UCF CRCV REU

Hannah Carey

Project: DistillScope: Analyzing the Scope and Limits of Diffusion-Based Dataset Distillation with Kunyang Li and Dr. Shah

What is Dataset Distillation?

Problem:

As AI advances, datasets and the amount of energy needed to train models grows massive.

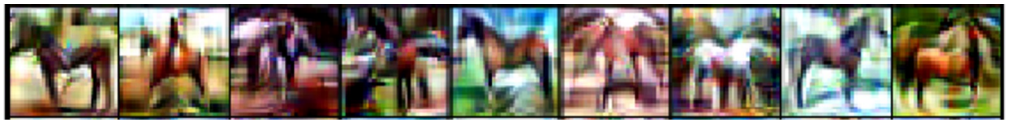
Goal:

Synthesize a dataset **much smaller** than the original real data with **similar performance.**



There are 5,000 images per class in CIFAR-10 dataset.

Distillation

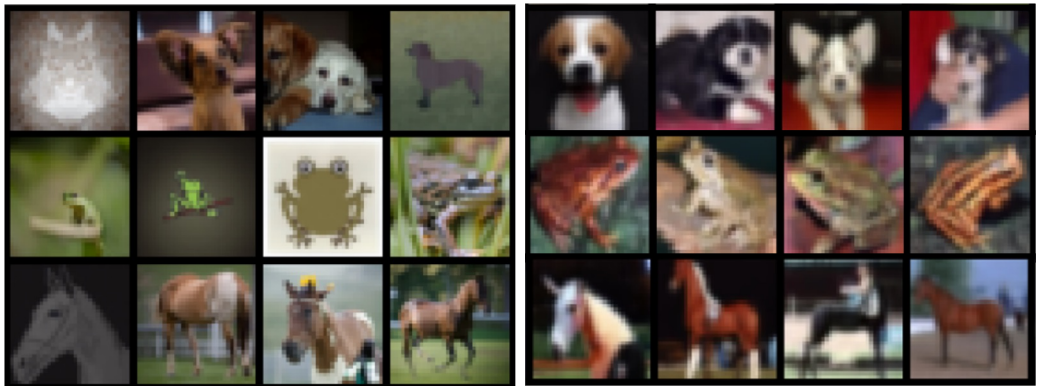


An alternate class of only 10 images; 0.2% the size of the original!

Optimization-based and Diffusion-based distillation

“What is Dataset Distillation Learning?” [2] focuses on theoretical analysis of traditional optimization-based distillation methods.

Our motivation: investigate for a similar theoretical understanding and comparison of diffusion-based distillation methods.



Key Takeaways

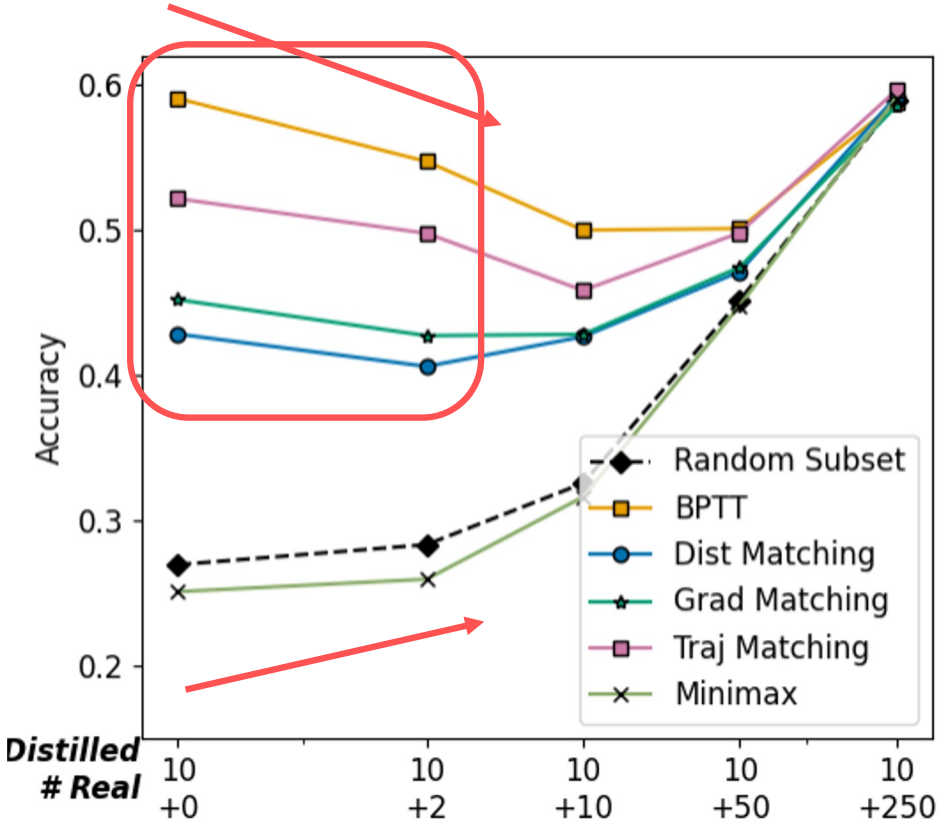
Diffusion-based distilled data:

...exhibits a **high compatibility with real data** when mixed in training.

...contains information about the **later stages of training**.

...provides **continuous learning signals** at the **same rate as real data**.

Mixing Real and Distilled Data



The performance of optimization-based data initially *suffers* when mixed with real data.

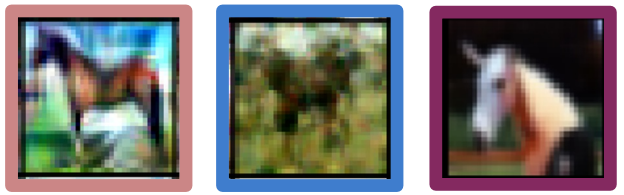
Diffusion-based data **mixes naturally** with real data, displaying a unique compatibility with real data.

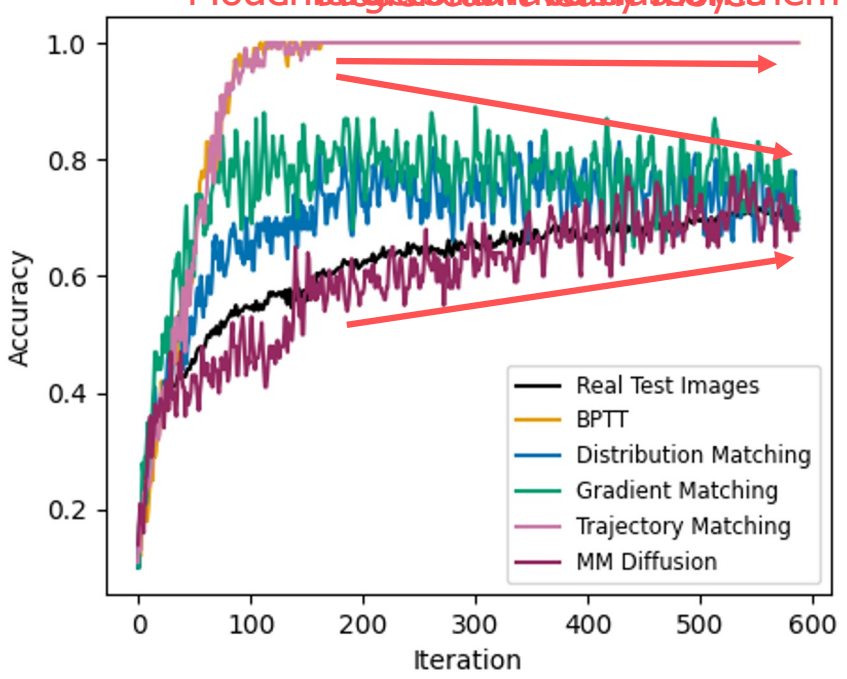
Data Recognition

As a model learns more, optimization-based data

1. **overgeneralizes**
2. **performs worse**

It also learns how to classify the diffusion-based data **better over time**, similar to the real data.





Thank you!