How Good are Inducing Points for Dataset Distillation ? (Student Abstract)

Shrutimoy Das

Indian Institute of Technology, Gandhinagar
shrutimoydas@iitgn.ac.in

Abstract

Dataset distillation methods learn a representative summary of the full dataset such that training on the distilled data is more efficient in terms of time and space. The current state-of-the-art methods exploit the correspondence between infinitely wide neural networks (NNs) and kernel ridge regression to design distillation methods that result in high-quality summaries of the data. In this work, we leverage the correspondence between infinitely wide networks and Gaussian Processes(GPs) for learning a distilled dataset. We investigate the feasibility of using the inducing points method for Gaussian Processes, as a data distillation method. While most of the existing dataset distillation methods are based on loss or gradient matching, our method looks at the function space approximation, facilitated by the NN-GP correspondence. Additionally, using recent theoretical results on GP regression and neural tangent kernels(NTKs), we also provide an upper bound on the size of the distilled data. We demonstrate the utility of inducing points as distilled data on a set of datasets empirically.

Introduction

One of the main reasons behind the success of deep learning models for a variety of tasks has been the availability of huge volumes of data. As a result, these models demand a huge computational burden both in terms of time as well as storage. This has led to a lot of research in the area of data subset selection for efficiently training neural networks. A closely related line of work has also looked at data distillation (DD) methods which was first introduced by (Wang et al. 2018). These algorithms aim to distill the entire dataset into a much smaller set of synthetic datapoints such that models trained on this smaller dataset have performance similar to models trained on the full dataset.

Dataset distillation algorithms received a further boost with the discovery of the Neural Tangent Kernel (NTK) (Jacot, Gabriel, and Hongler 2018), which revealed a correspondence between neural networks and the well-studied techniques of kernel ridge regression (KRR). This led to the development of DD algorithms that output high-quality synthetic data that perform as well as the full dataset (Loo et al. 2022). The study of infinitely wide neural networks (NNs) also led to the discovery of the correspondence between infinitely wide neural networks and Gaussian Processes (GPs). In this work, we combine the ideas from the inducing points method and the NN-GP correspondence to investigate the feasibility of using these inducing points as distilled data. While the previous DD methods aim to match the training loss or loss gradients, our proposed method aims to approximate the distribution of the underlying function. We present a simple algorithm that builds upon stochastic variational GP methods.

Motivation : Inducing points which are learned by optimizing a lower bound to the full marginal likelihood, contain enough information to approximate the posterior distribution of the underlying latent function space . It indicates that the optimized inducing points are intricately linked with the underlying function of a given dataset. This is in contrast to the loss or gradient matching methods, where the datasets are distilled to match the performance of a given learning algorithm (NNs or KRRs). Thus, the learned inducing points can be utilized as distilled data for learning the underlying function, and it is independent of the learning algorithm used for a given problem.

Our empirical results (for classification) show that the learned inducing points can be used to train just a simple one hidden layer neural network to give high test accuracies on the original dataset. This is advantageous as there is no need to train large complicated architectures on the distilled dataset. Thus, along with reducing the size of the training dataset, our method also facilitates training very simple networks for downstream tasks. Also, the most computationally involved component is the optimization of the lower bound to the full marginal likelihood. For the inducing points method, this takes 1 hour to train, which is much faster than training RFAD, which requires time ranging from $1 - 14$ hours, as mentioned in (Loo et al. 2022).

Latent Space Distillation

We present the pseudocode for our proposed method, which we term as Latent Space Distillation in Algorithm 1. Given the computational inefficiency of GPs for high-dimensional data, we learn low-dimensional representations of the datapoints and perform the sparse GP computations in this low-dimensional latent space. The dimensionality reduction method could either be PCA or a pretrained autoencoder.

Algorithm 1: Latent Space Distillation (LD)

Require: Data (X, y), NN Φ_θ, a dimensionality reduction method, A.

1: $X' \leftarrow A(X)$ /*low dimensional representation of X*/
2: INITIALIZE $f_0 \sim \mathcal{GP}(0, k = \Theta_{\text{NTK}}))$
3: INITIALIZE a set $\mathbf{Z}$ of m inducing points via k-MEANS
4: DEFINE a variational distribution $q(\mathbf{u} \mid \mathbf{m}, \mathbf{X}', \mathbf{Z})$, where $\mathbf{u}$ are the latent function values at $\mathbf{Z}$.
5: Define $\mathcal{L}_0$, an ELBO to the marginal log likelihood of f_0 w.r.t (X', Z, y, u)
6: Compute $\mathbf{Z}_{\text{OPT}} \leftarrow \text{MAXIMIZE}_{\mathbf{Z}}$ ELBO to get the optimal set of inducing points.
return $(\mathbf{Z}_{\text{OPT}}, \mathcal{GP}(\mathbf{Z}_{\text{OPT}}))$ as the distilled data for (X, y)

Given the d' dimensional representations $x' \in X'$, and the corresponding labels y, we define a centered Gaussian process (GP) prior over the latent function f, where $f(x') = y$. We use the NTK of an infinitely wide 2-layer neural network as the covariance function of the GP and maximize a lower bound on the marginal log likelihood (evidence lower bound, ELBO). This optimization step is done using trainable inducing points (initialized using a k-means based method). The learned inducing points are output as the distilled data, and the corresponding posterior means are taken as the labels.

Bounding the number of inducing points : Recent results in GP regression indicate that the number of inducing points required for the convergence of lower bound optimization step, can be bounded. Combined with results that show the similarity of the NTK of a 2 hidden layer network and the Laplace kernel, we can show that loosely $O(n^{\frac{d-1}{d(4+d)}})$ inducing points are sufficient for convergence of the ELBO maximization step.

Inducing Points Initialization

The inducing points are initialized using a k-means type method. Specifically, suppose we want to sample r points per class. Then, we perform a r-means clustering on the set of points for each class. The r cluster centers are then taken as the initial inducing points for that class.

Empirical NTK

In our experiments, we consider the ReLU NTK of an infinitely wide neural network with two layers. For any $x_1, x_2 \in \mathbb{R}^d$, let $\beta = \frac{\langle x_1, x_2 \rangle}{||x_1||_2||x_2||_2}$. Then, the NTK is computed as $\Theta_{ntk}(x_1, x_2) = ||x_1||_2||x_2||_2 k_{ntk}(\beta)$, where $k_{ntk}(\beta) = \frac{1}{\pi}(\sqrt{1-\beta^2} + 2\beta(\pi - \arccos\beta))$.

Results

In Table 1, we present the main results. All of the datasets consists of 10 classes. The column named `IPC` refers to the number of images from each class. Thus, $IPC = 10$ for `MNIST` means that we are sampling 10 images from each class, implying that the distilled dataset is of size 100. That is, we learn a set of 100 inducing points using the LD algorithm and use these 100 points for training a 1-hidden layer neural network, instead of the full 60000 points of $MNIST$. For each of the datasets, the inducing points were learned using 95% of the dataset. We report the test accuracies in the column named LD obtained on test points taken from the original dataset. The column named `RFAD to NN` is the current state-of-the-art (Loo et al. 2022). The reported results are averaged over three runs, with the standard deviations also reported.

Dataset	IPC	LD	RFAD to NN
MNIST	10	95.5 ± 0.1	**98.5 ± 0.1**
	50	96.3 ± 0.6	**98.8 ± 0.1**
Fashion-MNIST	10	**94.5 ± 0.4**	87.0 ± 0.5
	50	**96.4 ± 0.5**	88.8 ± 0.4
CIFAR 10	10	**84.6 ± 1.7**	66.3 ± 0.5
	50	**84.5 ± 1.4**	71.1 ± 0.4
SVHN	10	**93.8 ± 0.1**	74.9 ± 0.4
	50	**93.5 ± 0.2**	80.9 ± 0.3

Table 1: Here, we compare the results of LD with RFAD as the benchmark. 'IPC' refers to images per class while A_{Emb} refers to the low-dimensional embedding method used. We report our results for the best configuration of hyperparameters. For our results, we have aggregated over 3 runs of the experiments. The highest accuracies have been highlighted in bold.

Dataset	IPC	Random	DPP	k-Means
Fashion MNIST	100	92.4	92.9	94.8
CIFAR-10	100	83.6	82.7	84.6

Table 2: Downstream NN test accuracies (hard labels) for data distilled using different inducing point initialization methods.

In Table 2, we show the downstream test accuracies obtained with different initializations of the inducing points. Here, DPP refers to initialization using determinantal point process, using the NTK.

Conclusion and Future Work

In this paper, we investigated the efficacy of inducing points as distilled data for training neural networks. We observed empirically that the inducing points learned can be considered as high-quality distilled data. This set of distilled data almost matches and even gives state-of-the-art results on the datasets shown in the paper. One interesting observation is that simple neural networks are sufficient for training on such sets of distilled data. Since our proposed method computes the inducing points in the latent space, it would be interesting to see if these methods can be generalized to other data modalities as well. A tighter analysis of the size of the distilled data is left as a future direction of research.

Ethical Statement

This work focuses on computing a synthetic dataset for the purpose of improving efficieny. The datasets were taken from repositories that are available publicly and does not have any ethical conflicts.

Acknowledgements

The author would like to thank Prof. Anirban Dasgupta from IIT Gandhinagar and Progyan Das from Microsoft for the helpful discussions regarding this project. Shrutimoy is supported by the Prime Minister's Research Fellowship (PMRF) awarded by the Government of India.

References

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 8580–8589. Red Hook, NY, USA: Curran Associates Inc.

Loo, N.; Hasani, R.; Amini, A.; and Rus, D. 2022. Efficient Dataset Distillation Using Random Feature Approximation. arXiv:2210.12067.

Wang, T.; Zhu, J.; Torralba, A.; and Efros, A. A. 2018. Dataset Distillation. *CoRR*, abs/1811.10959.