기술기고문 온디바이스 AI 기반 미디어 기술

서브셋 매칭 알고리즘을 이용한 효율적인 데이터 응축 방법

An Efficient Dataset Condensation Method with a Subset Matching Algorithm

□ Linh Tam Tran, Sung-Ho Bae / Kyung Hee University

요 약

데이터셋 응축은 대량의 이미지 데이터셋에서 지식을 압축하여 더 작은 부분 집합으로 저장함으로써 학습 속도를 향상시키고 저장 공간 요구량을 줄이는 것을 목표로 한다. 그래디언트 매칭 기반 접근법에서는 원본 데이터셋과 합성 데이터셋에서 무작위로 미니배치를 샘플링한 후, 두 데이터셋의 그래디언트를 반복적으로 매칭하여 수렴할 때까지 학습을 진행한다. 그러나 이 과정은 모든 이미지의 정보를 전체적으로 집계하기 때문에 지역적인 텍스처 정보가 손실될 수 있다. 이러한 한계를 해결하기 위해, 우리는 원본 데이터셋을 여러 개의 부분 집합으로 분할한 후, 각 부분 집합 내에서 응축을 수행하는 새로운 부분 집합 기반 증류 방법을 제안한다. 이러한 전략은 지역적인 텍스처 정보를 더 잘 보존할 수 있도록 하여 보다 효과적인 지식 응축을 가능하게 한다. 다양한 데이터셋 조건에서 실험을 통해 제안하는 부분 집합 매칭 기반 데이터 응축 방법이 기존 방법보다 향상된 성능을 보임을 검증하였다.

I. Introduction

Deep Neural Networks (DNNs) have become the go-to solution for various computer vision tasks [1,2,3,4]. However, training deep networks effectively demands a vast amount of data and extensive GPU computation. To mitigate these challenges, Dataset Condensation (DC)[5,6,7] has been introduced as a method to compress large datasets into significantly smaller ones, reducing both training costs and storage requirements. Ideally, the condensed dataset should achieve performance comparable to the original, enabling efficient experimentation in resource-constrained environments.

Several DC methods have been introduced in recent years[5,6,7,8]. The pioneering work[5] formulates dataset condensation as a bi-level optimization problem, where synthetic images and network parameters θ are updated alternately. However, this process is highly computationally expensive, requiring millions of gradient steps for convergence, making it impractical for large-scale applications. To enhance condensation efficiency, recent studies have explored various acceleration techniques. These include minimizing the weight gradient differences between real and synthetic images[6,7,8], aligning the training trajectory[9], matching intermediate features[10], or approximating the data distribution[11].

Among the various approaches discussed, gradient matching-based methods[6,7,8] have shown promise for dataset condensation due to their class-wise matching, which enables efficient parallel processing. Traditional dataset condensation applies gradient matching by randomly sampling mini-batches from the entire dataset within each class. However, this standard approach tends to average out gradient directions, diminishing the effectiveness of distilling critical texture details-such as shapes, edges, and fine-grained features-during condensation. To overcome this limitation, we introduce a random subset matching strategy. By randomly partitioning both the real and synthetic datasets into subsets, we perform gradient matching within each subset independently. This method reduces interference between gradients representing different features, allowing texture details to be transferred more effectively from the real dataset to the synthetic one. We perform experiments with various budget settings and distillation methods to demonstrate the effectiveness of the proposed method.

Ⅱ. Related works

Dataset Condensation (DC), first introduced by Wang et al.[5], aims to generate a compact synthetic dataset that enables a network to achieve performance comparable to training on the full original dataset. Over time, various DC techniques have been developed, each focusing on different optimization objectives. These include gradient matching DC[6], trajectory matching[9], training feature matching[10], distribution matching[11], representative matching[12], attention matching[13], infinite-width kernel approximations[14], kernel ridge regression[15], soft label distillation[16,17], and the use of data augmentation DSA[7]. DC has been used in various domains such as continual learning[18,19], neural architecture search[20], federated learning[21,22], medical data[23], and privacy-preserving[24].

Gradient matching-based methods, including DC[6], DSA[7], and IDC[8], have achieved strong performance by aligning the gradients of real and synthetic data. However, these approaches typically perform gradient matching at the class level, which can cause gradient interference and limit the ability to capture fine-grained texture details. To address this limitation, we propose subset matching, a novel gradient-matching technique that divides both real and synthetic datasets into smaller subsets before performing gradient alignment. By refining the matching process at a more granular level, this method enhances the transfer of crucial image details, resulting in improved performance, especially when the Image per Class (IPC) budget is high.

Ⅲ. Preliminary on Dataset Condensation

DC aims to generate a compact synthetic dataset $\mathcal{S}$

from a larger dataset $\mathcal{T}$ such that a model trained on $\mathcal{S}$ achieves performance comparable to one trained on the full dataset. Given a distance metric $\mathcal{D}$ (e.g., Mean Squared Error) and a matching objective Φ, the general formulation of DC can be expressed as:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathcal{D}\left(\Phi\big(\mathcal{E}(\mathcal{S})\big), \Phi\big(\mathcal{E}(\mathcal{T})\big)\right)$$

In gradient matching-based approaches, our goal is to minimize the gradient difference between real samples and synthetic ones which can be expressed as:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathcal{D}\left(\nabla_{\theta\mathcal{L}_{\mathrm{CE}}}\big(\mathcal{E}(\mathcal{S})\big), \nabla_{\theta\mathcal{L}_{\mathrm{CE}}}\big(\mathcal{E}(\mathcal{T})\big)\right),$$

Where $\mathcal{E}(\cdot)$ is the differentiable augmentation and $\mathcal{L}_{\mathrm{CE}}$ denote Cross-Entrypy loss. In practice, we minimize the weight gradient, i.e., $\nabla_{\theta\mathcal{L}_{\mathrm{CE}}}(\cdot)$, by calculating the distance between gradients: one from the synthetic samples $\mathcal{S}$ and the other from mini-batches randomly sampled from $\mathcal{T}$ within the same class. This process is repeated across all classes, and the network's weights θ are updated on either $\mathcal{T}$ or $\mathcal{S}$ for a few iterations, forming an inner optimization loop that runs for j steps. The entire inner loop is then repeated n times, using either randomly initialized or pre-trained networks to capture diverse gradient signals.

Ⅳ. Proposed Subset Matching for Gradient Matching

Many gradient-matching-based methods[6,7,8] minimize the weight gradient difference between synthetic images and randomly sampled mini-batches of the real dataset. However, due to the significant difference in the number of real and synthetic dataset samples, dataset condensation tends to distill aggregate features, which leads to oversimplification in this conventional setting. This oversimplification may cause a loss of important local details. In particular, the gradient updates from real images, which cover a broad range of features, interfere with the updates for synthetic images, leading to inefficient learning. This may harm the preservation of critical local textures and shapes, which are essential for accurate object recognition.

To mitigate these problems, we propose to partition the original and synthetic images and perform matching within subsets to increase the local texture flowing to the synthetic images. Specifically, we initially split the original images and condensed images within each class into K subsets as follows:

$$\mathcal{T} = \mathcal{T}_{[1]} \cup \mathcal{T}_{[2]} \cup \cdots \cup \mathcal{T}_{[K]},$$

$$\mathcal{S} = \mathcal{S}_{[1]} \cup \mathcal{S}_{[2]} \cup \cdots \cup \mathcal{S}_{[K]}.$$

and $\mathcal{T}_{[1]} \cap \ldots \cap \mathcal{T}_{[K]} = \emptyset, \mathcal{S}_{[1]} \cap \ldots \cap \mathcal{S}_{[K]} = \emptyset$. It is noted that we partition the $\mathcal{T}$ and $\mathcal{S}$ only once and keep them fixed during training (i.e., an image belongs to a certain subset throughout the optimization). Then, we aim at jointly minimizing the gradient difference between $\mathcal{T}$ and $\mathcal{S}$, concurrently for each subset within each class. Thus, our total loss function is a combination of the original condensation loss $\mathcal{L}_{all}$ and the proposed subset matching loss $\mathcal{L}_{SM}$ which can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{all} + \lambda\mathcal{L}_{SM}$$
$$= \mathcal{D}\left(\Phi\left(\mathcal{E}(\mathcal{S})\right), \Phi\left(\mathcal{E}(\mathcal{T})\right)\right) +$$
$$\lambda\sum_{k=1}^{K}\mathcal{D}(\Phi(\mathcal{E}(\mathcal{S}_k)), \Phi(\mathcal{E}(\mathcal{T}_k)))$$

where λ controls the strength of the subset matching loss. In practice, we decay λ during training as $\lambda \leftarrow \lambda\left(1-\frac{e}{E}\right)$ where e and E denotes current iteration and total iteration, this strategy prioritizes subset matching in early iterations to distill more local texture initially and progressively shifts towards overall matching to optimize towards best global solution.

It is noted that our subset matching is different from SeqMatch[25]. SeqMatch splits only the synthetic S to subsets, our work differs in that we split not only synthetic images $\mathcal{S}$ but also original images $\mathcal{T}$ and perform gradient matching within the subset which is completely different from SeqMatch. Additionally, SeqMatch modifies the standard evaluation protocol to train with subsets of synthetic dataset sequentially, while we follow the conventional approach without any modifications.

Ⅴ. Experiments

1. Experiment setting

Dataset: We perform distillation using CIFAR-10[26], which has 50,000 images for training and 10,000 images for testing. There are 10 classes, each has a resolution of 32x32.

Architecture: We build upon prior research on dataset condensation[6,7,8] to design the architecture for our condensation process. In particular, we utilize a convolutional neural network (CNN) consisting of 3 layers, each with 128 channels, instance normalization, and RELU activation.

Condensation: For DC[6] and DSA[7] methods, we perform distillation for 1000 iterations. For IDC[8], we perform distillation for 200 iterations. By default, we split the original dataset and synthetic set into two subsets and used a grid search to determine the best value for λ.

Evaluation: We use the same evaluation protocol used in previous works for evaluation. For example, the networks are trained on synthetic images and evaluated on the test set. The networks are trained for 1000 epochs using SGD[27,28] optimizer with a learning rate of 0.01.

Storage budget: We condensed the image with a budget of 10 and 50 IPC when using DC and DSA. For IDC, we also consider 1 IPC scenario. This is because IDC is parameterized DC, which means that given the budget of 1 IPC, we generate 4 synthetic images, which allows us to split into subsets.

2. Main results

Since our framework is designed to be plug-and-play, we integrate it into several gradient-matching-based methods, including DC, DSA, and IDC. We perform distillation using the same hyperparameters used in previous works and summarize the results in

<Table 1> Performance comparison between original method and incorporating our method.

IPC	1	10	50
DC DC+Subset matching	- -	44.3±0.5 **44.6±0.6**	54.7±0.5 **55.1±0.5**
DSA DSA+Subset matching	- -	**52.3±0.3** 52.2±0.4	60.6±0.5 **61.1±0.4**
IDC IDC+Subset matching	44.2±0.3 **44.8±0.3**	63.5±0.4 **66.8±0.4**	73.2±0.4 **75.4±0.26**

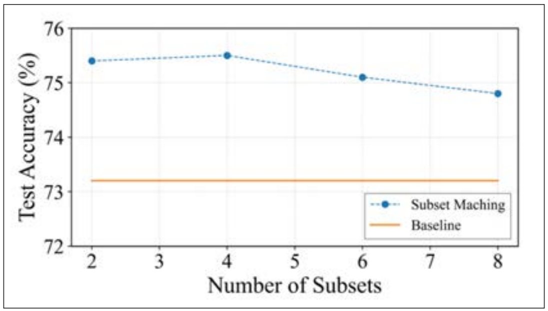
Table 1.

As shown in Table 1, incorporating subset matching loss improves the performance for most cases. Notably, the improvement is more significant when the IPC is larger. For example, using the IDC method for condensation, we achieve only 73.2% at IPC-50. However, by adding subset matching, the performance improves to 75.4%. The experimental results demonstrate the effectiveness of our subset-matching approach.

3. Ablation study

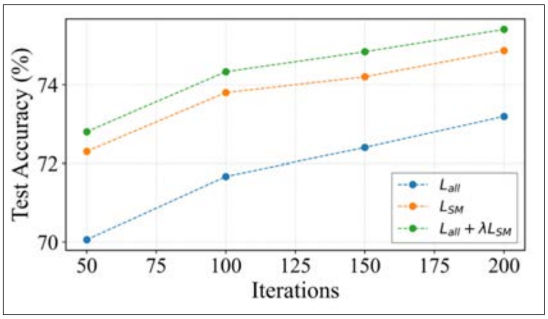
We perform an ablation study on the number of subsets and the effectiveness of subset matching loss. We use IDC to synthesize images with a budget of IPC=50.

Number of subsets: We conduct an ablation study on the impact of the number of subsets on final performance. We observe from Figure 1 that splitting both real and synthetic images into 2 to 4 subsets yields the best performance. However, further increasing the number of subsets degrades performance. We hypothesize that when subsets become too small, the model loses focus on critical texture information, which negatively impacts performance.



<Figure 1> Performance comparison between different number of subsets.

Effectiveness of subset matching: We validate the effectiveness of the proposed loss by comparing the performance of using only subset matching versus the conventional loss. As shown in Figure 2, employing only the subset matching loss results in higher performance than the conventional loss throughout the distillation process. Notably, with just 50 iterations, subset matching nearly reaches the same performance as the conventional method at 200 iterations. Furthermore, combining subset matching loss with the conventional loss yields the best performance, demonstrating the effectiveness of our proposed approach.



<Figure 2> Performance comparison between different losses.

Ⅵ. Conclusion

In this paper, we proposed a subset matching method to improve the performance of gradient matching-based method. Unlike traditional methods that align gradients at the class level, our approach divides both real and synthetic datasets into smaller subsets, allowing for more precise gradient matching. This strategy minimizes gradient interference and enhances the preservation of crucial texture details, such as shapes, edges, and fine-grained features. Through extensive experiments across various budget settings and distillation methods, we show that subset matching consistently enhances condensation performance, particularly when the IPC budget is high. Our results reveal the significance of fine-grained gradient alignment in retaining essential information during dataset condensation, contributing to the development of more efficient and effective synthetic datasets.

참고문헌

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, arXiv:1512.03385, 2015.

[2] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016).

[3] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015).

[4] Yamanaka, J., Kuwashima, S., Kurita, T.: Fast and accurate image super resolution by deep cnn with skip connection and network in network (2020).

[5] Wang, T., Zhu, J., Torralba, A., Efros, A.A.: Dataset distillation. CoRR abs/1811.10959 (2018).

[6] Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. In: International Conference on Learning Representations (2021).

[7] Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 12674-12685. PMLR (18?24 Jul 2021).

[8] Kim, J.H., Kim, J., Oh, S.J., Yun, S., Song, H., Jeong, J., Ha, J.W., Song, H.O.: Dataset condensation via efficient synthetic-data parameterization. In: Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 11102-11118. PMLR (17-23 Jul 2022).

[9] Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022).

[10] Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., You, Y.: Cafe: Learning to condense dataset by aligning features. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12186-12195 (2022).

참고문헌

[11] Zhao, B., Bilen, H.: Dataset condensation with distribution matching (2022). In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2023 (WACV).

[12] Y. Liu, J. Gu, K. Wang, Z. Zhu, W. Jiang, and Y. You, "Dream: Efficient dataset distillation by representative matching,". In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023.

[13] A. Sajedi, S. Khaki, E. Amjadian, L. Z. Liu, Y. A. Lawryshyn, and K. N. Plataniotis, "Datadam: Efficient dataset distillation with attention matching," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2023, pp. 17 097-17 107.

[14] T. Nguyen, R. Novak, L. Xiao, and J. Lee, "Dataset distillation with infinitely wide convolutional networks," in Advances in Neural Information Processing Systems, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[15] Y. Zhou, E. Nezhadarya, and J. Ba, "Dataset distillation using neural feature regression," in Advances in Neural Information Processing Systems, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022.

[16] I. Sucholutsky and M. Schonlau, "Soft-label dataset distillation and text dataset distillation," in International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1-8.

[17] O. Bohdal, Y. Yang, and T. M. Hospedales, "Flexible dataset distillation: Learn labels instead of images," CoRR, vol. abs/2006.08572, 2020.

[18] W. Masarczyk and I. Tautkute, "Reducing catastrophic forgetting with learning on synthetic data," 2020.

[19] A. Rosasco, A. Carta, A. Cossu, V. Lomonaco, and D. Bacciu, "Distilled replay: Overcoming forgetting through synthetic samples," in Continual Semi-Supervised Learning, F. Cuzzolin, K. Cannons, and V. Lomonaco, Eds. Cham: Springer International Publishing, 2022, pp. 104-117

[20] F. P. Such, A. Rawal, J. Lehman, K. Stanley, and J. Clune, "Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data," in Proceedings of the 37th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13-18 Jul 2020, pp. 9206-9216.

[21] J. Goetz and A. Tewari, "Federated learning via synthetic data," CoRR, vol. abs/2008.04489, 2020. [Online]. Available: https://arxiv.org/abs/ 2008.04489.

[22] Y. Zhou, G. Pu, X. Ma, X. Li, and D. Wu, "Distilled one-shot federated learning," CoRR, vol. abs/2009.07999, 2020. [Online]. Available: https://arxiv.org/abs/2009.07999.

[23] G. Li, R. Togo, T. Ogawa, and M. Haseyama, "Dataset distillation for medical dataset sharing," 2022.

[24] T. Dong, B. Zhao, and L. Lyu, "Privacy for free: How does dataset condensation help privacy?" in Proceedings of the 39th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17-23 Jul 2022, pp. 5378-5396.

[25] J. Du and et al, "Sequential subset matching for dataset distillation," in NeurIPS, 2023.

[26] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009

[27] Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing, 16*(5), 1190-1208.

[28] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, 177-186.

저자소개

Linh Tam Tran

- Sep. 2009 ~ Feb. 2014 : He received the bachelor's degree from the Department of Computer Science and Engineering, Ho Chi Minh City University of Technology, Vietnam
- Mar. 2016 ~ Feb. 2018 : He received the M.S. degree from Hongik University, Seoul, South Korea
- Sep. 2020 ~ present : He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Kyung Hee University, Yongin, South Korea.
- Interests: Neural architecture search and dataset condensation

배성호

- 2004.03 ~ 2011.02 : BS. Department of EE and CS, Kyung Hee University (dual major), Korea
- 2011.02 ~ 2016.08 : Ph. D., Department of Electrical Engineering, KAIST, Korea
- 2016.07 ~ 2017.08 : MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) Postdoc. Associate, MA, USA
- 2017.09 ~ present : Associate Professor, School of Computing, Kyung Hee University, Korea
- Interests: Generative AI, model compression, image processing/compression