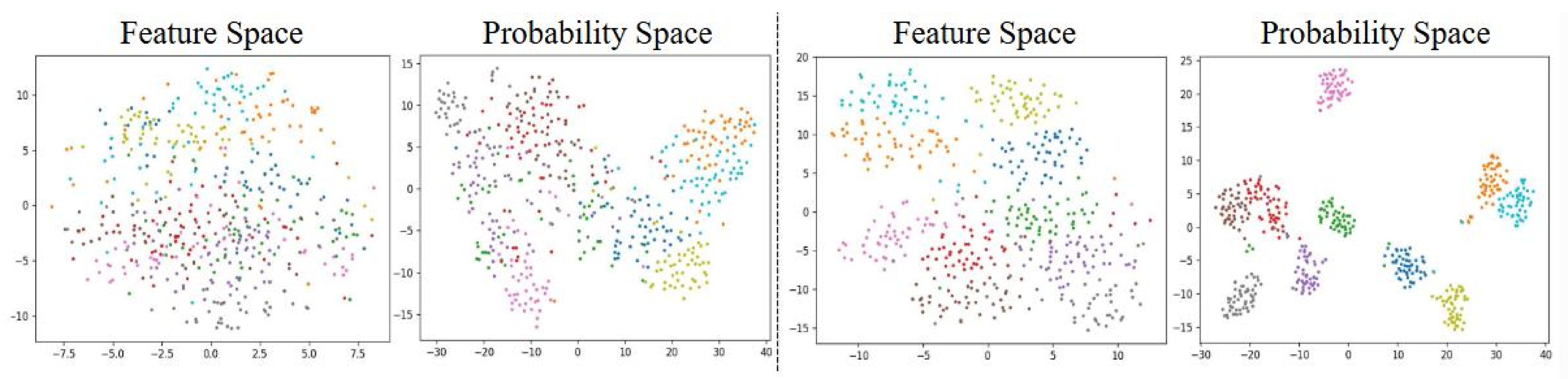
Going Beyond Feature Similarity: Effective Dataset distillation based on Class-aware Conditional Mutual Information

Xinhao Zhong[1], Bin Chen[1,†], Hao Fang[2], Xulin Gu[1], Shu-Tao Xia[2], Enhui-Yang[3]

[1]Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[2]Shenzhen International Graduate School, Tsinghua University [3]University of Waterloo
Email: xh021213@gmail.com

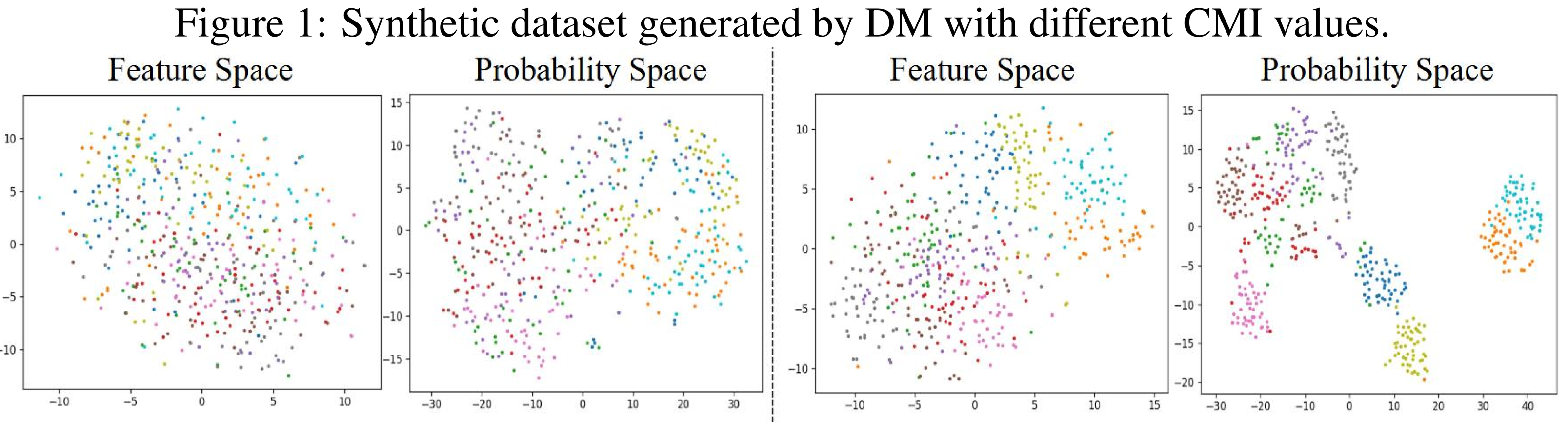
Motivation

- Dataset distillation aims to distillate all the information from the real dataset during optimization, the complexity of distilled information can make the synthetic datasets more challenging for models to learn.
- Empirical conditional mutual information (CMI) from information theory could serve as a class-aware complexity metric for measuring fine-grained condensed information within synthetic datasets.



(a) Synthetic dataset *w.o.* CMI constraint. (b) Synthetic dataset *w.* CMI constraint.

Figure 1: Synthetic dataset generated by DM with different CMI values.



(a) Synthetic dataset *w.o.* CMI constraint. (b) Synthetic dataset *w.* CMI constraint.

Figure 2: Synthetic dataset generated by DSA with different CMI values.

Main Contributions

- We provide an insight into the properties of different classes inherent in the synthetic dataset and point out that the generalization of the distilled data can be improved by optimizing the class-aware complexity of synthetic dataset quantified via CMI empirically and theoretically.
- Building on this perspective, we propose the CMI enhanced loss that simultaneously minimizes the distillation loss and CMI of the synthetic dataset in the feature space. This enables the class-aware complexity of synthetic dataset could be efficiently reduced, while distilled data becoming more focused around their class centers.
- Experimental results show that our method can effectively improve the performance of existing dataset distillation methods by up to 5.5%. Importantly, our method can be deployed as an plug-and-play module for all the existing DD methods with different optimization objectives.

GitHub: https://github.com/ndhg1213/CMIDD

Paper: https://openreview.net/forum?id=0no1Wp2R2j

The Proposed Method

Estimating the class-aware CMI for Synthetic Dataset

For a given input $\mathbf{s} \in \mathcal{S}$, the output feature $\mathbf{z}$ is a deterministic feature vector. We apply the softmax function to the feature vector $\mathbf{z} = (z^1, z^2, \dots, z^M)$ for an input sample $\mathbf{s} \in \mathcal{S}$. **The non-linear relationship between the input S and the output $\hat{Z}$ can be quantified by the conditional mutual information $I(S; \hat{Z} \mid Y)$. This can also be expressed as $I(S; \hat{Z} \mid Y) = H(S \mid Y) - H(S \mid \hat{Z}, Y)$, representing the difference between the uncertainty of S given both $\hat{Z}$ and Y and that of S given Y.**

- When a relatively diverse and large dataset (e.g., $\mathcal{T} \sim \mathbb{P}_X$) is used as the input to $f_{\theta^*}(\cdot)$, the corresponding output $\hat{Z}$ follows a more certain probability distribution produced by $f_{\theta^*}(\cdot)$, leading to a smaller CMI value.
- In contrast, since S is more challenging for randomly initialized networks to learn, its output $\hat{Z}$ often contains excessive confused information related to it, leading to a significant reduction in $H(S \mid \hat{Z}, Y)$.

Estimating the class-aware CMI for Synthetic Dataset

We employ the Kullback-Leibler (KL) divergence $D(P_S \| P_{\hat{Z}|y})$ to quantify the distance between P_S and the conditional distribution $P_{\hat{Z}|y}$ as follow:

$$I(S; \hat{Z} \mid Y = y) = \sum_{\mathbf{s} \in \mathcal{S}} P_{S|Y}(S = \mathbf{s} \mid y) \left[\sum_{i=1}^{M} P(\hat{Z} = i \mid \mathbf{s}) \times \ln \frac{P(\hat{Z} = i \mid \mathbf{s})}{P_{\hat{Z}|y}(\hat{Z} = i \mid Y = y)} \right] \quad (1)$$

$$= \mathbb{E}_{S|Y} \left[\left(\sum_{i=1}^{M} P_S[i] \ln \frac{P_S[i]}{P_{\hat{Z}|y}(\hat{Z} = i \mid Y = y)} \right) \middle| Y = y \right] \quad (2)$$

$$= \mathbb{E}_{S|Y} \left[D\left(P_S \| P_{\hat{Z}|y}\right) \mid Y = y \right], \quad (3)$$

Averaging $I(S; \hat{Z}|y)$ with respect to the distribution $P_Y(y)$ of Y, we can obtain the conditional mutual information between S and $\hat{Z}$ given Y as follow:

$$\mathrm{CMI}(\mathcal{S}) \triangleq I(S; \hat{Z} \mid Y) = \sum_{y \in [C]} P_Y(y) I(S; \hat{Z} \mid y). \quad (4)$$

To compute the $\mathrm{CMI}(\mathcal{S})$, we approximate $P(\mathbf{s}, y)$ by the empirical distribution of synthetic dataset $\mathcal{S}_y = \{(\mathbf{s}_1, y), (\mathbf{s}_2, y), \cdots, (\mathbf{s}_n, y)\}$ for any $y \in [C]$. Then the $\mathrm{CMI}_{\mathrm{emp}}(\mathcal{S})$ can be calculated as:

$$\mathrm{CMI}_{\mathrm{emp}}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{y \in [C]} \sum_{\mathbf{s}_j \in \mathcal{S}_y} \mathrm{KL}\left(P_{\mathbf{s}_j} \| Q^y_{\mathrm{emp}}\right), \quad (5)$$

$$\text{where } Q^y_{\mathrm{emp}} = \frac{1}{|\mathcal{S}_y|} \sum_{\mathbf{s}_j \in \mathcal{S}_y} P_{\mathbf{s}_j}, \text{ for } y \in [C]. \quad (6)$$

Dataset Distillation with CMI enhanced Loss

According to the above calculation of CMI, we propose the CMI enhanced Loss $\mathcal{L}$. Overall, it includes two parts:

$$\mathcal{L} = \mathcal{L}_{DD} + \lambda\, \mathrm{CMI}_{\mathrm{emp}}(\mathcal{S}). \quad (7)$$

The first term $\mathcal{L}_{DD}$ represents any loss function in previous DD methods, e.g., DM, DSA and MTT , $\lambda > 0$ is a weighting hyperparameter.

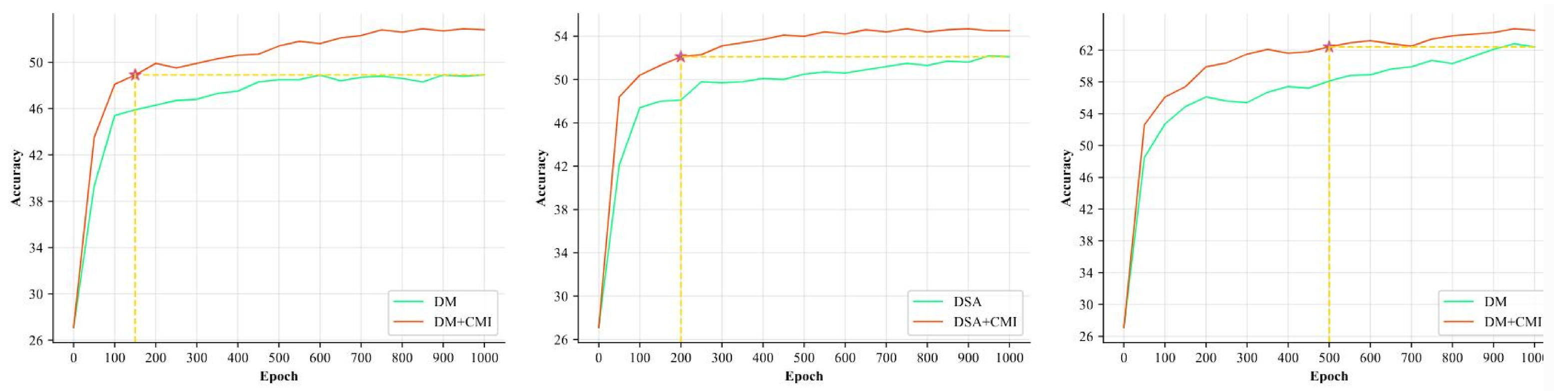
Experimental Results

Performance Enhancements as a Plug-and-Play Module

Table 1: Quantitative comparison with existing methods.

Method	SVHN		CIFAR10		CIFAR100	
IPC	10	50	10	50	10	50
MTT	79.9±0.1	87.7±0.3	65.3±0.4	71.6±0.2	39.7±0.4	47.7±0.2
MIM4DD	-	-	66.4±0.2	71.4±0.3	41.5±0.2	-
SeqMatch	80.2±0.6	88.5±0.2	66.2±0.6	**74.4±0.5**	41.9±0.5	**51.2±0.3**
MTT+CMI	**80.8±0.2**	**88.8±0.1**	**66.7±0.3**	72.4±0.3	**41.9±0.4**	48.8±0.2
Δ	(0.9↑)	(1.1↑)	(1.4↑)	(0.8↑)	(2.2↑)	(1.1↑)
DM	72.8±0.3	82.6±0.5	48.9±0.6	63.0±0.4	29.7±0.3	43.6±0.4
IID-DM	75.7±0.3	**85.3±0.2**	**55.1±0.1**	65.1±0.2	32.2±0.5	43.6±0.3
DM+CMI	**77.9±0.4**	84.9±0.4	52.9±0.3	**65.8±0.3**	**32.5±0.4**	**44.9±0.2**
Δ	(5.1↑)	(2.3↑)	(4.0↑)	(2.8↑)	(2.8↑)	(1.3↑)
IDM	81.0±0.1	84.1±0.1	58.6±0.1	67.5±0.1	45.1±0.1	50.0±0.2
IID-IDM	82.1±0.3	85.1±0.5	59.9±0.2	69.0±0.3	45.7±0.4	51.3±0.4
IDM+CMI	**84.3±0.2**	**88.9±0.2**	**62.2±0.3**	**71.3±0.2**	**47.2±0.4**	**51.9±0.3**
Δ	(3.3↑)	(4.8↑)	(3.6↑)	(3.8↑)	(2.1↑)	(1.9↑)
IDC	87.5±0.3	90.1±0.1	67.5±0.5	74.5±0.1	45.1±0.4	-
DREAM	87.9±0.4	90.5±0.1	69.4±0.4	74.8±0.1	**46.8±0.7**	52.6±0.4
PDD	-	-	67.9±0.2	76.5±0.4	45.8±0.5	53.1±0.4
IDC+CMI	**88.5±0.2**	**92.2±0.1**	**70.0±0.3**	**76.6±0.2**	46.6±0.3	**53.8±0.2**
Δ	(1.0↑)	(2.1↑)	(2.5↑)	(2.1↑)	(1.5↑)	-
Whole Dataset	95.4±0.2		84.8±0.1		56.2±0.3	

Training Efficiency



(a) The accuracy curve of adding CMI constraint to DM. (b) The accuracy curve of adding CMI constraint to DSA. (c) The accuracy curve of adding CMI constraint to MTT.

Figure 3: Applying CMI constraint brings stable efficiency improvements.

Conclusion

we present a novel conditional mutual information (CMI) enhanced loss for dataset distillation by analyzing and reducing the class-aware complexity of synthetic datasets. The proposed method computes and minimizes the empirical CMI of a pre-trained model, effectively addressing the challenges faced by previous dataset distillation approaches.