

Scalable Batch Acquisition for Deep Bayesian Active Learning

Anonymous Author(s)

Abstract—In deep active learning, it is especially important to choose multiple examples to markup at each step to work efficiently, especially on large datasets. At the same time, existing solutions to this problem in the Bayesian setup, such as BatchBALD, have significant limitations in selecting a large number of examples, associated with the exponential complexity of computing mutual information for joint random variables. We, therefore, present the Large BatchBALD algorithm, which gives a well-grounded approximation to the BatchBALD method that aims to achieve comparable quality while being more computationally efficient. We provide a complexity analysis of the algorithm, showing a reduction in computation time, especially for large batches. Furthermore, we present an extensive set of experimental results on image data both on toy datasets and larger ones such as CIFAR-100.

Index Terms—deep active learning, batch acquisition, computer vision

I. INTRODUCTION

In supervised machine learning tasks, the quality and volume of the training data play essential roles in achieving high performance. However, the process of data collection and labeling is often expensive, requiring a huge amount of time and resources [18], [22]. Therefore, active learning (AL) techniques, that choose the most informative samples for model training, minimizing the collection and annotation budget, are crucial in practice [20]. Active learning methods are successfully applied to the various types of data: tabular [28], image [7], text [25], audio [21], video [29] and others. Especially, AL can be helpful in the case of real-world data which requires involvement of subject-matter experts, namely medical [3] and manufacturing data [16].

In this work, we consider a pool-based active learning problem for an image classification task. It is assumed that there is a small amount of labelled data and a big unlabeled pool to select an object for annotation from. Selection procedure is carried out according to a certain criterion that is usually based on a so-called acquisition function. Acquisition function is maximized over the most informative samples in terms of model uncertainty measure or expected error. For example, as an acquisition function one uses variance reduction [9], entropy [23] or mutual information (also known as BALD) [10] maximization and others. Then, selected samples are labeled and added to the already labeled dataset for further training.

At each step of the active learning cycle, one or several pool samples can be selected for annotation. By selecting one sample to label at each step, one can greedily assemble an optimal set of labeled data for training. However, with a large number of objects, it is computationally inefficient to choose

only one object in each acquisition time period. In this case, it is better to select multiple objects from the pool at each active learning step.

Nevertheless, an acquisition of multiple objects at a time, leads to the problem of selecting similar training samples and data redundancy. Thus, the design of the BALD criterion does not take into account the interaction of samples within a batch, which results in the selection of similar samples. The inefficiency of the training dataset leads to model performance degradation and excessive use of resources. One of the state-of-the-art methods that partially copes with this problem is an extension of the BALD method, namely BatchBALD [11]. Its idea is to calculate mutual information in a batch manner using multiple network outputs as a joint random variable. This approach allows one to account for interactions between samples in a batch manner, preventing the selection of similar samples, but greatly increasing computational time and complexity.

In this work, we propose a new active learning acquisition function, *Large BatchBALD*, which is an approximation of the BatchBALD and is designed to deal with the weaknesses of the original, namely its high computational complexity. The main contributions of this work can be summarized as follows.

- We propose an active learning algorithm called Large BatchBALD, which is an approximation of the BatchBALD method. It is designed in the way that prevents taking similar images in one batch, the crucial property of BatchBALD. The details can be found in Section II.
- We analyze the complexity of the proposed algorithm (Section II-E), showing a reduction in running time compared to the original BatchBALD, especially for large batches.
- We provide an extensive experimental study showing the improved efficiency and successful performance of Large BatchBALD in active learning tasks compared to state-of-the-art approaches, see Section III.
- We additionally study how the stochastic sampling can help to further improve the results compared to the greedy approaches, see Sections II-D and III.

II. METHODOLOGY

A. Problem setting

In this paper, we consider a pool-based active learning problem statement for an image classification task. This implies that we have a small amount of labelled examples $\{x_i, y_i\}_{i=1}^N \in D_{\text{train}}$, $y_i \in \{0, 1, \dots, C\}$ and a much larger pool of unlabeled data $\{x_j\}_{j=1}^{N_{\text{pool}}} \in D_{\text{pool}}$. We consider a Bayesian

model M with parameters $\theta \sim p(\theta \mid D_{\text{train}})$. Here conditioning on D_{train} emphasizes that the model was trained using this dataset. Acquired samples from D_{pool} are selected as the ones that maximize a so-called acquisition function A :

$$x^* = \arg \max_{x \in D_{\text{pool}}} A(x \mid M, D_{\text{train}}). \quad (1)$$

It maps each example from the unlabeled pool to a numerical value using some information criteria or uncertainty measure. Acquired samples from the pool are selected for oracle labelling, and then these examples are added to the existing labelled data D_{train} . The target model is then trained on the currently labeled amount of data. This procedure is repeated throughout the active learning cycle and continues until the total budget of the algorithm is exhausted. At each step of the cycle, the acquisition function is recalculated. The quality of the model is evaluated on the test data D_{test} and compared at each step of the AL procedure.

There are various acquisition functions applicable in active learning. One of the basic ones is the Least Confident score:

$$A(x) = 1 - \max_{c \in C} p(y = c \mid x, \theta), \quad (2)$$

where θ are model parameters. That is, the choice of examples is based on whose likelihood the model is most uncertain.

In this work, we use both MC-dropout and deep ensembles approaches to estimate for the more accurate capturing of uncertainty for deep neural networks. In the case of deep ensembles, the same model with different initialization forms an ensemble, in the case of MC-dropout, dropout on inference is used to obtain a set of networks with different parametrization using different dropout masks. Formally, the final prediction can be written as follows:

$$\bar{p}_k(y = c \mid x, D_{\text{train}}) = \frac{1}{k} \sum_{i=1}^k p(y = c \mid x, \theta_i), \quad (3)$$

where θ_i is the model parameters for the i -th model and k is the number of models in a set. It can be either the number of network initializations in the case of deep ensemble, or the number of forward passes in the case of MC-dropout. Both MC-dropout and ensembling can be seen as instances of the general Bayesian formulation with model parameters being samples from the posterior distribution: $\theta_i \sim p(\theta \mid D_{\text{train}})$, $i = 1, \dots, k$.

While one can directly use the averaged predictive distribution (3) in least confident acquisition function (2), it might be beneficial to extract some additional information from the posterior on top of the predictive mean. Next, we will focus on the family of entropy-based acquisition functions that allow to achieve that.

B. Entropy-based acquisition functions

1) *Entropy*: The entropy maximization criterion is also one of the basic ways of selecting examples for active learning:

$$\begin{aligned} & H[y \mid x, D_{\text{train}}] \\ &= - \sum_{c=1}^C \bar{p}_k(y = c \mid x, D_{\text{train}}) \cdot \log \bar{p}_k(y = c \mid x, D_{\text{train}}). \end{aligned} \quad (4)$$

It again uses the predictive mean only by looking on the entropy of this distribution. This criterion is maximized for samples having similar probabilities predicted for the different classes.

2) *Bayesian Active Learning by Disagreement (BALD)*:

Starting from the entropy, one can construct criteria that look on the disagreement between the models in the Bayesian framework. The original BALD criterion [10] is formulated as the conditional mutual information $I(\theta, y \mid x, D_{\text{train}})$ between unknown (unobserved) output y and latent parameters θ , conditioned on input variable x and observed data D_{train} . Note, that the BALD criterion can be written in the y -space of mutual information and expressed as follows:

$$I_{\text{BALD}}(y; \theta) = H[y \mid x, D_{\text{train}}] - \mathbb{E}_{\theta \sim p(\theta \mid D_{\text{train}})} [H[y \mid x, \theta]], \quad (5)$$

where $H[y \mid x, D_{\text{train}}]$ is an entropy of model output y conditioned on data sample x and train data, $H[y \mid x, \theta]$ is an entropy of model output y conditioned on data sample x and sampled model latent model parameters $\theta \sim p(\theta \mid D_{\text{train}})$ which are integrated out by the expectation $\mathbb{E}_{\theta \sim p(\theta \mid D_{\text{train}})}$. In other words, it calculates the difference between the entropy of marginal predictive distribution and posterior mean conditional entropy. BALD intuition is that it seeks for data samples in whose outputs y the model is the most uncertain (leads to high marginal entropy), while being certain about individual model parameters (leads to confident predictions but highly diverse). In general, a continuous case BALD acquisition function is expressed as a KL divergence between $p(y_i; \theta)$ and $p(y_i)p(\theta)$:

$$I(y_i; \theta) = \int_{\theta} \int_{y_i} p(y_i; \theta) \log \frac{p(y_i; \theta)}{p(y_i)p(\theta)} dy_i d\theta. \quad (6)$$

In terms of batch active learning, when an acquisition batch consists of $b > 1$ data samples, then the BALD score is the sum of individual scores for each of the b items:

$$I_{\text{BALD}}(y_{1:b}; \theta) = \sum_{i=1}^b I(y_i; \theta). \quad (7)$$

This approach has a serious drawback, namely, it does not take into account pairwise interaction of the data samples in batches. As a result, BALD tends to acquire a batch of similar examples leading to suboptimal performance.

3) *BatchBALD*: To diversify samples in a batch, the BatchBALD acquisition function was proposed [11]. It is formulated as mutual information between a batch of observations and latent parameters, and can be expressed as:

$$\begin{aligned} I_{\text{BB}}(y_{1:b}; \theta) &= H[y_1, \dots, y_b \mid x_1, \dots, x_b, D_{\text{train}}] \\ &- \mathbb{E}_{\theta \sim p(\theta \mid D)} [H[y_1, \dots, y_b \mid x_1, \dots, x_b, \theta]], \end{aligned} \quad (8)$$

where $y_{1:b} = y_1, \dots, y_b$ is a joint random variable, b is the batch acquisition size. BatchBALD calculates mutual information between model output and model parameters but in a batch sense, that is, considering inter-variable correlation and taking a batch of outputs as a joint random variable. It avoids multiple accounting of variable interconnections and provides diverse data sampling. In terms of the continuous

general case, it is similar to BALD with the difference that $p(y_{1:b})$ is used instead of $p(y_i)$:

$$I(y_{1:b}; \theta) = \int_{\theta} \int_{y_{1:b}} p(y_{1:b}; \theta) \log \frac{p(y_{1:b}; \theta)}{p(y_{1:b})p(\theta)} dy_{1:b} d\theta. \quad (9)$$

While accounting nicely for the correlation between observation, BatchBALD criterion is often computationally expensive, especially for large batches (see Section II-E for the detailed complexity analysis). In the next section, we are going to provide its more computationally feasible alternative.

C. Large BatchBALD

One of possible generalizations of mutual information is a total correlation [31] between b random variables which, by definition, is calculated as:

$$C(y_{1:b}) = \int p(y_{1:b}) \log \frac{p(y_{1:b})}{\prod_{i=1}^b p(y_i)} dy_{1:b}. \quad (10)$$

It measures inter-variable dependencies, always positive and is nullified if and only if all the variables are independent of each other. Note that its form doesn't include model latent parameters θ .

Another possible form of total correlation [26] is an expression that uses mutual information of all possible variable subscripts:

$$C(y_{1:b}) = \sum_{i \neq j}^b I(y_i; y_j) + \sum_{i \neq j \neq k}^b I(y_i; y_j; y_k) + \dots + I(y_1; y_2; \dots; y_b). \quad (11)$$

The main idea of introducing of total correlation in this work is that it is exactly equals to the difference between BALD and BatchBALD acquisition functions:

$$\sum_{i=1}^b I(y_i; \theta) - I_{\text{BB}}(y_{1:b}; \theta) - C(y_{1:b}) = 0, \quad (12)$$

where $C(y_{1:b}) = C(y_1; y_2; \dots; y_b)$ is the mutual information of b random variables (i. e., generalization of mutual information of two variables), b is an acquisition batch size. A complete derivation of equation (12) can be found in the Appendix VI-A. Calculation of mutual information of all possible subscripts of data outputs is significantly time-consuming. Nevertheless, it can be neglected using total correlation approximation just by pairwise mutual information components:

$$\hat{C}(y_{1:b}) = \sum_{i \neq j}^b I(y_i; y_j). \quad (13)$$

In this sense, BatchBALD is equal to the difference between BALD and total correlation. Using the approximation of total correlation, we can write:

$$\begin{aligned} I_{\text{BB}}(y_{1:b}; \theta) &= \sum_{i=1}^b I(y_i; \theta) - C(y_{1:b}) \\ &\approx \sum_{i=1}^b I(y_i; \theta) - \sum_{i=1}^b \sum_{j \neq i}^b I(y_i; y_j). \end{aligned} \quad (14)$$

Note, that there are no latent parameters θ in the joint mutual information $C(y_{1:b})$. That means, that the difference, between BALD mutual information taking into account pair interactions and without it, is exactly the same as how the components y_i are correlated between each other. We call this approximation $I_{\text{LBB}}(y_{1:b}; \theta) := \sum_{i=1}^b I(y_i; \theta) - \sum_{i=1}^b \sum_{j \neq i}^b I(y_i; y_j)$ as Large BatchBALD (LBB). Importantly, LBB is significantly less computationally expensive compared to BatchBALD, see the complexity analysis in Section II-E.

D. Stochastic acquisition function extension

While BatchBALD and its modifications are efficient in obtaining diverse batches, one can propose alternative strategies to achieve a similar effect. One natural way is to step aside from greedy sampling and introduce stochasticity into the procedure. The idea is to convert the resulting scores to a distribution and then sample from it. In this case, we raise the scores to some power α and normalize the resulting values. Thus, the probability of selecting a sample x with an acquisition function equal to $A(x)$ from the unlabeled pool D_{pool} can be written as

$$p_{\alpha}(x) = \frac{A^{\alpha}(x)}{\sum_{x_i \in D_{\text{pool}}} A^{\alpha}(x_i)}. \quad (15)$$

In this work we consider such extension for the LBB and BALD algorithms, and call them Power Large BatchBALD (PLBB) and PowerBALD (PBALD) [12]. Thus, in the case of PLBB acquisition function is $A(x) = I_{\text{LBB}}(y; \theta)$, and in the case of PBALD it is $A(x) = I_{\text{BALD}}(y; \theta)$. Here y is the model output on the sample x and θ is the vector of model parameters. The magnitude of the power α in this case determines how much of a stochastic effect is present: with a smaller power the random effect is greater, with a greater power examples with larger scores are even more likely to be taken, and the random effect appears less.

E. Computational complexity

In this section, we discuss the computational complexity for the BALD, BatchBALD, and Large BatchBALD algorithms, see Table I. We also give some intuition how LBB, using BALD and pairwise mutual information, can significantly improve the complexity of the BatchBALD algorithm, especially noticeable in the case of large batches.

1) *BALD*: BALD time complexity is $\mathcal{O}((b+k) \cdot |D_{\text{pool}}|)$ and consists of calculating for each element of D_{pool} 2 components: $\mathcal{O}(b \cdot |D_{\text{pool}}|)$ is a cost to compute predictive distribution and $\mathcal{O}(k \cdot |D_{\text{pool}}|)$ is a cost to compute entropies in output space.

2) *BatchBALD*: To compute the exact joint entropies, we have to compute all possible configurations of the $p(y_1, \dots, y_b)$ and evaluate by averaging over $p(y_1, \dots, y_b | \theta)$. To compute approximate joint entropies, we have to sample possible configurations of the y_i from $p(y_1, \dots, y_b)$ stratified by $p(\theta)$ and evaluate $p(y_1, \dots, y_b)$ by averaging over $p(y_1, \dots, y_b | \theta)$.

BatchBALD complexity is $\mathcal{O}(b \cdot c \cdot \min\{c^b, m\} \cdot |D_{\text{pool}}| \cdot k)$, where c is the number of classes, k is the number of MC-dropout samples, and m is the number of MC-sampled

configurations of $y_{1:b-1}$, $|D_{\text{pool}}|$ is a volume of unlabeled pool data. It can be described as follows. On each of $i = 1 : b$ the acquisition steps, a new candidate x_i with $p(y_i | \theta)$ from $|D_{\text{pool}}|$ is greedily selected to the already formed batch $p(y_{1:i-1} | \theta)$ of elements $x_{1:i-1}$. This batch is already calculated and stored, elements $x_{1:i-1}$ are fixed, so the task is to calculate joint entropy between a new added point x_i and an existed batch in a one by one manner. In exact (means based on given draws of θ) joint entropy scenario, all possible combinations $y_{1:i-1}$ can be calculated exactly as c^i meaning c possible classes of each of i elements in a batch. As for approximated joint entropy scenario, if c^i value is big (in BatchBALD paper it is assumed after 5 acquired elements) then $p(y_{1:i-1})$ of $y_{1:i-1}$ is approximated using m MC-samples. In both cases, joint probability $p(y_1, \dots, y_i)$ is calculated by averaging over $p(y_1, \dots, y_i | \theta)$ (i. e., to find a probability density marginalizing over θ) with k MC-dropouts of θ . So, batch is selected in linear time, although joint probability still requires a lot of computation resources both in exact and approximate setting.

In a naive setting, complexity is $\mathcal{O}(c^b \cdot |D_{\text{pool}}|^b \cdot k)$ and can be described as follows. For every element of a batch of size b , from a data pool $|D_{\text{pool}}|$ with c possible classes, a new data sample is searched as a maximum of difference between joint entropy and conditional joint entropy. Difference with efficient implementation is that in efficient option $x_{1:i-1}$ is fixed and varies only x_i while in naive implementation all $x_{1:i}$ vary. $p(y_1, \dots, y_b)$ is also calculated by averaging over $p(y_1, \dots, y_b | \theta)$ with k MC-samples.

F. Large BatchBALD

Large BatchBALD time complexity is equal to the sum of BALD complexity and total correlation approximation complexity. BALD complexity, as noted above, is $\mathcal{O}((b+k) \cdot |D_{\text{pool}}|)$. Total correlation approximation complexity is $\mathcal{O}\left(2 \cdot \frac{|D_{\text{pool}}| \cdot |D_{\text{pool}} - 1|}{2} \cdot k \cdot c\right) = \mathcal{O}(|D_{\text{pool}}|^2 \cdot k \cdot c)$. That is, it consists of two tensor multiplication operations with dimensionality $[n, k, C] \times [n, k, C] = [n, n, k, C]$ and $[n, C] \times [n, C] = [n, n, C]$, where n is a processing batch size, see Appendix VI-B for details. It means, that Large BatchBALD complexity is $\mathcal{O}(|D_{\text{pool}}|^2 \cdot k \cdot c + |D_{\text{pool}}| \cdot (b+k))$.

The intuition of this method lies in the following. The given asymptotics denotes a linear dependence on the size of the batch, as in BALD. Thus, Large BatchBALD scales to the size of a batch consisting of hundreds of elements without significant costs. At the same time, the original BatchBALD algorithm works in a reasonable time only with batches consisting of tens of elements due to the calculation of mutual information of a joint random variable. In general, adding large batches in an active learning problem is a common practical scenario. With a huge pool of unlabeled data, adding a small amount of data up to a few tens will have little effect on the performance of the final model. Thus, it is computationally more efficient to be able to acquire large batches for annotation and training. Furthermore, LBB works

equally well with batches of tens of elements and already shows computational superiority in comparison with BatchBALD, see Table II.

III. EXPERIMENTS

For all datasets, the initial training set is balanced with respect to the number of images of each class. All datasets with the repetition option use each incoming image more than once with Gaussian noise applied, in our case we take 4 occurrences of each image in the datasets.

After each addition of new samples to the training set, the network is trained from scratch. All models use Glorot initialization. The parameters of MC-dropout experiments are similar to the [11] settings, deep ensembles experiments are performed with an ensemble of 5 models.

We measure the accuracy of the model prediction on a test dataset, depending on the amount of training data obtained by different algorithms. All results given are obtained as the average of the 5 runs, and the corresponding standard deviation is shown in the figures as filled error bars.

A. MNIST and its variations

1) *Experimental settings:* The first group of experiments is MNIST extensions: MNIST [15], Repeated MNIST (RMNIST) [11], Fashion MNIST (FMNIST) [32]. MNIST is a standard machine learning dataset suitable for active learning that consists of handwritten digit images, including 60,000 images from 10 classes. RMNIST is an extension of the MNIST dataset in which each image is repeated several times with a small Gaussian noise applied. FMNIST is a fashion product dataset containing 70,000 images of 10 classes.

Model architecture for MNIST, RMNIST and FMNIST datasets is taken similar as in [11] for the MC-dropout uncertainty case. For experiments with deep ensembles, the same architecture was adapted to use multiple initializations of the same network to form an ensemble. Comparing deep ensembles with MC-dropout, ensembles are more time-consuming than MC-dropout. While MC-dropout requires multiple forward passes on inference of the same network, for ensembles one needs to fully train multiple networks. Nevertheless, with deep ensembles, better model quality can be achieved due to better calibration of the resulting models [2]. On the mentioned datasets, we compared the performance of the AL algorithms for both ensembles and MC-dropout uncertainty estimates on acquisition batches of 10 images.

2) *Ensembles:* The results of the experiment on the MNIST dataset are shown in Fig. 1a. The Large BatchBALD algorithm is slightly better than the BALD algorithm, and as a BatchBALD approximation it is quite close to the original. As for the LBB and BALD extensions, namely PLBB and PBALD, they dominate among other algorithms, even outperforming the BatchBALD algorithm in both the ensemble and MC-dropout cases.

In RMNIST experiments, see Fig. 1b, BALD takes many similar images that do not introduce significant diversity into the training dataset, which is reflected in a loss of quality

BALD	BatchBALD	Large BatchBALD
$\mathcal{O}((b+k) \cdot D_{\text{pool}})$	$\mathcal{O}(b \cdot c \cdot \min\{c^b, m\} \cdot D_{\text{pool}} \cdot k)$	$\mathcal{O}(D_{\text{pool}} ^2 \cdot k \cdot c + D_{\text{pool}} \cdot (b+k))$

Table I: Complexity of BALD-based algorithms.

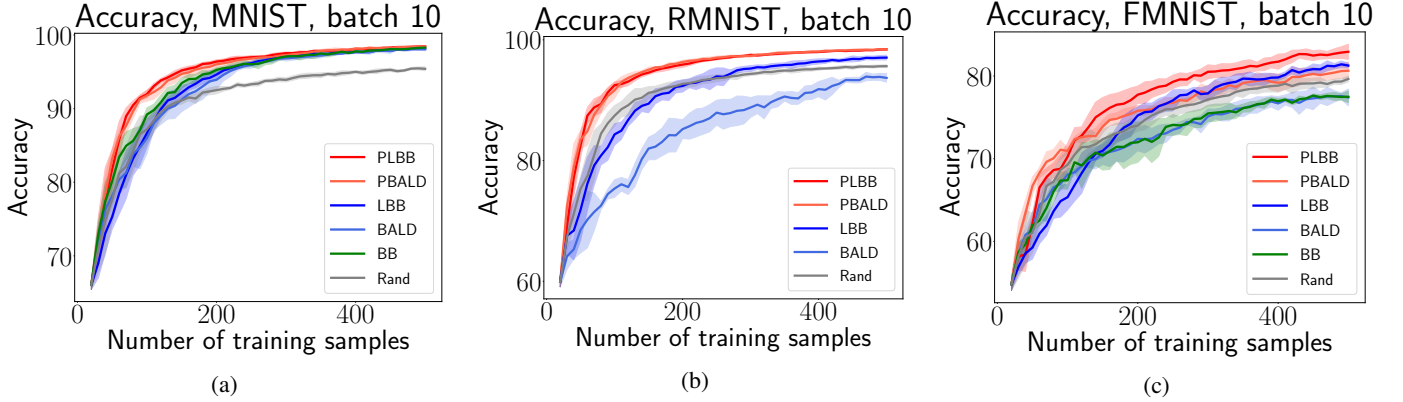


Figure 1: Test accuracy over acquired images, uncertainty estimates are based on deep ensembles. Datasets: (a) MNIST. (b) RMNIST. (c) FMNIST. LBB shows better performance than BALD, and the PLBB and PBALD extensions using added stochasticity lead among other algorithms.

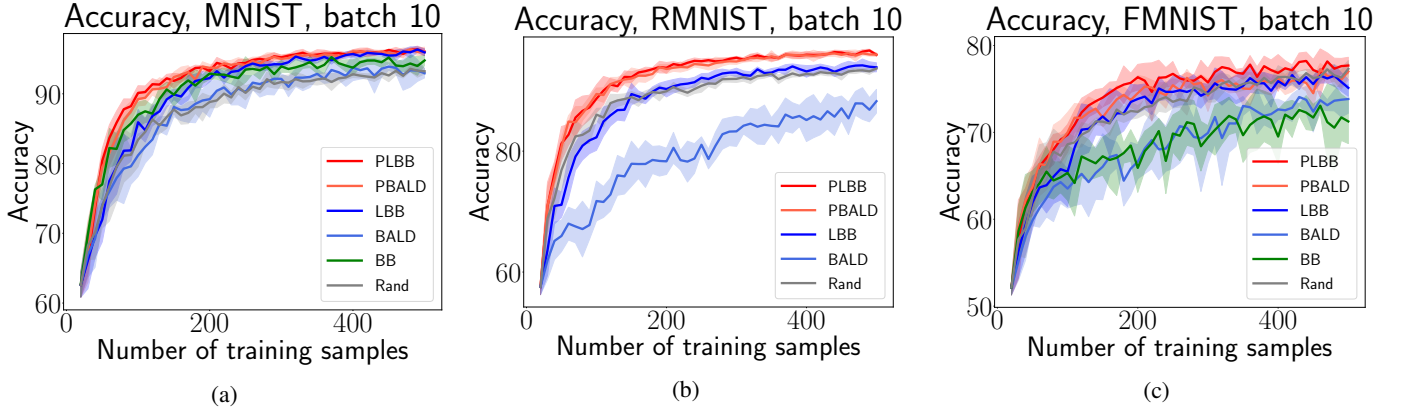


Figure 2: Test accuracy over acquired images, uncertainty estimates are based on MC-dropout. Datasets: (a) MNIST. (b) RMNIST. (c) FMNIST. LBB is clearly superior to BALD, and by a larger margin than in the case of deep ensembles.

compared to other algorithms. Large BatchBALD, on the other hand, as an approximation of BatchBALD performs much better than BALD, taking into account batch interconnections. It is still inferior to the random baseline at the beginning, but later outperforming it starting from a few hundred elements. In turn, PLBB and PBALD, combining informativity proportional to the LBB and BALD criterion scores, respectively, and the diversity obtained by sampling from the power distribution, their performance is quite close to each other and outperforms all other algorithms. Note, that on a dataset with a large pool, like RMNIST, and on large-batch experiments, BatchBALD becomes computationally infeasible.

Regarding the results on the FMNIST dataset, the Large BatchBALD algorithm shows the best quality compared to BALD, and the results of its PowerLBB and PowerBALD extensions are the best among other algorithms, with PLBB significantly outperforming PBALD, see Fig. 1c. This may be

due to the fact that PLBB has the additional data diversity contained in the LBB algorithm design, while PBALD has data diversity only due to sampling-driven randomness. Both mentioned algorithms are better than BatchBALD, which in turn, together with BALD, has comparable performance and worse results among competitors.

3) *MC-dropout*: Comparing the results obtained with MC-dropout and with deep ensembles, we see that the test accuracy of algorithms with MC-dropout is lower than with ensembles, as we mentioned earlier. At the same time, the margin between algorithms, for example, between LBB and BALD in all the figures is more clear in the figures obtained with MC-dropout. Thus, the Large BatchBALD algorithm is significantly better than BALD, see Fig. 2a, Fig. 2b, Fig. 2c. Also, in Fig. 2a the LBB algorithm even outperforms BatchBALD in quality, starting from 200 elements in the training set. The BALD algorithm, however, shows quality comparable to random

selection of samples. Furthermore, in Fig. 2b and Fig. 2c BALD performance is significantly worse than random selection of samples. Also, on all three figures, power extensions show the best accuracy among all algorithms. There are also figures for the bigger batch size of 20, see Appendix VI-C for details.

B. SVHN, RCIFAR-10, RCIFAR-100

CIFAR-10 and CIFAR-100 [13] are datasets of color images, each containing 60,000 images consisting of 10 and 100 classes, respectively. Repeated extensions of CIFAR datasets involve repeating each of the images multiple times in the dataset with a Gaussian noise applied, namely 4 times in our case, which increases proportionally the total size of each dataset. SVHN [17] is a Street View House Numbers dataset containing 70,000 images.

As a model, we used ResNet-18 [8] architecture with SGD optimizer with momentum = 0.9, weight decay = 0.0005, learning rate = 0.05. As a learning rate scheduler, we used MultiStepLR with gamma = 0.1, and milestones = 25, 40. The network was trained for 50 epochs keeping the model that showed the best quality on validation, the validation sample size consists of 5K examples.

Regarding the SVHN dataset, in the Fig. 3c calculated for a batch size 50, the LBB algorithm shows superiority over the BALD algorithm and the random baseline, while the BALD algorithm is inferior to the random baseline up to 2K examples. Additional randomization significantly improves the BALD displayed in the better performance of the algorithm, as shown by PBALD. At the same time, LBB is as good as, and in some places slightly better than, the PBALD algorithm without additional randomization. The leader among all algorithms is the randomized version of LBB, the PLBB algorithm.

Dataset Repeated CIFAR-100 (RCIFAR-100) is sufficiently challenging for the task of active learning due to the large number of classes and image repetitions, which increases the initial volume by 4 times. Moreover, such a dataset involves taking examples in large batches to get good performance in a reasonable amount of time. Note, that while the results based on RCIFAR-10, see Fig. 3a, show only slight superiority of the LBB algorithm over the BALD algorithm, on a more complex dataset like RCIFAR-100 the differences are already clear, see Fig. 3b. It shows that Large BatchBALD is more successful in quality than BALD and random baseline. As for the randomized versions of LBB and BALD, namely PLBB and PBALD, they improve the results of the original, with PLBB dominating in quality among the other algorithms on this dataset. For experimental results on the mentioned datasets for a larger batch sizes, see Appendix VI-C.

C. Algorithm runtime comparison

We present numerical execution times for the considered algorithms, namely BALD, PowerBALD, BatchBALD, Large BatchBALD, and Power Large BatchBALD, see Table II that is based on MNIST dataset. Execution results are obtained with deep ensembles constructed using a small convolutional network. The initial pool consists of 20 images, and the

unlabeled pool contains 49,880 images. Note that these results also support the claim that LBB, being an approximation of BatchBALD, is tens of times faster. Moreover, this difference becomes even more noticeable as the batch size increases, which can give a gain of tens of times. Thus, when working with batches of hundreds of items, it is evident that the calculation of the Large BatchBALD acquisition function is more feasible than that of the BatchBALD method.

IV. RELATED WORK

Uncertainty estimates in deep learning are often associated either with MC-dropout [6] or deep ensembles [14]. In the case of MC-dropout, one samples the dropout mask on inference to get an ensemble of models differently parameterized. In the case of deep ensembles, one trains a single model with different weight initialization. Speaking of classification, in both scenarios, the final prediction is the average of the softmax vectors from all the models in the ensemble. In the work [2] authors demonstrate the superior performance of ensembles over MC-dropout in image classification tasks. Nevertheless, we tested both of the approaches for dealing with uncertainty.

One of the most applicable baseline algorithms for active learning is Bayesian Active Learning by Disagreement (BALD) [10]. Its acquisition function is computed as mutual information between model output and its latent parameters. That is, BALD tries to find those examples in which the different models disagree, while each of the models is confident in its prediction. This approach is quite efficient when taking one example at each step for training.

In practice, with a large pool size, it is unprofitable to take a single example or even small batches, so it is important to be able to take batches of informative examples at each step of the AL loop. In practice, the top- k approach is most often used to take more than one example in the AL loop, where each step takes k examples with the highest values of the acquisition function. This approach has a serious drawback, namely, it does not take into account pairwise interaction of the data samples in batches that leads to acquiring a batch of similar examples and resulting performance degradation.

One possible solution is a batch modification of the BALD algorithm called BatchBALD [11]. The idea is to treat the mutual information in a batch manner as between a joint random variable (i. e., set of model outputs) and model parameters. In this scenario, points are added one at a time in a greedy manner, and the total mutual information is recursively recalculated. The diversity of acquired samples comes from accounting for the interactions in the batch between outputs. Nevertheless, BatchBALD has a significant drawback, namely, it takes a lot of working time [12]. In the original BatchBALD work, the standard choice is to take a batch of 10 elements, since complexity grows exponentially with batch size due to the joint entropy calculation. In practice, it is calculated directly for the first 5 samples in the batch, and the rest are sampled using MC-dropout.

Another way for diversification is presented in the already mentioned article [12]. Its authors get a diversity of chosen

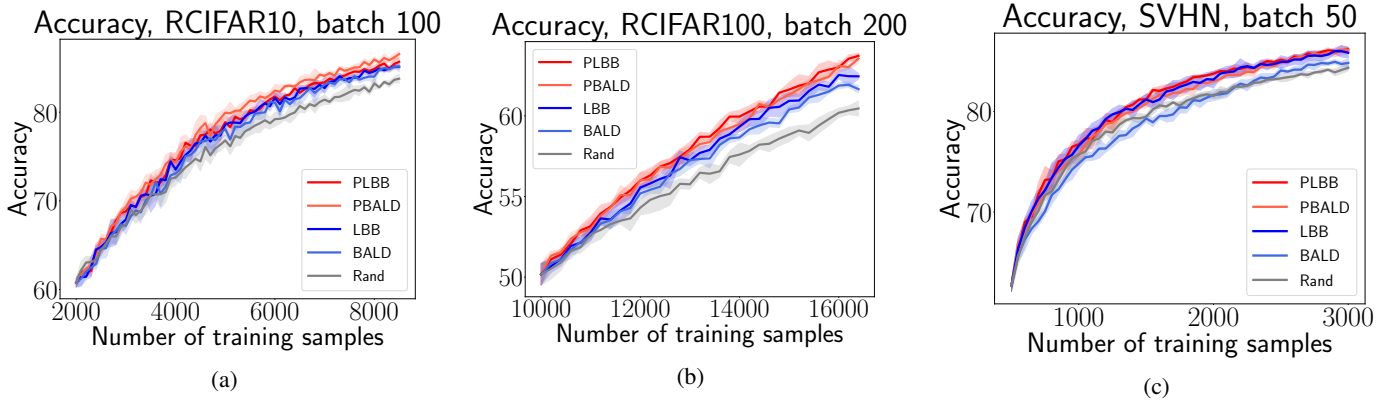


Figure 3: Test accuracy over acquired images, uncertainty estimates are based on deep ensembles. Datasets: (a) RCIFAR-10. (b) RCIFAR-100. (c) SVHN. LBB outperforms BALD, and also shows results close to the leading methods based on power distribution, namely PLBB and PBALD.

Batch size	BALD	PowerBALD	BatchBALD	Large BB	Power LBB
10	5.04 ± 0.51	5.29 ± 0.49	268.9 ± 10.37	18.18 ± 1.22	20.13 ± 2.61
20	5.13 ± 0.43	5.93 ± 1.43	838.85 ± 98.22	20.06 ± 5.31	18.56 ± 2.27

Table II: Algorithm runtime comparison. MNIST dataset, unlabeled pool of 49,980 images, uncert. estim. with deep ensembles.

examples at the expense of the stochasticity of the acquisition function. Their idea is to get scores from some algorithm, such as BALD, and then translate those scores into a distribution. Sampling from such a distribution, we get an acquisition batch. Thus, if we take all the scores obtained with BALD, raise to some power, normalize, and then sample, we obtain the PowerBALD algorithm, which is one of the comparative baselines in this paper. The basic idea is that a randomized sampling strategy is better than a greedy one, requires the same amount of time, and partially overcomes the data redundancy bottleneck.

To achieve the best quality of AL algorithm, it is often useful to focus only on uncertainty that is directly related to the quality of the algorithm. Thus, the authors of MOCU-based algorithms [33] propose to minimize only the classification error uncertainty as an acquisition function, taking into account the posterior, rather than the overall uncertainty, in contrast to BALD. As a result, the authors do not increase the probability of an already guessed classes, but extract controversial samples on the classification borders, taking into account the posterior. A similar idea was considered in the paper [4] where the authors proposed an acquisition function in a one-step look-ahead manner for regression on Gaussian processes [19]. The idea is to choose a new point to add so that the average variance over the space is reduced.

Another important issue in active learning, affecting the performance of the model, is the diversity of acquired data samples. Combining a Bayesian network and a Gaussian process with a known covariance function, the authors in [27] propose to obtain diverse data samples from the acquisition function as the maximum variance of the Gaussian process. Also, since the variance of the Gaussian process does not depend on the output of the network, each time sampling a

new point changes the total variance, which eliminates the need to retrain the network at each step. Another design of the acquisition function, which takes into account the diversity of samples, is based on the geometric properties of the data [24]. The idea is to add images with the greatest distance from the training set to find data that is still poorly represented by the training set.

Another natural but computationally expensive way to introduce diversity of AL samples into a training set is to use clustering. The authors in the paper [5] demonstrate an efficient data sampling with huge batch sizes by selecting samples from hierarchically clustered data in an ascending volume manner. Another current state-of-the-art work [1] uses k-means++ to achieve diverse acquired samples, along with an acquisition function built on the value of loss gradients relative to model parameters as the value of potential model change. In the work [30] the authors suggest using KNN classifier as the output layer of the network instead of softmax, due to better generalization ability to the unknown space.

V. CONCLUSIONS

To summarize, we propose the Large BatchBALD algorithm as an approximation of the BatchBALD method, using the BALD acquisition function and pairwise mutual information of model output components. The algorithm has comparable quality to the original method in terms of efficiency in avoiding taking similar images while computing the acquisition function several times faster, especially in the case of large batches. Thus, this active learning algorithm balances the uncertainty and diversity of the acquired samples and significantly reduces the acquisition time compared to the original BatchBALD. The resulting method is shown to be active for batch active learning in application to modern image datasets.

VI. APPENDIX

A. Main property proof

$$\begin{aligned}
& \sum_{i=1}^b I(y_i; \theta) - I(y_{1:b}; \theta) - C(y_{1:b}) \\
& \text{(by definition)} \\
& = \sum_{i=1}^b \int_{\theta} \int_{y_i} p(y_i; \theta) \log \frac{p(y_i; \theta)}{p(y_i)p(\theta)} dy_i \\
& \quad - \int_{\theta} \int_{y_{1:b}} p(y_{1:b}; \theta) \log \frac{p(y_{1:b}; \theta)}{p(y_{1:b})p(\theta)} dy_{1:b} \\
& \quad - \int_{y_{1:b}} p(y_{1:b}) \log \frac{p(y_{1:b})}{\prod_{i=1}^b p(y_i)} dy_{1:b}
\end{aligned}$$

(reduce the numerator and denominator)

$$\begin{aligned}
& = \sum_{i=1}^b \int_{\theta} \int_{y_i} p(y_i; \theta) \log \frac{p(y_i | \theta)}{p(y_i)} dy_i \\
& \quad - \int_{\theta} \int_{y_{1:b}} p(y_{1:b}; \theta) \log \frac{p(y_{1:b} | \theta)}{p(y_{1:b})} dy_{1:b} \\
& \quad - \int_{y_{1:b}} p(y_{1:b}) \log \frac{p(y_{1:b})}{\prod_{i=1}^b p(y_i)} dy_{1:b}
\end{aligned}$$

(factorization of joint probability due to independence of y_i conditioned on θ)

$$\begin{aligned}
& = \sum_{i=1}^b \int_{\theta} \int_{y_i} p(y_i; \theta) \log \frac{p(y_i | \theta)}{p(y_i)} dy_i \\
& \quad - \int_{\theta} \int_{y_{1:b}} p(y_{1:b}; \theta) \log \frac{\prod_{i=1}^b p(y_i | \theta)}{p(y_{1:b})} dy_{1:b} \\
& \quad - \int_{y_{1:b}} p(y_{1:b}) \log \frac{p(y_{1:b})}{\prod_{i=1}^b p(y_i)} dy_{1:b}
\end{aligned}$$

(log of product is equal to sum of logs and log fraction is expressed as a diff. of logs)

$$\begin{aligned}
& = \sum_{i=1}^b \int_{\theta} \int_{y_i} (p(y_i | \theta)p(\theta) \log p(y_i | \theta) \\
& \quad - p(y_i | \theta)p(\theta) \log p(y_i)) dy_i d\theta \\
& \quad - \sum_{i=1}^b \int_{\theta} \int_{y_{1:b}} (p(y_{1:b} | \theta)p(\theta) \log p(y_i | \theta) \\
& \quad - p(y_{1:b} | \theta)p(\theta) \log p(y_{1:b})) dy_{1:b} d\theta \\
& \quad - \int_{y_{1:b}} p(y_{1:b}) \log \frac{p(y_{1:b})}{\prod_{i=1}^b p(y_i)} dy_{1:b} \\
& \quad \left(\text{integr. out: } p(y_{1:b}; \theta) = \prod_{i=1}^b p(y_i | \theta)p(\theta) \right)
\end{aligned}$$

$$\begin{aligned}
& = \sum_{i=1}^b \int_{\theta} \int_{y_i} p(y_i | \theta)p(\theta) \log p(y_i | \theta) dy_i d\theta \\
& \quad - \sum_{i=1}^b \int_{y_i} p(y_i) \log p(y_i) dy_i \\
& \quad - \sum_{i=1}^b \int_{\theta} \int_{y_i} p(y_i | \theta)p(\theta) \log p(y_i | \theta) dy_i d\theta \\
& \quad + \int_{y_{1:b}} p(y_{1:b}) \log p(y_{1:b}) dy_{1:b} \\
& \quad - \int_{y_{1:b}} p(y_{1:b}) \log p(y_{1:b}) dy_{1:b} + \sum_{i=1}^b \int_{y_i} p(y_i) \log p(y_i) dy_i = 0.
\end{aligned}$$

which is an exact statement that was presented in the beginning.

B. Pairwise mutual information calculation

Here we give an explicit formula for calculating the approximation of the total correlation $\hat{C}(y_i; y_j)$ through pairwise mutual information:

$$\begin{aligned}
\hat{C}(y_i; y_l) & = \sum_{\hat{y}_i, \hat{y}_l} \left(\frac{1}{k} \sum_{j=1}^k p(\hat{y}_i | \hat{\theta}_j) p(\hat{y}_l | \hat{\theta}_j) \right) \\
& \quad \left(\log \left(\frac{1}{k} \sum_{j=1}^k p(\hat{y}_i | \hat{\theta}_j) p(\hat{y}_l | \hat{\theta}_j) \right) \right. \\
& \quad \left. - \log \left(\frac{1}{k} \sum_{j=1}^k p(\hat{y}_i | \hat{\theta}_j) \right) - \log \left(\frac{1}{k} \sum_{j=1}^k p(\hat{y}_l | \hat{\theta}_j) \right) \right),
\end{aligned}$$

where $i \neq l$, \hat{y} is a sample from $p(y)$, k is either the number of forward passes in the case of MC-dropout, or the number of models in an ensemble in the case of deep ensembles.

C. Additional results for bigger batches

Experimental results on a larger acquisition batch for the MNIST dataset are shown in Fig. 4a. Note that LBB and BALD have comparable accuracy here, as do PLBB and PBALD. Power extensions also dominate among other algorithms. Experiments for RMNIST with larger acquisition batch size, see Fig. 4b. Here we observe a similar picture with respect to LBB and BALD, as is that their extensions PLBB and PBALD show the best quality among all algorithms. Similar conclusions can be drawn from the FMNIST dataset Fig. 4c with acquisition batch size equal to 20. For MNIST, RMNIST, FMNIST figures on batch 20 with MC-dropout see Fig. 5a, Fig. 5b, Fig. 5c, respectively.

Regarding the size of the batch 100 on SVHN dataset, see the Fig. 6c, Large BatchBALD shows better quality compared to BALD. In turn, BALD has a similar performance to the random baseline. On the same figure, the PLBB algorithm slightly outperforms PBALD on the first few thousand elements, after which they have similar quality superior to the competitors. The results for RCIFAR-10 on batch 200 are shown in Fig. 6a. Also, on RCIFAR-100 with a larger batch the Large BatchBALD algorithm outperforming the BALD algorithm, see Fig. 6b.

REFERENCES

- [1] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *ICLR*, 2019. 7
- [2] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler. The power of ensembles for active learning in image classification. In *IEEE CVPR*, 2018. 4, 6
- [3] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71, 2021. 1
- [4] E. Burnaev and M. Panov. Adaptive design of experiments based on gaussian processes. In *Statistical Learning and Data Sciences*, pages 116–125. Springer, 2015. 7
- [5] G. Citovsky, G. DeSalvo, C. Gentile, L. Karydas, A. Rajagopalan, A. Rostamizadeh, and S. Kumar. Batch active learning at scale. *NeurIPS*, 34, 2021. 7
- [6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, PMLR, pages 1050–1059, 2016. 6
- [7] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *ICML*, PMLR, pages 1183–1192, 2017. 1
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 6
- [9] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *ICML*, page 417–424. ACM, 2006. 1
- [10] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, 1112.5745, 2011. 1, 2, 6
- [11] A. Kirsch, J. van Amersfoort, and Y. Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *NeurIPS*, volume 32, 2019. 1, 2, 4, 6
- [12] Andreas Kirsch, Sebastian Farquhar, Parmida Atighehchian, Andrew Jesson, Frederic Branchaud-Charron, and Yarin Gal. Stochastic batch acquisition for deep active learning. *CoRR*, 2106.12059, 2021. 3, 6
- [13] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, 2009. 6
- [14] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, volume 30, 2017. 6
- [15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4
- [16] Xiaoming Lv, Fajie Duan, Jia-Jia Jiang, Xiao Fu, and Lin Gan. Deep active learning for surface defect detection. *Sensors*, 20(6), 2020. 1
- [17] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on DL and Unsupervised Feature Learning 2011*, 2011. 6
- [18] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11), 2021. 1
- [19] Carl Edward Rasmussen. Gaussian processes for machine learning. MIT Press, 2006. 7
- [20] P. Ren, Y. Xiao, X. Chang, P. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021. 1
- [21] G. Riccardi and D. Hakkani-Tur. Active learning: theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(4):504–511, 2005. 1
- [22] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: A big data - ai integration perspective. *IEEE Trans. Knowl. Data Eng.*, 33(4):1328–1347, 2021. 1
- [23] P. Sebastiani and H. Wynn. Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society Series B*, 62:145–157, 2000. 1
- [24] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 7
- [25] A. Shelmanov, V. Liventsev, D. Kireev, N. Khromov, A. Panchenko, I. Fedulova, and D. V. Dylov. Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *IEEE BIBM*, pages 482–489, 2019. 1
- [26] Sunil Srinivasa. A review on multivariate mutual information. *Univ. of Notre Dame, Indiana*, 2(1), 2005. 3
- [27] Evgenii Tsymbalov, Sergei Makarychev, Alexander Shapeev, and Maxim Panov. Deeper connections between neural networks and gaussian processes speed-up active learning. *IJCAI*, page 3599–3605. AAAI Press, 2019. 7
- [28] Evgenii Tsymbalov, Maxim Panov, and Alexander Shapeev. Dropout-based active learning for regression. In *International conference on analysis of images, social networks and texts*, pages 247–258. Springer, 2018. 1
- [29] Carl Vondrick and Deva Ramanan. Video annotation and tracking with active learning. *NeurIPS*, 24, 2011. 1
- [30] F. Wan, T. Yuan, M. Fu, X. Ji, Q. Huang, and Q. Ye. Nearest neighbor classifier embedded network for active learning. *AAAI*, 35(11):10041–10048, 2021. 7
- [31] Satoshi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960. 3
- [32] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, 1708.07747, 2017. 4
- [33] G. Zhao, E. Dougherty, B. Yoon, F. J. Alexander, and X. Qian. Bayesian active learning by soft mean objective cost of uncertainty. In *AISTATS*, PMLR, pages 3970–3978, 2021. 7

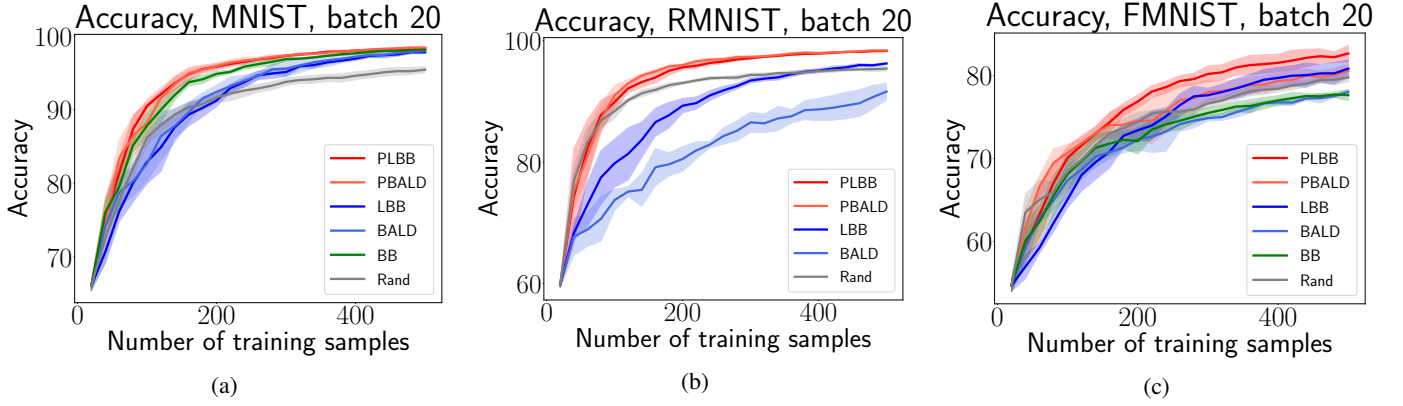


Figure 4: Test accuracy over acquired images, uncertainty estimates are based on deep ensembles. Datasets: (a) MNIST. (b) RMNIST. (c) FMNIST.

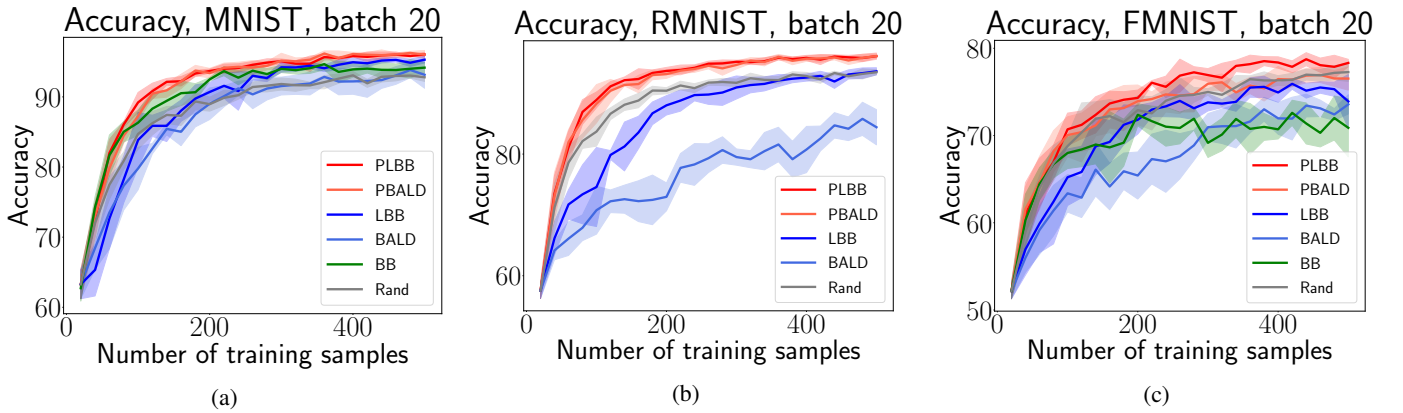


Figure 5: Test accuracy over acquired images, uncertainty estimates are based on MC-dropout. Datasets: (a) MNIST. (b) RMNIST. (c) FMNIST.

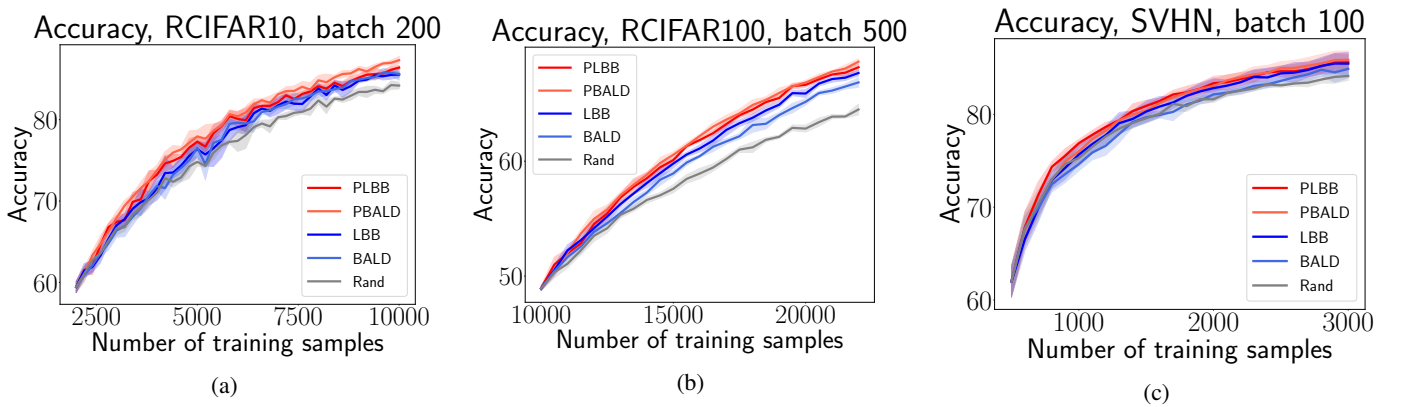


Figure 6: Test accuracy over acquired images, uncertainty estimates are based on deep ensembles. Datasets: (a) RCIFAR-10. (b) RCIFAR-100. (c) SVHN.