Dataset Condensation: Why Learn What You Already Know

Yun Kim [1] **Banseok Kim** [1]

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 ADS (Dataset Distillation via **A**ppropriate **D**ifficulty **S**upervision), a novel distillation framework that can be seamlessly integrated into existing approaches. ADS 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, ADS 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

Dataset distillation is the task of synthesizing a small dataset while retaining the informational richness of the original dataset. The original paper frames this as a meta-learning problem, where synthetic images are updated to minimize the loss on real data when a network is trained using the synthetic dataset. However, this approach is computationally impractical, as it requires an intractable amount of computation to back-propagate through the network's optimization graph. To address this challenge, surrogate objectives have been proposed. These surrogates embed information from the real data and update the synthetic images to align with the embedded representations. The embedding process can leverage feature maps, gradients, or training trajectories. These methods have demonstrated significant efficiency, achieving strong performance even without directly optimizing the meta-learning objective.

Despite these advances, dataset distillation fails to scale effectively to high image-per-class (IPC) settings. As shown in the graph above, performance plateaus after approximately 50 IPC, indicating that current methodologies do not scale to higher IPC levels. Further investigation reveals that the generated images contain redundant information, leading to repetitive representations. Ironically, although the goal of dataset distillation is to eliminate redundancies in the original dataset, the synthetic images themselves exhibit redundancy, causing saturation in high-IPC regions.

This lack of diversity significantly hampers the network's ability to generalize to unseen data, particularly in high-IPC settings, as illustrated in the graph. From these observations, we hypothesize that certain clusters of images are easily embedded into synthetic datasets, while other features of the original dataset are inherently difficult to learn. This issue has been noted in prior works, which have either exploited this property or attempted to mitigate it. Our approach focuses on the latter, while still leveraging the "easy learning" property during early training.

The core idea behind our methodology is that redundant features should not be generated, and hard-to-learn information from the original dataset must be embedded into the synthetic images. However, achieving this is challenging because synthetic images are typically generated in a single large batch, making it impossible to explicitly control which parts of the dataset are learned by individual images. To address this, we propose generating synthetic images in segments, enabling much greater control over the information embedded in each image.

But how do we decide which information to embed and which to exclude? Initially, we considered examining the feature space of generated images and grouping real images

[1]Yonsei University. Correspondence to: Yun Kim <yunkimmy@yonsei.ac.kr>, Banseok Kim <first2.last2@www.uk>.

Accepted to the 2^{nd} *Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Vienna, Austria. 2024.

within a threshold distance of the synthetic images. While feasible, this approach would require numerous design decisions and extensive hyperparameter tuning. Instead, we devised a simpler and more intuitive solution: training a network with synthetic images and using the trained network as a discriminator on the real dataset to determine whether an image's information is embedded in the synthetic set. This approach allows us to train different batches of synthetic images with different subsets of the real dataset, ensuring that diverse and rich features of the original dataset are captured without redundancy.

Our method, implemented on IDC, demonstrates a substantial performance improvement, particularly in high-IPC regions, where it even surpasses the performance of networks trained on the full original dataset. Even more notably, our method shows great robustness across multiple unseen architectures, marking a significant step toward the broader practicality of dataset distillation.

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 backpropagation 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 our work is based on IDC.

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 ADS 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.

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

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

4. Methodology

A. You *can* have an appendix here.