Batch Active Learning from the Perspective of Sparse Approximation

Maohao Shen[1*] Bowen Jiang[2*] Jacky Yibo Zhang[3*] Oluwasanmi Koyejo[3]

[1]Massachusettes Institute of Technology [2]University of Pennsylvania [3]Stanford University

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- Dataset label annotations require human expertise and can be costly.

Question

How to find an informative subset from the unlabeled dataset pool for label acquisition such that it can provide the most performance gain after including them into the training dataset?

- To do this, **we find a subset such that its corresponding training loss approximates its full data pool counterpart**.

Problem Formulation

We formulate the batch active learning as **a sparse approximation problem**. Given an ideal loss function with a labeled dataset D_l and an unlabeled dataset D_u:

$$\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\mathcal{D}_l}\ell(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{\theta})+\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}\ell(\boldsymbol{x}_j,\boldsymbol{y}_j^\star;\boldsymbol{\theta})$$

batch active learning finds a subset S of unlabeled data, such that the ideal loss function can be approximated as:

$$\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\mathcal{D}_l}\ell(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{\theta})+\sum_{\boldsymbol{x}_j\in\mathcal{S}}\ell(\boldsymbol{x}_j,\boldsymbol{y}_j^\star;\boldsymbol{\theta}),\quad \text{where } |\mathcal{S}|=b.$$

We generalize the batch active learning as below by considering a sparse and non-negative importance weight w_j for each unlabeled data:

$$\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\mathcal{D}_l}\ell(\boldsymbol{x}_i,\boldsymbol{y}_i;\boldsymbol{\theta})+\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}w_j\ell(\boldsymbol{x}_j,\boldsymbol{y}_j^\star;\boldsymbol{\theta}),\quad \text{where } \|\boldsymbol{w}\|_0=b$$

A good importance weight w is found when two unlabeled data loss functions are close to each other:

$$\tilde{L}_{\boldsymbol{w}}(\boldsymbol{\theta}):=\frac{1}{b}\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}w_j\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\boldsymbol{\theta})\quad\approx\quad\tilde{L}(\boldsymbol{\theta}):=\frac{1}{n_u}\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\boldsymbol{\theta})$$

Since true labels are unknown, we use an estimator of them based on the model trained on all labeled data.

Batch Active Learning from the Perspective of Sparse Approximation

$$\underset{\boldsymbol{w}\in\mathbb{R}_+^{n_u}}{\arg\min}\quad\mathbb{E}_{\mathscr{P}}[q(\tilde{L}-\tilde{L}_{\boldsymbol{w}})]\quad s.t.\quad\|\boldsymbol{w}\|_0=b,$$

where $\mathbb{E}_{\mathscr{P}}$ *stands for the expectation over* $\tilde{\boldsymbol{y}}_j\sim\mathscr{P}(\boldsymbol{x}_j)$ *for* $\forall j\in[n_u]$.

Reference

[1] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745, 2011.
[2] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In Advances in Neural Information Processing Systems, pages 7026–7037, 2019
[3] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489, 2017.
[4] Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. In Advances in Neural Information Processing Systems, pages 6359–6370, 2019.
[5] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671, 2019.

Scan Me for Full Paper

The Proposed Method

The original optimization is intractable, so we transform it into a finite-dimensional sparse optimization problem.
We derive an upper-bound that balances the trade-off between uncertainty (variance) and representation (bias) in a principled way:

$$\begin{aligned}\mathbb{E}_{\mathscr{P}}[q(\tilde{L}-\tilde{L}_{\boldsymbol{w}})]&=\mathbb{E}_{\mathscr{P}}[q(\tilde{L}-\mathbb{E}_{\mathscr{P}}[\tilde{L}]+\mathbb{E}_{\mathscr{P}}[\tilde{L}]-\mathbb{E}_{\mathscr{P}}[\tilde{L}_{\boldsymbol{w}}]+\mathbb{E}_{\mathscr{P}}[\tilde{L}_{\boldsymbol{w}}]-\tilde{L}_{\boldsymbol{w}})]\\&\le\underbrace{\mathbb{E}_{\mathscr{P}}[q(\tilde{L}-\mathbb{E}_{\mathscr{P}}[\tilde{L}])]+\mathbb{E}_{\mathscr{P}}[q(\tilde{L}_{\boldsymbol{w}}-\mathbb{E}_{\mathscr{P}}[\tilde{L}_{\boldsymbol{w}}])]}_{(i):\text{ variance}}+\underbrace{q(\mathbb{E}_{\mathscr{P}}[\tilde{L}]-\mathbb{E}_{\mathscr{P}}[\tilde{L}_{\boldsymbol{w}}])}_{(ii):\text{ approximation bias}}\end{aligned}$$

The bias term becomes immediately tractable:

$$\begin{aligned}(ii)&=q(\mathbb{E}_{\mathscr{P}}[\tfrac{1}{n_u}\textstyle\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\cdot)]-\mathbb{E}_{\mathscr{P}}[\tfrac{1}{b}\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}w_j\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\cdot)])\\&=q((\tfrac{1}{n_u}\textstyle\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}\mathbb{E}_{\mathscr{P}(\boldsymbol{x}_j)}[\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\cdot)])-(\tfrac{1}{b}\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}w_j\mathbb{E}_{\mathscr{P}(\boldsymbol{x}_j)}[\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\cdot)]))\end{aligned}$$

Given a decision $w_i > 0$, its label distribution will be concentrated on the true label offered by the oracle, with a zero corresponding variance:

$$\tilde{\boldsymbol{y}}_j\sim\mathscr{P}_{\boldsymbol{w}}(\boldsymbol{x}_j):=\begin{cases}\mathscr{P}(\boldsymbol{x}_j)&\text{if } w_j=0\\ \delta_{\boldsymbol{y}_j^\star}&\text{if } w_j>0\end{cases},\qquad\text{where } \boldsymbol{w}\in\mathbb{R}_+^{n_u}$$

We reach a more tractable sparse approximation:

$$\underset{\boldsymbol{w}\in\mathbb{R}_+^{n_u}}{\arg\min}\quad\|\boldsymbol{v}-\Phi\boldsymbol{w}\|_2^2-\alpha\sum_{\boldsymbol{x}_j\in\mathcal{D}_u}\mathbf{1}(w_j>0)\cdot\sigma_j^2+\beta\|\boldsymbol{w}-\mathbf{1}\|_2^2\quad s.t.\quad\|\boldsymbol{w}\|_0=b$$

where $\alpha>0$ *is to offer a trade-off between bias and variance, and* β *term is a regularizer. Moreover,* $\boldsymbol{v}:=\frac{1}{n_u}\sum_{j=1}^{n_u}\mathbb{E}_{\mathscr{P}(\boldsymbol{x}_j)}[\boldsymbol{g}_j(\tilde{\boldsymbol{y}}_j)]$, $\Phi:=\frac{1}{b}(\mathbb{E}_{\mathscr{P}(\boldsymbol{x}_1)}[\boldsymbol{g}_1(\tilde{\boldsymbol{y}}_1)],\dots,\mathbb{E}_{\mathscr{P}(\boldsymbol{x}_{n_u})}[\boldsymbol{g}_{n_u}(\tilde{\boldsymbol{y}}_{n_u})])$, *and* $\sigma_j=\frac{1}{n_u}\mathbb{E}_{\mathscr{P}(\boldsymbol{x}_j)}[\|\boldsymbol{g}_j(\tilde{\boldsymbol{y}}_j)-\mathbb{E}_{\mathscr{P}(\boldsymbol{x}_j)}[\boldsymbol{g}_j(\tilde{\boldsymbol{y}}_j)]\|_2]$, *where*

$$\boldsymbol{g}_j(\tilde{\boldsymbol{y}}_j):=\begin{cases}[\dots,(\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\boldsymbol{\theta}_i)-\bar{\ell}),\dots]_{i=1\dots m}^\top,&\bar{\ell}:=\frac{1}{m}\sum_{i=1}^m\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\boldsymbol{\theta}_i)\\ \nabla\ell(\boldsymbol{x}_j,\tilde{\boldsymbol{y}}_j;\boldsymbol{\theta}_0)\end{cases}$$

Each g_j is calculated by sampling posteriors in Bayesian settings or by gradient norms in non-Bayesian settings.

Optimization

We propose greedy and proximal iterative hard thresholding (IHT) optimization algorithms in solving the sparse approximation problem.

Greedy: greedily select the item than can minimize the loss function into a subset, until a given budget is met.
Proximal IHT: Iteratively doing (1) gradient descent to minimize the loss function and (2) projection to satisfy the sparsity constraint.

Time Complexity: If n is the number of data samples and b is the query batch size, greedy algorithm takes O(nb) in time and proximal IHT takes O(nlog(b)), lower than the SOTA method BADGE[5].

Experiments

We experiment our batch active learning framework on image classifications and adapt it to both Bayesian and non-Bayesian neural networks to demonstrate its flexibility. The model is reinitialized and retrained at the beginning of each iteration. It then queries a batch of unlabeled data and its test accuracy is evaluated on multiple random seeds. We implement proximal IHT and greedy optimizations for the sparse approximation.

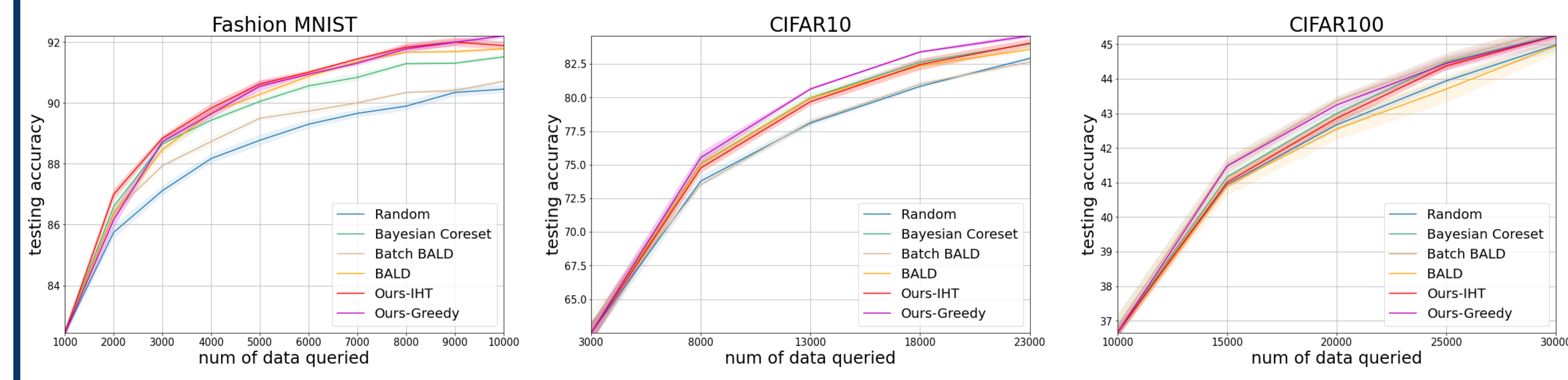

Figure 1: Active learning results on Bayesian models

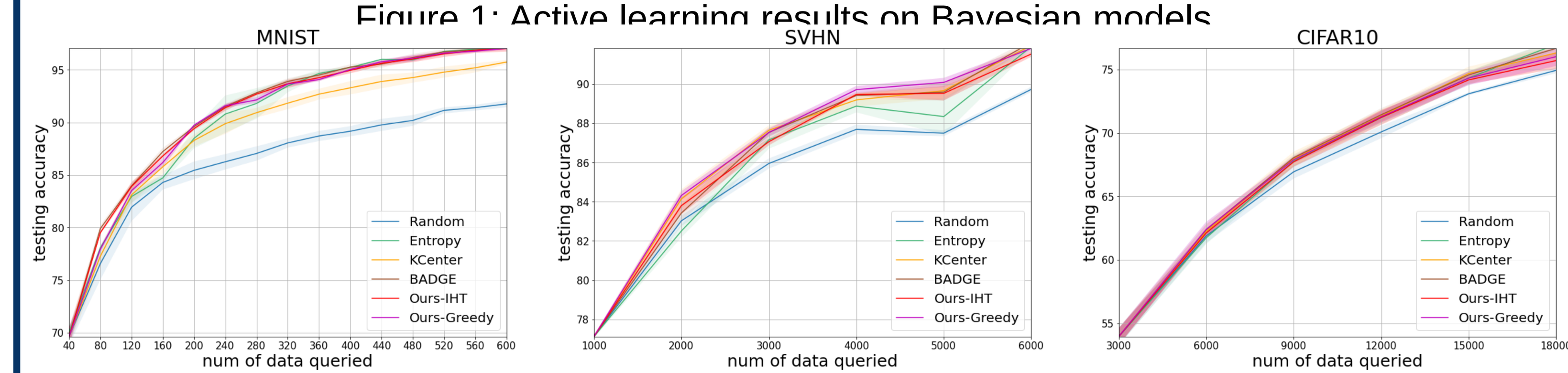

Figure 2: Active learning results on non-Bayesian models

Dataset	Method	Time (unit:s)	Dataset	Method	Time (unit:s)
SVHN	BADGE	732.18 ± 26.29	CIFAR10	BADGE	1207.19 ± 121.09
	Ours-Greedy	201.65 ± 3.54		Ours-Greedy	333.67 ± 3.53
	Ours-IHT	211.04 ± 10.65		Ours-IHT	174.28 ± 3.27
	KCenter	309.99 ± 0.81		KCenter	258.29 ± 1.75
	Entropy	16.46 ± 0.29		Entropy	11.76 ± 0.03
	Random	0.81 ± 0.02		Random	1.42 ± 0.02

Tabel 1: acquisition time on the first query iteration

We compare with Random, BALD[1], Batch BALD [2], and Bayesian Coreset[3] on Bayesian models; Random, Entropy, kCenter[4], and BADGE[5] on non-Bayesian models. **Results show that our methods achieve competitive performance with lower time complexity.**

Summary of Contributions

1. **We propose a flexible batch active learning framework from the perspective of sparse approximation**, adaptable for both Bayesian and non-Bayesian settings.
2. We realize this framework by deriving an upper bound to balance the trade-off between uncertainty and representation in a principled way.
3. We approximate the loss functions that lead to a finite-dimensional, sparsity-constrained, and discontinuous optimization problem.
4. We offer greedy and proximal IHT as two practical appro aches for solving the optimization problem.