ECE1512 Project A Report

Rémi Grzeczkowicz *MScAC Student*
University of Toronto
Student Number: 1010905399
remigrz@cs.toronto.edu

I. Introduction

II. Task 1

This part relies on the paper [2].

A. Part 1

(a) In this paper, the purpose of using Dataset distillation is to reduce the training cost while preserving the performance of the model.

(b) The advantages of their methodology over state-of-the-art are:
- It achieved unbiased representation of the real data distribution.
- It does not rely on rely on pre-trained network parameters or employ bi-level optimization.
- It has a reduced memory cost with a lower run time thanks to the fact that DataDAM does not use an inner-loop bi-level optimization.
- It outperformed other distillation methods except for one case where Matching Training Trajectory (MTT) performed better on CIFAR-10 with 10 Impage Per Class (IPC). MIT got an accuracy of $56.5\% \pm 0.7$ while DataDAM got $54.2\% \pm 0.8$.

(c) The novelty provided by this paper is the use of attention in data distillation. Indeed it has been used in knowledge distillation but never in dataset distillation.

(d) The methodology is as follows:
(1) Initialize a synthetic dataset $\mathcal{S}$ either using random noise or sampling from the original training dataset $\mathcal{T}$.
(2) For each class k a batch B_T^k of real images and a batch B_S^k of synthetic images are sampled from $\mathcal{T}$ and $\mathcal{S}$ respectively.
(3) Then a neural network ϕ_θ is employed to extract features from the images. The network have different layers, each creating a feature map. This multiple feature maps allow to capture low-level, mid-level and high-level representations of the data.
(4) Using the feature maps of each layer, the Spatial Attention Matching (SAM) module generates an attention map for real and synthetic images. The attention map is formulated as $A(f_{\theta,l}^{T_k}) = \sum_{i=1}^{C_l} |(f_{\theta,l}^{T_k})_i|^p$ where $(f_{\theta,l}^{T_k})_i$ is the i-th feature map in the lth layer, C_l is the number of channels and p is a parameter to adjust the weights of the feature maps.
(5) The attention maps for both datasets are then compared using the loss function $\mathcal{L}_{SAM}$.
(6) The output of the network for each dataset is also compared using the loss function $\mathcal{L}_{MMD}$ based on the Maximum Mean Discrepancy (MMD).
(7) The total loss is then given by $\mathcal{L} = \mathcal{L}_{SAM} + \mathcal{L}_{MMD}$.
(8) Then $\mathcal{S}$ is updated such as $\mathcal{S} = arg\min_{\mathcal{S}} \mathcal{L}$.

(e) DataDAM could be used in machine learning for continual learning by providing an efficient memory management method by storing the synthetic data in the memory instead of the real data. This allows for a better memory usage and a lower computational cost. DataDAM could also be used for neural architecture search. Indeed, instead of training many architectures on the full dataset, those architectures could trained on the distilled dataset, leading to a faster search.

B. Part 2

This part relies on the paper [1].

III. Task 2

A. Part 1

(a)

References

[1] Zekai Li, Ziyao Guo, Wangbo Zhao, Tianle Zhang, Zhi-Qi Cheng, Samir Khaki, Kaipeng Zhang, Ahmad Sajed, Konstantinos N Plataniotis, Kai Wang, et al. Prioritize alignment in dataset distillation. *arXiv preprint arXiv:2408.03360*, 2024.

[2] Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Datadam: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17097–17107, 2023.