Dataset Distillation via Difficulty Aligned Sequential Matching

Firstname1 Lastname1 [*1] **Firstname2 Lastname2** [*1 2] **Firstname3 Lastname3** [2] **Firstname4 Lastname4** [3] **Firstname5 Lastname5** [1] **Firstname6 Lastname6** [3 1 2] **Firstname7 Lastname7** [2] **Firstname8 Lastname8** [3] **Firstname8 Lastname8** [1 2]

Abstract

Dataset distillation has emerged as a prominent area of research, aiming to compress large, original datasets into compact, lightweight synthetic datasets. Currently, three main methodologies address this problem; however, they all face a critical limitation. While synthetic images perform well under low image-per-class (IPC) settings, their performance quickly plateaus at higher IPC settings and, in some cases, even falls behind that of coreset selection methods. Our analysis reveals that this phenomenon is primarily due to a lack of diversity in the generated images. This redundancy degrades performance, particularly in high IPC settings. To address this fundamental issue, we propose DASM (Dataset Distillation via **D**ifficultly **A**ligned **S**equential **M**atching), a novel distillation framework that can be seamlessly integrated into existing approaches. DASM facilitates the generation of synthetic images that retain high performance across a wide range of IPC settings, effectively mitigating saturation. When incorporated with existing methods, DASM achieves SOTA performance in high IPC settings, even surpassing the performance of networks trained on the full dataset. Moreover, our method shows a very strong capability to generalize to unseen architectures, paving the way to make dataset distillation significantly more practical.

1. Introduction

In the current era of data-driven and power-hungry AI, dataset distillation provides a powerful method of dataset compression for light-weight storage and efficient training. Dataset distillation is the task of generating a small number of synthetic images that preserve the ability to train models comparably to the original real dataset. The synthetic images can then be used instead of the full dataset for various tasks such as continual learning, privacy preservation, or neural architecture search.

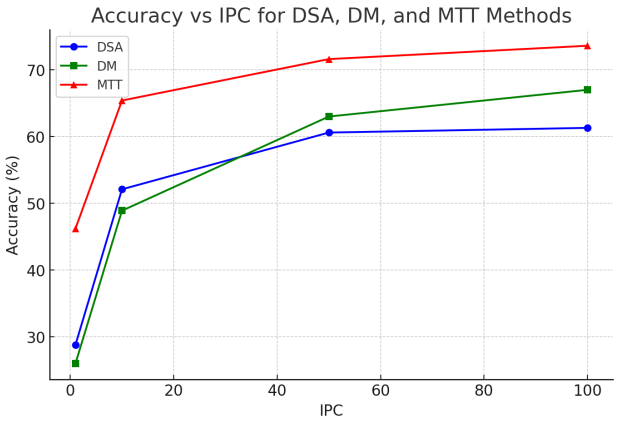

Figure 1. Performance of Dataset Distillation Under Various IPC Settings

The general formulation of dataset distillation treats the pixel values of synthetic images as parameters for update. The synthetic images are optimized to align attributes of real images such as feature maps, gradients, or training-trajectories. This procedure can be seen as embedding the characteristics of the real dataset into the synthetic images. These methods allow real datasets to be compressed to datasets as small as one synthetic image-per-class (IPC). For example, the MNIST dataset can be distilled to just ten synthetic images while showing 91.7% test accuracy. For larger datasets, low IPC synthetic datasets fails to fully capture the training ability of real datasets. Hence, higher IPC synthetic datasets are essential. Despite this, existing dataset distillation methods fails to scale effectively to high IPC settings.

As shown in figure 1, performance plateaus for higher IPC

[*]Equal contribution [1]Department of XXX, University of YYY, Location, Country [2]Company Name, Location, Country [3]School of ZZZ, Institute of WWW, Location, Country. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>, Firstname2 Lastname2 <first2.last2@www.uk>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

settings. Our analysis reveals that current approaches distill synthetic images that reflect mainly the representative features of the real dataset. Specifically, the alignment of synthetic images is typically skewed toward low-loss (easy) samples, with insufficient emphasis on high-loss (difficult) samples. This phenomenon is also stated in concurrent works, addressing the difficulty of embedding features from difficult samples. A naive approach can be filtering out the easy samples from the real dataset prior to distillation, using metrics such as forgetting scores or loss values from a network trained on the real dataset. However, this approach resulted in a big performance degradation. This lead to the conclusion that both easy and difficult features are essential for providing quality synthetic datasets.

In this paper, we propose a novel framework Difficulty-Aligned Sequential Matching (DASM), that sequentially uses easy to hard samples to distill synthetic images . The core idea is we divide the synthetic dataset into segments to generate synthetic images sequentially. Each segment is distilled using different parts of the real dataset, dynamically adjusted based on a model trained on all previously distilled segments. Our method, when integrated into existing approaches, demonstrates the ability to capture diverse semantics, encompassing both easy and hard features. Experiments on widely used benchmarks demonstrates a substantial performance improvement, particularly in high-IPC regions and large scale datasets. Also, our method shows great robustness across multiple unseen architectures, marking a significant step toward the broader practicality of dataset distillation. We summarize our contributions as follows:

- We identify and analyze why existing dataset distillation methodologies fail to scale to high IPC settings, attributing this limitation to the easy-sample bias inherent in current approaches.
- To address this challenge, we introduce DASM (Dataset Distillation via **D**ifficulty-**A**ligned **S**equential **M**atching), a novel framework that facilitates the generation of synthetic images capturing both easy and hard features.
- We demonstrate the effectiveness of our method through extensive experiments, which show that our method can be incorporated into existing works boosting performance and cross-architecture generalization.

2. Related Works

Dataset distillation compresses real images into synthetic images by optimizing the distance between the features of real and synthetic images. Current research primarily focuses on two strategies: gradient/trajectory matching and distribution matching.

Gradient/Trajectory Matching: The core principle of these methods is that synthetic images should produce back-propagation signals similar to those of real images. DC emphasizes matching one-step gradients, while MTT advances this concept by matching entire training trajectories. Although trajectory matching methods are often regarded as the current SOTA framework for dataset distillation, our experiments reveal that gradient matching methods outperform them on large-scale datasets and high-IPC settings. Furthermore, gradient matching methods demonstrate superior generalization across multiple architectures, which is why we mainly implement our method to gradient matching frameworks.

Distribution Matching: Another approach focuses on aligning the feature maps of real and synthetic data. Early methods such as CAFE and DM adopted this strategy but exhibited inferior performance compared to gradient or trajectory matching frameworks. More recent works propose a novel take on distribution matching, generating synthetic images by aligning the batch normalization statistics of real images. However, these methods are highly memory-intensive, requiring approximately 40 times more memory to store soft labels than is needed for synthetic images. Moreover, synthetic images generated using this approach often perform worse than randomly selected images. Based on these findings, we conclude that this line of work is not well-suited for dataset distillation. Consequently, our focus remains on traditional gradient/trajectory matching methods.

3. Preliminary

Dataset distillation is a task of generating a batch of synthetic images $\mathcal{S}$ such that networks trained with $\mathcal{S}$ perform similar to networks that are trained with the original training set $\mathcal{R}$. However, directly optimizing with respect to the generalization performance is very expensive, so surrogate objectives are solved instead to minimize distance between the real and synthetic dataset.

$$R : \text{Real Dataset} \tag{1}$$

$$S : \text{Synthetic Dataset} \tag{2}$$

$$S^* = \arg\min_S \mathcal{D}(S, R) \tag{3}$$

The distance between the datasets can be measured in term of feature maps, gradients, or training trajectories.

The DASM framework can be implemented into all three surrogate objective methods, but as stated earlier, we work on gradient matching methods due to its superior performance and cross architecture generalization. In particular, we build upon IDC, where gradient matching is performed with downsampled synthetic images.

$$S^* = \arg\min_S \mathcal{D}\left(\nabla_\theta \mathcal{L}(\theta(\mathcal{A}^*(S))), \nabla_\theta \mathcal{L}(\theta(\mathcal{A}(R)))\right) \quad (4)$$

$\mathcal{A}$ represents differential siamese augmentation, and $\mathcal{A}^*$ is upsampling followed by $\mathcal{A}$.

4. Methodology

References

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.