CRCV REU Week 8

Hannah Carey

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

Dataset Distillation

Goal: Derive a small dataset from a much larger one.

Traditional focus: Optimization-based approaches
(gradient, trajectory , distribution matching).

Key insights (from the paper “What is Dataset Distillation Learning?” [1]**):**

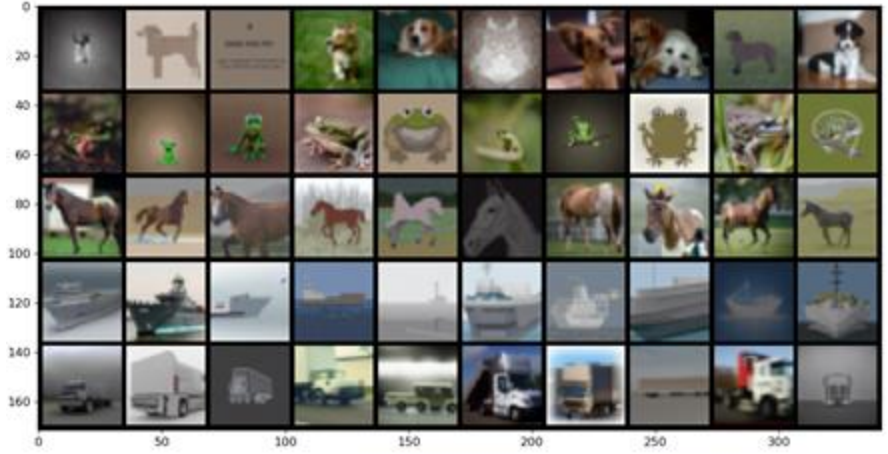
- Distilled data ≠ Real data
- Compress early training dynamics
- Individual samples contain semantic meaning



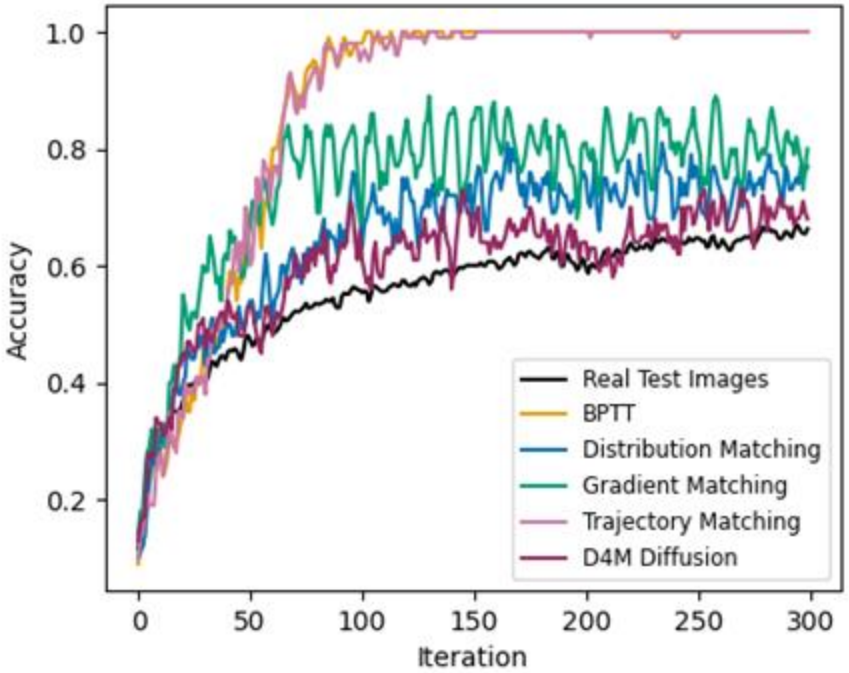
BPTT distilled

D4M Diffusion for CIFAR-10

D4M-generated distilled dataset evaluation, issue of domain gap



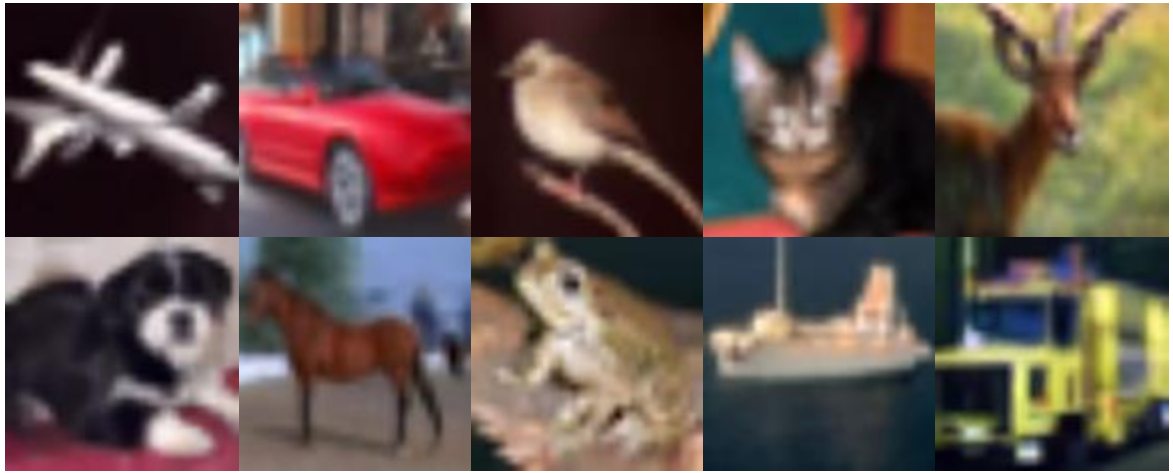




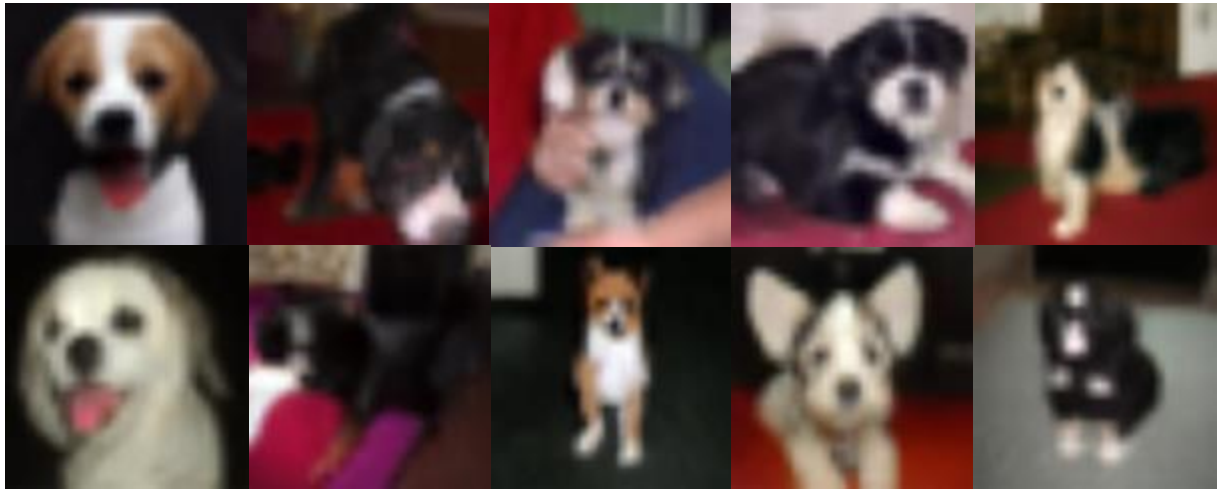
Minimax Diffusion for CIFAR-10

Generated 10 IPC dataset using Minimax Diffusion for CIFAR-10

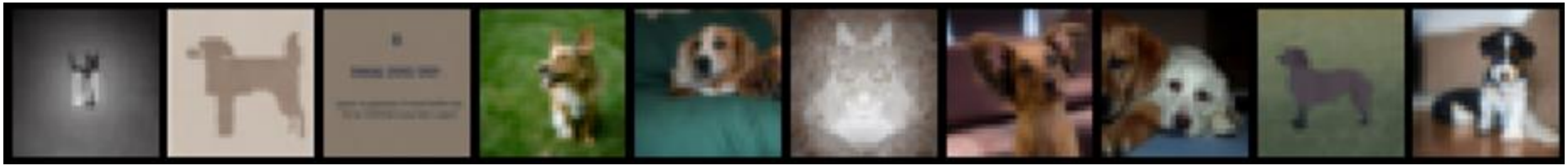
First appearances: more representative, diversity uncertain



Minimax Diffusion for CIFAR-10: dog label



D4M distilled data: dog label



Next steps

Finish all evaluation metrics for CIFAR-10 diffusion distilled datasets - using Minimax and Influence-Guided. (early next week)

Working on poster: aim to communicate findings on lower resolution dataset

Future work in higher resolution (later next week).