Accelerating Dataset Distillation via Model Augmentation

Zhang, L., Zhang, J., Lei, B., Mukherjee, S., Pan, X., Zhao, B., Ding, C., Li, Y. and Xu, D.

CVPR, 2023 (Highlight)

Background: data distillation

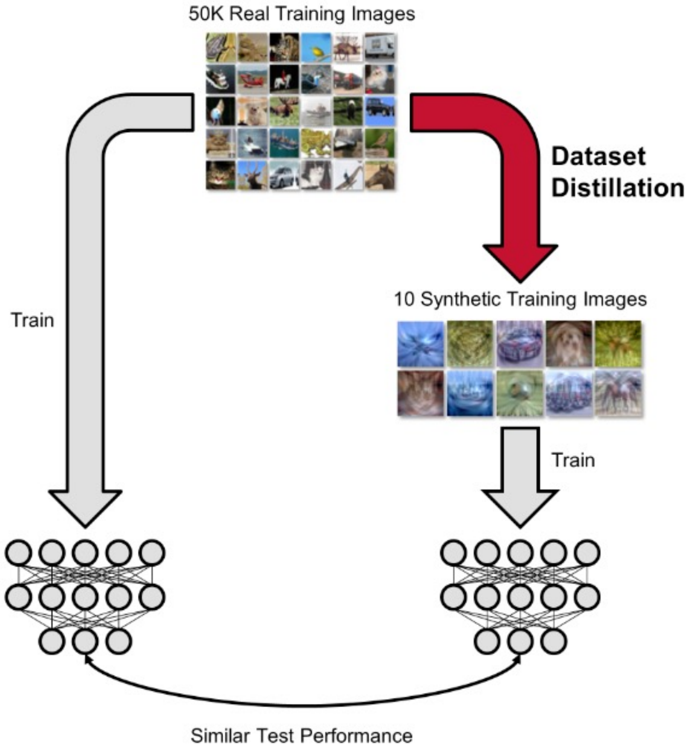
- Deep learning rely on large dataset
  - Lots of computational resources
  - Time-consuming training process

- Data distillation (DD) or data condensation [1]
  - Generate a small but informative synthetic data

- Matches the network gradients on
  - synthetic dataset & real dataset

$$\underset{\mathcal{S}}{\text{maximize}} \sum_{t=0}^{\mathcal{T}} \operatorname{Cos}\left(\nabla_\theta \ell\left(\theta_t ; \mathcal{S}\right), \nabla_\theta \ell\left(\theta_t ; \mathcal{T}\right)\right)$$

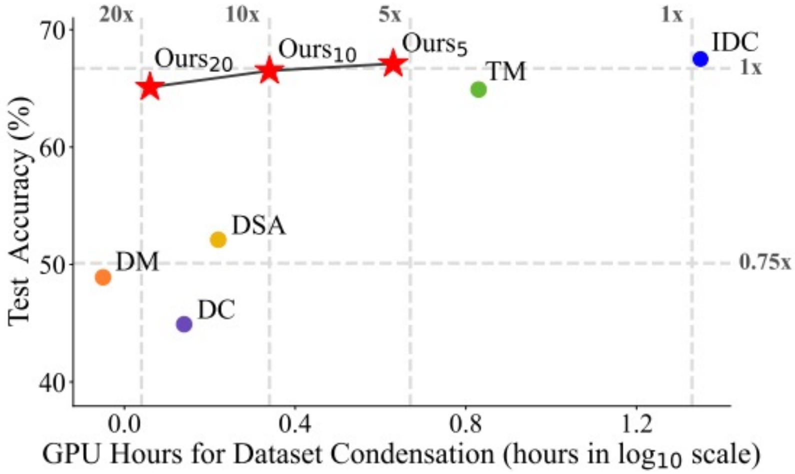
$$w.r.t. \quad \theta_{t+1}=\theta_t-\eta \nabla_\theta \ell\left(\theta_t ; \mathcal{S}\right)$$



[1] https://georgecazenavette.github.io/mtt-distillation/

Background: limitation in data distillation

- Data distillation process is expensive
  - Although model training on a small synthetic data is fast
  - SOTA method (IDC) takes 30 hours to condense 50,000 CIFAR-10 images to 500 synthetic images on one RTX-2080 GPU.
  - That equals to train 60 ConvNet-3 models on the original dataset.
  - The cost will rapidly increase for large-scale datasets e.g. ImageNet-1K.



Background: why DD is expensive

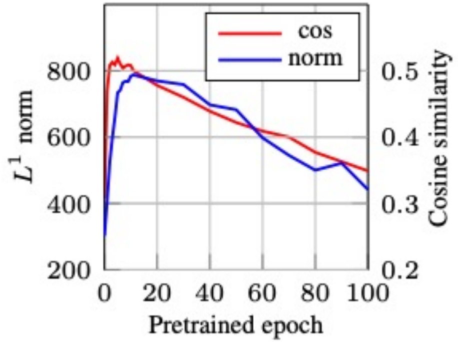
- They focus on generalizability
  - They requires optimizing the synthetic set over thousands of differently initialized network.
  - IDC: requires 2000 randomly initialized models
  - TM: requires 200 pre-trained expert models
  - Intuition: training the synthetic data with diverse models leads to better generalization performance
- Question 1:
  - How to design the candidate pool of models to learn from synthetic data?
- Question 2:
  - Can we learn a good synthetic set using only a few models?

Overview

- Question 1:
  - How to design the candidate pool of models to learn from synthetic data?
- Answer 1:
  - Early-stage models are more efficient for gradient matching based dataset condensation methods
- Question 2:
  - Can we learn a good synthetic set using only a few models?
- Answer 2:
  - Yes! (weight perturbation on selected early-stage models)

Method: early-stage models

- Gradient guidance from randomly initialized networks:
    - Insufficient and unstable
    - Requires many epochs or a large number of models
- Well-trained models have small gradients [1].



- Model augmentation
    - Utilize pre-trained information
    - Ensemble is helpful: pre-train a small set of models
    - Early stage models have large gradients → alleviating gradient vanishing challenge

[1] Dataset Condensation via Efficient Synthetic-Data Parameterization, ICML 2022.

Method: weight perturbation

- Data augmentation: perturbing training data to induce diversity

- Weight perturbation to perturb early-stage network weights
  - Diversify the feature space

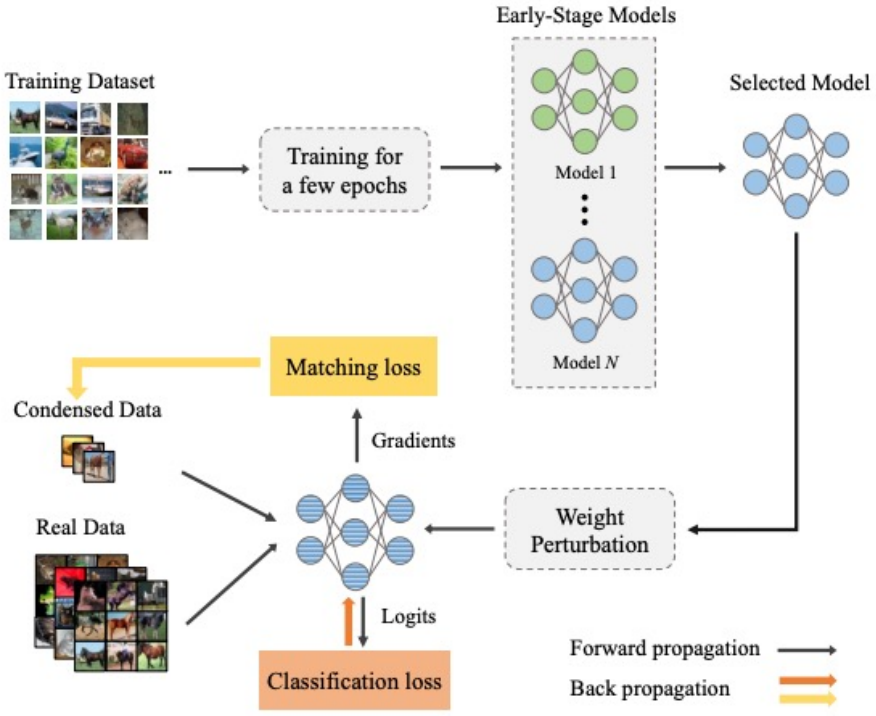
$$\min_{\mathcal{S}} D\left(\nabla_\theta \ell\left(\hat{\theta};\mathcal{S}\right), \nabla_\theta \ell\left(\hat{\theta};\mathcal{T}\right)\right)$$

$$w.r.t.\ \hat{\theta} \rightarrow \theta^{\mathcal{T}} + \alpha \times \mathbf{d},$$

$$\mathbf{d}_{l,j} \leftarrow \frac{\mathbf{d}_{l,j}}{\|\mathbf{d}_{l,j}\|_F} \|\mathbf{w}_{l,j}\|_F$$

Method

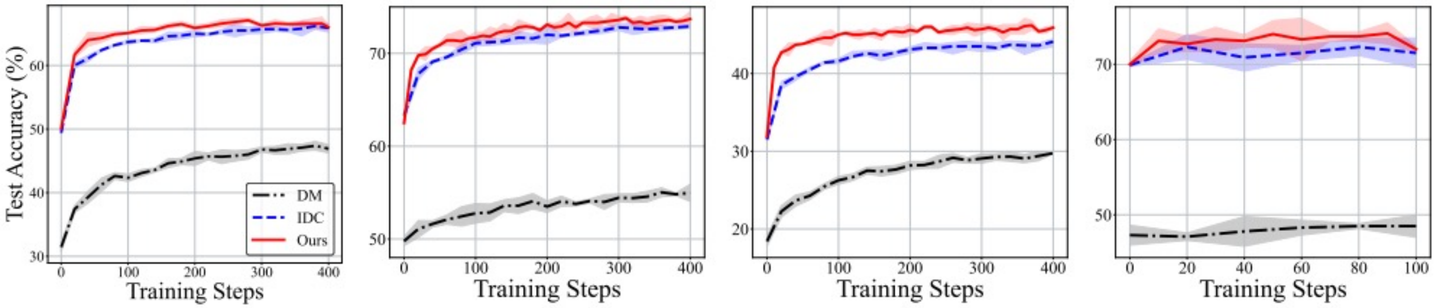
- Early-stage models + weight perturbation



Results

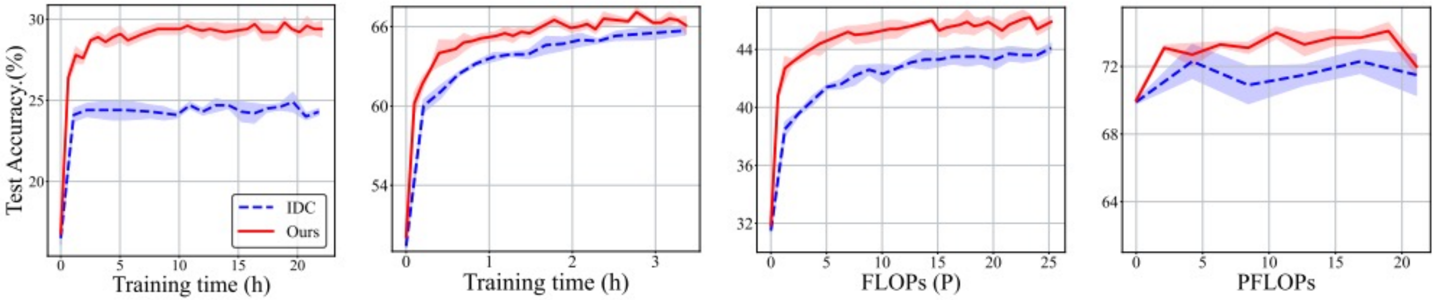
Dataset	Method	Img/Cls 1	Img/Cls 10	Img/Cls 50	Speed Up	Acc. Gain
CIFAR-10	Full Dataset	88.1	88.1	88.1	-	-
	IDC [27]	50.6 (21.7h)	67.5 (22.2h)	74.5 (29.4h)	1.00×	1.00×
	CAFE [56]	30.3	46.3	55.5	-	0.54×
	DSA [62]	28.2 (0.09h)	52.1 (1.94h)	60.6 (11.1h)	85.0×	0.71×
	DM [63]	26.0 (0.25h)	48.9 (0.26h)	63.0 (0.31h)	89.0×	0.69×
	TM [4]	46.3 (6.35h)	65.3 (6.69h)	71.6 (7.39h)	3.57×	0.94×
	Ours_5	49.2 (4.44h)	67.1 (4.45h)	73.8 (6.11h)	**4.90×**	**0.99×**
	Ours_{10}	48.5 (2.22h)	66.5 (2.23h)	73.1 (3.05h)	**9.77×**	**0.97×**
CIFAR-100	Full Dataset	56.2	56.2	56.2	-	-
	IDC [27]	25.1 (125h)	45.1 (127h)	-	1.00×	1.00×
	CAFE [56]	12.9	27.8	37.9	-	0.56×
	DSA [62]	13.9 (0.83h)	32.3 (17.5h)	42.8 (221.1h)	78.9×	0.63×
	DM [63]	11.4 (1.67h)	29.7 (2.64h)	43.6 (2.78h)	61.4×	0.55×
	TM [4]	24.3 (7.74h)	40.1 (9.47h)	47.7 (-)	14.7×	0.92×
	Ours_5	29.8 (25.1h)	45.6 (25.6h)	52.6 (42.00h)	**4.97×**	**1.10×**
	Ours_{10}	29.4 (12.5h)	45.2 (12.8h)	52.2 (21.00h)	**9.96×**	**1.09×**
	Ours_{20}	29.1 (6.27h)	44.1 (6.40h)	52.1 (10.50h)	**19.9×**	**1.07×**

Results



(a) CIFAR10 (Img/Cls=10) (b) CIFAR10 (Img/Cls=50) (c) CIFAR100 (Img/Cls=10) (d) ImageNet-10 (Img/Cls=10)

Figure 3. Performance comparison across a varying number of training steps.



(a) CIFAR-100 (Img/Cls=1). (b) CIFAR-10 (Img/Cls=10) (c) CIFAR100 (Img/Cls=10) (d) ImageNet-10 (Img/Cls=10)

Figure 4. Performance comparison across varying training time and FLOPs.

Q & A