CRCV REU Week 7

Hannah Carey

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

Plan Overview

1. Reproduce paper results “What is Dataset Distillation?” on optimization-based distillation methods
2. Expand to diffusion based CIFAR 10 images
3. Generating diffusion distilled dataset based on Imagenet subset (higher res)

Goals:

Investigation of distillation methods on higher-res images

Adaptation of previous analytical methods to these higher-res images

Propose new analytical methods specific to diffusion-based distillation

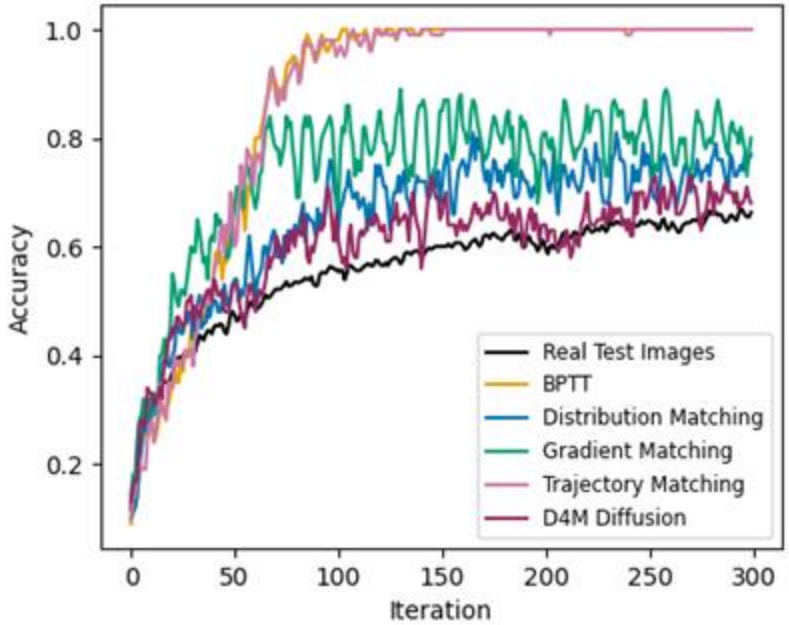
This week

Working on generating datasets

CIFAR-10:

- Optimization:
  - NCFM ⇒ Issues with codebase (abandoned ??)
  - Distribution matching
- Diffusion:
  - D4M ⇒ Domain gap issue
  - Minimax Diffusion ⇒ adapting code for generating less classes

Recognition analysis



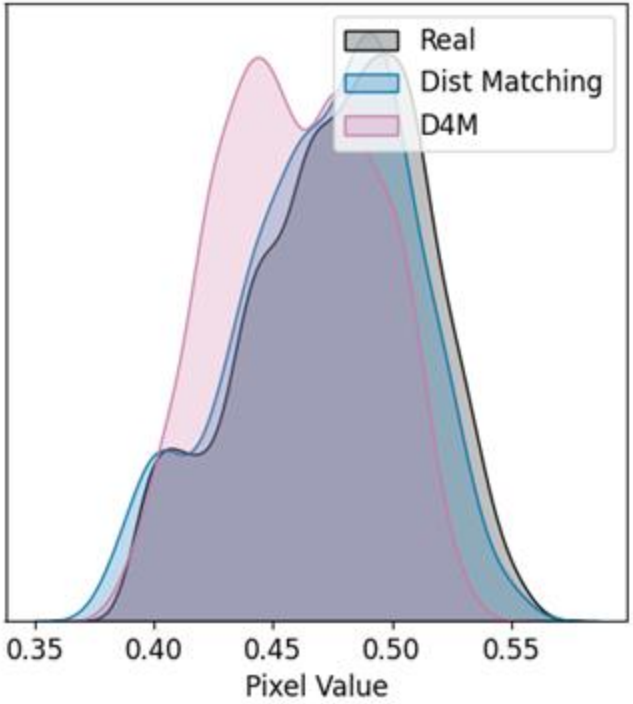
- Even D4M obtains a higher accuracy earlier on
- Despite domain gap and other issues, this hints that D4M is capturing some earlier training dynamics
- Similar to Distribution Matching

Pixel intensities

Original figure from the paper displayed pixel intensities of a single image

Adapted to represent average values across all images in dataset

Continues to highlight issue of domain gap for D4M dataset



Next steps

Cifar:

- Finish changing code for Minmax Diffusion
- Generate dataset
- Re-run metrics with new MM data instead of D4M

ImageNet:

- Currently have 2 distilled datasets with 50 IPC
- Need to upscale methods to handle larger images