Critical Gap Between Generalization Error and Empirical Error in Active Learning

Yusuke Kanebako
Ricoh Company, Ltd.
yuusuke.kanebako@jp.ricoh.com

Abstract

Conventional research papers on Active Learning (AL) have conducted evaluations based on the assumption that a large amount of annotated data is available for evaluating model performance apart from the data selected by AL. This evaluation method is not realistic for the setting where AL learns models with few annotation costs. If a large amount of annotated data is available, it should be used for both evaluation and training, not only for evaluation. Therefore, in a realistic model construction using AL, generalization error in the actual production environment should be estimated by cross-validation only using the data selected by AL. However, the data selected by AL tend to be a biased dataset because the data are selected based on some criteria. Therefore, there is a gap between the actual generalization error and the empirical error when conducting cross-validation on the AL-selected data. In addition, if validation is performed using only the selected dataset by AL, it is possible to fail to realize this fatal gap. In this paper, we show that cross-validation using selected data in conventional AL methods either overestimate or underestimate model performance. As a result, we show a significant difference between generalization error and empirical error from cross-validation.

1. Introduction

Deep Learning achieves high performance on a variety of tasks by learning large amounts of data. However, in actual machine learning projects, it is sometimes difficult to build a large dataset from the beginning of the project. In such cases, the dataset is built gradually, and the performance of the model is evaluated and its application to the actual production environment is verified. However, if the annotation itself is an expensive task that requires expert knowledge or if the amount of annotated data produced per unit of time is low, it is desirable to achieve maximum model performance at minimum annotation cost. In this respect, various studies have shown the effectiveness of Active Learning (AL) [7, 10, 18–20]. AL is a technique for predicting which data will be most effective in improving the performance of a model from a set of unlabeled data and annotating only the selected data. The most basic AL technique uses the uncertainty in the model's predictions. If the model's prediction uncertainty is high, it is probable that the data has features that have not yet been learned by the model, and learning this data can be expected to improve performance. The effectiveness of classical AL methods has been demonstrated for Deep Learning tasks [1, 5], and AL methods designed for Deep Learning have also been proposed [8]. In addition, a method is proposed for selecting data points from a dataset such that they cover the entire data distribution, not just based on uncertainty or the behavior of the predictions computed by the model [17]. These papers focus on efficiently collecting data for training and demonstrate that higher performance can be achieved with fewer data points compared to a random selection from the dataset.

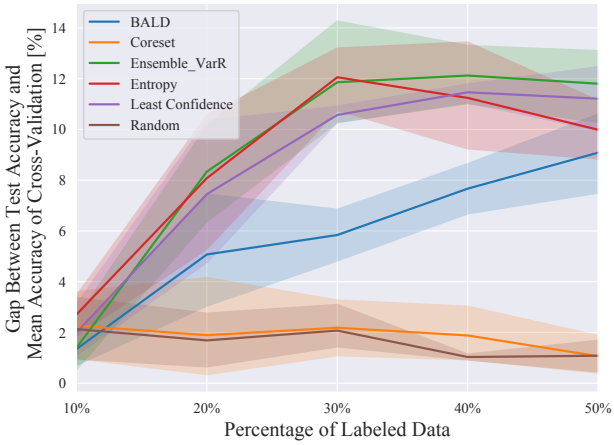


Figure 1. Gap between predicted generalization error from cross-validation using AL-selected data and generalization error using test data.

However, in the problem setup of the AL paper, there is

an unlabeled dataset to be used for training, and the data to be annotated are selected based on the designed acquisition function. Apart from this, a large amount of annotated data is used to evaluate the training results. In the CIFAR-10 [9], which is often used for AL validation in image classification tasks, 50,000 images are available for training and 10,000 images are available for testing. After training the model, we evaluate it on the 10,000 test images. In reality, there are limited situations where an unlabeled dataset exists as a training dataset, and a large amount of annotated evaluation data is available. It is natural that we use the annotated evaluation data for model training if it is available. In a realistic AL situation, large amounts of annotated evaluation data are not available, and cross-validation is performed on a dataset selected by the AL method or a set of randomly sampled dataset for evaluation. This will give an estimate of the generalization performance that the model can achieve in a real production environment. Dataset will continue to be collected until this estimated performance meets the target performance. However, it is noted that the dataset selected by AL generally result in a biased dataset set [3,4,11]. Therefore, cross-validation using the dataset selected by AL should be overestimate or underestimate the performance of the model. Suppose cross-validation using a biased dataset results in a low estimate of the model's performance. In that case, more data will be collected than necessary, while a high estimate of the model's performance will result in lost opportunities if the target performance is not achieved during actual operation.

In this paper, we clarify the problem of using a large amount of annotated evaluation data for validation, which has been overlooked in the validation of conventional AL methods. We show that the dataset collected by conventional AL methods is highly biased and that the generalization error estimated by cross-validation using these datasets has a large gap with the actual generalization error. We also propose a simple AL procedure to overcome this gap. Specifically, we use only the data selected by AL for training the model while using for validation all of the random datasets initially selected for AL model building. This is expected to result in a distribution of the data for evaluation that is close to that of the population, and as a result, it is possible to estimate forecast results that are close to those in actual operation.

The contributions of this paper can be summarized as follows.

- We clarified the problems in the problem setting and evaluation methods of existing AL papers.
- We showed that cross-validation using data selected by existing AL methods underestimates the generalization error.
- To solve the above problem, we proposed an AL procedure in which the first randomly selected data is used only for model training of the first AL step, and after that, only for model evaluation.

2. Related Works

2.1. Active Learning

The typical problem setting for AL is pool-based AL [10], where a dataset of unlabeled data is prepared and methods are used to select data that is effective for improving model performance. This is known to be effective when there is a cost associated with labeling [18]. One of the typical method for designing the acquisition function, which is the data selection policy, is to use the uncertainty in the model's predictions. Least Confidence [10], which selects the data with the smallest maximum model prediction probability, and methods that select data with high entropy in the model prediction probability distribution are simple methods [19]. Furthermore, methods have been proposed that select data with high entropy of predictions and low expected entropy of predictions regarding the posterior distribution of model parameters [5, 7]. Additionally, methods have been proposed that determine prediction uncertainty using ensemble of multiple models [1, 20]. While these methods can select data that contributes to the improvement of model performance, it has been pointed out that the selected data may be biased [3, 11]. Moreover, in [4], after pointing out the bias in data, it solves problems during learning by estimating statistical bias in data and introducing corrective weights to remove the bias.

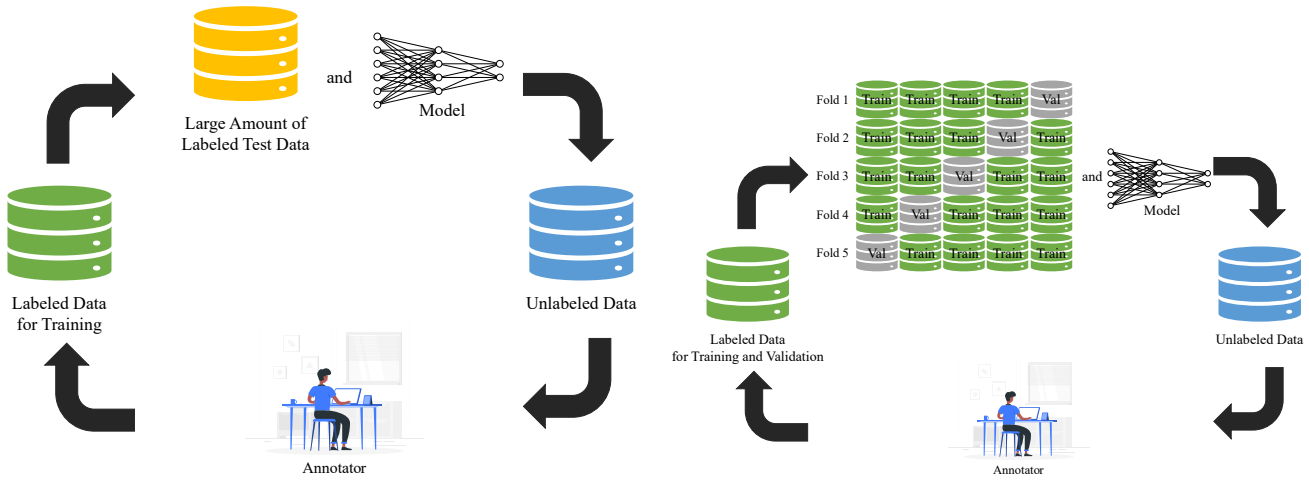
On the other hand, methods have been proposed that select data to provide a cover for all data points in the prepared unlabeled dataset, regardless of the model's behavior, using the geometric shape of the data points [17]. It is a method that tries to find a set of data points such that the distance from the nearest cover point to any data point is minimized. Therefore, it can be expected that the bias in data is smaller compared to AL methods that use model uncertainty.

Furthermore, in the early stages of AL, there are few data points, and the model may be overfitted when learning. Papers have pointed out the importance of regularization during model learning with AL [13, 15].

2.2. Dataset Optimization

In [12], a method is proposed that minimizes the expected future collection costs, which includes not only the annotation cost but also the penalty in case the designer fails to achieve the target. In [16, 22], it has been confirmed that by identifying important data from the training dataset based on gradient information and removing unnecessary data, test performance can be improved.

These papers evaluate algorithms and models under the



(a) Conventional Evaluation Procedure of Active Learning

(b) Evaluation procedure in This Paper

Figure 2. Comparison of the conventional AL procedure and the performance validation procedure in this paper.

assumption that a large amount of annotated data already exists; in that AL's motivation is to minimize annotation cost, realistically, if one has access to a large set of annotated data, it should be reasonable to use it for training. Therefore, the traditional AL problem setup is not realistic.

3. Gap Between Generalization Error and Empirical Error

In this paper, we do not have access to a large amount of annotated data when evaluating the model and consider estimating the generalization error by cross-validation from a set of datasets selected by the AL methodology. We show here that the datasets collected by AL are biased, especially based on the uncertainty of the model's predictions. We also show that estimating the generalization error by cross-validation using a biased dataset results in a large deviation from the actual generalization error.

3.1. Problem Settings

The objective in general supervised learning is to minimize the loss by a model f_θ with parameter θ for input data x and label y sampled from a population data distribution D. Let $l(y, f_\theta(x))$ denote the loss function, and define the generalization error $R(f_\theta)$ as follows:

$$R(f_\theta) = \mathbb{E}_{(x,y)\sim D}[l(f_\theta(x)), y] \tag{1}$$

In this paper, we consider a pool-based AL as a problem setup for AL. From a U unlabeled set $\mathcal{U} = \{x^{(u)}\}_{u=1}^{U}$, $\mathcal{L} = \{(x, y)^{(l)}\}_{l=1}^{L}$ is the annotated dataset. The empirical error $\hat{R}(f_\theta)$ of training the model f_θ using this dataset $\mathcal{L}$ is defined as follows:

$$\hat{R}(f_\theta) = \frac{1}{L}\sum_{l=1}^{L} l(f_\theta(x_l), y_l) \tag{2}$$

In the conventional AL problem setting, learning of the model f_θ is carried out by minimizing the Eq. (1) using the dataset $\mathcal{L}$. To check the generalization performance of the learned model, we evaluate the performance of the AL method by calculating the generalization error of Eq. (1) under the assumption that a large amount of annotated data is available. We consider this evaluation method to be impractical due to the motivation to minimize the annotation cost of AL. Therefore, we use cross-validation to predict the generalization error of the model. A comparison with conventional AL is shown in Fig. 2. Cross-validation is performed by dividing the annotated dataset $\mathcal{L}$ into K folds. The average empirical error $R^{CV}(f)$ of the empirical error $\hat{R}_k(f_{\theta_k})$ calculated by the training evaluation in each fold is defined as follows.

$$R^{CV}(f) = \frac{1}{K}\sum_{k=1}^{K} \hat{R}_k(f_{\theta_k}) \tag{3}$$

Here, if the dataset $\mathcal{L}$ selected in AL follows the data distribution D of the population, then Eq. (1) and Eq. (3) will asymptote with the number of data points. However, since the dataset selected based on model uncertainty in AL is a biased dataset, Eq. (1) and Eq. (3) are significantly different. This is shown in the next section.

3.2. Dataset Bias in Uncertainty Sampling

We assume that the pooled dataset $\mathcal{U}$ is i.i.d and its distribution is uniform. The data x^* to be selected based on uncertainty sampling with respect to entropy H can be written

as follows:

$$x^* = \underset{x \in \mathcal{U}/\mathcal{L}}{\text{argmax}} H(\hat{y}|x) \tag{4}$$

Here, $\hat{y} = f_\theta(x)$ is the prediction by the model trained on the data selected by AL. If ideal Uncertainty Sampling could be achieved in the AL framework, the sampled data would be preferentially selected from the classification boundaries of image classes. Therefore, data collected in the early to mid phases of AL will be more biased than in the case of random sampling or data selected geometrically from a distribution of data, as in the case of Coreset [17]. In addition, since the usual image classification uses cross-entropy as the loss function, the data selected based on the Eq. (4) criterion will be lossy data for the model at that point in time. Thus, the dataset $\mathcal{L}$ selected by AL is biased and hard training sample for the model. Therefore, when cross-validating with a dataset $\mathcal{L}$, we must carefully set up the data partitioning to accurately estimate the generalization error. This is also generally applicable to other uncertainty-based methods. For example, in MC Dropout [5], the posterior probability $p(y|x)$ of the model's predictive distribution is estimated using a model $f_{\theta'}$ with CNN weights Dropout [21] as follows:

$$p(y|x) = \frac{1}{T} \sum_{t=1}^{T} p(f_{\theta'}) \tag{5}$$

The sample with the largest entropy $H(p(y|x)$ is selected. In addition, Ensemble-VarR [1] trains N models, where f_m is the number of predictions to each class category, and the variance ratio is defined as follows to calculate uncertainty.

$$v = 1 - \frac{f_m}{N} \tag{6}$$

As shown above, a hard training sample is selected for the model in each AL method.

3.3. Cross-validation Procedure NOT Using Data Selected by AL for Validation

As shown in Sec. 3.2, the data distribution selected by AL is very different from that of randomly sampled data. Even if generalization performance is confirmed by cross-validation using these data, it is not possible to accurately estimate generalization performance because it is different from the actual population. For this reason, we present a simple solution in which the data initially sampled for AL are used only for validation. In the usual AL based on model uncertainty, it is necessary to randomly sample data in the first phase and train the model once on the data. After that, the model is used to sample data that are considered valid for learning in the AL procedure. In this case, the data selected in AL are used only for training, and only the data randomly sampled in the first stage are used for evaluation. This approach allows for the selection of learning data that are effective for the model's generalization performance. Additionally, the model can be evaluated on data that follow the actual population, eliminating the gap between generalization error and empirical error that occurs in all methods. The gap between generalization error and empirical error is eliminated in all methods, as shown in Fig. 1.

4. Experiments Settings

4.1. Dataset

In this experiment, we evaluated three datasets, CIFAR-10/100 [9] and SVHN [14].

4.2. Networks and Training Details

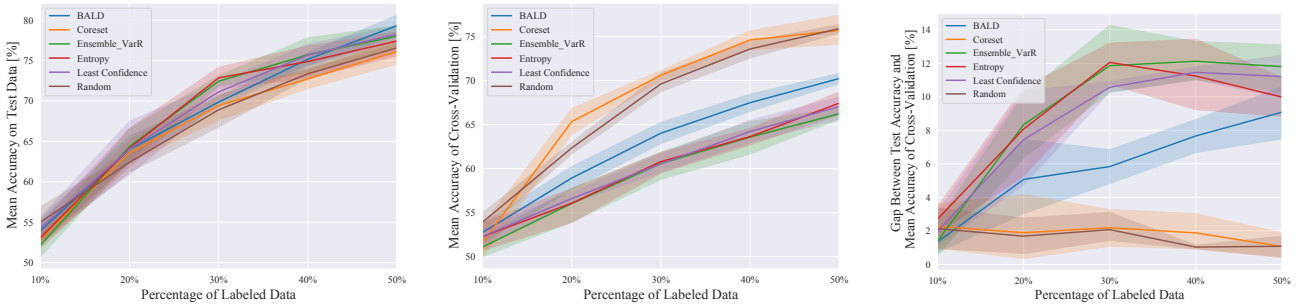
In this experiment, the network architecture is ResNet-18 [6], the batch size is set to 128, and training is carried out over 100 epochs. The learning rate is set to 0.001, β_1 to 0.9, and β_2 to 0.999. In addition, to confirm the importance of regularization in AL [13, 15], we also perform learning using Random Augmentation [2]. Random Augmentation parameters are N (the number of transformations) set to 1 and M (index of the magnitude) set to 2.

4.3. AL Methods

We use the existing AL methods from the Uncertainty-base method: Least Confidence [10], Entropy [19], BALD [5], Ensemble-VarR [1]. In addition, we use Coreset [17] and random sampling as methods that do not depend on the model's uncertainty. BALD performs 5 inference runs with a 20% dropout of the weights of fully connected layers of ResNet. Ensemble-VarR ensembles 5 models with the same settings as the models and training conditions used in this experiment, changing only the initial random seed of the models. The 5 models are used only for data selection from the pooled data.

4.4. Active Learning Settings and Evaluation

This section describes the experimental setup for AL. In this experiment, unless otherwise noted, we first randomly select 10 % of the data from the unlabeled dataset. First, the model is trained on the randomly selected data, and then the data are selected according to the acquisition function of each AL method. The sampling volume during the data selection step of AL is set to a number that corresponds to 10% of initial unlabeled dataset size at the beginning of AL. For example, if there are 50,000 unlabeled data before AL starts, 5,000 should be selected in each AL data selection step. In this experiment, generalization performance is confirmed by cross-validation. The number of data splits for cross-validation is set to 5, and the entire data selected by AL is randomly split. As a result of cross-validation,



(a) Test Data Accuracy (b) Mean Accuracy of Cross-Validation (c) Absolute Difference Between 3a and 3a

Figure 3. Test data accuracy, mean of cross-validation accuracy, and the gap between test data accuracy and cross-validation accuracy in CIFAR-10.

5 models are trained. The model with the smallest loss to the validation data segmented by cross-validation is used as the model for selecting the data in AL. In Coreset, data selection calculations are performed not in the original data space, but in the latent variable space consisting of the intermediate features of the models. As with other methods, the model used to calculate the latent variable space is the model with the smallest loss for the data to be validated in the cross-validation. In this experiment, we calculate the average performance of the AL models in cross-validation, the generalization performance of each dataset to the test data as a reference value, and the gap between the average performance in cross-validation and the generalization performance on the test data. The experiments for each AL method are averaged over 5 trials with different random initialization seeds.

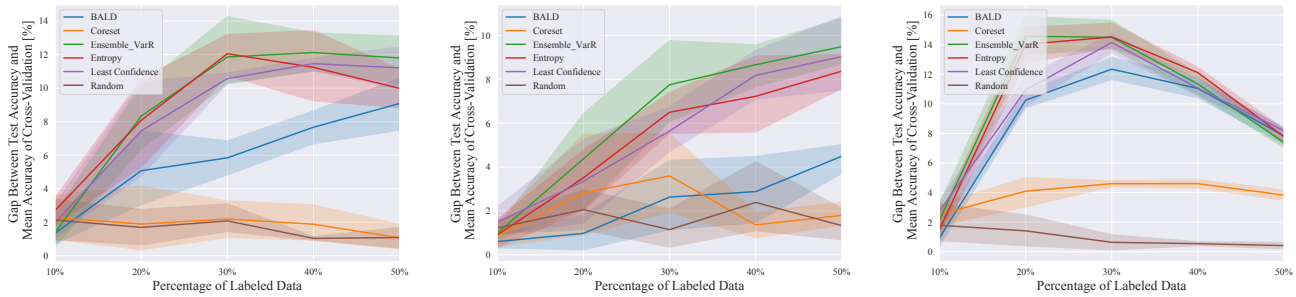
As an experiment to improve the gap between cross-validation and generalization performance, we will conduct an experiment in which only randomly selected data will be used for validation. In this case, the first model is trained on randomly selected data, and then the data are selected according to each AL method. After that, all the randomly selected data are used for validation, and training is performed only on the data selected by AL. Therefore, cross-validation is not performed because the data for validation is fixed. Here, we examine the impact of reducing the number of randomly selected data to 10%, 5%, and 1% of the unlabeled dataset. Also in this experiment, the sampling volume is set to a number that corresponds to 10% of initial unlabeled dataset size.

5. Results

5.1. Generalization Performance and Cross-Validation Accuracy in AL

Fig. 3a indicates the result of test data accuracy, Fig. 3b indicates the result of mean cross-validation accuracy in the data selected by AL, and Fig. 3c indicates the absolute difference between test data accuracy and cross-validation accuracy. BALD achieved the highest accuracy for the test data at the 50% data acquisition, while Coreset performed the lowest accuracy. In addition, Random and Coreset achieved the highest average accuracy in cross-validation using the data selected by AL, while the other Uncertainty-based methods performed significantly worse results. As a result, there is a large difference in cross-validation and generalization performance, as shown in Fig. 3c. This is because the data distribution sampled by the uncertainty-based method differs from the actual population, as described in Sec. 3.2. On the other hand, the Random selection and Coreset methods do not cause such a gap because the data are selected without bias. The point we wish to argue in this paper is that the results of the test data of Fig. 3a are originally unobservable results, and the generalization performance should be confirmed based on the results of Fig. 3b. However, the results shown in Fig. 3b underestimate the average accuracy of cross-validation compared to the true generalization performance in the Uncertainty-based method. Consequently, if AL is performed while referring to these results, it may lead to the collection of more data than necessary in actual applications.

Fig. 4 indicates the gap between test data accuracy and cross-validation accuracy in CIFAR-10/100 and SVHN. Similar to the results shown in Fig. 3c, the gap between the cross-validation results and the generalization performance occurred for CIFAR-100 and SVHN as well. For SVHN, the gap between the cross-validation results and generalization performance becomes smaller when the amount of data selected for AL exceeds 30%, which is thought to be due to the characteristics of the SVHN dataset. Fig. 5 shows the average value of the minimum coverage distance of the selected data points in each AL step using Coreset. In CIFAR-10/100, the minimum coverage distance does not start to decrease even when the percentage of pooled data selected in the AL step reaches 50%, but the minimum coverage distance starts to decrease from 30% in SVHN. This indicates



(a) CIFAR-10 (b) CIFAR-100 (c) SVHN

Figure 4. The gap between test data accuracy and cross-validation accuracy in CIFAR-10/100 and SVHN.

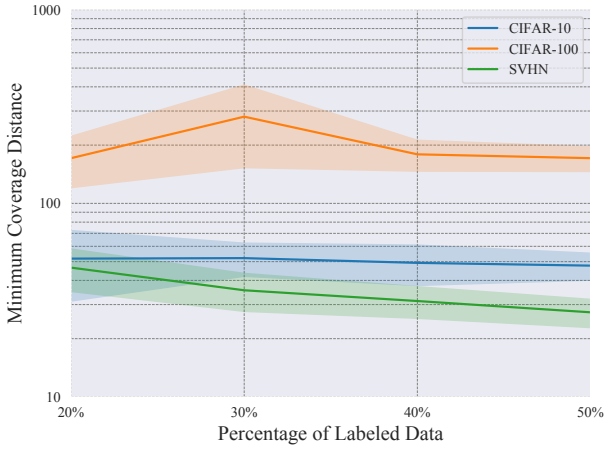
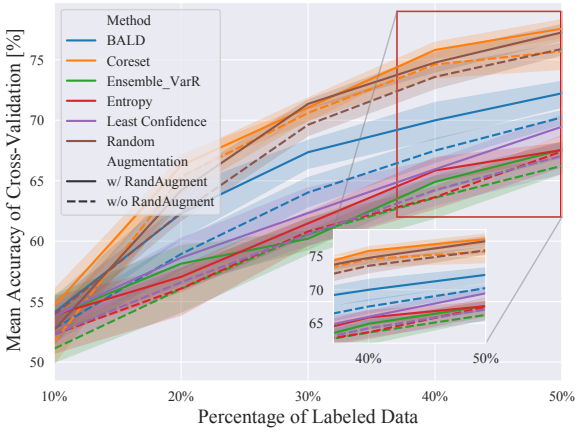
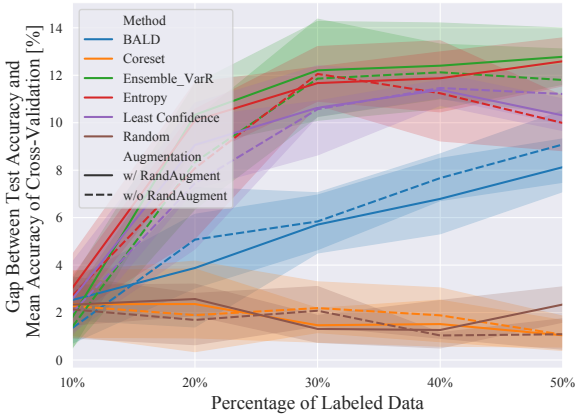


Figure 5. Average of the minimum coverage distance at each AL step in the core set sampling. Calculate the minimum coverage distance together with the already selected data for the pooled data at each AL step and compute the average of the minimum coverage distances.



(a) Mean Accuracy of Cross-Validation

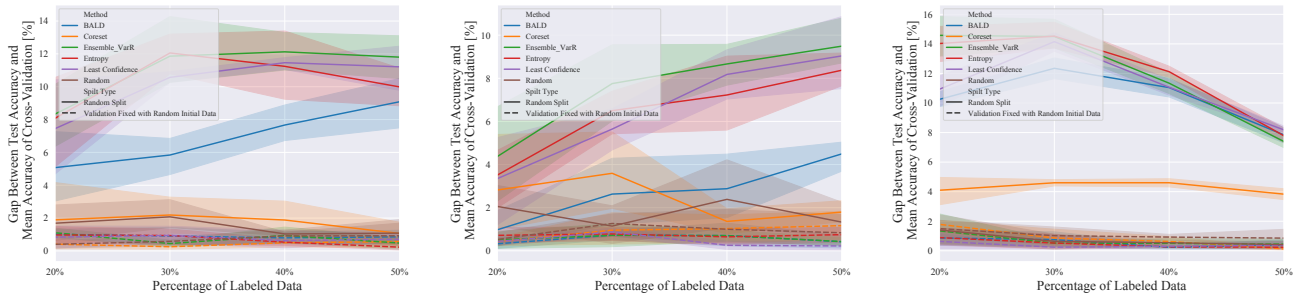


(b) Gap Between Mean of Cross-Validation Accuracy and Test Data Accuracy

Figure 6. Gap between cross-validation and generalization performance when Random Augmentation is applied in CIFAR-10. Although the average percentage of correct answers for cross-validation has improved for both AL methods, the gap between the cross-validation and generalization performance has not been eliminated.

that SVHN lacks data diversity compared to CIFAR-10/100 and that the AL phase is able to construct a dataset that encompasses the entire dataset from an early stage. Therefore, as shown in Fig. 7c, the Uncertainty-based method is also able to construct a set that encompasses the entire dataset, which is thought to lower the gap between cross-validation and generalization performance.

Fig. 6 shows the results of training with Random Augmentation to confirm the effect of regularization in phases with a small number of AL data. As shown in Fig. 6a, the addition of Random Augmentation improves the cross-validation results for each AL method. The results show that with Random Augmentation, the same performance can be achieved with approximately 10% less data than without Random Augmentation. However, as shown in Fig. 6b, there is still a large gap between cross-validation and generalization performance, and this problem has not been resolved by regularization.



(a) CIFAR-10 (b) CIFAR-100 (c) SVHN

Figure 7. Comparison of the results of cross-validation with random splits on the data selected in AL with the results of using only the first randomly selected data in AL.

5.2. Experiments using ONLY Randomly Selected Data for Validation

Fig. 7 shows the results of using only the first randomly selected data as validation data before the data selection step of AL. The solid line shows the results of the random cross-validation shown in Fig. 4, and the dotted line shows the results of the validation using only the first randomly selected data. It can be seen that the gap between cross-validation and generalization performance is suppressed for both AL methods and dataset. This indicates that by fixing the data for validation to data randomly selected from the unlabeled dataset, the bias of the data for validation is less than in the case of random cross-validation, and the validation can be performed in accordance with the population. This result also indicates that the Uncertainty-based method is biased in its selection of data.

In the previous experiments, 10% of the training data for each dataset was sampled randomly at the beginning, but the results for different amounts of data sampled at the beginning are shown in Fig. 8. It can be seen that the gap between cross-validation and generalization performance increases as the amount of data sampled first decreases from 10% to 5% and 1%. However, even when the amount of data initially sampled is 1%, the gap between cross-validation and generalization performance is no larger than when cross-validation is performed by randomly splitting the data, indicating that the gap between cross-validation and generalization performance has been significantly improved. The number of images for training and validation is 50,000 in CIFAR-10/100 and 73,257 in SVHN, so 1% of each dataset corresponds to 500 images and 732 images, respectively. Even with this small amount of data, it is possible to correctly evaluate the model in the framework of AL if data that correctly follows the population is selected as the data for validation.

6. Conclusions

This paper discusses that the evaluation methods used in conventional AL papers are not realistic, because they require a large amount of annotated data. We showed that, in a realistic setting, cross-validation based on the data selected by AL does not correctly estimate generalization performance. We showed that AL with uncertainty biases the selected data, resulting in a large gap between the cross-validation results and generalization performance. We also showed that a simple solution to this problem is to use only randomly selected data in the first stage of AL for validation.

7. Limitations

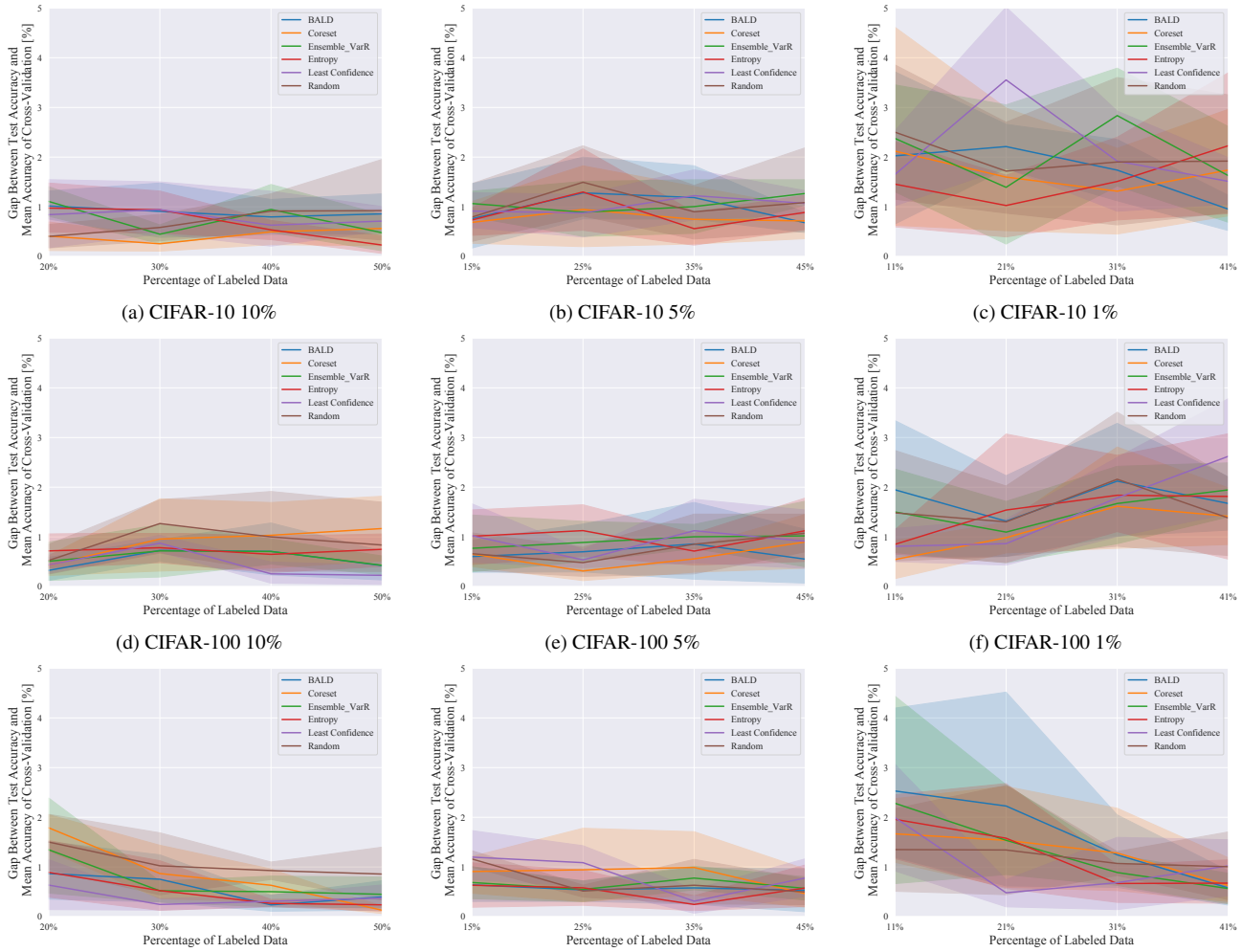
In this paper, we consider a pooled-based AL problem setup. The assumption is that the population and the pooled data distribution are uniformly distributed without any bias toward a particular class. In actual problems, there is no guarantee that the population is uniformly distributed, nor is there any guarantee that the pooled dataset is unbiased. Although this paper has proposed a realistic learning and evaluation procedure for AL under the assumption of uniform distribution, it is necessary to consider evaluation methods for a variety of problem settings and tasks.

References

[1] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 4

[2] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. 4

[3] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International*



(a) CIFAR-10 10% (b) CIFAR-10 5% (c) CIFAR-10 1%

(d) CIFAR-100 10% (e) CIFAR-100 5% (f) CIFAR-100 1%

(g) SVHN 10% (h) SVHN 5% (i) SVHN 1%

Figure 8. Results for different amounts of data randomly selected at the beginning of AL and set as data for validation in CIFAR-10/100 and SVHN.

Conference on Machine Learning, ICML '08, page 208–215, New York, NY, USA, 2008. Association for Computing Machinery. 2

[4] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021. 2

[5] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR, 06–11 Aug 2017. 1, 2, 4

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[7] Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011. 1, 2

[8] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. *BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning*. Curran Associates Inc., Red Hook, NY, USA, 2019. 1

[9] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 2, 4

[10] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, pages 3–12. ACM/Springer, 1994. 1, 2, 4

[11] David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992. 2

[12] Rafid Mahmood, James Lucas, Jose M. Alvarez, Sanja Fidler, and Marc T. Law. Optimizing data collection for machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, November 2022. 2

[13] Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. Towards robust and reproducible active learning using neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 223–232, June 2022. 2, 4

[14] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 4

[15] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza (Reza) Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12237–12246, June 2022. 2, 4

[16] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, 2021. 2

[17] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 1, 2, 4

[18] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 1, 2

[19] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii, Oct. 2008. Association for Computational Linguistics. 1, 2, 4

[20] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA, 1992. Association for Computing Machinery. 1, 2

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 4

[22] Kailas Vodrahalli, Ke Li, and Jitendra Malik. Are all training examples created equal? an empirical study. *CoRR*, abs/1811.12569, 2018. 2