NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

Rethinking Dataset Distillation for Classification:

Do Distilled Sets Outperform Coresets?

Trisha Mittal

Akshay Mehra

Josh Kimball

Dolby Laboratories, Inc.

NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

Data-Centric ML

Move away from scaling data, models and compute. Focus on the quality, structure and utility of training data instead. The core question is how to identify this condensed set of data.

Dataset Distillation

Creates a small synthetic training set that compresses information from the original dataset.

Coreset Selection

Selects a small, representative subset of real samples from the original dataset.

NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

Identifying Gaps in Literature

Evaluation Protocol Mismatch

Hard-label ERM and teacher-supervised training made it unclear where gains came from.

Weak Baselines

Strong coreset methods are often absent, and pruning/selection is portrayed as inherently limited.

Practical Gaps

Large scale accuracy, construction cost, and distribution coverage were rarely evaluated together.

NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

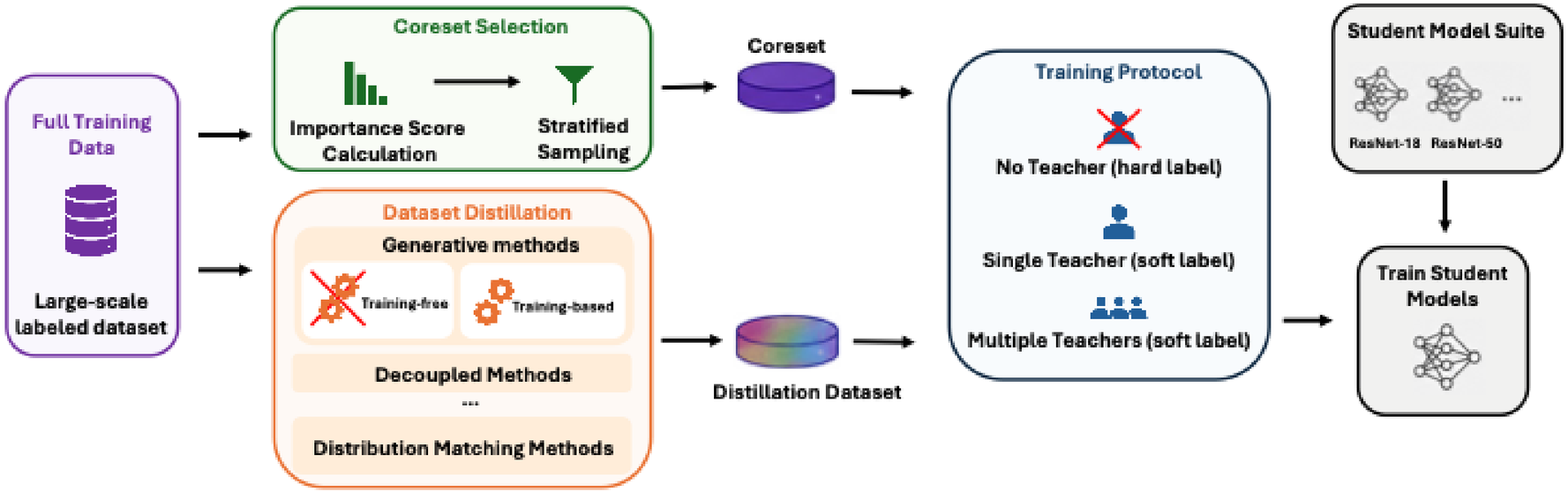
Research Question

> **Do dataset distillation methods truly outperform coreset selection when evaluated under a common experimental framework w.r.t. construction costs and performance accuracy?**

NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

Experimental Setup



Datasets

ImageNet-1k, ImageNet-100, ImageNette

Methods

7 Dataset Distillation Methods
4 Coreset Selection Methods

Evaluation Protocols

Hard Label (No Teacher)
Soft Label (Single Teacher)
Soft Label (Multiple Teacher)

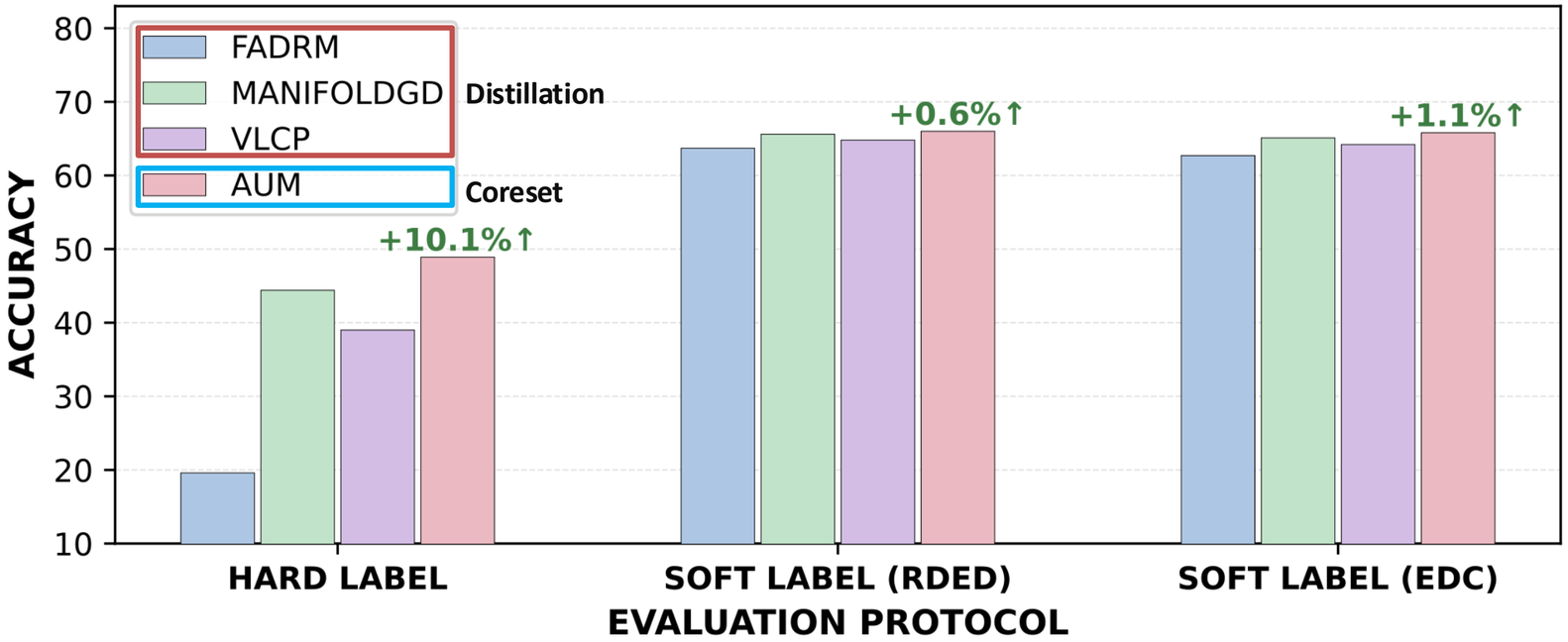
NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

Results

TAKEAWAY 1

Leading distillation methods are comparable to or worse than coreset methods at scale across datasets and coreset/distilled set size.

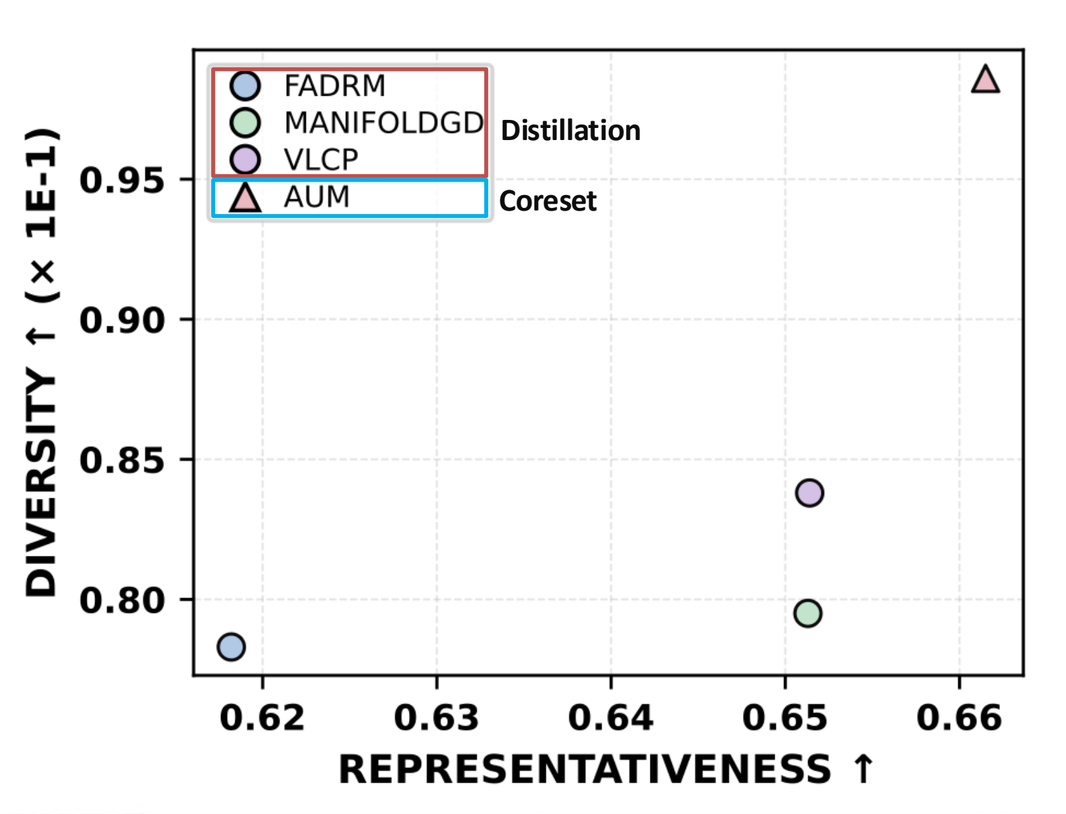


Dataset: ImageNet-1k
Condensed Set: 50 Images/Class (IPC)

NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

Results



Dataset: ImageNet-1k
Condensed Set: 50 Images/Class (IPC)

TAKEAWAY 2

Across datasets, coresets are more representative of the original data and more diverse within each class than distilled sets.

This helps explain why coresets stay so competitive, while distilled sets tend to repeat a narrower set of visual patterns.

NEURAL INFORMATION PROCESSING SYSTEMS

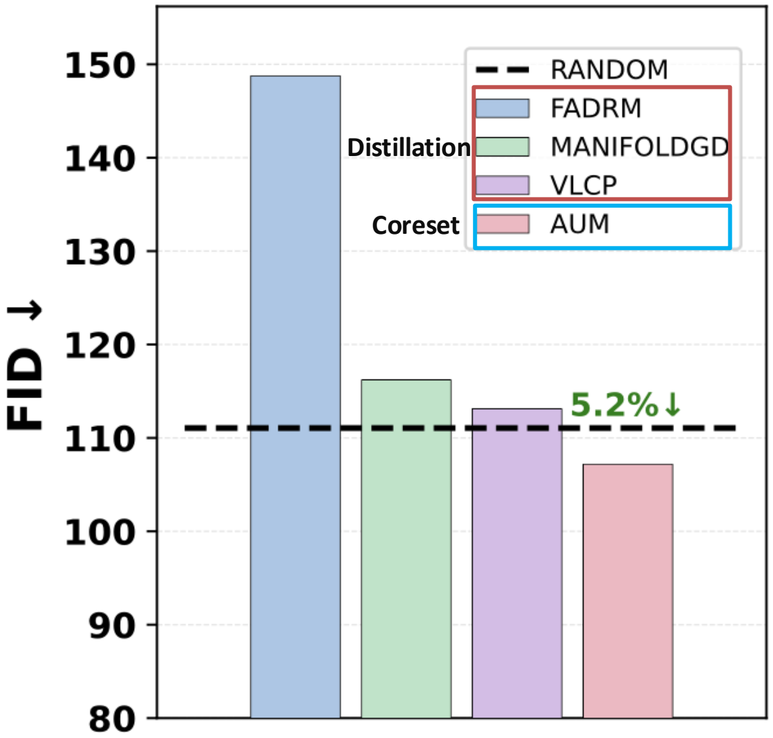
Dolby

Results

TAKEAWAY 3

Coresets keep the natural variety of real images and reach the lowest FID, while distilled sets tend to favor a few canonical views.





Dataset: **ImageNet-1k**
Condensed Set: **50 Images/Class (IPC)**

NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

Results

	Method	Time (hrs)	
		10 IPC	50 IPC
Distillation	ManifoldGD [6]	12.8	51.7
	MGD [13]	11.8	50.7
	DiT [46]	1.4	7.0
	FADRM+ [20]	2.5	12.5
Coreset	AUM [5]	6.0	6.0
	Concepts [21]	3.0	3.0

Dataset: ImageNet-1k
Condensed Set: 10 IPC, 50 IPC

TAKEAWAY 4

Distillation cost grows with IPC (up to about 4× from 10 to 50 IPC), while coreset selection stays constant, making it more practical as condensed sets get larger.

NEURAL INFORMATION PROCESSING SYSTEMS

Dolby

Conclusion

Under standardized evaluation protocols, state-of-the-art dataset-distillation methods are comparable to or worse than coresets on large-scale datasets, despite substantially higher construction costs.

Coresets also provide better coverage of the original data distribution, making them a competitive and efficient alternative.

Poster: Thursday Dec 10, Session #6, Sydney.

Trisha.Mittal@dolby.com

Akshay.Mehra@dolby.com

Paper Link