Summarising BatchBALD Research Paper

SRIP Task 5

Aditya Nandy
Second Year Undergraduate, Electrical Engineering
IIT Kharagpur

1. Introduction

Deep learning models are based on data acquisition and labelling them efficiently. We can randomly collect and label a large dataset one by one but it is computationally expensive. Hence, it is necessary to take only the most informative points and label them. The goal to AL is to minimise the number of points to be labelled.

Informativeness can be assessed by an *acquisition function*, like *model uncertainty* or *mutual information*. In active learning, batches of data points are acquired and labelled to reduce number of times the model is retrained. In *batch acquisition* we take the top b points which have the highest *acquisition score*.

$$\left\{\boldsymbol{x}_1^*, ..., \boldsymbol{x}_b^*\right\} = \underset{\{\boldsymbol{x}_1,...,\boldsymbol{x}_b\}\subseteq\mathcal{D}_{\text{pool}}}{\arg\max} \; a\left(\{\boldsymbol{x}_1, ..., \boldsymbol{x}_b\}, \mathrm{p}(\boldsymbol{\omega} \mid \mathcal{D}_{\text{train}})\right).$$

2. BALD

BALD is an acquisition function which estimates the mutual information between model predictions and parameters. It acquires individual data points and immediately retrains the model which takes substantial amount of time. BALD sums up over the individual scores and then finds the top b scoring data points.

$$a_{\text{BALD}}\left(\{\boldsymbol{x}_1, ..., \boldsymbol{x}_b\}, \mathrm{p}(\boldsymbol{\omega} \mid \mathcal{D}_{\text{train}})\right) = \sum_{i=1}^{b} \mathbb{I}(y_i ; \boldsymbol{\omega} \mid \boldsymbol{x}_i, \mathcal{D}_{\text{train}}),$$

3. BatchBALD

BatchBALD is an extension of BALD, where we estimate mutual information between multiple data points and model parameters.

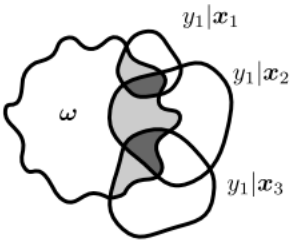
$$a_{\text{BatchBALD}}\left(\{\boldsymbol{x}_1, ..., \boldsymbol{x}_b\}, \mathrm{p}(\boldsymbol{\omega} \mid \mathcal{D}_{\text{train}})\right) = \mathbb{I}(y_1, ..., y_b ; \boldsymbol{\omega} \mid \boldsymbol{x}_1, ..., \boldsymbol{x}_b, \mathcal{D}_{\text{train}}).$$

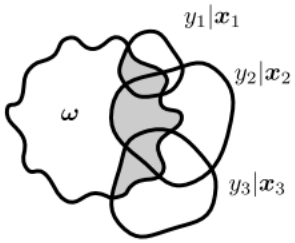
It takes overlaps into account and can get more diverse cover from there.

$$\mathbb{I}(y_1, ..., y_b ; \boldsymbol{\omega} \mid \boldsymbol{x}_1, ..., \boldsymbol{x}_b, \mathcal{D}_{\text{train}}) = \mathbb{H}(y_{1:b} \mid \boldsymbol{x}_{1:b}, \mathcal{D}_{\text{train}}) - \mathbb{E}_{\mathrm{p}(\boldsymbol{\omega}\mid\mathcal{D}_{\text{train}})} \mathbb{H}(y_{1:b} \mid \boldsymbol{x}_{1:b}, \boldsymbol{\omega}, \mathcal{D}_{\text{train}})$$
$$= \mu^*\left(\bigcup_i y_i\right) - \mu^*\left(\bigcup_i y_i \setminus \boldsymbol{\omega}\right) = \mu^*\left(\bigcup_i y_i \cap \boldsymbol{\omega}\right)$$

4. Intuition BALD and BatchBALD

BALD sums over the *acquisition scores* which leads it to overestimate the mutual information. However, BatchBALD takes overlap between the variables which helps to acquire a better cover of model parameters.





$$\sum_i \mathbb{I}(y_i ; \boldsymbol{\omega} \mid \boldsymbol{x}_i, \mathcal{D}_{\text{train}}) = \sum_i \mu^*(y_i \cap \boldsymbol{\omega})$$

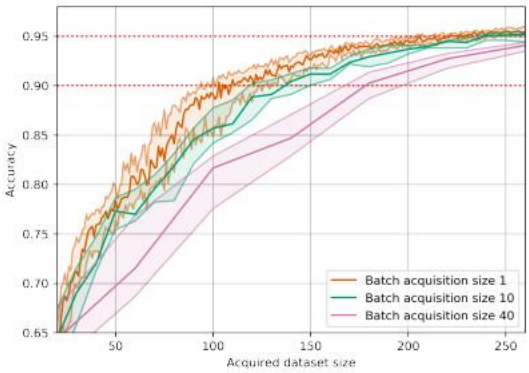
(a) **BALD**

$$\mathbb{I}(y_1, \ldots, y_b ; \boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_b, \mathcal{D}_{\text{train}}) = \mu^*\left(\bigcup_i y_i \cap \boldsymbol{\omega}\right)$$

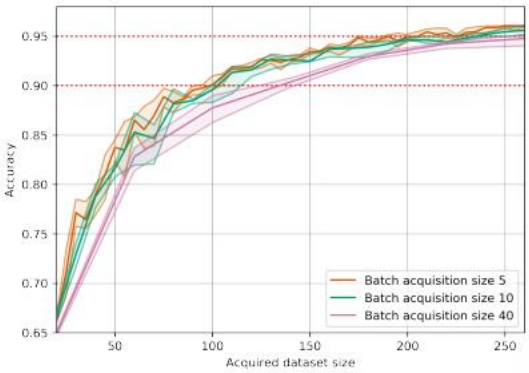
(b) **BatchBALD**

5. Experiments

The BatchBALD outperforms BALD in many aspects. It maintains its performance with increasing acquisition size as shown below for the MNIST Dataset:



(a) **BALD**



(b) **BatchBALD**

6. Limitations

- **Unbalanced datasets:** BALD and BatchBALD don't work well when test set is unbalanced as they don't follow the density of the dataset and try to learn all classes.
- **Unlabelled data:** BatchBALD does not take into account any information from unlabelled dataset
- **Noisy estimator:** A lot of noise is generated from variational approximation for training BNN's. Sampling join entropies also add noise, reducing which will improve model performance.