Dataset Distillation: A Data-Efficient Learning Framework

Swapnil Patel
University of Toronto
swap.patel@mail.utoronto.ca
https://github.com/Swapnil949/ECE1512_2024F_ProjectRepo_SwapnilPatel

***Abstract*— The entire project source can be found at: Github Repository.**

1. Introduction

Dataset Distillation is an emerging technique in machine learning designed to compress large datasets into smaller, synthetic datasets that retain the critical information needed to train neural networks effectively. This approach addresses the growing demand for computational efficiency by significantly reducing memory requirements and training times without compromising model performance. By distilling the essential features of a vast dataset, researchers can train models on a compact, representative subset, thus enabling rapid prototyping and exploration of deep learning architectures. As dataset sizes continue to expand, methods such as dataset distillation offer promising pathways for scalable and resource-efficient machine learning.

Dataset distillation has several applications such as *Neural Architecture Search (NAS)*, *Privacy Preservation*, *Federated Learning*, *Memory-efficient Continual Learning*, and *Data sharing*.

Neural Architecture Search (NAS). In NAS, dataset distillation provides a compact proxy dataset that approximates the performance of training on a full-scale dataset [7]. This enables faster and more resource-efficient model evaluation across architectures, allowing for rapid exploration and selection of optimal models without the computational expense of using the full dataset.

Privacy Preservation. For privacy preservation, distilled datasets reduce exposure to sensitive information by retaining only essential features rather than raw, identifiable data [2]. This allows machine learning models to be trained securely in privacy-sensitive domains (e.g., healthcare and finance) and enhances data sharing compliance in settings like federated learning.

2. Dataset Distillation with Attention Matching

The paper "DataDAM: Efficient Dataset Distillation with Attention Matching" [5] presents a novel dataset distillation approach aimed at reducing training dataset sizes while retaining essential information. The primary challenge addressed by the authors is the high computational cost associated with deep learning models when using full-sized datasets.

In this work, the authors generate a synthetic dataset by employing `Spatial Attention Mapping (SAM)` to capture attention maps from real and synthetic data across various layers within a family of randomly initialized neural networks. This approach alleviates the substantial memory demands typically encountered in state-of-the-art methods. The use of randomly initialized neural network instead of Pre-trained models also makes their solution more versatile and improves cross-architecture generalization results.

Additionally, they introduce a `complementary loss` as a regularization technique to align the last-layer feature distributions between the real and synthetic datasets. Fig. 1 illustrates the novel methodology proposed by the authors.

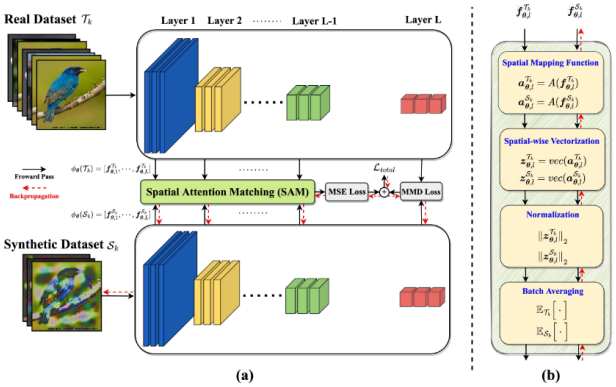

Figure 1. DataDAM Method [5]

With this new method, authors are able to beat existing state-of-the-art solutions by improving performance by 6.5% for CIFAR100 and 4.1% for ImageNet-1k. They also achieve up to a 100x reduction in training costs for learning synthetic dataset.

2.1. MNIST Dataset

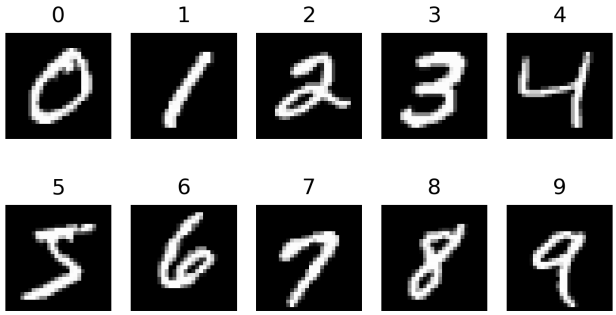

Figure 2. MNIST Dataset [1]

The MNIST dataset is a widely used collection of handwritten digits that is commonly used for training and testing machine learning and computer vision algorithms. MNIST stands for the "Modified National Institute of Standards and Technology" database. It was created by modifying the original NIST dataset, which contained a much larger and more diverse set of handwritten characters, to focus specifically on handwritten digits.

The MNIST dataset contains 28x28-pixel grayscale images of handwritten digits (0 through 9), along with corresponding labels indicating which digit each image represents [1]. There are 60,000 training images and 10,000 testing images in the MNIST dataset, making it a popular benchmark for various image classification tasks.

2.1.1. ConvNet-3.

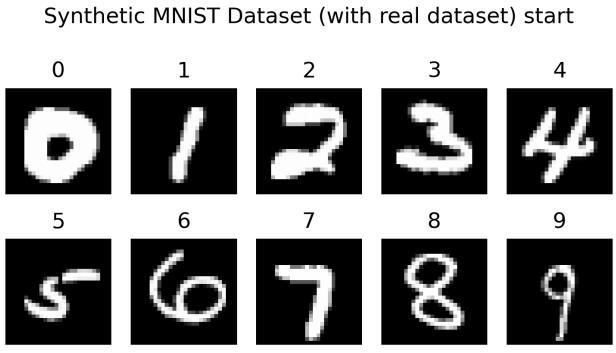

Figure 3. Sample of Synthetic MNIST Dataset created from real images (starting image)

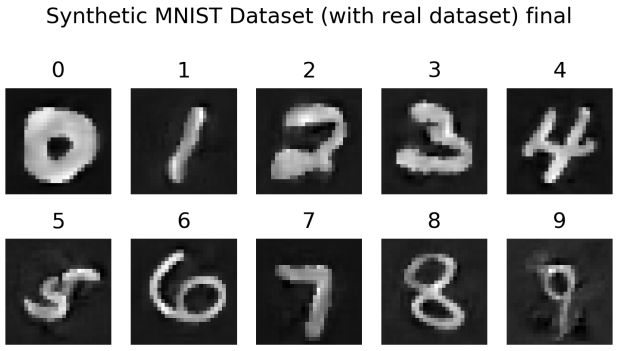

Figure 4. Sample of Synthetic MNIST Dataset created from real images (final image)

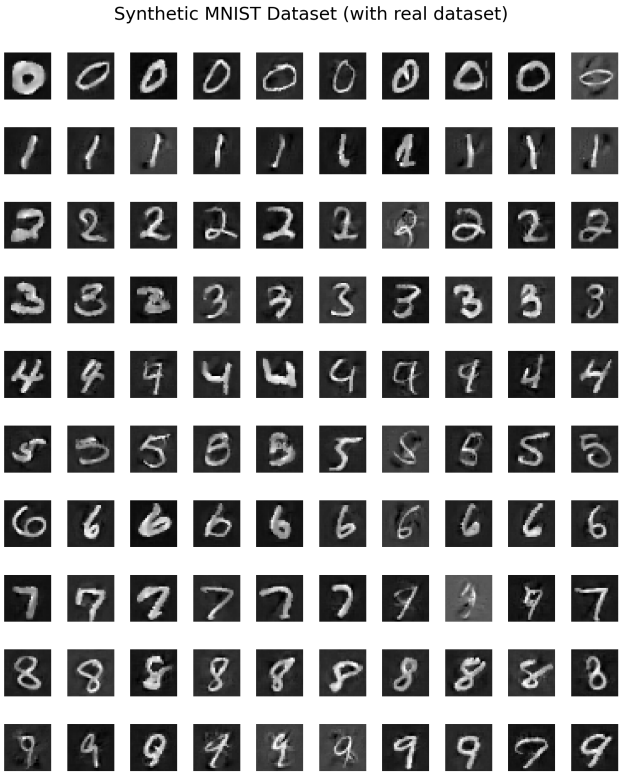

Figure 5. Synthetic MNIST Dataset created from real images

2.1.2. Synthetic Dataset using real images.

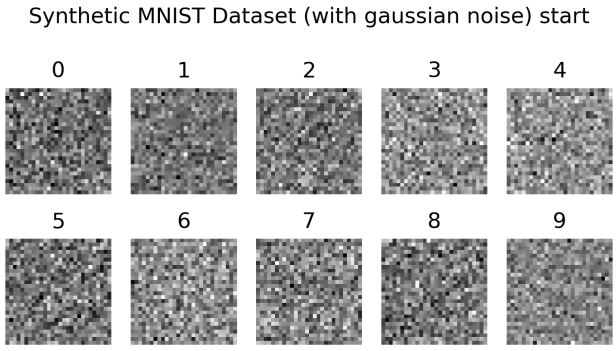

Figure 6. Sample of Synthetic MNIST Dataset created from Gaussian noise (starting image)

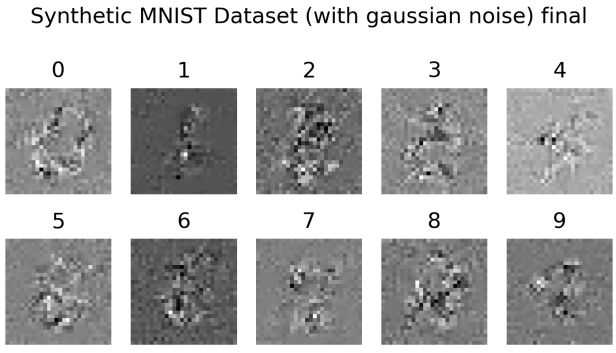

Figure 7. Sample of Synthetic MNIST Dataset created from Gaussian noise (final image)

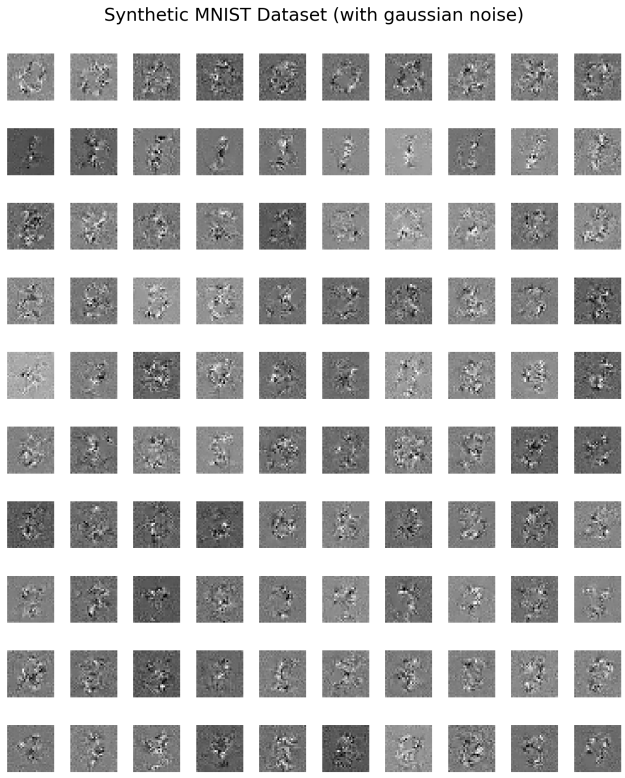

Figure 8. Synthetic MNIST Dataset created from Gaussian noise

2.1.3. Synthetic Dataset using Gaussian noise.

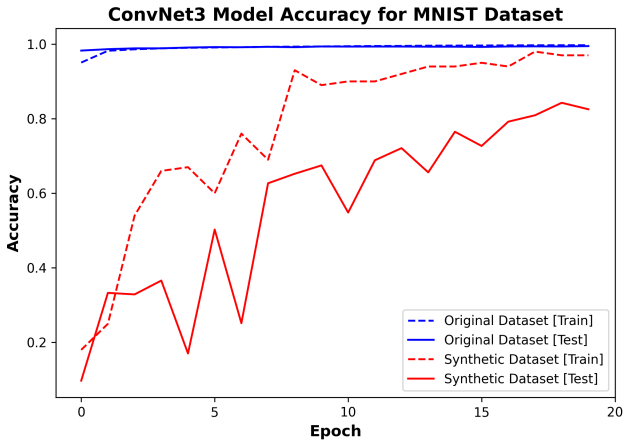

Figure 9. ConvNet-3 Model Trained using Original and Synthetic dataset

2.1.4. ConvNet-3 using Synthetic Dataset.

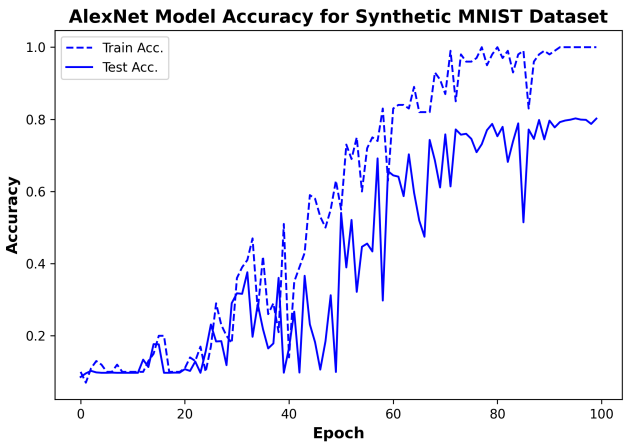

Figure 10. AlexNet Model Trained using Synthetic dataset

2.1.5. Cross-architecture Generalization - AlexNet.

2.2. MHIST Dataset

The MHIST dataset, short for ”Minimalist Histopathology Image Screening Test,” is a specialized dataset commonly used for training and evaluating machine learning models in the field of digital pathology. It was developed to address specific challenges in histopathological image analysis, focusing particularly on the differentiation between types of colorectal polyp tissues.

The MHIST dataset contains 224x224-pixel RGB images of hematoxylin and eosin (H&E) stained histopathology slides, with labels distinguishing between two classes: hyperplastic (non-cancerous) and adenomatous (precancerous) polyps. It consists of 3,152 image patches divided into 2,456 training images and 696 testing images, making it a valuable resource for binary classification tasks in medical image analysis [6].

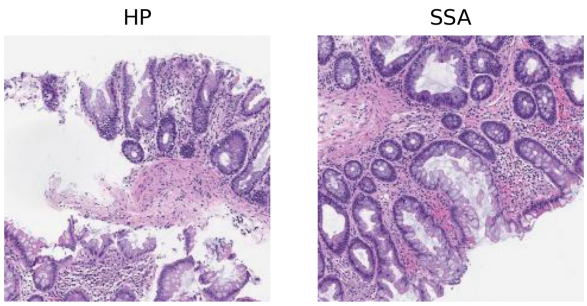

Figure 11. MHIST Dataset [6]

Figure 12. Sample of Synthetic MHIST Dataset created from real images (starting image)

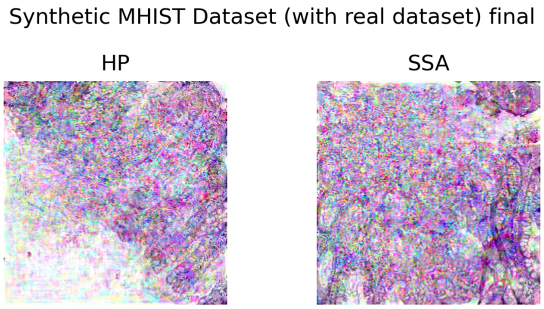

Figure 13. Sample of Synthetic MHIST Dataset created from real images (final image)

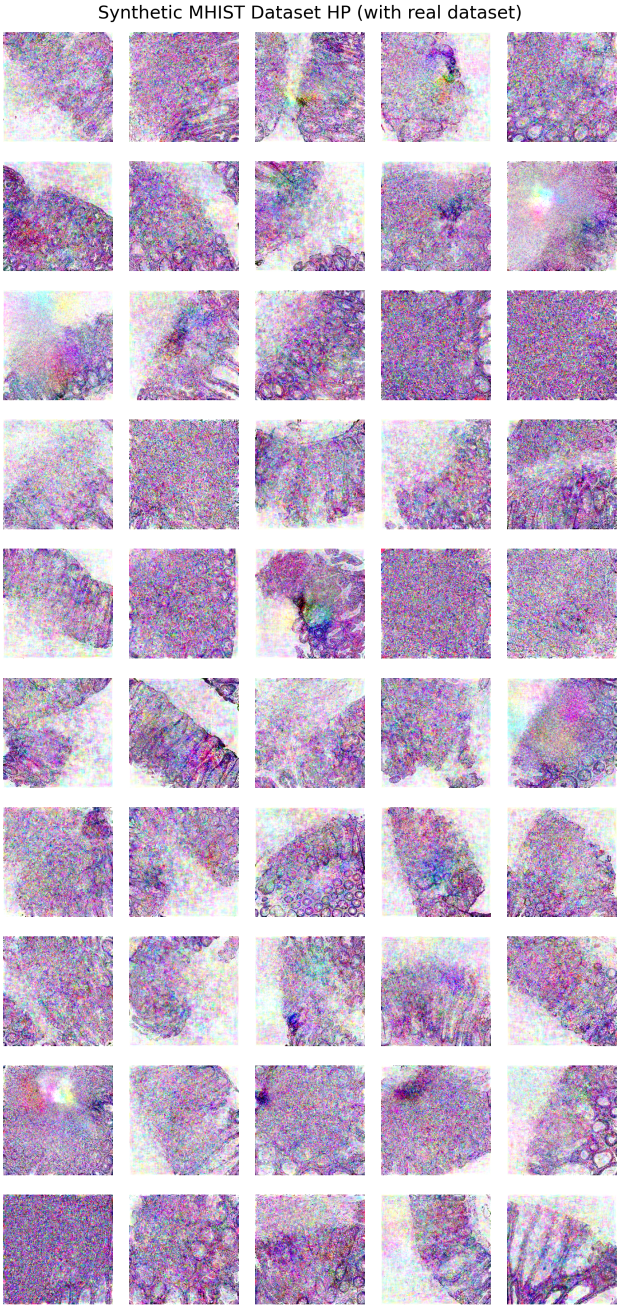

Figure 14. Synthetic MHIST Dataset created from real images [HP]

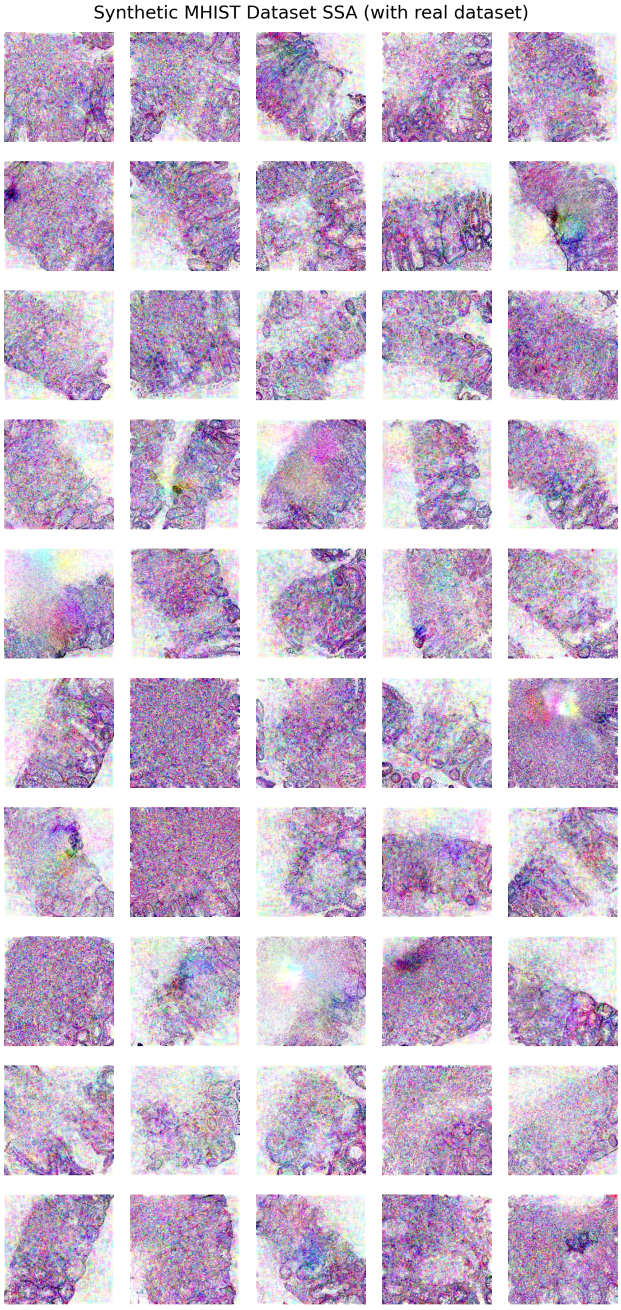

Figure 15. Synthetic MHIST Dataset created from real images [SSA]

2.2.1. Synthetic Dataset using real images.

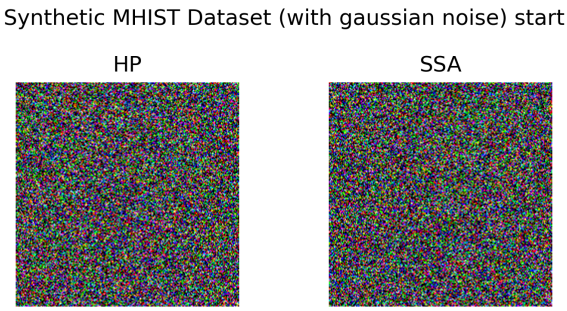

Figure 16. Sample of Synthetic MHIST Dataset created from Gaussian noise (starting image)

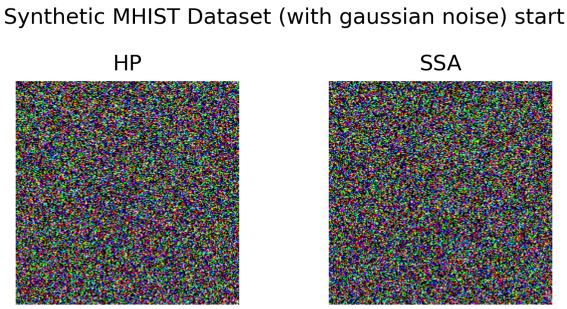

Figure 17. Sample of Synthetic MHIST Dataset created from Gaussian noise(final image)

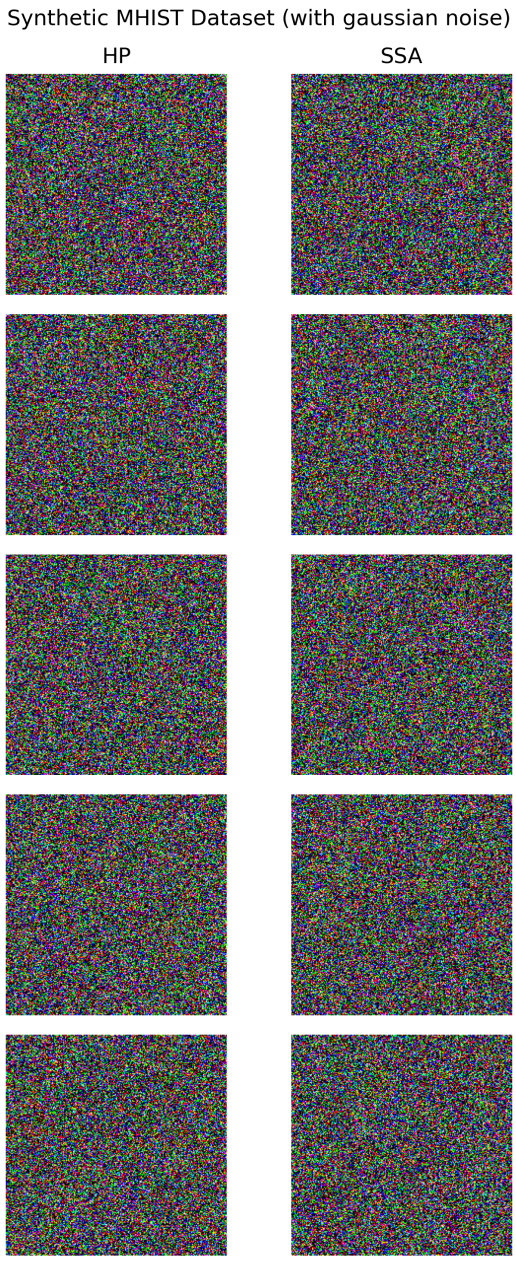

Figure 18. Synthetic MHIST Dataset created from Gaussian noise

2.2.2. Synthetic Dataset using Gaussian noise.

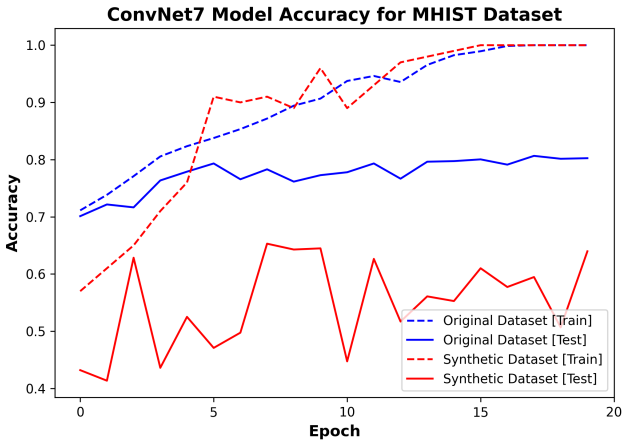

Figure 19. ConvNet-7 Model Trained using Original and Synthetic dataset

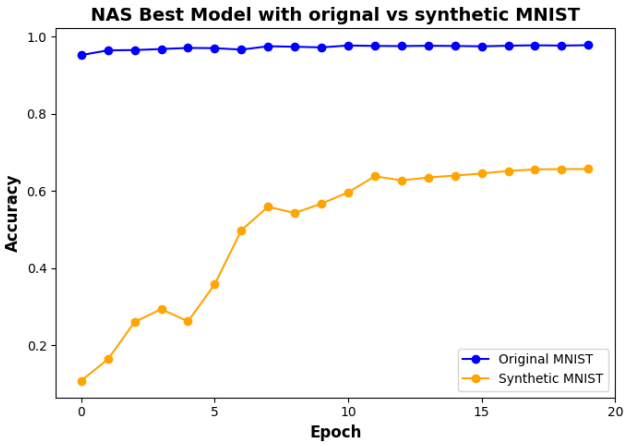

Figure 21. NAS best model with original vs synthetic MNIST dataset

2.2.3. ConvNet-7 using Synthetic Dataset.

3. Prioritize Alignment in Dataset Distillation

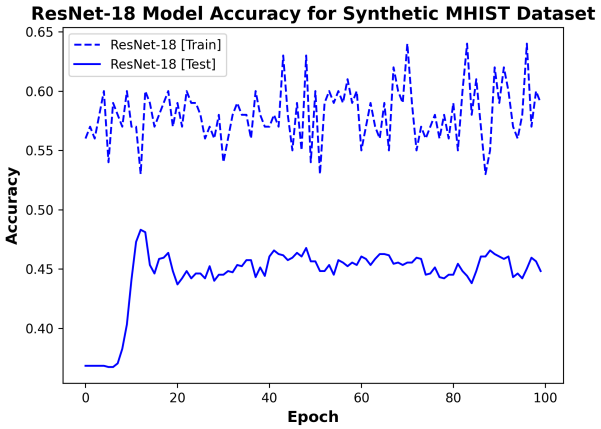

Figure 20. ResNet-18 Model Trained using Synthetic dataset

2.2.4. Cross-architecture Generalization - ResNet-18.

2.3. Data Distillation Application

NAS with original dataset took 1607.3 seconds. with synthetic dataset took 2.8 seconds.

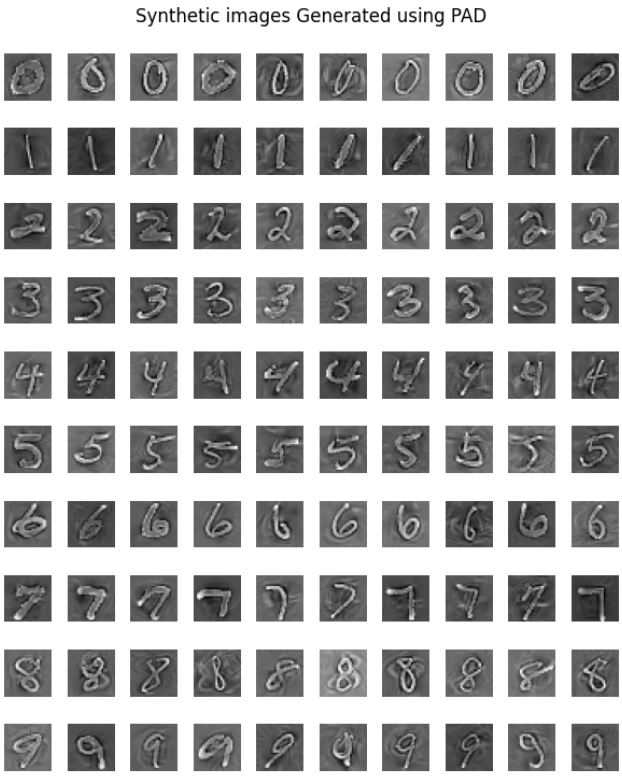

Figure 22. Synthetic MNIST Dataset using PAD[4]

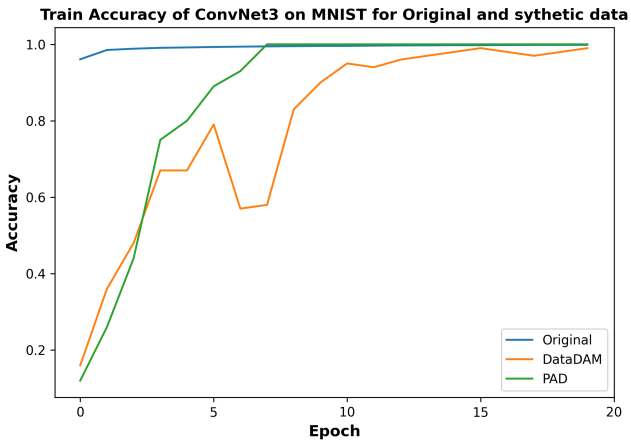

Figure 23.

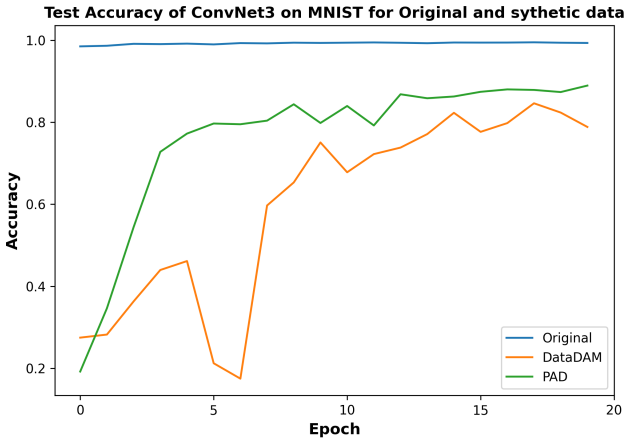

Figure 24.

4. Conclusion

5. References

References

[1] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[2] Tian Dong, Bo Zhao, and Lingjuan Lyu. *Privacy for Free: How does Dataset Condensation Help Privacy?* 2022. arXiv: 2206.00240 [cs.CR]. URL: https://arxiv.org/abs/2206.00240.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[4] Zekai Li et al. *Prioritize Alignment in Dataset Distillation*. 2024. arXiv: 2408.03360 [cs.LG]. URL: https://arxiv.org/abs/2408.03360.

[5] Ahmad Sajedi et al. *DataDAM: Efficient Dataset Distillation with Attention Matching*. 2023. arXiv: 2310.00093 [cs.CV]. URL: https://arxiv.org/abs/2310.00093.

[6] Jerry Wei et al. "A Petri Dish for Histopathology Image Analysis". In: *International Conference on Artificial Intelligence in Medicine*. Springer. 2021, pp. 11–24.

[7] Colin White et al. *Neural Architecture Search: Insights from 1000 Papers*. 2023. arXiv: 2301.08727 [cs.LG]. URL: https://arxiv.org/abs/2301.08727.