Deep Bayesian Active Learning with Hybrid Query Strategies

Zeping Li
School of Informatics
The University of Edinburgh
Edinburgh, EH8 9AB
lzp632697@gmail.com

Mari Ashiga
School of Informatics
The University of Edinburgh
Edinburgh, EH8 9AB
m.ashiga@sms.ed.ac.uk

Shuheng Yang
School of Informatics
The University of Edinburgh
Edinburgh, EH8 9AB
s.yang-97@sms.ed.ac.uk

Abstract

Active learning is an effective technique for reducing expert supervision with annotating costs by selectively querying unannotated data points. It has been shown powerful in improving learning efficiency for neural networks, especially Bayesian neural networks, an architecture suitable for small datasets. However, traditional active learning usually query data points based solely on uncertainty or diversity, which can cause myopic decision boundaries or low query efficiency. To test whether query strategies considering both uncertainty and diversity can further accelerate learning, we develop 54 hybrid query strategies (3 uncertainty metrics * 3 diversity metrics * 6 combination methods), and compare them with pure query strategies on the MINST and CIFAR-10 datasets in a Bayesian convolutional neural network framework. We find query strategy performances are adjusted by the task difficulty and number of queried points. Hybrid query strategies shine in the difficult task with limited queried points. Their advantages decrease but still exist with more queried points.

1 Introduction

Active Learning (AL) is a technique that selectively annotates data points for training. It aims to achieve acceptable performances with less supervision by proper query strategies. A recent direction in AL is to combine it with neural networks (Ren et al., 2022). On the one hand, neural networks are notorious for the dependence of big datasets, but manually annotating them can be costly. AL can alleviate this problem. On the other hand, the scalability to high dimensional data like images is a challenge for traditional AL methods (Tong, 2001), which can be mitigated by neural networks.

Among current neural network architectures, Bayesian neural networks (BNNs) (MacKay, 1992) are a strong candidate. Classical neural networks overfit quickly on small datasets, but BNNs are robust to overfitting. Besides, BNNs encode the uncertainty in parameters (epistemic uncertainty) explicitly, thus can offer a better estimate of prediction uncertainty that helps to query new points in AL. Gal et al. (2017) have built a Bayesian convolutional neural network (BCNN) with AL for image classification. They find the model perform better than deterministic CNNs over different query strategies. However, their, and most previous deep AL studies (e.g., Sener & Savarese (2018); Ducoffe & Precioso (2018)) use query strategies only considering uncertainty or diversity of unannotated data points, which can meet problems in specific situations.

A burgeoning direction in deep AL is to consider information from both sides (Zhdanov, 2019; Yin et al., 2017). However, to our best knowledge, there have been no studies systematically comparing pure and hybrid query strategies in deep AL, not to mention BNNs. To fill this gap, this study aims to compare those query strategies in a BCNN framework. Our main contributions are:

1. We propose two computationally efficient diversity metrics, and develop 54 hybrid query strategies using them and other existing uncertainty and diversity metrics. These strategies are compared with pure query strategies on the MNIST and CIFAR-10 datasets.

2. We show that the task difficulty and number of queried points adjust the performances of query strategies. In the simple task (MNIST), uncertainty-based query strategies are enough to accelerate learning. In the difficult task (CIFAR-10) with limited queried points, diversity-based query strategies are more efficient among pure query strategies. The pattern is reversed with more queried points. Hybrid query strategies can further accelerate learning in the difficult task. Their advantages decrease but still exist with more queried points.

2 Datasets and task

All models are trained on the MNIST dataset (LeCun & Cortes, 1998) first to preliminarily compare query strategies. A further experiment on the CIFAR-10 (Krizhevsky, 2009) dataset is applied to test whether their performances are adjusted by the task difficulty and number of queried points.

MNIST is a dataset of handwritten digits. It consists of 70,000 one-channel images of size 28x28 pixels, each representing a digit from 0 to 9. Each image is transformed to a tensor and normalized before training. They are split into a training set of 60,000 images and a test set of 10,000 images.

CIFAR-10 dataset consists of 60,000 three-channel images of size 32x32 pixels, each representing one of the ten object classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. Similarly, each image is transformed to a tensor and normalized channel-wise before training. They are split into a training set of 50,000 images and a test set of 10,000 images. This dataset is more challenging for image classification than the MNIST dataset due to its complexity and variability.

The main task is to compare the performances of BCNNs with different query strategies on image classification. The evaluation metric is the mean final test accuracy. Test accuracy as a function of query times is also recorded to analyse whether the findings are stable during the learning process.

3 Methodology

3.1 Bayesian convolutional neural networks

BCNNs put priors over network parameters. Inferring the posterior is usually computationally intractable. Fortunately, Gal & Ghahramani (2016) show a common technique, dropout (Srivastava et al., 2014), is equivalent to sampling from a special variational distribution, a mixture of Bernoulli mixture and Gaussian mixture, in BCNNs. So training a classical CNN with dropout is equivalent to conducting a variational inference for a BCNN. And keeping dropout at the test stage is equivalent to using samples from the variational posterior to make predictions.

3.2 Query strategies

The core step of AL is to select the most valuable data batch[1] from the unannotated data pool. So query strategies (or called acquisition functions) can directly influence AL's performance. They can be roughly divided into three branches: Uncertainty-based, diversity-based and their hybrid versions.

3.2.1 Uncertainty-based query strategies

Uncertainty-based query strategies typically rank the unannotated points based on uncertainty and choose the top K most uncertain ones. There are two types of uncertainty: aleatoric and epistemic uncertainty (Nguyen et al., 2019). The former comes from the inherent noise in data (e.g., two points with the same features can belong to different classes). The latter comes from the finite sample sizes (e.g., parameter estimates usually fluctuate less with larger sample sizes). Deterministic neural networks can only capture the former but BNNs can capture both where the latter is reflected on the variance in samples from the posterior. In this study, we consider three uncertainty metrics: entropy,

[1]Unless specified, "batch" means the group of unannotated points selected in one query and "batch size" means the size of that group in this study.

Bayesian active learning by disagreements (BALD) and variation ratio. The uncertainty is larger when the metric value is larger.

Entropy For classification tasks, entropy (Shannon, 1948) is a natural uncertainty metric since the prediction is a categorical distribution. The predictive entropy of an unannotated point $\boldsymbol{x}$[2] is

$$E(\boldsymbol{x}) = -\sum_{c} p(\boldsymbol{x} \in c|\boldsymbol{x}, M) log p(\boldsymbol{x} \in c|\boldsymbol{x}, M)$$

Here $p(\boldsymbol{x} \in c|\boldsymbol{x}, M)$ is the predictive probability that the point belongs to class c under M, the model after training. In a BCNN, it is $\frac{1}{T}\sum_{t=1}^{T} p_t(\boldsymbol{x} \in c|\boldsymbol{x}, M)$, where T is the forward propagation number and $p_t(\boldsymbol{x} \in c|\boldsymbol{x}, M)$ is the result of the t^{th} forward propagation. Each propagation should generate a different result due to dropout. To scale the range of entropy to [0, 1], which is important when combining metrics, we divide it by $log(N)$ where N is the number of possible classes.

BALD Entropy can be overconfident on small datasets for classical CNNs. This is partly mitigated by BCNNs. But a more proper uncertainty metric for BCNNs is BALD (Houlsby et al., 2011):

$$B(\boldsymbol{x}) = E(\boldsymbol{x}) - \frac{1}{T}\sum_{t=1}^{T} E_t(\boldsymbol{x})$$

Here $E_t(\boldsymbol{x})$ is the entropy calculated from the t^{th} forward propagation. From the informatic perspective, $B(\boldsymbol{x})$ is the mutual information between $\boldsymbol{x}$'s class and current model parameters, i.e., how much the entropy of parameters will decrease after observing $\boldsymbol{x}$'s class. Intuitively, $B(\boldsymbol{x})$ will be high if the first term is high and the second term is low. This can happen when each forward propagation makes a confident class prediction ($\arg\max_c p(\boldsymbol{x} \in c|\boldsymbol{x}, M)$), but the predictions vary much among forward propagations. We divide BALD by $log(N)$ to scale its range.

Variation ratio Variation ratio (Freeman, 1965) is heuristic metric to measure the dispersion of categorical distributions. It is calculated by

$$V(\boldsymbol{x}) = 1 - \max_{c} p(\boldsymbol{x} \in c|\boldsymbol{x}, M)$$

Since BCNNs themselves somewhat realize ensemble modeling, variation ratio serves as a commitee-based metric (Seung et al., 1992). Specifically, a BCNN can be seen as a family of deterministic CNNs which share the same architecture but have different parameter values. Each forward propagation with dropout is drawing a CNN and making a prediction using it. T forward propagations lead to T class predictions, and $V(\boldsymbol{x})$ measures how consistent the T predictions are. In other words, $p(\boldsymbol{x} \in c|\boldsymbol{x}, M)$ here is not the mean of probabilistic predictions, but the mean of class predictions, i.e., how many members of the T drawn CNNs agree that $\boldsymbol{x}$ belongs to c. We divide it by $1 - \frac{1}{N}$ to scale its range.

Only considering uncertainty is not enough. Such strategies can be fooled by adversarial examples (Ducoffe & Precioso, 2018). That is, they may only focus on points around the current decision boundary but ignore the true data distribution. Figure 1 is an example. It illustrates a binary classification task with a triangle decision boundary. Unfortunately, the initial training set mainly contains points around the left edge. Uncertainty-based query strategies will continuously query points around that edge if most points are distributed within and around the triangle. Since AL usually does not query many points, it is likely the model will finally be confident that the true decision boundary is that edge. In other words, uncertainty-based query strategies can stuck in a myopic decision boundary in difficult tasks, i.e., complex true decision boundaries. Diversity-based query strategies can mitigate this problem.

3.2.2 Diversity-based query strategies

The definition of diversity itself is "diverse". It can refer to how representative the selected points are of the whole unannotated pool. Corresponding query strategies usually construct a core set as a surrogate (Sener & Savarese, 2018; Geifman & El-Yaniv, 2017). Uniformly choosing points is

[2]Strictly speaking, $\boldsymbol{x}$ indicates the features or feature embeddings of a point. But we do not distinguish them for simplicity.

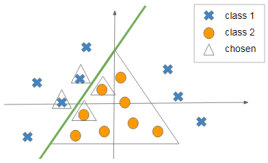


Figure 1: An example of a myopic decision boundary.

somewhat also such a query strategy since it implicitly matches the distribution. Another definition is how different the selected points are. Corresponding query strategies select points far away from each other (Yin et al., 2017). They prefer the batch covering the entire embedding space (Calculating the diversity in the original space is inefficient for images. CNNs can project them to a low-dimensional space, which we call the embedding space) but ignores the actual distribution. The third definition is the difference between the annotated and unannotated points. Corresponding query strategies will select points far from the annotated ones (Gissin & Shalev-Shwartz, 2019; Shui et al., 2020). We follows this definition since the relevant metrics are computationally efficient. Diversity-based query strategies derived from the former two definitions usually require sequential traversals over the unannotated pool. Those derived from the third definition traverses over the annotated points, which are usually much less, and can be conducted in parallel. In this study, we consider three diversity metrics: Discriminator score, posterior variance and minimum distance. We propose the latter two inspired by previous studies. The diversity is larger when the metric value is larger.

Discriminator score The core idea of discriminator score (Gissin & Shalev-Shwartz, 2019) is to introduce another classification model. It is trained on all points to distinguish whether a point is annotated. Then for a given point, it can generate an "unannotated score"

$$D(\boldsymbol{x}) = p(\boldsymbol{x} \in unannotated | \boldsymbol{x}, M_1)$$

Here M_1 is the discriminator model after training. To solve the problem of imbalanced samples, we re-sample the annotated points until their number is equal to the pool size when training M_1.

Posterior variance Posterior variance is based on Gaussian process (GP; Rasmussen & Williams 2006). GP defines the beliefs over function values in a continuous domain, which is the embedding space in this study. Such a function can map each point to a real value. The prior belief is that the values of any points follow the same Gaussian distribution. After observing some points' values, we can adjust the value belief of other points. For an given point, the posterior belief of its value is a Gaussian distribution with a smaller variance. The decrease in variance is larger when it is "closer" to the points whose values are observed with specific kernel functions. Li & Guo (2013) use this decrease as a diversity metric to calculate the representativeness of point of the pool. Instead, we define the diversity of an unannotated point as the posterior variance of our value belief after observing the values of all annotated points. To calculate it, we first choose the radio basis kernel function to measure the "similarity" between any two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2d})$$

Here exp is the exponential function, $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ is the square of Euclidean distance between the two points and d is the embedding dimension. $2d$ is chosen heuristically since our embedding data is normalized before calculation. The expectation of the distance square between two randomly selected points is $2d$. Then the posterior variance of an unannotated point $\boldsymbol{x}$ is calculated using

$$P(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - k_{\boldsymbol{x},\boldsymbol{Y}} K(\boldsymbol{Y})^{-1} k_{\boldsymbol{x},\boldsymbol{Y}}^T$$

Here $k(\boldsymbol{x}, \boldsymbol{x})$ is 1. T and $^{-1}$ are the transpose and inverse symbols. $\boldsymbol{Y}$ indicates the annotated points. We use $\boldsymbol{y}_i$ to denote the i^{th} annotated point. If there are n annotated points, $k_{\boldsymbol{x},\boldsymbol{Y}}$ is a n-dimensional vector whose i^{th} element is $k(\boldsymbol{x}, \boldsymbol{y}_i)$ and $K(\boldsymbol{Y})$ is a $n * n$ matrix whose (i,j) element is $k(\boldsymbol{y}_i, \boldsymbol{y}_j)$.

Minimum distance Minimum distance is a heuristic metric for the distance between one point and a group of points. It is inspired by the work of Sener & Savarese (2018), in which they use the metric to add points to a core set greedily. In our version, the minimum distance of an unannotated point $\boldsymbol{x}$ is

$$M(\boldsymbol{x}) = \min_i \|\boldsymbol{x} - \boldsymbol{y}_i\|$$

Here $\|\boldsymbol{x} - \boldsymbol{y}_i\|$ is the Euclidean distance between $\boldsymbol{x}$ and the i^{th} annotated point. To scale its range, we divide the minimum distances of all unannotated points by the maximum value among them during one query.

The limitations of diversity-based query strategies are also obvious. First, the calculation of diversity metrics is computationally costly. Even our metrics are more efficient than most previous ones, they still face the problem of growing annotated point number during the learning process. Second, diversity metrics are not instructive in simple tasks, i.e., simple true decision boundaries. Consider a binary classification task where the true decision boundary is just a line. Querying few points around it is enough to achieve high model performance. Diversity-based query strategies can only be effective if the data is extremely unbalanced. Otherwise, they will continuously query points far away from the current boundary, lowering the learning efficiency. And even in difficult tasks, we still expect uncertainty-based query strategies will perform better with enough queried points, since there are sufficient annotated points to reflect the true data distribution, though they are not chosen to fulfill it deliberately.

To some extent, the tradeoff between uncertainty and diversity is the tradeoff between exploitation and exploration. Since the disadvantages of them can be overcame by each other, we hypothesize combing the two metrics together can further accelerate learning.

3.3 Combination methods

Once we decide the two metrics, the next step is to integrate them. In this study, we consider three combination methods: Weighted arithmetic mean, weighted geometric mean and two stage query.

Weighted arithmetic mean Since all metrics are scaled to [0, 1], an intuitive way is to calculate the weighted arithmetic mean:

$$H(x) = \alpha U(\boldsymbol{x}) + (1 - \alpha) D(\boldsymbol{x})^3$$

The weight parameter α controls the importance of the uncertainty metric. Its range is [0, 1]. The metric degenerates to the uncertainty metric when $\alpha = 1$, and to the diversity metric when $\alpha = 0$.

The weight parameter is constant during the learning process above. But as discussed, the importance of diversity should decrease with more points queried. Thus, we also test a time-decayed variant:

$$H(\boldsymbol{x}) = (1 - exp(-\beta t)) U(\boldsymbol{x}) + exp(-\beta t) D(\boldsymbol{x})$$

Here t represents the times of query having been done (the minimum is 1 including the current one). The decay parameter β indicates how fast the weight of diversity decreases. We set its range to [0, 1].

Despite of its simplicity, weighted arithmetic mean is not popular in deep AL (Yin et al., 2017; Shui et al., 2020). One reason might be this combination is physically implausible since each metric has its own unit. Consider an example of adding time and temperature. This is somewhat meaningless.

Weighted geometric mean Weighted geometric mean can create a more physically plausible metric and has been attempted by Li & Guo (2013). This calculation is

$$H(\boldsymbol{x}) = U(\boldsymbol{x})^{\alpha} D(\boldsymbol{x})^{(1-\alpha)}$$

Here the role and range of the weight parameter α are the same.

Similarly, we test a time-decayed variant:

$$H(\boldsymbol{x}) = U(\boldsymbol{x})^{(1-exp(-\beta t))} D(\boldsymbol{x})^{exp(-\beta t)}$$

The role and range of the decay parameter β are also the same.

Two stage query The two stage query does not combine the metrics explicitly. The core idea is to use them successively. Suppose we will select m points in one query. This method should first select γm points based on one metric, then choose m points from them based on the other metric (Ash et al., 2021; Zhdanov, 2019). Since we can calculate the uncertainty and diversity for each unannotated point, the selection at the two stages are simply choosing top γm and m points.

[3] D here represents diversity rather than the discriminator score.

The two stage query in our study also has two variants: uncertainty-first and diversity-first. As the names suggest, the order of metrics used is uncertainty-diversity in the former variant but diversity-uncertainty in the latter variant. There is a hyperparameter γ. It is an integer ranging from 1 to the pool size divided by m. The methods will purely use the first metric when γ is 1, and purely use the second metric when γ is the pool size divided by m.

4 Experiments

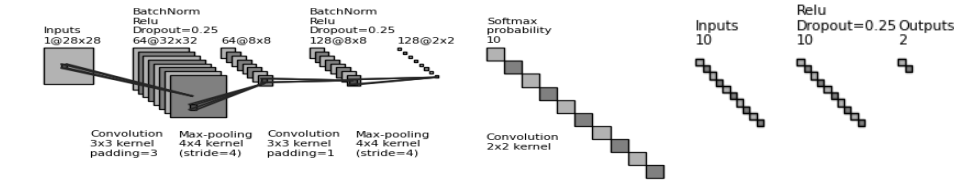
4.1 MNIST experiment

One motivation of this experiment is to reproduce the results of Gal et al. (2017), which find the performance order of uncertainty-based query strategies is: variation ratio > BALD $\approx$ entropy. Another motivation is to compare different query strategies in this simple task.

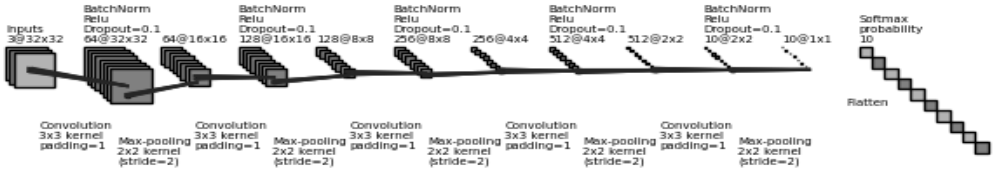
Gal et al. (2017) use the default CNN architecture in Keras (Chollet et al., 2015), but we try a more complicated architecture. This is because the calculation of the metrics depends on model outputs. Their values and the relevant comparison should be more reliable when the model performs better on the task. We attempt some classical architectures like VGG and ResNet, but they quickly overfit on the small dataset. The adopted architecture finally is called the Network in Network (NiN) (Lin et al., 2013). It discards the fully-connected layers and has been shown powerful in mitigating overfitting.

4.1.1 Implementation details

The model architecture in this experiment is shown in Figure 2(a). Since MNIST dataset is simple, we apply an NiN with 5 layers (excluding batch normalization and dropout). Each image is finally projected to a 10*1*1 tensor. This tensor is flattened to a 10-dimensional vector, then passed to the softmax function to generate the probabilistic prediction.



(a) MNIST (b) Discriminator



(c) CIFAR-10

Figure 2: Model architecture for the two experiments

The output layer is also the embedding layer. There are two benefits. First, each component of the vector shows the "support" of the image to the corresponding class. When their magnitudes are similar, diversity-based query strategies tend to choose equal number of images from each class based on current decision boundary. It helps get balanced samples, which is important in classification tasks. Second, if a constant is added to each component, the softmax output will not change. But diversity-based query strategies can capture the difference. The embeddings are normalized before calculating the diversity.

At the training stage, 20 images, 2 from each class, are randomly selected as the initial training set. 1000 images are then randomly selected as the validation set. The rest images in the original training set serves as the unannotated data pool. We set a batch size of 128 ("batch size" here indicates the number of points sampled in one stochastic gradient descent step), a epoch number of 50 and a learning rate of 0.001 using the Adam optimizer (Kingma & Ba, 2015).

After each training loop (i.e., 50 epochs), the model queries 10 points from the pool (to further reduce computational cost, 2000 points are randomly selected from the original pool as the actual pool) based on the query strategy. The number of forward propagation to calculate uncertainty is 100. After adding the points to the training set, the model is re-initialized and re-trained. In each experiment, the model conducts 20 times of query and test accuracy after each query is recorded.

We run 5 experiments for each query strategy and average the results. Each experiment is controlled by a random seed, which is generated by a higher-level random seed specified before all experiments. This ensures the results of different query strategies are comparable and reproducible.

For the discriminator score metric, we built a feedforward neural network as the discriminator, whose architecture is shown in Figure 2(b). Its parameter settings are the same as those of the main BCNN.

For the first four combination methods, we conduct a grid search from 0.1 to 0.9, with a step size of 0.1 for hyperparameter tuning. The value producing the highest mean final validation accuracy (all details are the same, except replacing the test set with the validation set) is chosen. For the latter two combination methods, we try 5 values of γ: 2, 3, 4, 5, 6, and choose the value in the same way.

The pure query strategies, together with the uniform query strategy are treated as baselines. So there are 7 baselines and 3 (uncertainty metrics) * 3 (diversity metrics) * 6 (combination methods) = 54 hybrid query strategies to be compared[4].

4.1.2 Results and discussion

The mean final test accuracy and accuracy curves are shown in Table 1 and Figure 3(a). For the hybrid query strategies, we only show the results of the best one. Performances of other hybrid query strategies can be found in the Appendix.

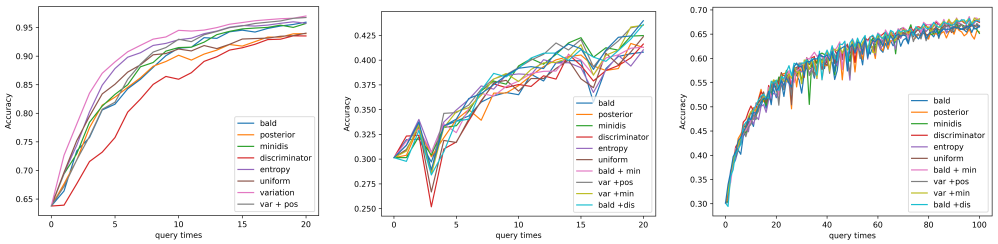
Table 1: Final test accuracy(%) in the two experiments

Query strategy	MNIST	CIFAR-10	CIFAR-10_extended
Entropy	95.75	41.04	67.91
BALD	95.93	40.80	66.47
Variation ratio	97.01	41.30	67.67
Discriminator	93.52	41.62	67.43
Posterior var	93.93	41.28	65.43
Min distance	95.70	42.48	65.15
Uniform	94.01	42.36	66.67
Time-decayed geometric mean of var ratio and posterior var	96.75		
Time-decayed geometric mean of var ratio and min distance		43.60	67.62
Time-decayed arithmetic mean of var ratio and posterior var		43.52	68.18
Constant geometric mean of BALD and min distance		43.53	66.64
Two stage diversity first of BALD and discriminator		43.98	66.63

Our experiment basically reproduces the results of Gal et al. (2017) on the uncertainty-based query strategies. The performance order is still: variation ratio > BALD $\approx$ entropy. Besides, our model is more efficient since the uniform baseline reaches the accuracy of 94% using 220 images, while the accuracy is only about 88% with the same number of images in their study.

Surprisingly, all uncertainty-based query strategies outperform the uniform baseline, but only one diversity-based query strategy beats it. Considering the uniform baseline is also a diversity-based query strategy in general, diversity information may not be helpful in guiding query in this task. The further evidence comes from the best hybrid query strategy. It is the time-decayed weighted geometric mean of variation ratio and posterior variance. The hyperparameter β is 0.6, indicating the weight of posterior variance quickly decreases with query times. In the meantime, its performance

[4]The experimental codes and results are available at https://github.com/RL-LBAM/DBALHQ



(a) MNIST (b) CIFAR-10 main (c) CIFAR-10 extended

Figure 3: Test accuracy (%) as a function of query times in the two experiments

is only comparable to the best baseline, and this pattern is pretty stable during the learning process. All of those support uncertainty-based query strategies are enough to accelerate the learning in this simple task.

However, MNIST is notorious for its simplicity and low-variability. Diversity-based query strategies are theoretically not suitable for such datasets. So the advantage of hybrid query strategies in combining different information may not be obvious in such situations. Thus, It is worth comparing the query strategies on a more complicated task.

4.2 CIFAR-10 experiment

The motivation of this experiment is to test whether the task difficulty adjusts the performances of different query strategies, and whether hybrid query strategies can perform better in a difficult task.

4.2.1 Implementation details

The model architecture in this experiment is shown in Figure 2(c). Considering the CIFAR-10 dataset is more challenging, we increase the layer number from 5 to 10 and correspondingly decrease the dropout probability to accelerate training. We also increase the initial training set size to 100 images, 10 from each class. All other settings are the same as those in the MNIST experiment.

After observing the results, we decide to run an extended experiment to see whether the advantages of hybrid query strategies will continue when more points are queried. This is because the final accuracy is relatively low and the points queried are still few compared to the dataset size.

We choose the top 4 hybrid query strategies whose performances all surpass the best baseline. Along with 7 baselines, we compare 11 query strategies in the extended experiment. We increase the query times to 100 and number of points per query to 50. All other settings are the same, except we directly use the hyperparameter values tuned from the main experiment to decrease the computational cost.

4.2.2 Results and discussion

The mean final test accuracy and accuracy curves are shown in Table 1, Figure 3(b) and Figure 3(c). We show the results of top 4 hybrid query strategies since their performance differences are small.

The accuracy in the main experiment shows the larger difficulty of CIFAR-10. The model is more complex and uses more images, but only reaching accuracy lower than the half of the MNIST experiment, with more fluctuant accuracy curves. Meanwhile, the advantage of diversity-based query strategies in difficult tasks with limited queried points is supported. The best uncertainty-based query strategy is only comparable to the worst diversity-based query strategy. The best diversity-based query strategy achieves a similar final accuracy to the uniform baseline, indicating the diversity based on the first and third definitions can be equally effective. Besides, the hybrid query strategies shine in this situation. Several hybrid query strategies (see the Appendix) outperform the best baseline. The pattern is relatively stable during the learning process.

For the extended experiment, the order of uncertainty-based and diversity-based query strategies reverse again. But the advantage of uncertainty-based query strategies is not obvious now, since its best member is only comparable to the best diversity-based query strategy. The best uncertainty-

based and diversity-based query strategies both outperform the uniform baseline, indicating the two information can both accelerate learning. The best query strategy is still hybrid, though its advantage decreases. From the accuracy curves, we can observe the performances of hybrid query strategies fluctuate less and are usually better than those of baselines. The accuracy curves are overall more chaotic, which may be due to task difficulty and the decline of relative importance of diversity.

Overall, the results support diversity-based query strategies are more efficient among pure query strategies in the difficult task with limited queried points, though this pattern is reversed with more queried points. Hybrid query strategies outperform pure query strategies in the difficult task. Their advantages decrease but still exist when more points are queried.

5 Related work

Calculating uncertainty and diversity respectively can separate their effects, but requiring explicit assumptions of combination. Ash et al. (2020) introduce the gradient embeddings, for which the embedding of an unannotated point is the gradient of the last layer after passing it to the neural network with current class prediction. A fixed-size determinantal point process (Kulesza & Taskar, 2011) is then used to select the data batch from the pool. The process prefers points with high uncertainty (i.e., large gradient norm) when the batch size is small, but points different from each other when the batch size is large. They show this query strategy consistently performs well or better than pure query strategies regardless to batch sizes or architectures. The method integrates uncertainty and diversity into an unified framework, and avoid hyperparameter tuning for combination methods. Similarly, Kirsch et al. (2019) extend BALD to a batch-level metric. They consider the mutual information between model parameters and data batch to decrease repeated information, and show it performs better than BALD with larger batch sizes. The method implicitly encodes the diversity information to avoid querying points who provide the same uncertainty information. These two query strategies can be understood as "soft" versions of uncertainty-first two stage query.

Yin et al. (2017) explicitly treat the tradeoff between uncertainty and diversity as a exploitation-exploration problem. They use entropy as the uncertainty metric and a diversity metric based on the second definition. The two metrics are combined using constant weighted arithmetic mean. To ensure the exploration will decrease with more queried points, they first choose m points based on the uncertainty metric, then choose the rest points based on the hybrid metric. m will increase with query times until a threshold is reached. They compare it with some baselines on the MNIST dataset, and find it is always better than the pure query strategies. Interestingly, they also observe the uncertainty-based query strategies underperform the diversity-based query strategies at the early stage of learning, but surpass them when more points are queried. This is similar to what we find in the CIFAR-10 dataset. There are other studies (Hsu & Lin, 2015; Liu et al., 2018) directly modeling the switch between uncertainty and diversity in a reinforcement learning framework.

6 Future directions

We consider two directions to explore in the future. The first is to decrease the training time of deep AL. Following Gal et al. (2017), we re-initialize the models after each query to isolate the effects of query strategies. This is time-consuming and is unaffordable with deeper architectures and larger datasets. We have tried to train the models only on newly queried points without re-initialization, but the performances sharply decrease. We speculate this is because the queried batches are dependent in this situation, breaking the assumption of stochastic gradient descent. One possible solution may be simply increasing batch size of query, or decreasing epochs of training like early stopping. A more reasonable solution may be using AL to finetune models which only allows few parameters to be updated rather than train models from scratch.

The second is to deign more ingenious hybrid query strategies. The best hybrid query strategies in the experiments are usually time-decayed. And we empirically show the relative importance of diversity decreases with query times. Switching from diversity to uncertainty after the queried times reach a threshold may be more flexible. And as mentioned, modeling the switch between uncertainty and diversity in a reinforcement learning framework may also help increase query strategy efficiency.

References

Ash, Jordan T., Zhang, Chicheng, Krishnamurthy, Akshay, Langford, John, and Agarwal, Alekh. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ryghZJBKPS.

Ash, Jordan T., Goel, Surbhi, Krishnamurthy, Akshay, and Kakade, Sham M. Gone fishing: Neural active learning with fisher embeddings. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8927–8939, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4afe044911ed2c247005912512ace23b-Abstract.html.

Chollet, François et al. Keras. https://keras.io, 2015.

Ducoffe, Melanie and Precioso, Frédéric. Adversarial active learning for deep networks: a margin based approach. *CoRR*, 2018. URL http://arxiv.org/abs/1802.09841.

Freeman, L.C. *Elementary Applied Statistics: For Students in Behavioral Science*. Wiley, 1965. ISBN 9780471277804. URL https://books.google.co.uk/books?id=hRUjAAAAMAAJ.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1050–1059. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

Gal, Yarin, Islam, Riashat, and Ghahramani, Zoubin. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1183–1192. PMLR, 2017. URL http://proceedings.mlr.press/v70/gal17a.html.

Geifman, Yonatan and El-Yaniv, Ran. Deep active learning over the long tail. *CoRR*, 2017. URL http://arxiv.org/abs/1711.00941.

Gissin, Daniel and Shalev-Shwartz, Shai. Discriminative active learning. *CoRR*, 2019. URL http://arxiv.org/abs/1907.06347.

Houlsby, Neil, Huszár, Ferenc, Ghahramani, Zoubin, and Lengyel, Máté. Bayesian active learning for classification and preference learning. *CoRR*, 2011.

Hsu, Wei-Ning and Lin, Hsuan-Tien. Active learning by learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pp. 2659–2665. AAAI Press, 2015. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9636.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Kirsch, Andreas, van Amersfoort, Joost, and Gal, Yarin. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 7024–7035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html.

Krizhevsky, Alex. Learning multiple layers of features from tiny images. Technical report, 2009.

Kulesza, Alex and Taskar, Ben. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 1193–1200. Omnipress, 2011. URL https://icml.cc/2011/papers/611_icmlpaper.pdf.

LeCun, Yann and Cortes, Corinna. MNIST handwritten digit database. 1998. URL http://yann.lecun.com/exdb/mnist/.

Li, Xin and Guo, Yuhong. Adaptive active learning for image classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pp. 859–866. IEEE Computer Society, 2013. doi: 10.1109/CVPR.2013.116. URL https://doi.org/10.1109/CVPR.2013.116.

Lin, Min, Chen, Qiang, and Yan, Shuicheng. Network in network. *CoRR*, 2013. URL http://arxiv.org/abs/1312.4400.

Liu, Ming, Buntine, Wray L., and Haffari, Gholamreza. Learning how to actively learn: A deep imitation learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 1874–1883. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1174. URL https://aclanthology.org/P18-1174/.

MacKay, David JC. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Nguyen, Vu-Linh, Destercke, Sébastien, and Hüllermeier, Eyke. Epistemic uncertainty sampling. In *Discovery Science - 22nd International Conference, DS 2019, Split, Croatia, October 28-30, 2019, Proceedings*, volume 11828 of *Lecture Notes in Computer Science*, pp. 72–86. Springer, 2019. doi: 10.1007/978-3-030-33778-0_7. URL https://doi.org/10.1007/978-3-030-33778-0_7.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X. URL https://www.worldcat.org/oclc/61285753.

Ren, Pengzhen, Xiao, Yun, Chang, Xiaojun, Huang, Po-Yao, Li, Zhihui, Gupta, Brij B., Chen, Xiaojiang, and Wang, Xin. A survey of deep active learning. *ACM Comput. Surv.*, 54(9):180:1–180:40, 2022. doi: 10.1145/3472291. URL https://doi.org/10.1145/3472291.

Sener, Ozan and Savarese, Silvio. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Seung, H. Sebastian, Opper, Manfred, and Sompolinsky, Haim. Query by committee. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992, Pittsburgh, PA, USA, July 27-29, 1992*, pp. 287–294. ACM, 1992. doi: 10.1145/130385.130417. URL https://doi.org/10.1145/130385.130417.

Shannon, Claude E. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shui, Changjian, Zhou, Fan, Gagné, Christian, and Wang, Boyu. Deep active learning: Unified and principled method for query and training. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1308–1318. PMLR, 2020. URL http://proceedings.mlr.press/v108/shui20a.html.

Srivastava, Nitish, Hinton, Geoffrey E., Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. *Journal Machine Learning Research*, 15(1):1929–1958, 2014. doi: 10.5555/2627435.2670313. URL https://dl.acm.org/doi/10.5555/2627435.2670313.

Tong, Simon. *Active learning: theory and applications*. PhD thesis, Stanford University, USA, 2001. URL https://searchworks.stanford.edu/view/4762444.

Yin, Changchang, Qian, Buyue, Cao, Shilei, Li, Xiaoyu, Wei, Jishang, Zheng, Qinghua, and Davidson, Ian. Deep similarity-based batch mode active learning with exploration-exploitation. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pp. 575–584. IEEE Computer Society, 2017. doi: 10.1109/ICDM.2017.67. URL https://doi.org/10.1109/ICDM.2017.67.

Zhdanov, Fedor. Diverse mini-batch active learning. *CoRR*, 2019. URL http://arxiv.org/abs/1901.05954.

Appendix

Table 1: Constant weighted arithmetic mean

Uncertainty	Diversity	MNIST Acc.	CIFAR Acc.
Entropy	Discriminator	96.28	41.22
	Posterior Var	96.35	40.70
	Min distance	96.20	41.80
BALD	Discriminator	95.83	41.68
	Posterior Var	95.93	41.13
	Min distance	96.21	43.05
Var ratio	Discriminator	96.38	42.00
	Posterior Var	96.50	42.39
	Min distance	96.35	42.96

Table 2: Time-decayed weighted arithmetic mean

Uncertainty	Diversity	MNIST Acc.	CIFAR Acc.
Entropy	Discriminator	96.30	42.68
	Posterior Var	96.51	41.99
	Min distance	96.55	41.42
BALD	Discriminator	95.84	40.16
	Posterior Var	95.87	42.45
	Min distance	96.02	42.87
Var ratio	Discriminator	96.62	40.14
	Posterior Var	96.58	43.52
	Min distance	96.60	42.21

Table 3: Constant weighted geometric mean

Uncertainty	Diversity	MNIST Acc.	CIFAR Acc.
Entropy	Discriminator	96.36	41.73
	Posterior Var	96.55	40.34
	Min distance	96.26	42.63
BALD	Discriminator	96.06	39.14
	Posterior Var	96.40	38.38
	Min distance	95.71	43.53
Var ratio	Discriminator	96.59	42.43
	Posterior Var	94.62	42.66
	Min distance	96.19	42.21

Table 4: Time-decayed weighted geometric mean

Uncertainty	Diversity	MNIST Acc.	CIFAR Acc.
Entropy	Discriminator	95.64	40.54
	Posterior Var	96.09	40.34
	Min distance	96.26	43.14
BALD	Discriminator	96.15	40.27
	Posterior Var	96.05	40.98
	Min distance	95.65	43.27
Var ratio	Discriminator	96.73	41.66
	Posterior Var	96.75	41.45
	Min distance	96.59	43.60

Table 5: Uncertainty first two stage query

Uncertainty	Diversity	MNIST Acc.	CIFAR Acc.
Entropy	Discriminator	93.72	41.31
	Posterior Var	94.22	41.71
	Min distance	94.34	42.06
BALD	Discriminator	93.71	41.01
	Posterior Var	94.24	40.72
	Min distance	94.40	42.94
Var ratio	Discriminator	93.80	39.59
	Posterior Var	94.26	42.01
	Min distance	94.78	42.57

Table 6: Diversity first two stage query

Uncertainty	Diversity	MNIST Acc.	CIFAR Acc.
Entropy	Discriminator	93.52	41.39
	Posterior Var	94.23	41.34
	Min distance	94.27	41.88
BALD	Discriminator	94.25	43.98
	Posterior Var	94.23	40.28
	Min distance	94.25	43.37
Var ratio	Discriminator	94.39	41.01
	Posterior Var	94.41	42.91
	Min distance	94.28	42.09