MIM4DD: Mutual Information Maximization for Dataset Distillation

Yuzhang Shang[1], **Zhihang Yuan**[2], **Yan Yan**[1*]
[1]Department of Computer Science, Illinois Institute of Technology
[2]Huomo AI
yshang4@hawk.iit.edu, zhihang.yuan@huomo.ai, yyan34@iit.edu

A Appendix

A.1 In-variance of Mutual Information

Theorem 1 (In-variance of Mutual Information): Mutual information is invariant under reparametrization of the marginal variables. If $X' = F(X)$ and $Y' = G(Y)$ are homeomorphisms (*i.e.*, $F(\cdot)$ and $G(\cdot)$ are smooth uniquely invertible maps), then

$$I(X, Y) = I(X', Y'). \tag{A.1}$$

Proof. If $X' = F(X)$ and $Y' = G(Y)$ are homeomorphisms (smooth and uniquely invertible maps), and $J_X = \|\frac{\partial X}{\partial X'}\|$ and $J_Y = \|\frac{\partial Y}{\partial Y'}\|$ are the Jacobi determinants, then

$$\mu'(x', y') = J_X(x')J_Y(y')\mu(x, y) \tag{A.2}$$

and similarly for the marginal densities, which gives

$$\begin{aligned} I(X', Y') &= \iint dx'dy'\mu'(x', y') \log \frac{\mu'(x', y')}{\mu'_x(x')\mu'_y(y')} \\ &= \iint dxdy\mu(x, y) \log \frac{\mu(x, y)}{\mu_x(x)\mu_y(y)} \\ &= I(X, Y). \end{aligned} \tag{A.3}$$

More details can be found in [10].

Discussion on Theorem 1.

Our objective is to maximize the Mutual Information (MI) between the synthetic dataset and the real dataset (Eq. 3), a task that is numerically unfeasible. To overcome this challenge, we present this theorem. It allows us to transform the target problem at the data level (Eq. 3) into a more manageable problem at the feature map level (Eq. 9). Given that each layer's mapping $\mathbf{W}^k : \mathbb{R}^{d_{k-1}} \longmapsto \mathbb{R}^{d_k} (k = 1, ..., K)$ in the network (as per Eq. 4, 5, and 6) can be treated as smooth, uniquely invertible maps, we can achieve the goal of maximizing the mutual information between the two datasets. This is done by maximizing the mutual information between two sets of down-sampled feature maps.

A.2 Datasets and Implementation Details

A.2.1 Datasets

MNIST [12] is a dataset for handwritten digits recognition that is widely used for validating image recognition models. It contains 60,000 training images and 10,000 testing images with the size of 28×28.

*Corresponding author

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

CIFAR10/100 [11] are two datasets consist of tiny colored natural images with the size of 32×32 from 10 and 100 categories, respectively. In each dataset, 50,000 images are used for training and 10,000 images for testing.

A.2.2 Implementation Details.

In the experiments, we optimize synthetic sets with 1/10/50 Images Per Class (IPC) across all three datasets, using a three-layer Convolutional Network (ConvNet) identical to those used in [23, 19, 3]. The ConvNet comprises three consecutive blocks of 'Conv-InstNorm-ReLU-AvgPool.' Each convolutional layer has 128 channels, and AvgPool represents a 2×2 average pooling operation with stride 2. The synthetic images' initial learning rate is 0.1, which is halved at the 1,800th and 2,800th iterations. The training is stopped after 5,000 iterations. To test the ConvNet's performance on the synthetic dataset, we train the network on synthetic sets for 300 epochs and assess the performance using five randomly initialized networks. The network's initial learning rate is 0.01. As per [3], we conduct five experiments and report the mean and standard deviation across the five networks. The default batch size is 256, and λ in Eq.17 is 0.8. The effect of λ is explored in Sec.3.3.

A.3 Synthetic Samples Visualization.

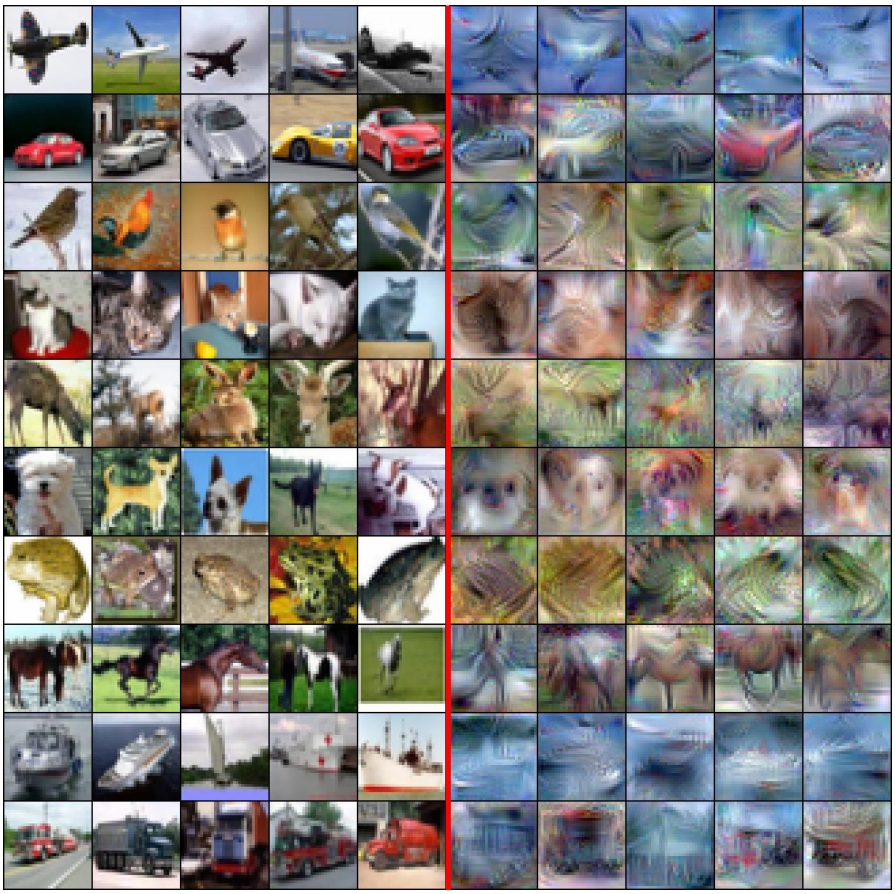

Figure 1: **(Left)** Samples from CIFAR10; **(Right)** Samples from Synthetic dataset based on CIFAR10. We observe that the heterogeneity in the generated images enhanced, benefited from the contrastive learning loss (Loss $\mathcal{L}_{\texttt{MIM4DD}}$ in Eq.17).

B Related Work

Dataset Distillation (DD) is firstly introduced by Wang *et al.* [20], in which they optimize the distilled images using gradient-based hyperparameter optimization [14]. The key problem is to

optimize the specific-designed metrics of networks on real and synthetic datasets to update the optimizable images. Subsequently, several works significantly improve the results by designing different metrics. For example, Bohdal *et al.*and Sucholutsky *et al.* [2, 17] use distance between networks' soft labels; Zhao *et al.* [23] define the gradients of networks as metric; Zhao *et al.* [22] further adopts augmentations to enhance the alignment ability; Wang *et al.* [19] utilize distance of network feature maps as metric; and Cazenavette [3] propose long-range trajectory to construct the metric function. Lee *et al.* [13] propose Dataset Condensation with Contrastive Signals (DCC) by modifying the loss function to enable the DC methods to effectively capture the differences between classes. On the other hand, researchers take DD as a bi-level optimization problem. For example, Zhou *et al.* [24] employ a closed-form approximation for the unrolled inner optimization; Deng *et al.* [6] revisits the optimization framework in [20] and observe that the inclusion of a momentum term in inner optimization can significantly enhance performance, leading to state-of-the-art results in certain settings.

DD is essentially a compression problem that emphasizes on maximizing the preservation of information contained in the data. We argue that well-defined metrics which measure the amount of shared information between variables in information theory are necessary for success measurement, but are never considered by previous works. Therefore, we propose to introduce a well-defined metric in information theory, mutual information (MI), to guide the optimization of synthetic datasets.

Contrastive Learning and Mutual Information. The fundamental idea of all contrastive learning methods is to draw the representations of positive pairs closer and push those of negative pairs farther apart within a contrastive space. Several self-supervised learning methods are rooted in well-established ideas of MI maximization, such as Deep InfoMax [9], Contrastive Predictive Coding [15], MemoryBank [21], Augmented Multiscale DIM [1], MoCo [8] and SimSaim [5]. These are based on NCE [7] and InfoNCE [9] which can be seen as a lower bound on MI [16]. Meanwhile, Tian *et al.* [18] and Chen *et al.* [4] extend the contrastive concept into the realm of Knowledge Distillation (KD), pulling and pushing the representations of teacher and student.

The formulation of our method for DD, `MIM4DD` also absorbs the core idea (*i.e.*, constructing the informative positive and negative pairs for contrastive loss) of the existing contrastive learning methods, especially the contrastive KD methods, CRD [18] and WCoRD [4]. However, our approach has several differences from those methods: (i) our targeted MI and formulated numerical problem are totally different; (ii) our method can naturally avoid the cost of MemoryBank [21] for the exponential number of negative pairs in CRD and WCoRD, thanks to the small size of the synthetic dataset in our task. Given that the size of the synthetic dataset M typically ranges from $0.1 - 1\%$ of the size of the real dataset N, the product $M \cdot N$ is significantly smaller than $N \cdot N$ (*i.e.*, $M \cdot N \ll N \cdot N$).

Difference with DCC [13]. Recently, Lee et al. [13] introduced Dataset Condensation with Contrastive Signals (DCC), modifying the loss function to allow Dataset Condensation methods to effectively discern differences between classes. However, several distinctions exist between DCC and our method: **(i)** They are motivated differently. Our approach is predicated on information degradation, while DCC hinges on class diversity. **(ii)** From the perspective of contrastive learning, the view, positive and negative samples differ considerably. Our approach can be implemented at the feature map level, thanks to the introduced Theorem 1, while DCC can only be deployed at the gradient level. **(iii)** The performance of our method significantly surpasses that of DCC.

C Codes

Codes can be found anomalously in Supplement.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019.

[2] Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Flexible dataset distillation: Learn labels instead of images. *arXiv preprint arXiv:2006.08572*, 2020.

[3] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, 2022.

[4] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *CVPR*, 2021.

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[6] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. In *NeurIPS*, 2022.

[7] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[9] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[10] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 2004.

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[13] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *ICML*, 2022.

[14] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *ICML*, 2015.

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[16] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019.

[17] Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. In *IJCNN*, 2021.

[18] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2021.

[19] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *CVPR*, 2022.

[20] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[21] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[22] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*, 2021.

[23] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *ICLR*, 2021.

[24] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *arXiv preprint arXiv:2206.00719*, 2022.