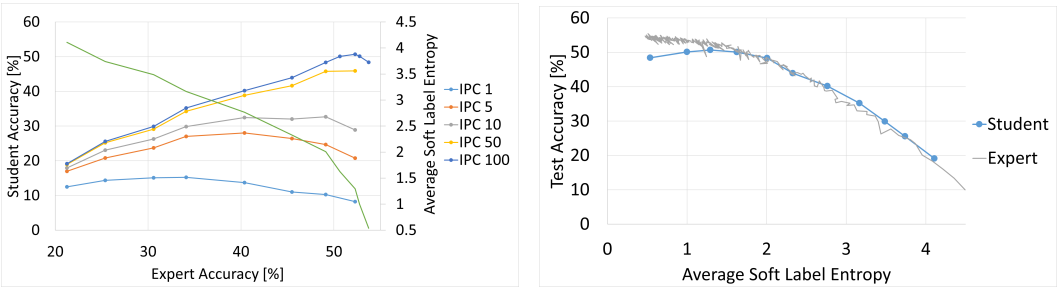


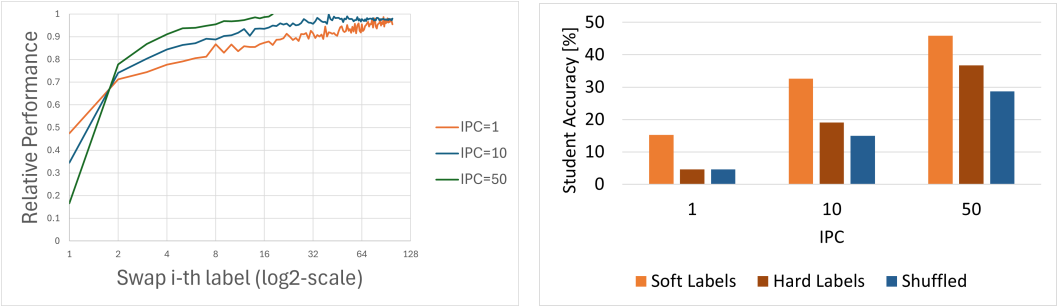
Figure 1: The patterns in the plot resemble those from Figure 2 of the original paper. Good student accuracy relies on a balance between expert accuracy and soft label entropy.

2.3 Structured Information

To show the presence and importance of structured information in the soft labels, the authors experiment with swapping the i-th top label, sorted by softmax values, with the last label. This is based on the assumption that higher softmax value labels contain more useful information than lower ones. By swapping each label with the last label we can see how the relative performance (the swapped student accuracy divided by the original student accuracy) changes, showing which labels contain

the most useful information. As entropy remains constant when labels are swapped, we hope to isolate the impact of structural information from any regularisation effect. Figure 2a shows our results for this experiment. The student model's performance suffers when labels with high softmax values are swapped, with the largest effect seen for the highest values, supporting the authors' claim that top labels contain important information for training the student model.

Figure 2b shows an extension to the i-th label swap experiment, where rather than swapping the i-th top label with the last label, the top label is kept fixed and the remaining 99 labels are randomly shuffled. The experiment was repeated five times per IPC value. The average student accuracy is presented alongside accuracies for students trained on the expert soft labels and hard labels. Students trained on shuffled labels performed worse than both those trained on soft and on hard labels. This further supports the claim made by the authors that the structured information in the soft labels is important. The entropy between the soft labels and the shuffled labels is kept the same, but the structured information is lost, resulting in a reduced performance.



(a) (b)

Figure 2: Results for the i-th label swap experiment (left) and random shuffle experiment (right).

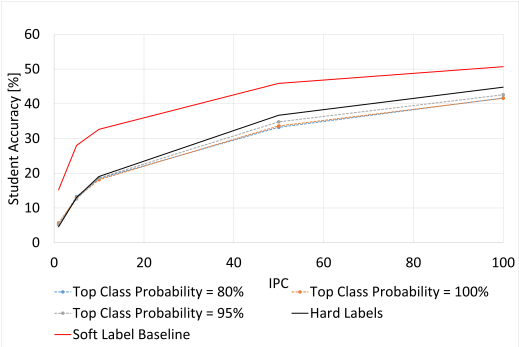
Inspired by Müller et al. (2020), we devised another experiment investigating the structured information of soft labels. The hard labels were assigned 80%, 90%, or 95% of the probability mass, with the remaining 99 classes receiving the remaining probability mass distributed among them randomly. The results for models trained with generated labels, hard labels, and expert soft labels are shown in Figure 3a. The graph shows that the models trained on the generated labels performed similarly or slightly worse than the model trained on hard labels. These results further support the claim that it is the structure of the soft labels that is important.

2.4 Optimal Expert Epoch

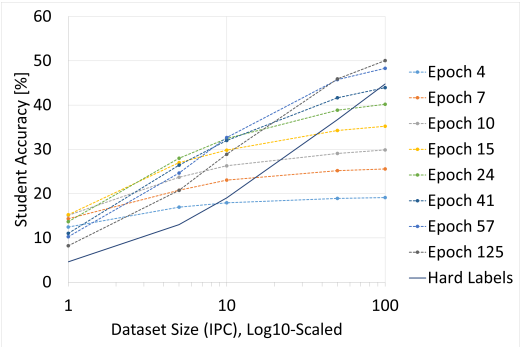
The authors propose a Pareto-optimal front for data-efficient learning by selecting the expert epochs which yield the best student performance for a certain value of IPC. The same experiment was reproduced and its results can be seen in Figure 3b. It shows that for smaller IPC values, an earlier stopped expert will result in better student performance.

2.5 Increasing the Uncertainty of Low Entropy Soft Labels

Building on the authors' experiments we hypothesised that low entropy soft labels across individual training images in the distilled dataset may be harmful for the student, even when the optimal expert epoch was chosen. To test this we identify the subset of $\mathcal{D}_{dist}$ with the lowest entropy and mix normalised uniform noise into them, artificially increasing their entropy while ensuring it stays a valid probability distribution. The results in Figure 4 do not support the hypothesis that introducing uncertainty into data, that the expert is most confident in, benefits the student model. However, we note that IPC = 1 is fairly robust to the noise, likely due to the lack of information stored in the soft labels obtained from such an early stopped expert (Qin et al., 2025). Additionally, while the student accuracy drops for IPC = 10 we do not observe a particularly large decrease in performance, even as the amount data altered is increased. This supports the idea that higher entropy labels are most important for student learning, but increasing the entropy of low entropy labels artificially does not benefit performance due to the lack of any structural information.

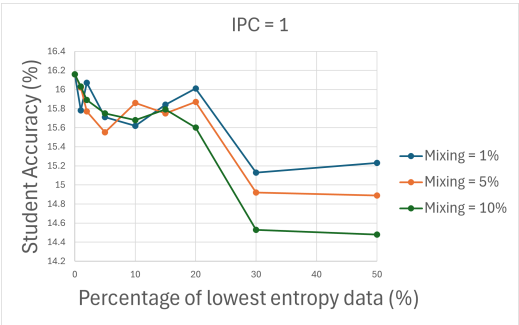


(a) Experiment with student models trained with generated labels.



(b) Test accuracy of students trained on early-stopped experts, showing support for a Pareto-optimal front

Figure 3



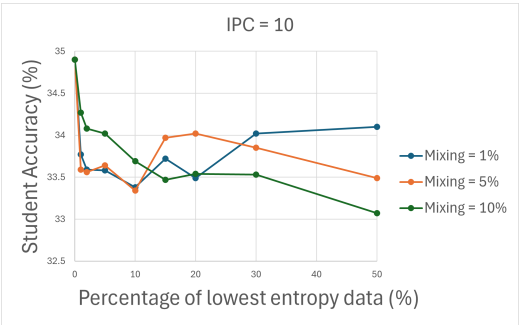


Figure 4: Student performance degrades when noise is mixed into subsets of the distilled data with the lowest entropy soft labels.

3 Conclusion

The results shown in Qin et al. (2025) were successfully reproduced for the CIFAR-100 dataset. Our results were comparable to the findings of the original paper, leading us to a similar conclusion regarding the importance of soft labels and in particular the structure of the soft labels. Additional experiments were carried out to further test the importance of soft labels, the structure of the soft labels, and the impact of entropy. These experiments support the authors' claim that the soft label structure created by an expert aids student training. Further, they show that soft labels without such a structure are no better than hard labels, and that high entropy without structure is not beneficial to the student.

References

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help?, 2020. URL https://arxiv.org/abs/1906.02629.

Tian Qin, Zhiwei Deng, and David Alvarez-Melis. A label is worth a thousand images in dataset distillation, 2025. URL https://arxiv.org/abs/2406.10485.

Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset Distillation: A Comprehensive Review . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(01):150–170, January 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3323376. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2023.3323376.