Consistency-based Semi-supervised Active Learning: Towards Minimizing Labeling Cost

Mingfei Gao[1⋆], Zizhao Zhang[2], Guo Yu[3], Sercan Ö. Arık[2], Larry S. Davis[1], and Tomas Pfister[2]

[1]University of Maryland [2]Google Cloud AI [3]University of Washington

Abstract. Active learning (AL) combines data labeling and model training to minimize the labeling cost by prioritizing the selection of high value data that can best improve model performance. In pool-based active learning, accessible unlabeled data are not used for model training in most conventional methods. Here, we propose to unify unlabeled sample selection and model training towards minimizing labeling cost, and make two contributions towards that end. First, we exploit both labeled and unlabeled data using semi-supervised learning (SSL) to distill information from unlabeled data during the training stage. Second, we propose a consistency-based sample selection metric that is coherent with the training objective such that the selected samples are effective at improving model performance. We conduct extensive experiments on image classification tasks. The experimental results on CIFAR-10, CIFAR-100 and ImageNet demonstrate the superior performance of our proposed method with limited labeled data, compared to the existing methods and the alternative AL and SSL combinations. Additionally, we also study an important yet under-explored problem – "When can we start learning-based AL selection?". We propose a measure that is empirically correlated with the AL target loss and is potentially useful for determining the proper starting point of learning-based AL methods.

Keywords: Active Learning, Semi-supervised Learning, Consistency-based Sample Selection.

1 Introduction

Deep learning models are improved when trained with more labeled data [19]. A standard deep learning procedure involves constructing a large-scale labeled dataset and optimizing a model on it. Yet, in many real-world scenarios, large-scale labeled datasets can be very costly to acquire, especially when expert annotators are required, as in medical diagnosis. An ideal framework would integrate data labeling and model training to improve model performance with minimal amount of labeled data.

⋆ Work done while the author was an intern at Google; now at Salesforce Research. Email: mgao@cs.umd.edu

Active learning (AL) [2] assists the learning procedure by judicious selection of unlabeled samples for human labeling, with the goal of maximizing the model performance with minimal labeling cost. We focus on practically-common pool-based AL, where an unlabeled data pool is given initially and the AL mechanism iteratively selects batches to label in conjunction with training.

Learning-based AL methods select a batch of samples for labeling with guidance from the previously-trained model and then add these samples into the labeled dataset for the model training in the next cycle. Existing methods generally start with a randomly sampled labeled set. The size of the starting set affects learning-based AL performance – when the start size is not sufficiently large, the models learned in subsequent AL cycles are highly-biased which results in poor selection, a phenomenon commonly known as the *cold start problem* [26,23]. When cold start issues arise, learning-based selection yields samples that lead to lower performance improvement than naive uniform sampling [26].

To improve the performance at early AL cycles when the amount of labeled data is limited, it is important to address cold-start and ensure high performance later on with low labeling cost. Along this line of research, one natural idea for pool-based AL is integration of abundant unlabeled data into learning using semi-supervised learning (SSL) [55,48]. Recent advances in SSL [5,50,51,45,46] has demonstrated the vast potential of utilizing unlabeled data for learning effective representations. Although "semi-supervised AL" seems natural, only a small portion of AL literature has focused on it. Past works that use SSL for AL [14,36,55,39] treated SSL and AL independently without considering their impact on each other. We on the other hand, hypothesize that a better AL selection criterion should be in coherence with the corresponding objectives of unlabeled data in SSL to select the most valuable samples. A primary reason is that SSL already results in embodiment of knowledge from unlabeled data in a meaningful way, thus AL selection should reflect the value of additionally collected labeled data on top of such embodied knowledge. Based on these motivations, we propose an AL framework that integrates SSL to AL and also a selection metric that is highly related to the training objective.

The proposed AL framework is based on an insight that has driven recent advances in SSL [5,50,51] – a model should be consistent in its decisions between a sample and its meaningfully distorted versions, obtained via appropriate data augmentation. This motivates us to introduce an AL selection strategy: *a sample along with its distorted variants that yields low consistency in predictions indicates that the SSL model may be incapable of distilling useful information from that unlabeled sample – thus human labeling is needed.*

Overall, **our contributions** are summarized as follows:

1. We propose to unify model training and sample selection with a semi-supervised AL framework. The proposed framework outperforms the previous AL methods and the baselines of straightforward SSL and AL combinations.
2. We propose a simple yet effective selection metric based on sample consistency which implicitly balances sample uncertainty and diversity during se-

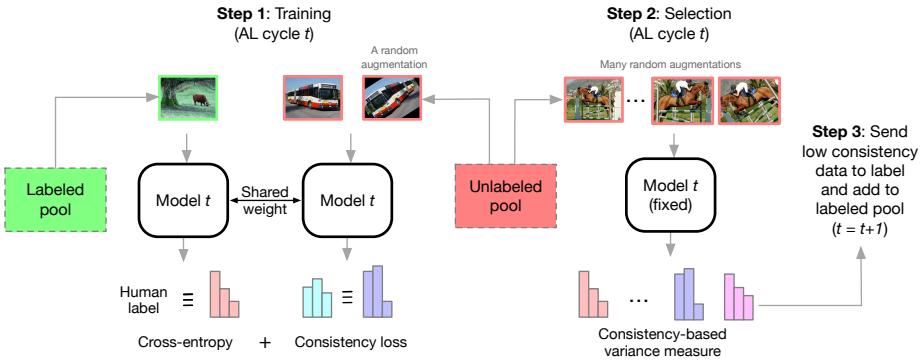

Fig. 1. Illustration of the proposed framework at t^{th} AL cycle. During training, both labeled and unlabeled data are used for the model optimization, with cross-entropy loss encouraging correct predictions for the labeled samples and consistency-based loss encouraging consistent outputs between unlabeled samples and their augmentations. During sample selection, the unlabeled samples and their augmentations are evaluated by the model obtained from the training stage. Their outputs are measured by our proposed consistency-based metric. The samples with low consistency scores are selected for labeling and sent to the labeled pool

lection. With comprehensive analyses, we demonstrate the rationale behind the proposed consistency-based sampling.

3. We propose a measure that is potentially useful for determining the proper start size to mitigate cold start problems in AL.

2 Related Work

2.1 Active learning

There exists a broad literature on AL [12,11,2,8]. Most AL methods can be classified under three categories: uncertainty-based methods, diversity-based methods and methods based on model performance change. Most uncertainty-based methods use *max entropy* [30,31] and *max margin* [37,3,25] criteria due to their simplicity. Some others use distances between samples and the decision boundary [49,6]. Most uncertainty-based methods use heuristics, while recent work [53] directly learns the target loss of inputs jointly with the training phase and shows promising results. Diversity-based methods select diverse samples that span the input space maximally [34,32,22,39]. There are also methods that consider uncertainty and diversity in conjunction [21,15,52]. The third category estimates the future model status and selects samples that encourage optimal model improvement [38,40,16].

2.2 Semi-supervised active learning

Both AL and SSL aim to improve learning with limited labeled data, thus they are naturally related. Yet, only a few recent works have considered combining AL and SSL. In [14], joint application of SSL and AL is considered for speech understanding, and significant error reduction is demonstrated with limited labeled speech data. Their AL selection criteria is based on a confidence score obtained from the posterior probabilities of the decoded text. Rhee *et al.* [36] propose an semi-supervised AL system which demonstrates superior performance in the pedestrian detection task. Zhu *et al.* [55] combine AL and SSL using Gaussian fields. Sener *et al.* [39] also consider SSL during AL cycles. However, in their setting, the performance improvement is marginal when adding SSL in comparison to their supervised counterpart, potentially due to the sub-optimal SSL method and combination strategy. Recently, Sinha *et al.* propose VAAL in [44], where a variational autoencoder and an adversarial network are learned using both labeled and unlabeled samples to infer the representativeness of unlabeled samples during the sampling process. Although, unlabeled data is not used for model training. The concurrent AL works [42][47] also consider integrating SSL, but their selection procedures are independent from the model training. We demonstrate that our proposed method unifying AL selection with SSL training is superior than the straightforward-combination strategy.

2.3 Agreement-based active learning

Agreement-based methods, also referred as "query-by-committee", base the selection on the opinions of a committee which consists of independent AL metrics or models [41,7,33,24,4,9]. Our method is related to agreement-based AL where samples are determined based on the conformity of different metrics or models. Specifically, our method selects samples that mostly disagree with the predictions of their augmentations.

3 Consistency-based Semi-supervised AL

We consider the setting of pool-based AL, where an unlabeled data pool is available for selection of samples to label. To minimize the labeling cost, we propose a method that unifies selection and model updates. The proposed semi-supervised AL is depicted in Fig. 1.

Most conventional AL methods base model learning only on the available labeled data, ignoring the useful information in the unlabeled data. While, we incorporate a semi-supervised learning (SSL) objective at training phases of AL cycles. The target model M_t at AL selection cycle t is learned by minimizing an objective loss function of the form $\mathcal{L}_l + \mathcal{L}_u$, where L_l and L_u indicate supervised and unsupervised losses, respectively. $\mathcal{L}_l$ is the supervised learning objective, such as the standard cross-entropy loss for classification. The proposed semi-supervised AL framework is presented in Algorithm 1. For $\mathcal{L}_u$, we adopt the

Algorithm 1 A semi-supervised learning based AL framework

Require: Unlabeled data pool $\mathcal{D}$, the total number of steps T, selected sample batch set B, AL batch size K, start size $K_0 \ll |\mathcal{D}|$
$B_0 \leftarrow$ uniformly sampling from $\mathcal{D}$ with $|B_0| = K_0$
$U_0 \leftarrow \mathcal{D} \backslash B_0$
$L_0 \leftarrow \{(x, \mathcal{J}(x)) : x \in B_0\}$, where $\mathcal{J}(x)$ stands for the assigned label of x.
for $t = 0, \ldots, T-1$ **do**
(training) $M_t \leftarrow \arg\min_M \left\{ \frac{1}{|L_t|} \sum_{(x,y)\in L_t} \mathcal{L}_l(x, y, M) + \frac{1}{|U_t|} \sum_{x \in U_t} \mathcal{L}_u(x, M) \right\}$
(selection) $B_{t+1} \leftarrow \arg\max_{B \subset U_t} \{\mathcal{C}(B, M_t),\ s.t.\ |B| = K\}$
(labeling) $L_{t+1} \leftarrow L_t \cup \{(x, \mathcal{J}(x)) : x \in B_{t+1}\}$
(pool update) $U_{t+1} \leftarrow U_t \setminus B_{t+1}$
end for
$M_T \leftarrow \arg\min_M \left\{ \frac{1}{|L_T|} \sum_{(x,y)\in L_T} \mathcal{L}_l(x, y, M) + \frac{1}{|U_T|} \sum_{x \in U_T} \mathcal{L}_u(x, M) \right\}$
return M_T

recent successful advances in SSL [1,5,54,50], that are based on minimizing the notion of sensitivity to perturbations with the idea of inducing "consistency", i.e., imposing similarity in predictions when the input is perturbed in a way that would not change its perceptual content. For consistency-based SSL, the common choice for the loss is

$$\mathcal{L}_u(x, M) = D(P(\hat{Y} = \ell | x, M), P(\hat{Y} = \ell | \tilde{x}, M)), \tag{1}$$

where D is a distance function such as KL divergence [51], or L2 norm [28,5], M indicates the model and $\tilde{x}$ denotes a distortion (augmentation) of the input x.

The design of the selection criteria is crucial while integrating SSL into AL. The unsupervised objective exploits unlabeled data by encouraging consistent predictions across slightly-distorted versions of each unlabeled sample. We hypothesize that *labeling samples that have highly-inconsistent predictions should be valuable, because these samples are hard to be minimized using* $\mathcal{L}_u$. Human annotations on them can ensure a correct label, to be useful for supervised model training at next cycle. The samples that yield the large performance gains with SSL would not be necessarily the samples with the highest uncertainty, as the most uncertain data could be out-of-distribution examples, and including them in training might be misleading. Based on the intuitions, we argue that, for semi-supervised AL, valuable samples are the ones that demonstrate highly unstable predictions given different input distortions, i.e., the samples that a model can not consistently classify as a certain class.

To this end, we propose a simple metric to quantify the inconsistency of predictions over a random set of data augmentations given a sample:

$$\mathcal{E}(x, M) = \sum_{\ell=1}^{J} \mathrm{Var}\left[P(\hat{Y} = \ell | x, M),\ P(\hat{Y} = \ell | \tilde{x}_1, M),\ ...,\ P(\hat{Y} = \ell | \tilde{x}_N, M) \right], \tag{2}$$

where J is the number of response classes and N is the number of perturbed samples of the original input data x, $\{\tilde{x}_1, ..., \tilde{x}_N\}$, which can be obtained by standard augmentation operations.

For batch selection, we jointly consider K samples and aim to choose the subset B such that the aggregate metric $\mathcal{C}(B, M) = \sum_{x \in B} \mathcal{E}(x, M)$ is maximized, i.e. the high inconsistency samples can be selected to be labeled by humans.

4 Experiments

In this section, we present experimental results of our proposed method. First, we compare our method to naive AL and SSL combinations, to show the effectiveness of our consistency based selection when all the methods are trained in a semi-supervised way. Second, since most recent AL methods still use only labeled data to conduct model training, we compare our method to recent AL methods and show a large improvement, motivating future research for semi-supervised AL. Third, we present qualitative analyses on several important properties of the proposed consistency based sampling.

4.1 Experimental setup

Datasets. We demonstrate the performance of our method on CIFAR-10, CIFAR-100 [27] and ImageNet [13] datasets. CIFAR-10 and CIFAR-100 have 60K images in total, of which 10K images are for testing. CIFAR-10 consists of 10 classes and CIFAR-100 has 100 classes. ImageNet is a large-scale image dataset with 1.2M images from 1K classes.

Implementation details. Different variants of SSL methods encourage consistency loss in different ways. In our implementation, we focus on the recently-proposed method, Mixmatch [5], which proposes a specific loss term to encourage consistency of unlabeled data. For comparison with selection baselines on CIFAR-10 and CIFAR-100, we use Wide ResNet-28 [35] as the base model and keep the default hyper-parameters for different settings following [5]. In each cycle, the model is initialized with the model trained in the previous cycle. 50 augmentations of each image are obtained by horizontally flipping and random cropping, but we observe that 5 augmentations can produce comparable results. To perform a fair comparison, different selection baselines start from the same initial model. The initial set of labeled data is randomly sampled and is uniformly distributed over classes. When comparing with advanced supervised AL methods, we follow [44] for the settings of start size, AL batch size and backbone architecture (VGG16 [43]). We adopt an advanced augmentation strategy, RandAugment [10], to perform augmentation of unlabeled samples on ImageNet.

Selection baselines. We consider three representative selection methods. *Uniform* indicates random selection (no AL). *Entropy* is widely considered as an uncertainty-based baseline in previous methods [39,53]. It selects uncertain samples that have maximum entropy of its predicted class probabilities. *k-center* [39] selects representative samples by maximizing the distance between a selected

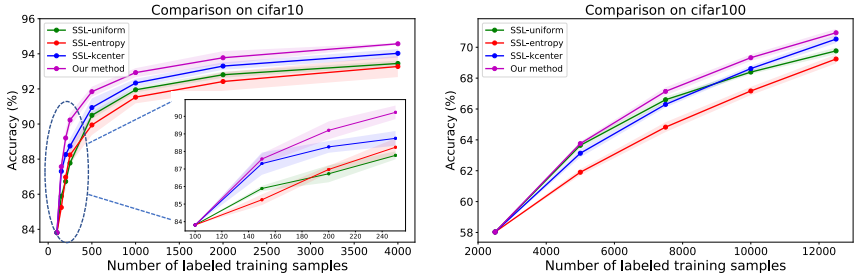

Fig. 2. Model performance comparison with different sample selection methods on CIFAR-10 and CIFAR-100. Solid lines indicate the averaged results over 5 trials. Shadows represent standard deviation

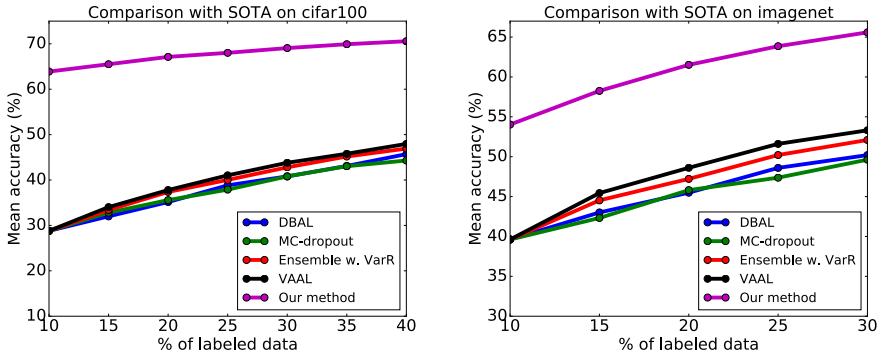

Fig. 3. Comparison with recent AL methods on CIFAR-100 and ImageNet. Our results on CIFAR-100 and ImageNet are averaged over 5 and 3 trials, respectively

sample and its nearest neighbor in the labeled pool. We use the features from the last fully connected layer of the target model to compute sample distances.

4.2 Comparison with selection baselines under SSL

To demonstrate the effectiveness of our method over the straightforward AL and SSL combinations, we focus on comparing with different selection methods in SSL framework. Fig. 2 and Table 1 show that when integrated with SSL training, our method outperforms baselines by a clear margin: on CIFAR-10, with 250 labeled images, our method outperforms *uniform* (passive selection) by $\sim 2.5\%$ and outperforms *k-center*, by $\sim 1.5\%$. As the number of labels increases, it is harder to improve model performance, but our method outperforms the *uniform* selection with 4K labels using only 2K labels, halving the labeled data requirements for the similar performance. Given access to all the labels (50K) for the

Table 1. Comparison of different sampling methods on CIFAR-10. Note that all the methods are under the SSL setting and start with 100 labeled samples. When the number of labeled samples reaches 250, AL batch size K is set to be 250 and doubled afterwards. The reported results are averaged over 5 trials

Methods	# of labeled samples in total				
	250	500	1000	2000	4000
Uniform	87.78±0.23	90.50±0.21	91.95±0.15	92.81±0.17	93.45±0.16
Entropy	88.24±0.51	89.95±0.58	91.53±0.35	92.42±0.53	93.28±0.61
k-center	88.75±0.42	90.94±0.53	92.34±0.24	93.30±0.21	94.03±0.25
Ours	**90.23±0.39**	**91.84±0.29**	**92.93±0.26**	**93.78±0.38**	**94.57±0.06**

entire training set, a fully-supervised model achieves an accuracy of 95.83% [5]. Our method with 4K (8%) examples achieves about 30% more error compared to the fully supervised method. CIFAR-100 is a more challenging dataset as it has the same amount of training images of CIFAR-10, but 10× more categories. On CIFAR-100, we observe a consistent outperformance over baselines of our method at all AL cycles.

There is typically a trade-off between using a large and a small AL batch sizes. A large batch size will lead to insufficient usage of active learning given a limited budget. However, selecting a small batch of samples would lead to more AL cycles, which is computationally expensive. We conduct experiments on CIFAR-10 following the setting in Fig. 2 using reasonable AL batch sizes. Results show that when consuming 200 labels in total, our methods obtain comparable performance (89.5%, 89.2% and 89.3%) with AL batch size set to be 25, 50 and 100, respectively.

4.3 Comparison with supervised AL methods

We have shown that our method clearly outperforms the straightforward AL and SSL combinations in Sec. 4.2. As mentioned, most AL methods focus on learning with only labeled samples. Consequently, it is worth showing the overall gap between our proposed framework and the existing methods to emphasize the benefit of the proposed framework. We choose the following recent methods as baselines: MC-Dropout [17], DBAL [18], Ensembles w. VarR [4] and VAAL [44] and compare with them on CIFAR-100 and ImageNet. The results of the baselines are reprinted from [44].

As can be seen from Fig. 3, our method significantly outperforms the existing supervised AL methods at all AL cycles on both datasets. Specifically, when 40% images are labeled, our method improves the best baseline (VAAL) by 22.62% accuracy on CIFAR100 and by 12.28% accuracy on ImageNet. The large improvements are mostly due to effective utilization of SSL at AL cycles.

Moreover, the performance of our method over the supervised models combined with the selection baselines in the scenario of very few labeled samples is of interest. As shown in Table 2, our method significantly outperforms the

Table 2. Comparison between our method (trained in SSL) and our baselines that trained in supervised setting with very few labeled samples on CIFAR-10. All methods start from 100 labeled samples. The following columns are results of different methods with the same selection batch size. The reported results are over 5 trials

Setting	Methods	# of labeled samples in total			
		100	150	200	250
Supervised	Uniform	41.85	46.13±0.38	51.10±0.60	53.45±0.71
	Entropy		46.05±0.34	50.15±0.79	52.83±0.82
	k-center		48.33±0.49	50.96±0.45	53.77±1.01
Semi-supervised	Ours	83.81	**87.57±0.31**	**89.20±0.51**	**90.23±0.49**

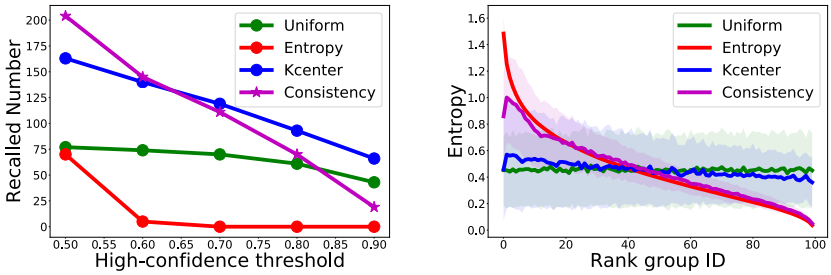

Fig. 4. Left: Number of overconfident mis-classified samples in top 1% samples ranked by different methods. Overconfident samples are defined as those having the highest class probability larger than threshold. Right: the average entropy of unlabeled samples ranked by different selection metrics. The ranked samples are divided into 100 groups for computing average entropy. Shadows represent standard deviation

methods which only learn from labeled data at each cycle. When 150 samples in total are labeled, our method outperforms *kcenter* by 39.24% accuracy.

4.4 Analyses of consistency-based selection

To build insights on the superior performance of our AL selection method, we analyze different attributes of the selected samples, which are considered to be important for AL. Experiments are conducted on CIFAR-10.

Uncertainty and overconfident mis-classification. Uncertainty-based AL methods query the data samples close to the decision boundary. However, deep neural networks yield poorly-calibrated uncertainty estimates when the raw outputs are considered – they tend to be overconfident even when they are wrong [20,29]. Entropy-based AL metrics would not distinguish such overconfident mis-classifications, thus may result in sub-optimal selection. Fig. 4 (left) demonstrates that our *consistency*-based selection is superior in detecting high-confident mis-classification cases compared to *entropy*-based selection. Fig. 4(right)

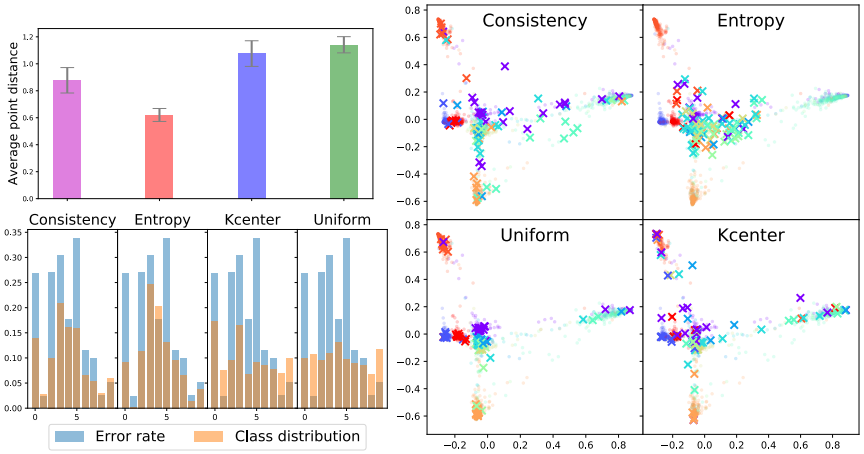

Fig. 5. Average distance between samples (top-left): the average pair-wise L_2 distance of top 1% unlabeled samples ranked by different selection metrics. Per-class error rate vs. the class distribution of the selected samples are shown in bottom-left. Diversity visualization (right): Dots and crosses indicate unlabeled (un-selected) samples and the selected samples (top 100), respectively. Each color represent a ground truth class

shows the uncertainty in the selected samples with different methods, quantified using entropy. When different AL selection methods are compared, *uniform* and *k-center* methods do not base selection on uncertainty at all, whereas *consistency* tends to select highly-uncertain samples but not necessarily the top ones. Such samples might contribute to our superior performance compared to *entropy*.

Sample diversity. Diversity has been proposed as a key factor for AL [52]. *k-center* is a diversity based AL method, preferring to select data points that span the whole input space. Towards this end, Fig. 5 (right) visualizes the diversity of samples selected by different methods. We use principal component analysis to reduce the dimensionality of embedded samples to a two-dimensional space. *Uniform* chooses samples equally-likely from the unlabeled pool. Samples selected by *entropy* are clustered in certain regions. On the other hand, *consistency* selects data samples as diverse as those selected by *k-center*. The average distances between top 1% samples selected by different methods are shown in Fig. 5 (top-left). We can see that *entropy* chooses samples relatively close to each other, while *consistency* yield samples that are separated with much larger distance which are comparable to samples selected by *uniform* and *k-center*.

Class distribution complies with classification error. Fig. 5 (bottom-left) shows the per-class classification error and the class distribution of samples selected by different metrics. Samples selected by *entropy* and *consistency* are correlated with per class classification error, unlike the samples selected by *uniform* and *k-center*.

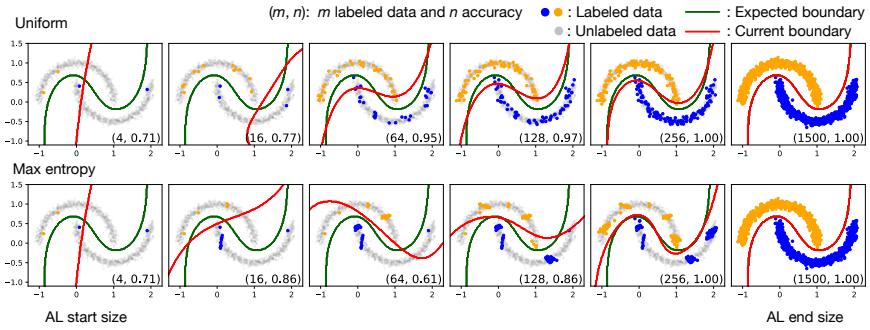

Fig. 6. Illustration of cold-start problems for uncertainty-based AL. When the learned decision boundary is far away from the expected boundary (the boundary when all labels are available for the entire training set), e.g. the second and third columns, the selected samples by uncertainty-based AL is biased, leading to sub-optimal performance

5 When can we start learning-based AL selection?

Based on the studies above, our proposed semi-supervised AL framework demonstrates clear advantages. While towards minimizing the labeling cost, a challenging issue, cold start failure, may occur when only a extreme small labeled set is available, which leads to sub-optimal AL performance. The proper study of this problem is essential for scenarios especially when labels are very expensive or challenging to collect.

5.1 Cold-start failure

When the size of the initial labeled dataset is too small, the learned decision boundaries could be far away from the real boundaries and AL selection based on the model outputs could be biased. To illustrate the problem, Fig. 6 shows the toy two-moons dataset using a simple support vector machine model with an RBF kernel, trained in supervised setting to learn the decision boundary [35]. As can be seen, the naive uniform sampling approach achieves better predictive accuracy by exploring the whole space. On the other hand, the samples selected by *max entropy* concentrate around the poorly-learned boundary.

Next, we study the cold start phenomenon for our proposed semi-supervised AL method. We focus on CIFAR-10 with small labeled initial sets, shown in Fig. 7. Using uniform sampling to select different starting sizes, AL methods achieve different accuracies. For example, the model starting with $K_0 = 50$ data points clearly under-performs the model starting with $K_0 = 100$ samples, when both models reach 150 labeled samples. It may be due to the cold start problem encountered when $K_0 = 50$. On the other hand, given a limited labeling budget, naively choosing a large start size is also not practically desirable, because it may lead to under-utilization of learning-based selection. For example, starting with $K_0 = 100$ labeled samples has better performance than starting from 150

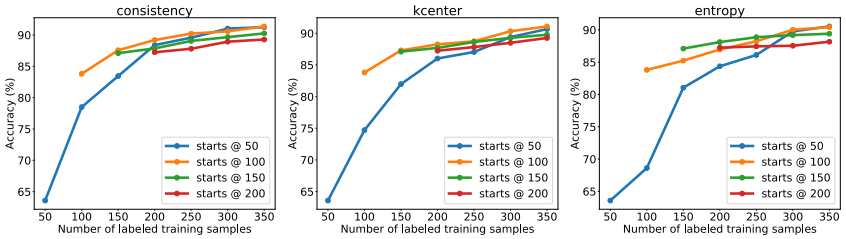

Fig. 7. Comparison of different sampling methods trained with SSL framework on CIFAR-10 when AL starts from different number of labeled samples

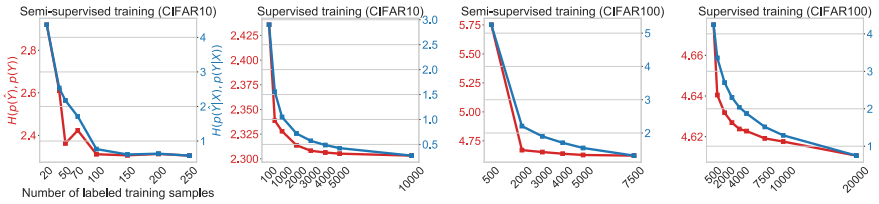

Fig. 8. Empirical risk (i.e. the target loss) on the entire training data (in blue) and cross-entropy between $p(\hat{Y})$ and $p(Y)$ show strong correlations in both semi-supervised and supervised settings

or 200, since we have more AL cycles in the former case given the same label budget. The semi-supervised nature of our learning proposal encourages the practice of initiating learning-based sample selection from a much smaller start size. However, the initial model can still have poorly-learned boundary when started with extremely small labeled data. If there is a sufficiently large validation dataset, this problem can be relieved by tracking validation performance. However, in practice, such a validation set typically doesn't exist. These motivate us to conduct an exploration to systematically infer a proper starting size.

5.2 An exploratory analysis in start size selection

Recall from the last step of Algorithm 1, if T is set such that $U_T = \emptyset$, i.e., if the entire dataset is labeled, then the final model M_T is trained to minimize the purely supervised loss $\mathcal{L}_l$ on the total labeled dataset L_T. Consider the cross-entropy loss for any classifier $p(\hat{Y}|X)$, which we call the *AL target loss*:

$$\mathcal{L}_l\left[L_T, p(\hat{Y}|X)\right] = -\frac{1}{|L_T|}\sum_{(x,y)\in L_T} \log p(\hat{Y} = y|X = x). \tag{3}$$

Note that the goal of an AL method can be viewed as minimizing the AL target loss of the entire training set L_T [55] with the small subset of labeled data

available. In any intermediate AL step, we expect our model can minimize the target loss. If the model do a poor job in approximating and minimizing Eq. 3 (cold start problem occurs), the quality of the samples selected in the subsequent AL cycles could be consequently poor. Therefore, it is crucial to assess the performance of the currently-learned model in minimizing the criterion in Eq. 3. However, since the labeled data set L_t at cycle t is a strict subset of the entire training set L_T, it is impossible to simply plug the most recently-learned model $\hat{Y}$ in Eq. 3 for direct calculation.

To this end, we approximate the target loss based on the following proposition (see proof in the supplementary material), which gives upper and lower bounds on the expected loss:

Proposition 1. *For any given distribution of Y, and any learned model $\hat{Y}$, we have*

$$\begin{aligned} H\left[p(Y), p(\hat{Y})\right] - H[p(X)] \leq R_H\left[p(\hat{Y}|X)\right] &= \mathrm{E}_X\left\{H\left[p(Y|X), p(\hat{Y}|X)\right]\right\} \\ &\leq H\left[p(Y), p(\hat{Y})\right] - H[p(X)] - \log \hat{Z}, \end{aligned} \tag{4}$$

where $H[p, q]$ is the cross-entropy between two distributions p and q, $H[p(X)]$ is the entropy of the random variable X, and $\hat{Z} = \min_{x,y} p(X = x|\hat{Y} = y)$.

Proposition 1 indicates that the AL target loss, *i.e.*, $R_H\left[p(\hat{Y}|X)\right]$, can be both upper and lower bounded. In particular, both bounds involve the quantity $H[p(Y), p(\hat{Y})]$, which suggests that $H[p(Y), p(\hat{Y})]$ could potentially be tracked to analyze $R_H[p(\hat{Y}|X)]$ when different numbers of samples are labeled. Unlike the unavailable target loss on the entire training set, $H[p(Y), p(\hat{Y})]$ does not need all data to be labeled. In fact, to compute $H[p(Y), p(\hat{Y})]$, we just need to specify a distribution for Y, which could be assumed from prior knowledge or estimated using all of the labels in the starting cycle.

As shown in Fig. 8, we observe a strong correlation between the target loss and $H[p(Y), p(\hat{Y})]$, where Y is assumed to be uniformly distributed. In practice, a practitioner can trace the difference of $H[p(Y), p(\hat{Y})]$ between two consecutive points and empirically stop expanding the start set when the difference is within a pre-defined threshold. Particularly, in SSL setting, 100 or 150 labeled samples may be used as start set on CIFAR-10, as the value of $H[p(Y), p(\hat{Y})]$ essentially ceases decreasing, which coincides with the oracle stopping points if we were given access to the target loss. In contrast, a start size of 50 may not be favorable since the difference of $H[p(Y), p(\hat{Y})]$ between the points of 50 and 20 are relatively large. A similar pattern in the supervised learning setting is also shown in Fig. 8.

6 Weaknesses of our method

We explore how well our AL selection method would perform with supervised learning using only labeled samples. Following [53], we start with 1000 labeled

Table 3. Comparison of different sampling methods in the supervised setting on CIFAR-10. All methods start from 1000 labeled samples. The reported results are over 5 trials

Methods	# of labeled samples in total				
	1000	1500	2000	2500	3000
Uniform	72.93	75.38±0.17	77.46±0.3	78.79±0.38	80.81±0.28
Entropy		76.31±0.18	79.50±0.29	81.30±0.31	82.67±0.55
k-center		74.25±0.29	77.56±0.30	79.50±0.20	81.70±0.32
Ours		76.63±0.17	79.39±0.31	80.99±0.39	82.75±0.26

samples on CIFAR-10. As shown in Table 3, after 4 AL cycles ($B = 500$, totaling 3000 labels), *uniform*, *k-center*, *entropy* and our method (*consistency*) achieve accuracy of 80.81%, 81.70%, 82.67% and 82.75%, respectively. It shows that *consistency* sampling performs comparable with the baseline metrics without significant improvement. This discourages the direct application of our selection metric in the supervised setting. Mixmatch is mainly used as the target model in this work and we experiment with two more SSL methods (see results in the supplementary material). However, comprehensive analyses with extensive SSL methods are desirable to further understand the advantages/disadvantages of our approach. As an exploratory analysis, we propose a measure that is shown to be strongly correlated with the AL target loss, but exact determination of the optimal start size is yet to be addressed.

7 Conclusion

We present a consistency-based semi-supervised AL framework and a simple pool-based AL selection metric to select data for labeling by leveraging unsupervised information of unlabeled data during training. Our experiments demonstrate that our semi-supervised AL method outperforms the state-of-the art AL methods and also alternative SSL and AL combinations. Through quantitative and qualitative analyses, we show that our proposed metric implicitly balances uncertainty and diversity when making selection. In addition, we study and address the practically-valuable and fundamentally-challenging question – "When can we start learning-based AL selection?". We present a measure to assist determining proper start size. Our experimental analysis demonstrates that the proposed measure correlates well with the AL target loss (i.e. the ultimate supervised loss on all labeled data), thus is potentially useful to evaluate target models without extra labeling effort. Overall, semi-supervised AL opens new horizons for training with very limited labeling budget, and we highly encourage future research along this direction to further analyze SSL and cold-start impacts on AL.

Acknowledgment. Discussions with Giulia DeSalvo, Chih-kuan Yeh, Kihyuk Sohn, Chen Xing, and Wei Wei are gratefully acknowledged.

References

1. Athiwaratkun, B., Finzi, M., Izmailov, P., Wilson, A.G.: There are many consistent explanations of unlabeled data: Why you should average. ICLR (2019)
2. Balcan, M.F., Beygelzimer, A., Langford, J.: Agnostic active learning. Journal of Computer and System Sciences **75**(1), 78–89 (2009)
3. Balcan, M.F., Broder, A., Zhang, T.: Margin based active learning. In: International Conference on Computational Learning Theory (2007)
4. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification (2018)
5. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)
6. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: ICML (2003)
7. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine learning **15**(2), 201–221 (1994)
8. Cortes, C., DeSalvo, G., Mohri, M., Zhang, N.: Agnostic active learning without constraints. In: ICML (2019)
9. Cortes, C., DeSalvo, G., Mohri, M., Zhang, N., Gentile, C.: Active learning with disagreement graphs. In: ICML (2019)
10. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. arXiv preprint arXiv:1909.13719 (2019)
11. Dasgupta, S., Hsu, D.: Hierarchical sampling for active learning. In: ICML (2008)
12. Dasgupta, S., Hsu, D.J., Monteleoni, C.: A general agnostic active learning algorithm. In: NIPS (2008)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
14. Drugman, T., Pylkkonen, J., Kneser, R.: Active and semi-supervised learning in asr: Benefits on the acoustic and language models. arXiv preprint arXiv:1903.02852 (2019)
15. Elhamifar, E., Sapiro, G., Yang, A., Shankar Sasrty, S.: A convex optimization framework for active learning. In: CVPR (2013)
16. Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: Active learning with expected model output changes. In: ECCV (2014)
17. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML (2016)
18. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: ICML (2017)
19. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`
20. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
21. Guo, Y.: Active instance sampling via matrix partition. In: NIPS (2010)
22. Hasan, M., Roy-Chowdhury, A.K.: Context aware active learning of activity recognition models. In: CVPR (2015)
23. Houlsby, N., Hernández-Lobato, J.M., Ghahramani, Z.: Cold-start active learning with robust ordinal matrix factorization. In: ICML (2014)

24. Iglesias, J.E., Konukoglu, E., Montillo, A., Tu, Z., Criminisi, A.: Combining generative and discriminative models for semantic segmentation of ct scans via active learning. In: Proc Conference on Information Processing in Medical Imaging (2011)
25. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: CVPR (2009)
26. Konyushkova, K., Sznitman, R., Fua, P.: Learning active learning from data. In: NIPS (2017)
27. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
28. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. ICLR (2017)
29. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NIPS (2017)
30. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine learning proceedings 1994, pp. 148–156. Elsevier (1994)
31. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR'94. pp. 3–12 (1994)
32. Mac Aodha, O., Campbell, N.D., Kautz, J., Brostow, G.J.: Hierarchical subquery evaluation for active learning on a graph. In: CVPR (2014)
33. McCallumzy, A.K., Nigamy, K.: Employing em and pool-based active learning for text classification. In: ICML (1998)
34. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: ICML (2004)
35. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. In: NeurIPS (2018)
36. Rhee, P.K., Erdenee, E., Kyun, S.D., Ahmed, M.U., Jin, S.: Active and semi-supervised learning for object detection with imperfect data. Cognitive Systems Research **45**, 109–123 (2017)
37. Roth, D., Small, K.: Margin-based active learning for structured output spaces. In: ECML (2006)
38. Roy, N., McCallum, A.: Toward optimal active learning through monte carlo estimation of error reduction. ICML (2001)
39. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. ICLR (2018)
40. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: NIPS (2008)
41. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proc Workshop on Computational learning theory (1992)
42. Siméoni, O., Budnik, M., Avrithis, Y., Gravier, G.: Rethinking deep active learning: Using unlabeled data at model training. arXiv preprint arXiv:1911.08177 (2019)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
44. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. arXiv preprint arXiv:1904.00370 (2019)
45. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020)
46. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
47. Song, S., Berthelot, D., Rostamizadeh, A.: Combining mixmatch and active learning for better accuracy with fewer labels. arXiv preprint arXiv:1912.00594 (2019)
48. Tomanek, K., Hahn, U.: Semi-supervised active learning for sequence labeling. In: ACL (2009)

49. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. JMLR **2**(Nov), 45–66 (2001)
50. Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. International Joint Conferences on Artifical Intelligence (2019)
51. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019)
52. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. IJCV **113**(2), 113–127 (2015)
53. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: CVPR (2019)
54. Zhang, Z., Zhang, H., Arik, S.O., Lee, H., Pfister, T.: Distilling effective supervision from severe label noise. In: CVPR (2020)
55. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: ICML Workshops (2003)