Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.31931

Collaborative Learning of Large-Scale Dataset Distillation and Filtering for Efficient AI Model Training

Yunchao Lei

College of Engineering and Information Technology, Adelaide University, South Australia, Australia
18476693318@163.com

Abstract. In order to solve the endogenous contradiction between the data scale dividend and the diminishing marginal effect of computing power in large-scale deep learning, this paper proposes a collaborative learning framework for large-scale Dataset distillation and Filtering (DF-CoLearn). By constructing a dynamic feedback closed loop based on bi-level optimization and mutual information maximization, the Pareto optimality between training efficiency and model generalization ability is realized, which provides a new theoretical perspective and technical path for green and efficient AI model training.

Keywords: Dataset distillation, Data filtering, Collaborative learning, Bilevel optimization, Mutual information maximization

1. Introduction

In the journey of artificial intelligence towards universal cognition, deep learning models are undergoing a profound transition from quantitative change to qualitative change. For a long time, the academic community firmly believes in the "law of scale", and believes that the expansion of the parameter scale means the improvement of the intelligence level [1-3].

However, as the model volume exceeds the trilla-level parameter threshold, this extensive growth model relying on computing power accumulation is facing the severe challenge of diminishing marginal effect [4]. In order to alleviate the above sharp contradictions, researchers have carried out arduous exploration on the core proposition of "data reduction". However, as shown in Figure 1, examining the existing results, although both have their own advantages, they are both trapped by their own endogenous defects, and it is difficult to independently undertake the task of large-scale and efficient training.

DOI: 10.54254/2755-2721/2026.31931

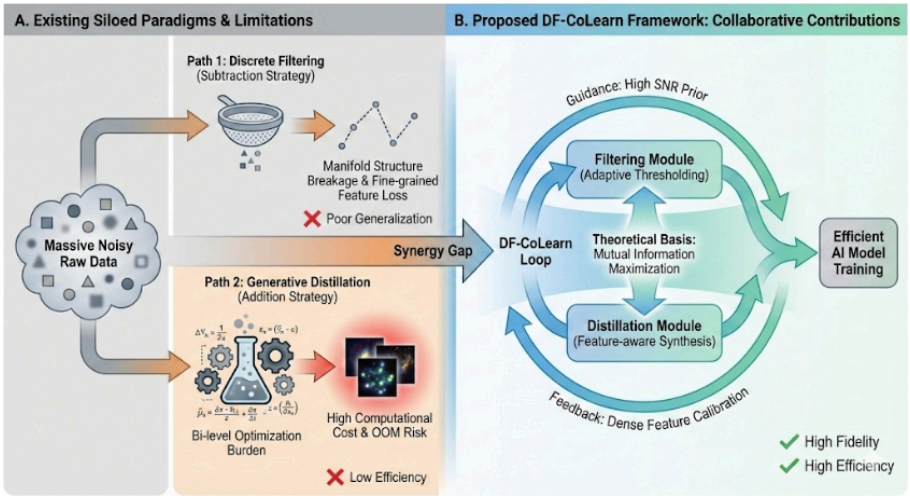

Figure 1. Insufficient existing research

The core set selection is essentially a subtraction strategy based on discrete sampling. However, it often fails due to the loss of fine-grained features [5,6]. Compared with direct filtering, data distillation is an additive strategy based on generative adversarial. Unfortunately, this generative paradigm often falls into the computational quack of bileval optimization when facing large-scale datasets, which largely offsets the efficiency dividend brought by data compression [7].

To sum up, most of the existing studies regard the two as either/or alternatives, and few scholars try to build a collaborative mechanism that can integrate the advantages of the two and complement their shortcomings. Based on the above pain points, this study proposes a DF-CoLearn framework. The framework aims to break the limitation of traditional single strategy and establish a dynamic interactive and spiral-ascending optimization closed loop. By introducing the mutual information maximization theory, the objective function of cooperative optimization is constructed, and the necessity of cooperative mechanism over single strategy is clarified from the mathematical principle, which lays a solid theoretical foundation for data efficiency research. Through detailed comparative experiments, it is strongly confirmed that the framework can maintain the excellent generalization performance of the model with very low data retention rate.

2. Methodology

In this paper, large-scale data set processing is formalized as a constrained bi-level optimization problem [8], as shown in FIG. 2, and then a gradient-based dynamic interactive algorithm is derived from the perspective of information theory [9]. Let the original large-scale dataset be:

$$\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{T}|} \tag{1}$$

Where $x \in \mathcal{X} \subset R^d$, $y \in \mathcal{Y}$. Our goal is to generate a minimal synthetic dataset:

$$\mathcal{S} = \{(s_j, y_j)\}_{j=1}^{|\mathcal{S}|} \tag{2}$$

Where $|\mathcal{S}| \ll |\mathcal{T}|$, so that the model $\mathcal{S}$ trained on $\theta_{\mathcal{S}}$ can achieve the minimum generalization error on the real data distribution. However, direct distillation from the full $\mathcal{T}$ is computationally too expensive and susceptible to noise interference. Therefore, we introduce a

DOI: 10.54254/2755-2721/2026.31931

learnable filter mask vector $m \in [0,1]^{|\mathscr{T}|}$, which is used to define the filtered subset distribution $\mathscr{P}_{\mathscr{T}}(m)$. We define the overall objective function of DF-CoLearn as follows.

$$\min_{\mathscr{S},m} \underbrace{E_{(x,y)\sim\mathbb{T}}\left[\mathscr{L}\left(f_{\theta^*(\mathscr{S},m)}(x),y\right)\right]}_{\text{Task Performance}} - \lambda \underbrace{I\left(\mathscr{S};\mathscr{T}\right)}_{\text{Information Synergy}} \tag{3}$$

$$\text{s.t.} \quad \theta^*(\mathscr{S},m) = \arg\min_{\theta} \mathscr{L}_e(\mathscr{S},\theta) + \gamma\left|m\right|_1$$

Among them, the first term aims to minimize the validation loss and ensure the generalization performance. The second term $I\left(\mathscr{S};\mathscr{T}\right)$ is the mutual information regularization term, which aims to maximize the information overlap between synthetic data and screening data, forcing m to select samples that can most effectively support the generation of $\mathscr{S}$ [10].

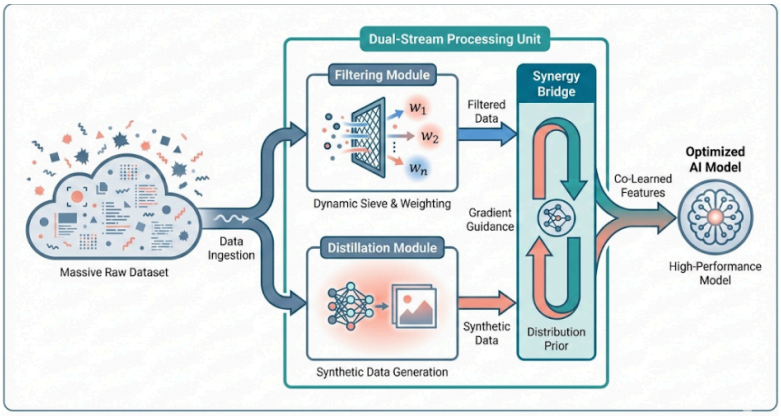

Figure 1: DF-CoLearn Framework Architecture – Iterative Closed-Loop Process for Robust AI Training

Figure 2. Overall architecture diagram of DF-CoLearn

To solve this problem, we define the "hard coefficient" $\kappa_i^{(t)}$ of the sample (x_i, y_i) at iteration t. Not only the prediction error is considered, but also the rate of change of gradient norm is introduced to capture the contribution of samples to the decision boundary:

$$\kappa_i^{(t)} = \alpha \cdot \mathscr{L}\left(f_{\theta_t}(x_i), y_i\right) + (1-\alpha) \cdot \left|\nabla_\theta \mathscr{L}\left(f_{\theta_t}(x_i), y_i\right)\right|_2 \tag{4}$$

Based on this, an adaptive truncation function $\Phi(\cdot)$ is constructed to generate the soft mask. The goal of the distillation module is to generate a synthetic dataset $\mathscr{S}$. As shown in Figure 3, the traditional gradient matching loss function is modified to a weighted form:

DOI: 10.54254/2755-2721/2026.31931

$$\mathscr{L}_{\mathscr{G}_{\mathscr{M}}}(\mathscr{S}, m) = \sum_{c=1}^{C} \left| \frac{1}{|\mathscr{S}|} \sum_{(s,y)\in\mathscr{S}} \nabla_{\theta} \mathscr{L}(s, y) - \frac{1}{\sum m_j} \sum_{(x_j,y_j)\in\mathscr{T}} m_j \cdot \nabla_{\theta} \mathscr{L}(x_j, y_j) \right|_2^2 \tag{5}$$

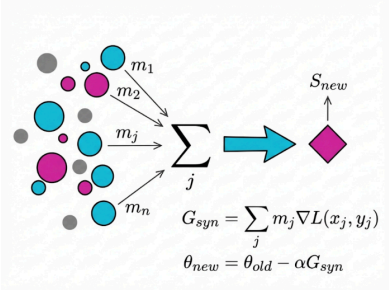

Figure 3. Schematic diagram of the microscopic mechanism of gradient cooperation

To solve the inclusion joint optimization problem, we design an alternating update strategy based on gradients. Randomly initialize the set $\mathscr{S}^{(0)}$, and initialize the mask $m^{(0)} = 1$. Fix the mask m and update the synthesized image using stochastic gradient descent as follows.

$$\mathscr{S}^{(+1)} \leftarrow \mathscr{S}^{(\,)} - \epsilon_S \nabla_{\mathscr{S}} \mathscr{L} \quad \left(\mathscr{S}^{(\,)}, m^{(k)}\right) \tag{6}$$

The temporary model $\mathscr{S}^{(+1)}$ is trained using the current $\theta'_{\mathscr{S}}$. Fix the synthetic set $\mathscr{S}$ and calculate the loss feedback of the original data under $\theta'_{\mathscr{S}}$. Update the mask with the idea of meta-gradient [11] to reward the original samples whose gradient direction is the same as that of the validation set:

$$m^{(k+1)} \leftarrow m^{(k)} - \epsilon_m \nabla_m \mathscr{L} \quad \left(\theta^*(\mathscr{S}, m), \mathscr{T}\right) \tag{7}$$

The termination condition is to reach a preset number of iterations or performance convergence.

3. Experiments

In order to verify the generalization robustness of the algorithm under different semantic complexity, CIFAR-10/100 and ImageNet-1K are selected as benchmark datasets [12]. The comparison methods include: Herding [13]; Data distillation (Dataset Condensation, DSA, Matching Training Trajectories, TESLA) [14-16].

The experiments were conducted on NVIDIA A100 cluster, using standard ConvNet and ResNet-AP architectures. The evaluation metric is anchored as Top-1 test set accuracy (%) versus relative training speedup.

Table 1 shows the results of the lateral comparison under different Settings of the number of images per class. In the extreme compression scenario with IPC=1, DF-CoLearn achieves a 3.6%

DOI: 10.54254/2755-2721/2026.31931

improvement over the second-best model TESLA. The advantages are further amplified in the CIFAR-100 high-dimensional classification task. Under the ResNet-18 architecture, the proposed method achieves an accuracy of 28.4% when IPC=10, which confirms the excellent performance of the framework in dealing with large-scale heterogeneous data in the real world.

Table 1. Comparative analysis of top-1 accuracy (%)

Method	Type	CIFAR-10 (IPC=1)	CIFAR-10 (IPC=10)	CIFAR-10 (IPC=50)	CIFAR-100 (IPC=1)	CIFAR-100 (IPC=10)
Random	Subset	14.4 ± 1.2	26.0 ± 1.1	43.4 ± 1.0	4.2 ± 0.3	14.6 ± 0.5
Herding	Coreset	21.5 ± 1.0	31.6 ± 0.8	55.8 ± 0.7	8.4 ± 0.4	17.3 ± 0.6
DC + DSA	Distill	28.3 ± 0.5	52.1 ± 0.4	60.6 ± 0.5	12.7 ± 0.3	32.1 ± 0.4
MTT	Distill	46.3 ± 0.8	65.3 ± 0.7	71.6 ± 0.6	24.3 ± 0.6	40.1 ± 0.5
TESLA	Distill	48.5 ± 0.6	66.4 ± 0.5	72.8 ± 0.5	24.8 ± 0.5	41.7 ± 0.4
DF-CoLearn	Ours	52.1 ± 0.4	69.8 ± 0.3	75.4 ± 0.3	27.6 ± 0.4	45.2 ± 0.3
Improvement	-	+3.6%	+3.4%	+2.6%	+2.8%	+3.5%

The results of the step-by-step peeling experiments are shown in Table II. Pipelining only "filtering-then distillation" is better than pure distillation, but still worse than DF-CoLearn. This shows that the gradient information of the synthetic data successfully calibrates the discriminative threshold of the filter, allowing the model to dynamically mine those "hard examples" that are critical to the decision boundary. Compared to distillation alone, the full framework improves by 4.5%. This increment does not come from the increase in the amount of data, but purely from the information purification effect brought by mutual information maximization.

Table 2. Ablation studies of component contributions

Configuration	Filtering (m)	Distillation (S)	Synergy Loop	Accuracy (%)	Remark
Baseline A	×	×	×	26.0	Random Selection
Baseline B	√	×	×	34.2	Filtering Only (Limited by discrete nature)
Baseline C	×	√	×	65.3	Distillation Only (Suffers from noise)
Variant D	√	√	×	67.1	Pipeline (Filter then Distill, no feedback)
DF-CoLearn	√	√	√	69.8	Full Collaborative Loop

4. Conclusion

By reconstructing the underlying logic of data reduction, this study confirms that "distillation" and "filtering" are not zero-sum games at the level of information theory, but have deep complementary coupling. The successful construction of the DF-CoLearn framework not only solves the problem of feature collapse in large-scale data synthesis at the technical level, but also illustrates the necessity of improving the information density of data through "dynamic collaborative feedback" at the theoretical level. The empirical data show that the proposed method effectively maintains the decision boundary robustness of the model in complex semantic scenes while greatly reducing the training cost, and completes the paradigm transformation from "violent data stacking" to "high-quality data intelligence".

Future work will focus on the transfer and adaptation of this collaborative paradigm in the pre-training stage of multimodal large model, and further explore the theoretical explanation framework based on neural tangent kernel, in order to define the physical boundary of data efficiency in a broader dimension.

DOI: 10.54254/2755-2721/2026.31931

References

[1] Lin, H. Y. (2022). Large-scale artificial intelligence models. Computer, 55(05), 76-80.

[2] Yang, L., Qi, C., Lin, X., Li, J., & Dong, X. (2019). Prediction of dynamic increase factor for steel fibre reinforced concrete using a hybrid artificial intelligence model. Engineering Structures, 189, 309-318.

[3] Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research [research frontier]. IEEE computational intelligence magazine, 5(4), 13-18.

[4] Bhardwaj, E., Alexander, R., & Becker, C. (2025). Limits to AI Growth: The Ecological and Social Consequences of Scaling. arXiv preprint arXiv: 2501.17980.

[5] Xu, X., Liang, T., Zhu, J., Zheng, D., & Sun, T. (2019). Review of classical dimensionality reduction and sample selection methods for large-scale data processing. Neurocomputing, 328, 5-15.

[6] Tian, Q., Sun, W., Zhang, L., Pan, H., Chen, Q., & Wu, J. (2023). Gesture image recognition method based on DC-Res2Net and a feature fusion attention module. Journal of Visual Communication and Image Representation, 95, 103891.

[7] Qiao, Y., Wilson, A., & Zhang, Z. (2023). A Lightweight Ensemble Model Based on Knowledge Distillation and Distributed Data Parallelism for Predicting User Advertising Return on Investment. Journal of Information, Technology and Policy, 1-15.

[8] Qi, S., Wang, R., Zhang, T., Yang, X., Sun, R., & Wang, L. (2024). A two-layer encoding learning swarm optimizer based on frequent itemsets for sparse large-scale multi-objective optimization. IEEE/CAA Journal of Automatica Sinica, 11(6), 1342-1357.

[9] Raginsky, M., & Rakhlin, A. (2011). Information-based complexity, feedback and dynamics in convex programming. IEEE Transactions on Information Theory, 57(10), 7036-7056.

[10] Estévez, P. A., Tesmer, M., Perez, C. A., & Zurada, J. M. (2009). Normalized mutual information feature selection. IEEE Transactions on neural networks, 20(2), 189-201.

[11] Simon, C., Koniusz, P., Nock, R., & Harandi, M. (2020, August). On modulating the gradient for meta-learning. In European conference on computer vision (pp. 556-572). Cham: Springer International Publishing.

[12] Singh, N. D., Croce, F., & Hein, M. (2023). Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. Advances in Neural Information Processing Systems, 36, 13931-13955.

[13] Guo, C., Zhao, B., & Bai, Y. (2022, July). Deepcore: A comprehensive library for coreset selection in deep learning. In International Conference on Database and Expert Systems Applications (pp. 181-195). Cham: Springer International Publishing.

[14] Gao, B., Zhao, B., Gowda, S. N., Xing, X., Yang, Y., Hospedales, T., & Clifton, D. A. (2025). Enhancing Generalization via Sharpness-Aware Trajectory Matching for Dataset Condensation. arXiv preprint arXiv: 2502.01865.

[15] Zhao, B., & Bilen, H. (2023). Dataset condensation with distribution matching. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 6514-6523).

[16] Lee, S., Chun, S., Jung, S., Yun, S., & Yoon, S. (2022, June). Dataset condensation with contrastive signals. In International Conference on Machine Learning (pp. 12352-12364). PMLR.