STRUCT

Dataset Distillation by Matching Training Trajectories

CVPR 2022 (Oral)

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, Jun-Yan Zhu

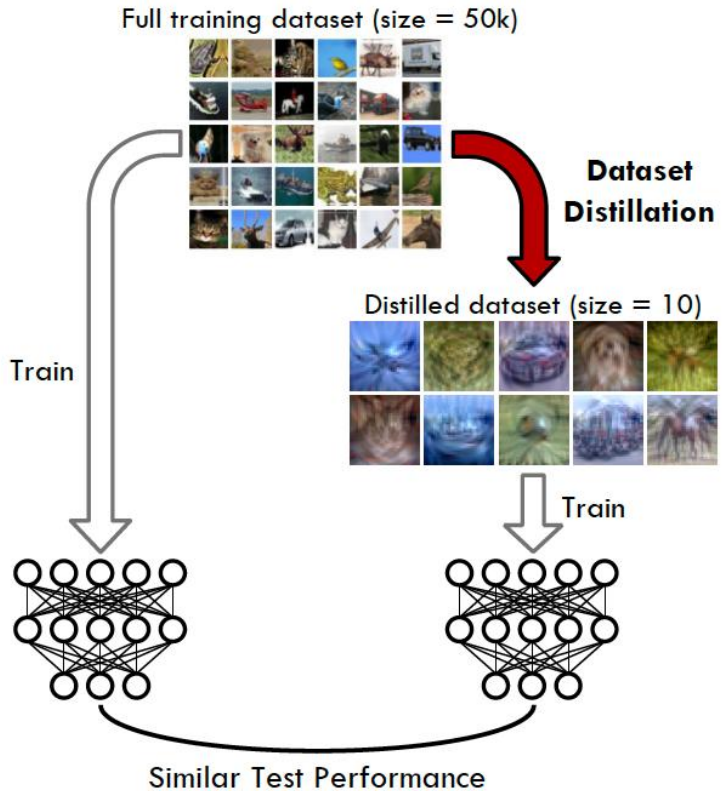
Presented by Yuzhang Hu
2023.2.26

Outline

- Authorship
- Background
- Method
- Experiment
- Conclusion

Background

Dataset Distillation



- Generate a small dataset from a full dataset
- Similar test performance trained on the distilled one

Background

- Traditional Training

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \ell(\mathbf{x}_t, \theta_t),$$

- Dataset Distillation Target

$$\theta_1 = \theta_0 - \boxed{\tilde{\eta}} \nabla_{\theta_0} \ell(\boxed{\tilde{\mathbf{x}}}, \theta_0)$$

- Optimization Process

$$\tilde{\mathbf{x}}^*, \tilde{\eta}^* = \arg\min_{\tilde{\mathbf{x}}, \tilde{\eta}} \mathcal{L}(\tilde{\mathbf{x}}, \tilde{\eta}; \theta_0) = \arg\min_{\tilde{\mathbf{x}}, \tilde{\eta}} \ell(\mathbf{x}, \theta_1) = \arg\min_{\tilde{\mathbf{x}}, \tilde{\eta}} \ell(\mathbf{x}, \theta_0 - \tilde{\eta} \nabla_{\theta_0} \ell(\tilde{\mathbf{x}}, \theta_0)),$$

Background

Similar Methods

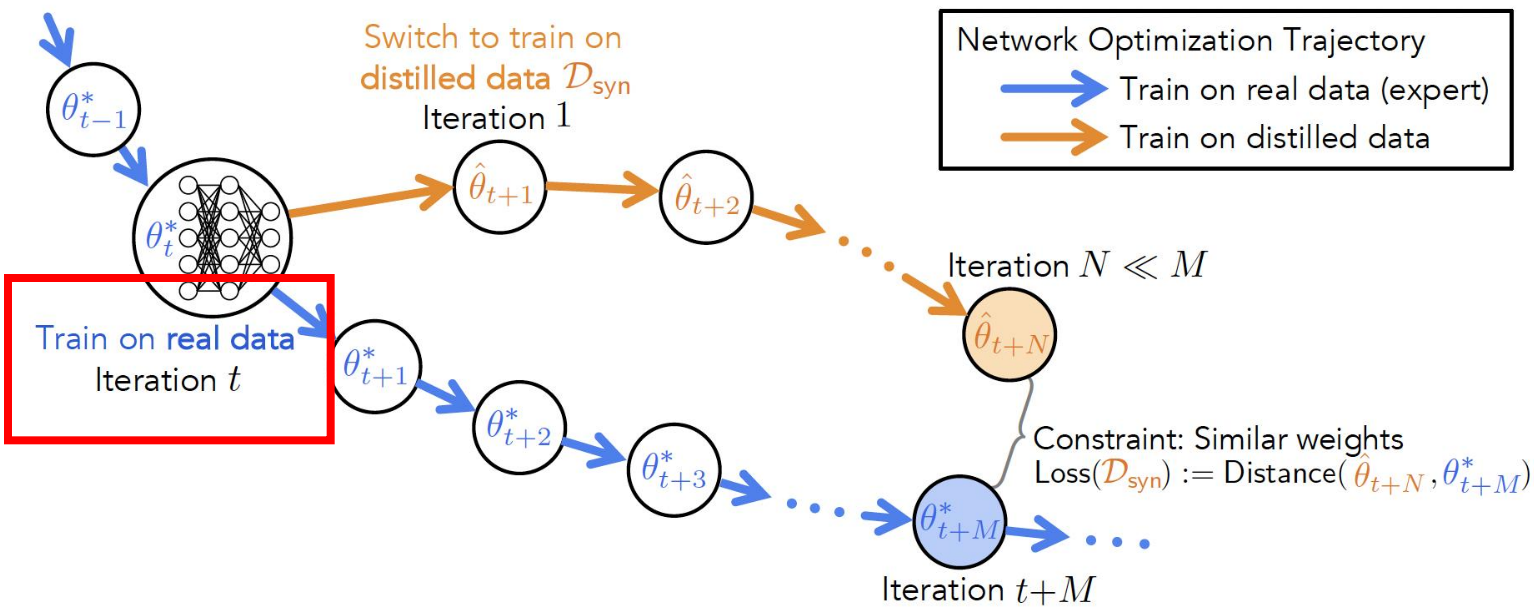
- Imitation Learning
  - Learn a good policy by observing multiple expert demonstrations
- Coreset and Instance Selection
  - Select a subset of the entire training dataset

Background

Example Distilled Image

Method

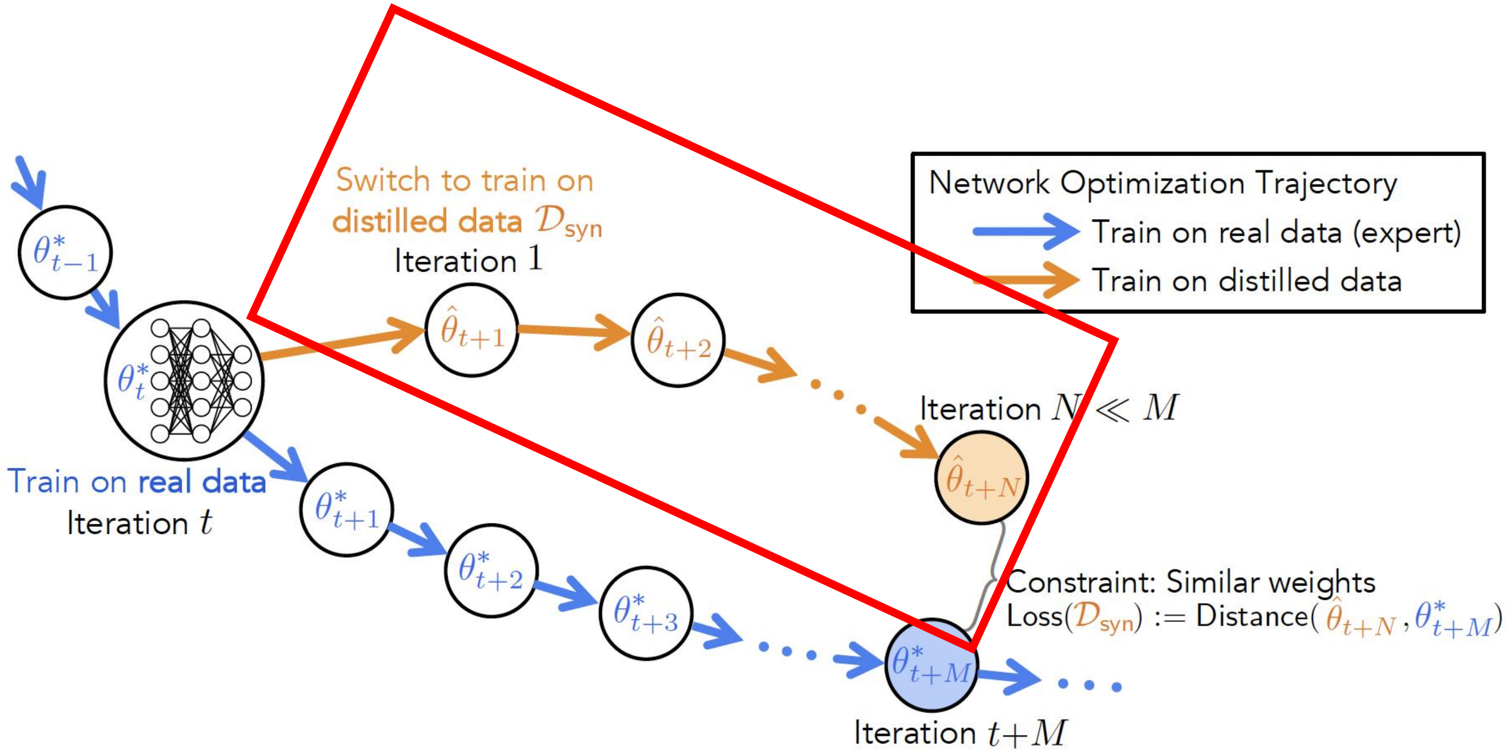
Overall Pipeline
(each distillation iteration)



Step1: sample expert trajectory and start point t

Method

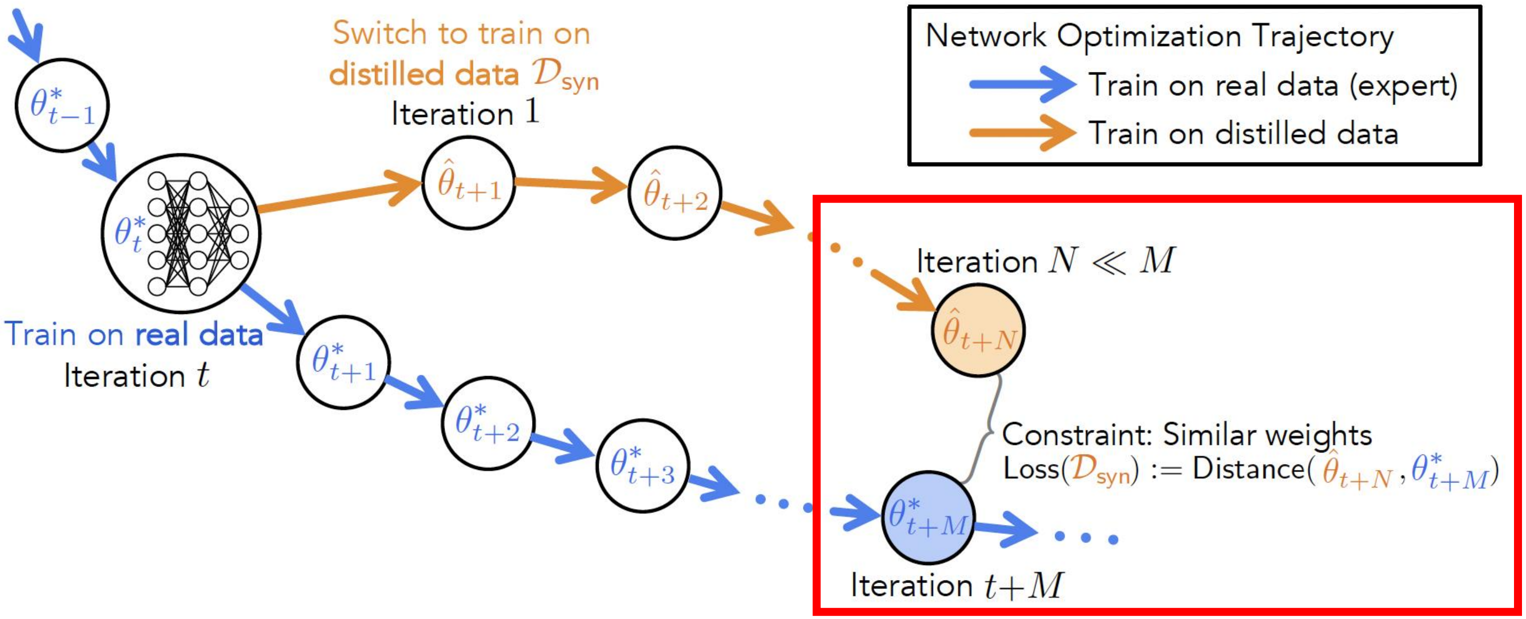
Overall Pipeline



Step2: train the model from the start point on distilled dataset

Method

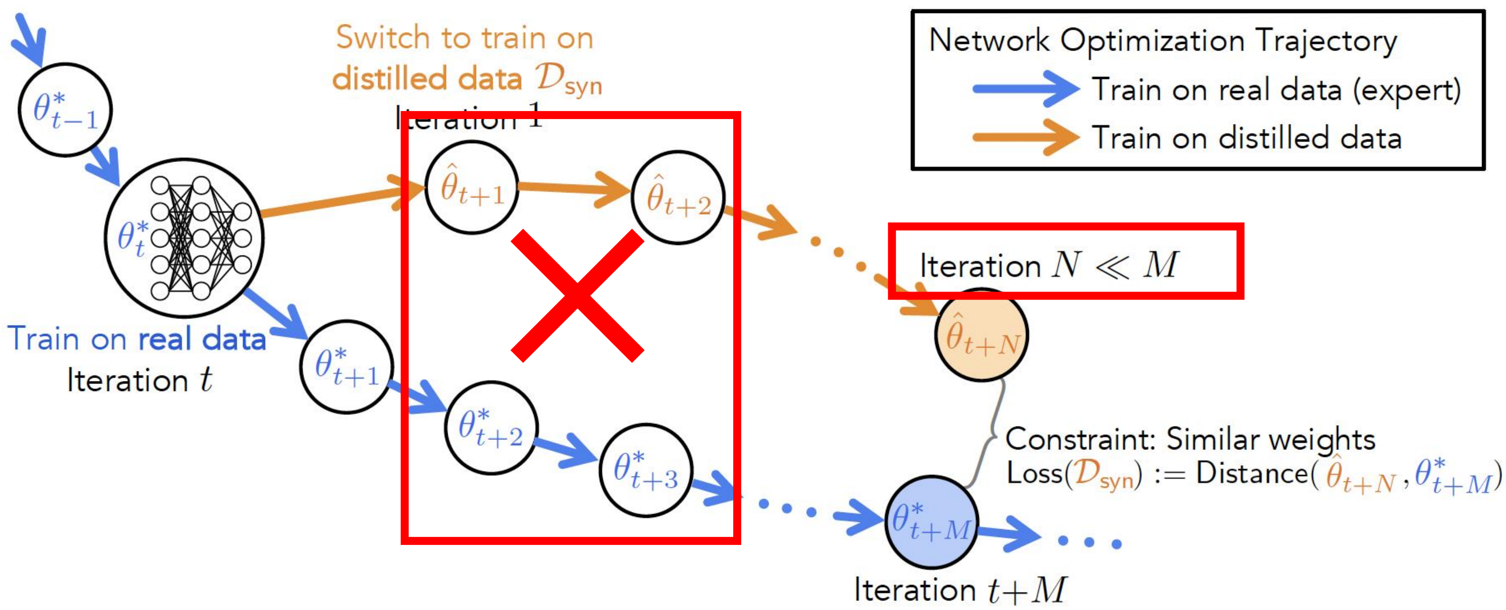
Overall Pipeline



Step3: compute loss between two models and update distilled dataset

Method

More Details



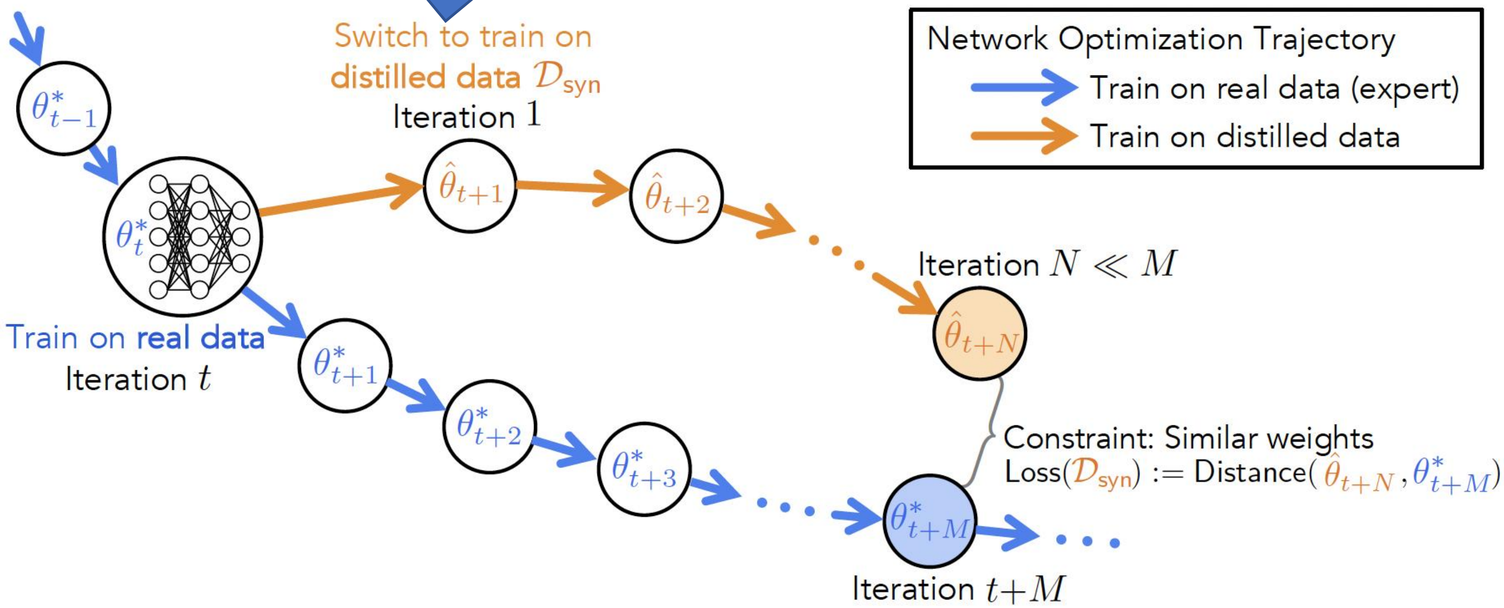
- Avoid being short-sighted and focusing on single steps
- Modeling the full trajectory is difficult to optimize

Method

Memory Constraint

$$b_{t+n} \sim \mathcal{D}_{\text{syn}}$$

$$\hat{\theta}_{t+n+1} = \hat{\theta}_{t+n} - \alpha \nabla \ell(\mathcal{A}(b_{t+n}); \hat{\theta}_{t+n}).$$



- Distill one class at a time→ expert trajectories are trained on all classes simultaneously
- Sampling a new mini-batch every distillation step→ redundant information be distilled into multiple images

Method

Experiment

Quantitative Evaluation

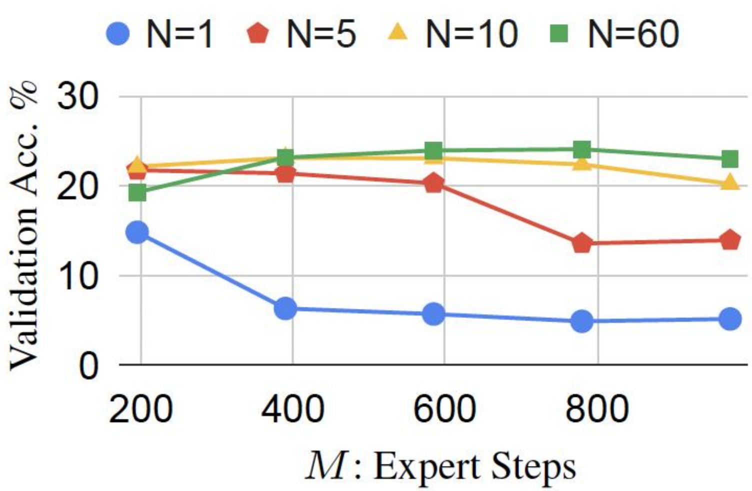
- CIFAR-10: 32x32, 5000 images/class
- CIFAR-100: 32x32, 500 images/class
- Tiny ImageNet: 64x64, 500 images/class

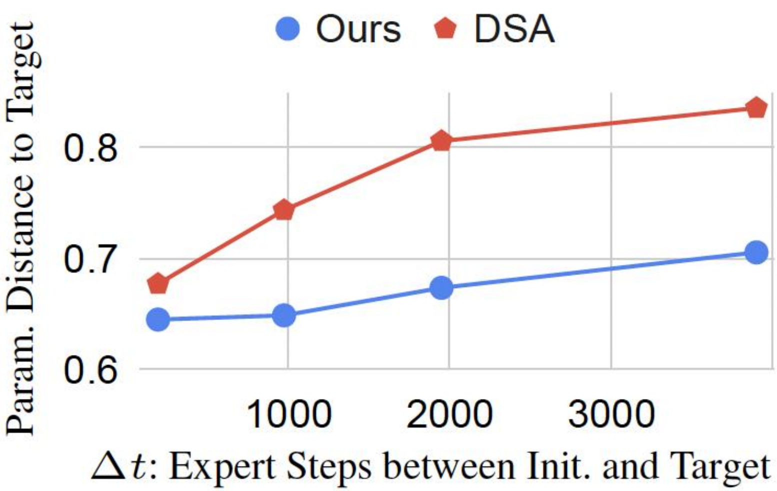
	Img/Cls	Ratio %	Coreset Selection			Training Set Synthesis								Full Dataset
			Random	Herding	Forgetting	DD† [44]	LD† [2]	DC [47]	DSA [45]	DM [46]	CAFE [43]	CAFE+DSA [43]	Ours	
CIFAR-10	1	0.02	14.4 ± 2.0	21.5 ± 1.2	13.5 ± 1.2	-	25.7 ± 0.7	28.3 ± 0.5	28.8 ± 0.7	26.0 ± 0.8	30.3 ± 1.1	31.6 ± 0.8	**$46.3 \pm 0.8^*$**	
	10	0.2	26.0 ± 1.2	31.6 ± 0.7	23.3 ± 1.0	36.8 ± 1.2	38.3 ± 0.4	44.9 ± 0.5	52.1 ± 0.5	48.9 ± 0.6	46.3 ± 0.6	50.9 ± 0.5	**$65.3 \pm 0.7^*$**	84.8 ± 0.1
	50	1	43.4 ± 1.0	40.4 ± 0.6	23.3 ± 1.1	-	42.5 ± 0.4	53.9 ± 0.5	60.6 ± 0.5	63.0 ± 0.4	55.5 ± 0.6	62.3 ± 0.4	**71.6 ± 0.2**	
CIFAR-100	1	0.2	4.2 ± 0.3	8.4 ± 0.3	4.5 ± 0.2	-	11.5 ± 0.4	12.8 ± 0.3	13.9 ± 0.3	11.4 ± 0.3	12.9 ± 0.3	14.0 ± 0.3	**$24.3 \pm 0.3^*$**	
	10	2	14.6 ± 0.5	17.3 ± 0.3	15.1 ± 0.3	-	-	25.2 ± 0.3	32.3 ± 0.3	29.7 ± 0.3	27.8 ± 0.3	31.5 ± 0.2	**40.1 ± 0.4**	56.2 ± 0.3
	50	10	30.0 ± 0.4	33.7 ± 0.5	30.5 ± 0.3	-	-	-	42.8 ± 0.4	43.6 ± 0.4	37.9 ± 0.3	42.9 ± 0.2	**$47.7 \pm 0.2^*$**	
Tiny ImageNet	1	0.2	1.4 ± 0.1	2.8 ± 0.2	1.6 ± 0.1	-	-	-	-	3.9 ± 0.2	-	-	**8.8 ± 0.3**	
	10	2	5.0 ± 0.2	6.3 ± 0.2	5.1 ± 0.2	-	-	-	-	12.9 ± 0.4	-	-	**23.2 ± 0.2**	37.6 ± 0.4
	50	10	15.0 ± 0.4	16.7 ± 0.3	15.0 ± 0.3	-	-	-	-	24.1 ± 0.3	-	-	**28.0 ± 0.3**	

- Coreset Selection: Select a subset of the entire training dataset
  - Only use existing training data
- Training Set Synthesis: Synthesize training data from real data
  - Type1: Match each training step
  - Type2: Match the start and end point

Experiment

Short-Range vs. Long-Range Matching





- Left: long-range achieves better accuracy than short-range
- Right: long-range scheme better approximate real data training

Experiment

One image per class visualization

- Remain high-fidelity for each class with obvious feature
- The first to be capable of distilling higher-resolution images

1st row: Monarch SecoAfrican Elephant, Jellyfish, Kimono, Lampshade,
2nd row: Organ, Pizza, Pretzel, Teapot, Teddy

Conclusion

- Directly optimizing the synthetic data to induce similar network training dynamics as the real data
- Balance between short-range single-step matching and computational intensity of optimizing over training process

Thanks!