Dataset Condensation

Likith S G
1RVU23CSE236

RV UNIVERSITY
Go, change the world
an initiative of RV EDUCATIONAL INSTITUTIONS

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING (SoCSE)
RV University
Bengaluru.

OCTOBER-2024

Dataset Condensation

Submitted by

Likith S G
1RVU23CSE236

Under the guidance of

Dr. Shabbeer Basha S H

Center for Visual Computing and Sustainable Intelligence (CVCSI)

Submitted in partial fulfillment of the requirements for
the summer internship conducted during
July-August 2024, as part of the
BTech CSE(Hons.) program at the
School of Computer Science and Engineering.

प्रज्ञा: ★ धीरा:

RV
UNIVERSITY
Go, change the world

an initiative of RV EDUCATIONAL INSTITUTIONS

CERTIFICATE

This is to certify that the work presented in this report, entitled "***Dataset Condensation***", has been carried out by ***Likith S G (USN NO:1RVU23CSE236)*** as part of the summer internship during July-August 2024. The work was completed under my supervision and guidance at the School of Computer Science and Engineering, RV University, Bengaluru, India.

I confirm that the content of this report is the student's original work and has not been submitted elsewhere for the award of any degree or diploma.

Dr. Shabbeer Basha S H
Associate Professor
School of Computer Science and Engineering,
RV University, Bengaluru.

(Signature)

Center Head
Dr. Shabbeer Basha S H
Associate Professor
School of Computer Science
and Engineering, (SoCSE)
RV University, Bengaluru.

(Signature)

Program Head
Dr. Sudhakar K N
Program Director
School of Computer Science
and Engineering, (SoCSE)
RV University,
Bengaluru.

(Signature)

Bengaluru,
October, 2024

CERTIFICATE

I hereby certify that the work presented in this report, entitled “***Dataset Condensation***”, has been carried out by me during my summer internship in July-August 2024. This work was conducted under the supervision of Dr. Shabbeer Basha S H, Associate Professor, School of Computer Science and Engineering, RV University, Bengaluru, India.

I confirm that the work embodied in this report is my own and has not been submitted for the award of any degree or diploma.

Likith S G
1RVU23CSE236
School of Computer Science and Engineering,
RV University, Bengaluru.

Bengaluru,

October, 2024

Acknowledgment:

We would like to extend our sincere thanks to RV University for providing the vital resources and support necessary for this research. Special appreciation goes to **Dr. Shabbeer Basha S H** and **Shobha H R** for their exceptional mentorship and guidance. The accomplishment of this endeavor was made possible by their all the time assistance and important suggestions.

Introduction:

In the current world of huge amounts of data and the growth of machine learning algorithms being trained on these data takes up days or even weeks of training time. This problem is large, and as data grows even further it becomes much harder for humans to train their models through Dataset Condensation: A Solution for This Problem.

In simple terms, dataset condensation is to create smaller versions of original datasets and use them without losing prediction performance. This technique helps reduce training time and resources required but still keeping the same accuracy. Established methods like **Trajectory Matching**, **Kernel Ridge Regression (KRR)**, and **Gradient Matching** have been widely used in this area.

- Trajectory Matching: aligns the training dynamics of models on synthetic data with those trained on real datasets by replicating their training paths.
- Kernel Ridge Regression (KRR): helps manage non-linear data by using a kernel function to map features into a higher-dimensional space, where relationships become easier to model.
- Gradient Matching: ensures the gradients of the synthetic dataset match those of the original dataset during training, maintaining similar learning behavior

We use mutual information to identify and select the most informative samples from the original dataset. This method guarantees that the chosen data points include all of the labels' important characteristics in its sample. Once these key samples are chosen, reinforcement learning is applied to optimize and create synthetic datasets. The reinforcement learning process focuses on maintaining the core attributes of the original data, allowing models to train effectively on smaller, distilled datasets while preserving performance.

Objectives:

The difficulties of training machine learning models on big datasets are the focus of this research. More resources are used and training takes longer with larger datasets. We aim to study a new direction, dataset distillation, producing small-scale synthetic datasets that are able to mimic the performance of models trained on original large-scale real-world datasets. This is very important because it may cut down on the need for huge amounts of computation and massive data sizes (training time).

This research also includes the development of efficient machine learning models with the use of synthetic datasets to achieve optimization. We created models such as MPS1, which combines a convolutional neural network with reinforcement learning to build synthetic datasets in this study. The second model we have developed, MPS1_with_GAN utilizes Generative Adversarial Networks (GANs) to enhance the synthetic data quality even more. These models aim at creating synthetic datasets whose result is on par with the performance of models trained on the entire real datasets.

The first thing that has to be solved when the models are created is to see how well the models perform in a real life situation. It compares different models especially MPS2 considering accuracy, training time, and resource utilization. We evaluate MPS2 over multiple datasets: MNIST, Fashion-MNIST and CIFAR-10 to see how it fares with respect to conventional methods such as Gradient Matching. You are intentionally trying to balance between computational efficiency and model accuracy.

Methodology:

i. Research Design: The research utilizes an experimental framework focused on creating smaller, synthetic datasets from original large datasets. The aim is to determine if these smaller datasets can be used to train models efficiently without substantial loss of accuracy. Three distinct models—MPS1, MPS1_with_GAN, and MPS2—are developed and tested.

ii. Data Selection: In the research, it used four well known datasets: MNIST,Fashion-MNIST,CIFAR-10, and CIFAR-100 as test data. These datasets are well established within machine learning, and span a range of complexity to examine the success of dataset distillation methods. We sample 50 images per class for the distillation process for each dataset. The original datasets are used as benchmarks for assessing the models' performance.

iii. Model Development :

- **MPS1 Model:** This model is a convolutional neural network (CNN) designed for classification tasks. It uses reinforcement learning to generate synthetic datasets. The reinforcement learning agent optimizes the selection of representative samples from the original dataset, reducing the dataset size without losing important features.
- **MPS1_with_GAN:**It builds upon MPS1 by including GAN (Generative Adversarial Network). The role of the GAN here is to generate a higher-quality version of those synthetic datasets by constructing data distributions more accurately, which effectively increases the generalization capability of the model with respect to the synthetic set.
- **MPS2 Model:** MPS2 refines the earlier models by adding early stopping mechanisms to prevent overfitting and improve computational efficiency. MPS2 is designed to handle multiple datasets, including more complex ones like CIFAR-10 and CIFAR-100.
- **MPS2U Model:** MPS2U was designed to enhance training efficiency, accuracy, and versatility across multiple datasets, focusing on reducing training time while maintaining high performance. MNIST and FashionMNIST benefitted from fewer epochs due to their simpler structures. CIFAR-10 and CIFAR-100, with their increased complexity, required more epochs to adequately capture intricate patterns. A Convolutional Neural Network (ConvNet) was utilized to train on synthetic datasets from scratch. The architecture was optimized for effective learning and performance across the selected datasets.

- **Mutual Information:**

The Mutual Information (MI)-Based method built upon MPS2U's principles, incorporating mutual information and reinforcement learning for enhanced sample selection. By employing mutual information, the method identified the top 1,000 samples with the highest scores, ensuring that only the most informative samples were used for training.

Reinforcement Learning Integration: An RL agent was included to optimize sample selection based on model performance, creating a dynamic and iterative improvement process.

Adaptation of MPS2U: The mutual information script was a direct adaptation of MPS2U, retaining its structural advantages while adding the capability of refined sample selection through mutual information.

iv. Training and Implementation: Each model is implemented using Python and relevant machine learning libraries such as TensorFlow and PyTorch. The models are trained on both the original and synthetic datasets. The training process involves optimizing the synthetic datasets using reinforcement learning (in MPS1) or GANs (in MPS1_with_GAN). MPS2 also incorporates early stopping mechanisms to minimize overfitting and reduce training time.

v. Performance Evaluation: The models are evaluated based on key metrics:

- Accuracy: compare the accuracy of the models trained with synthetic vs original datasets. The exercise we would perform here is the models should still be able to adhere to their accuracy, even though with a reduced dataset.
- Training Time: The time required to train each model is measured and compared. The expectation is that models trained on synthetic datasets will take significantly less time to train than those using full datasets.
- Computational Efficiency: The use of computational resources, including GPU usage and memory, is monitored to assess the efficiency of the distillation techniques.

ix. Comparison with Existing Techniques: The results from the models developed in this research are compared with traditional dataset distillation techniques, particularly Gradient Matching. This comparison focuses on the trade-offs between computational efficiency, accuracy, and training time, providing insights into the advantages and limitations of each method.

Implementation

- **MPS2U Implementation**:

  This was an adaptation of the previous script MPS2 which had employed reinforcement learning to create synthetic dataset.

  ConvNet Training: A ConvNet was implemented to train on synthetic datasets, with its architecture designed to adapt easily to various dataset complexities.

  Visualization of Synthetic Data:A directory function was added to save and visualize generated synthetic samples, facilitating analysis of model outputs.

Robust Input Handling: An improved input handling function was developed to preprocess and format data correctly for the ConvNet, enhancing the overall training pipeline.

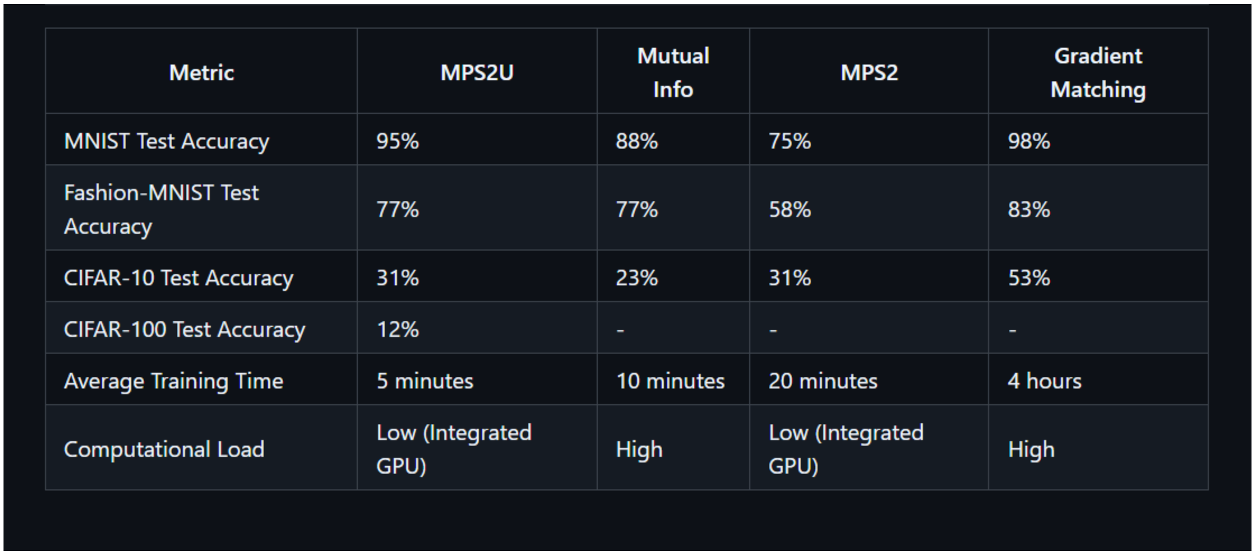
- **Mutual Information-Based Implementation:**

  Mutual Information Scoring:A scoring mechanism was introduced to rank samples based on mutual information, leading to the selection of the top 1,000 samples for training.

  Reinforcement Learning:An RL agent optimized the sample selection process, enhancing training efficiency and effectiveness.

  Integration with TensorFlow:The mutual information approach built upon MPS2U's framework, maintaining TensorFlow’s advantages for efficient computation.

Results



Metric	MPS2U	Mutual Info	MPS2	Gradient Matching
MNIST Test Accuracy	95%	88%	75%	98%
Fashion-MNIST Test Accuracy	77%	77%	58%	83%
CIFAR-10 Test Accuracy	31%	23%	31%	53%
CIFAR-100 Test Accuracy	12%	-	-	-
Average Training Time	5 minutes	10 minutes	20 minutes	4 hours
Computational Load	Low (Integrated GPU)	High	Low (Integrated GPU)	High

Figure 1: Performance metrics comparison of MPS2U and Mutual Info across four datasets: MNIST, FashionMNIST, CIFAR-10, and CIFAR-100. Benchmark algorithms MPS2 and Gradient Matching.

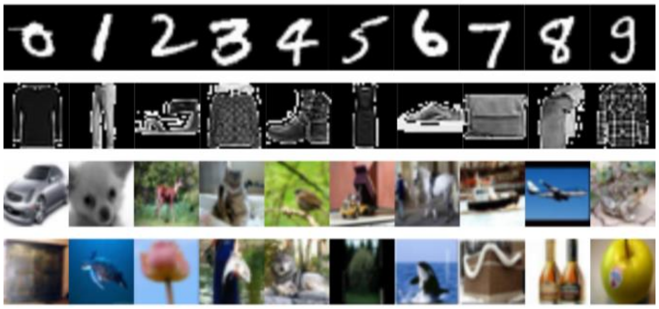


Figure: Visualization of synthetic datasets generated by MPS2U for MNIST, FashionMNIST, CIFAR-10, and CIFAR-100.

- Performance Metrics (MPS2U):

  MNIST: Achieved 95% accuracy with 50 ipc.

  FashionMNIST: Reached 77% accuracy with 50 ipc.

  CIFAR-10: Recorded 31% accuracy with 50 ipc.

  CIFAR-100: Attained 12% accuracy with 50 ipc.

  Average Training Time: Approximately 5 minutes for all four datasets.

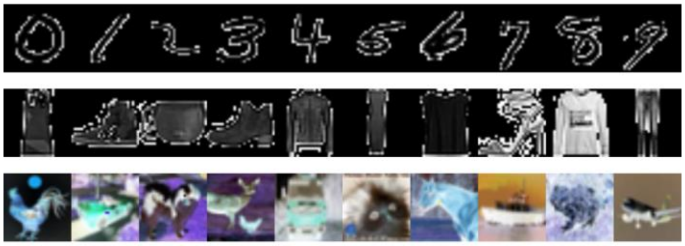


Figure: Visualization of synthetic datasets generated by Mutual Info for MNIST, FashionMNIST, and CIFAR-10.

- Performance Metrics (Mutual Info):

  MNIST: Achieved 88% accuracy with 50 ipc.

  FashionMNIST: Maintained 77% accuracy with 50 ipc.

  CIFAR-10: Obtained 23% accuracy with 50 ipc.

  Average Training Time: Approximately 10 minutes for all three datasets.

  While the Mutual Info script demonstrated notable accuracy, particularly on MNIST, it was slightly more computationally intensive than MPS2U, as reflected in the longer training times.

Conclusion

Through models like MPS2U and MPS2U with Mutual Information, we were able to achieve a significant amount of accuracy with very low computational power. Reinforcement learning played an important role in generating synthetic data and increased model efficiency by huge amounts.

Future work could focus on increasing accuracy on complex datasets like CIFAR-10 and CIFAR-100. We can also use a better implementation of Mutual Information to select the best data points to get a condensed dataset.

References

[1] Cazenavette, G., Wang, T., Torralba, A., Efros, A.A. and Zhu, J.-Y., 2022. Dataset distillation by matching training trajectories. Available at: https://doi.org/10.48550/arXiv.2203.11932 [22 March 2022].

[2] Zhao, B., Mopuri, K.R. and Bilen, H., 2021. Dataset condensation with gradient matching. Available at: https://doi.org/10.48550/arXiv.2006.05929 [8 March 2021].

[3] Singh, R. and Vijaykumar, S., 2023. Kernel ridge regression inference. Available at: https://doi.org/10.48550/arXiv.2302.06578 [19 October 2023].

[4] Li, G., Togo, R., Ogawa, T. and Haseyama, M., 2023. Dataset distillation using parameter pruning. Available at: https://doi.org/10.48550/arXiv.2209.14609 [21 August 2023].

[5] Ding, M., Xu, Y., Rabbani, T., Liu, X., Gravelle, B., Ranadive, T., Tuan, T.-C. and Huang, F., 2024. Calibrated dataset condensation for faster hyperparameter search. Available at: https://arxiv.org/abs/2405.17535 [27 May 2024].

[6] Liu, D., Sepulveda, N. and Zheng, M., 2018. Artificial neural networks condensation: A strategy to facilitate adaptation of machine learning in medical settings by reducing computational burden. Available at: https://doi.org/10.48550/arXiv.1812.09659 [23 December 2018].

[7] Rather, I.H. and Kumar, S., 2024. Generative adversarial network based synthetic data training model for lightweight convolutional neural networks. Multimedia Tools and Applications, 83(1), pp.6249-6271. Available at: https://doi.org/10.1007/s11042-023-15747-6 [Accessed 10 August 2024].

[8] Zhao, B., Mopuri, K.R. & Bilen, H., 2021. Dataset Condensation with Gradient Matching. In International Conference on Learning Representations. Available at: https://openreview.net/forum?id=mSAKhLYLSsl.

[9] Visual computing(VICO). University of Edinburgh 2023. Dataset Condensation. GitHub repository. Available at: https://github.com/VICO-UoE/DatasetCondensation