Review of: "Enhancing Dataset Distillation via Label Inconsistency Elimination and Learning Pattern Refinement"

Alberto Fernández[1]

1 Universidad de Granada, Spain

Potential competing interests: No potential competing interests to declare.

This paper presents a solution for the ECCV-2024 Data Distillation Challenge, which aims to create smaller synthetic datasets that enable machine learning models to perform comparably to models trained on the full datasets. The authors propose a modified version of the Difficulty-Aligned Trajectory Matching (DATM) method, named M-DATM, designed to address issues related to label inconsistency and the difficulty of learning hard patterns, especially on challenging datasets like Tiny ImageNet. M-DATM achieves high performance by removing soft labels to ensure label consistency and adjusting the trajectory matching range to focus on simpler patterns. Experimental results demonstrate that M-DATM achieves superior results, ranking first in the ECCV-2024 challenge on the CIFAR-100 and Tiny ImageNet datasets.

The research community has a growing interest in dataset distillation, which condenses large datasets into smaller synthetic ones without sacrificing model performance. This work addresses two key challenges in dataset distillation—label consistency and optimal learning difficulty—that have hindered the effectiveness of existing approaches. The improvements in M-DATM could provide a robust baseline for future dataset distillation work, making it particularly relevant for researchers in model efficiency and scalable AI solutions.

The work is fine, and the results are competitive. However, there are some issues that could be addressed prior to publication:

1. In the introduction, the research objectives could be framed more clearly; the introduction implicitly discusses goals but does not explicitly state them as research questions or hypotheses.

2. The paper provides a good overview of relevant methods in dataset distillation, particularly around gradient matching and trajectory matching, setting up DATM as a state-of-the-art approach. However, the background on challenges specific to label inconsistency and trajectory matching could be expanded to provide a stronger basis for the proposed modifications.

3. Despite the detailed descriptions of the proposal, a structured pseudo-code or algorithm summary is missing, which would further aid reproducibility. Additionally, the paper does not discuss hyperparameter tuning in detail.

4. The removal of soft labels to eliminate inconsistency is a novel solution to the issue of label misalignment in dataset distillation, and the paper effectively highlights the need for this approach. In order to support the actual novelty, authors may further differentiate M-DATM from closely related techniques, like other trajectory matching or gradient-based

methods.

5. The experimental framework focuses exclusively on image datasets; testing on additional data types could strengthen claims of generalizability. Moreover, using additional metrics (e.g., F1 score, recall) would add robustness to the analysis.

6. Results show consistent improvements over DATM, whereas the ablation study effectively illustrates the value of each modification in M-DATM. In spite of the former, further statistical validation of the results (e.g., significance testing) would add confidence in the findings, particularly when comparing performance across matching ranges.

7. The discussion on why focusing on early trajectory information (simpler patterns) improves synthetic dataset quality could be expanded to include a theoretical perspective.

8. Finally, please add specific directions for future research, such as exploring adaptive matching ranges based on dataset characteristics or expanding M-DATM to multi-modal datasets.