CONSISTENCY-BASED SEMI-SUPERVISED ACTIVE LEARNING: TOWARDS MINIMIZING LABELING BUDGET

Anonymous authors
Paper under double-blind review

ABSTRACT

Active learning (AL) aims to integrate data labeling and model training in a unified way, and to minimize the labeling budget by prioritizing the selection of high value data that can best improve model performance. Readily-available unlabeled data are used to evaluate selection mechanisms, but are not used for model training in conventional pool-based AL. To minimize the labeling budget, we unify unlabeled sample selection and model training based on two principles. First, we exploit both labeled and unlabeled data using semi-supervised learning (SSL) to distill information from unlabeled data that improves representation learning and sample selection. Second, we propose a simple yet effective selection metric that is coherent with the training objective such that the selected samples are effective at improving model performance. Our experimental results demonstrate superior performance with our proposed principles for limited labeled data compared to alternative AL and SSL combinations. In addition, we study the AL phenomena of 'cold start', which is becoming an increasingly more important factor to enable optimal unification of data labeling, model training and labeling budget minimization. We propose a measure that is found to be empirically correlated with the AL target loss. This measure can be used to assist in determining the proper start size.

1 INTRODUCTION

Deep learning models are improved when trained with more labeled data (Goodfellow et al., 2016). A standard deep learning development 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. The costs are exacerbated especially when expert annotators are required, as in medical diagnosis and loan prediction, or some labels correspond to rare cases, as in self-driving cars. An ideal framework would integrate data labeling and model training in a meaningful way to push model performance with minimal amount of labeled data.

Active learning (AL) (Balcan et al., 2009) aims to optimize the learning procedure by judicious selection of the samples for which to obtain labels, with the goal of maximizing the ultimate model performance with minimal labeling budget. As studied in Ghorbani & Zou (2019); Toneva et al. (2019), data samples contribute to learning differently, thus prioritizing data for labeling is expected to boost performance. In this paper, we focus on pool-based AL where an unlabeled data pool is given initially and the AL mechanism iteratively selects batches to get labels for in conjunction with training. As the name 'learning-based AL selection' suggests, each batch is selected with the guidance from the previously-trained model, is labeled, and then integrated into the existing labeled dataset on which the model is trained in the subsequent step.

Maximization of performance with minimal labeled data requires properly leveraging model learning and AL sample selection, especially in the early phases of AL cycles. One of the natural observations for pool-based AL is the opportunity to utilize the unlabeled pool in a semi-supervised learning (SSL) way for model training (Zhu et al., 2003; Tomanek & Hahn, 2009). While using the unlabeled pool as the candidates in the AL selection phase, SSL objectives can be integrated to leverage performance by learning meaningful data representations from the unlabeled pool. We empirically demonstrate that the AL selection criterion should be redesigned to be in coherence with

the SSL objectives to select the most valuable samples. First, unsupervised losses could alter the learned representation and decision manifolds significantly (Oliver et al., 2018) and the AL sample selection should reflect that. Second, SSL already results in the embodiment of knowledge from unlabeled data in a meaningful way; thus AL selection should reflect the extra value of the labeled data on top of it.

In the absence of any labeled data, a common practice to initiate AL is to uniformly select a small starting subset of data for labeling. Learning-based AL selection is then used in subsequent cycles. The size of the starting subset affects the AL performance – when the start size is not sufficiently large, the models learned in subsequent AL cycles are highly-skewed and result in biased selection, a phenomenon commonly known as the *cold start problem* (Konyushkova et al., 2017; Houlsby et al., 2014). When cold start occurs, learning-based selection yields samples that lead to lower performance improvement than using the naive uniform sampling in subsequent AL cycles (Konyushkova et al., 2017). Increasing the start size alleviates the cold start problem, but it is suboptimal to increase the start size arbitrarily as that consumes a larger portion of labeling budget before learning-based AL selection is utilized. Our proposed semi-supervised AL method aims to minimize the labeling budget and encourages the practice of initiating learning-based sample selection from a much smaller start size. It is important to choose a start size that is large enough in avoiding cold start, yet sufficiently small to minimize the labeling budget.

Contributions: The goal in this paper is to push the limits of performance with minimal labeled data in pool-based AL scenario by exploiting an unlabeled data pool for model training. Our method is based on the insight that have driven the recent advances in SSL (Berthelot et al., 2019; Verma et al., 2019; Xie et al., 2019): a model should be consistent in its decisions between a sample and its meaningfully-distorted versions. Such insights motivate us to introduce a hypothesis for AL: a sample that yields low consistency in predictions might indicate that the model is incapable of distilling useful information from the sample and thus human labeling is necessary to provide more meaningful supervision. Correspondingly, we propose a simple yet effective selection metric that is in coherent with the consistency-related information in SSL. Experiments demonstrate that our proposed metric can consistently outperform previous methods when they are considered with SSL. With various quantitative and qualitative analyses, we demonstrate the rationale behind why such a selection criteria are highly effective in AL. In addition, in an exploratory analysis we propose a measure that can be used to assist in determining the proper start size in order to mitigate the chance that cold start occurs.

2 Related Work

Extensive research has been done in AL (Dasgupta et al., 2008; Dasgupta & Hsu, 2008; Balcan et al., 2009; Cortes et al., 2019a). Traditional AL methods can be roughly classified into three categories: uncertainty-based methods, diversity-based methods and expected model change-based methods. Among uncertainty-based ones, methods based on *max entropy* (Lewis & Catlett, 1994; Lewis & Gale, 1994) and *max margin* (Roth & Small, 2006; Balcan et al., 2007; Joshi et al., 2009) are popular for their simplicity. Some other uncertainty-based methods measure distances between samples and the decision boundary (Tong & Koller, 2001; Brinker, 2003). Most uncertainty-based methods use heuristics, while the recent work (Yoo & Kweon, 2019a) directly learn the target loss of inputs jointly with the training phase and shows promising results. Diversity-based methods aim to select diverse samples that span the input space the most (Nguyen & Smeulders, 2004; Mac Aodha et al., 2014; Hasan & Roy-Chowdhury, 2015; Sener & Savarese, 2018). There are also methods that consider both uncertainty and diversity (Guo, 2010; Elhamifar et al., 2013; Yang et al., 2015). The third category of approaches estimate the future model status and select samples that encourage optimal model improvement (Roy & McCallum, 2001; Settles et al., 2008; Freytag et al., 2014).

Both AL and SSL aim to improve learning with limited labeled data, thus they are naturally related. Only a few works have considered combining AL and SSL in different tasks. In Drugman et al. (2019), joint application of SSL and AL is considered for speech understanding, and significant error reduction is demonstrated with limited labeled speech data. For AL, their selection criteria is based on a confidence score that quantifies the observed probabilities of words being correct. Rhee et al. (2017) propose an active semi-supervised learning system which demonstrates superior performance in the pedestrian detection task. Zhu et al. (2003) combine AL and SSL using Gaussian fields and validate their method on synthetic datasets. Sener & Savarese (2018) also consider SSL

Algorithm 1 A semi-supervised learning based AL framework

Require: Unlabeled data pool $\mathcal{D}$, the total number of steps T, 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

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 suboptimal SSL method.

Agreement-based methods, also referred as "query-by-committee", base the selection decisions on the opinions of a committee which consist of independent AL metrics or models (Seung et al., 1992; Cohn et al., 1994; McCallumzy & Nigamy, 1998; Iglesias et al., 2011; Beluch et al., 2018; Cortes et al., 2019b). 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 data that mostly disagrees with the outputs of its augmentations.

3 Consistency-based Semi-supervised AL

3.1 Proposed Method

We consider the setting of pool-based AL, where an unlabeled data pool is available for selection of samples to get labels for. To minimize the labeling budget to achieve a certain performance, we propose a novel method that unifies selection and model updates, overviewed in Algorithm 1. The proposed method has two key novel aspects.

Conventional AL methods base model learning only on the available labeled data, which is an opportunity cost given the useful information in the unlabeled data. Our first novelty is incorporating the semi-supervised learning (SSL) framework in each training phase of AL. Specifically, as shown in Algorithm 1, each model M_t is learned by minimizing an objective loss function of the form $\mathcal{L}_l + \mathcal{L}_u$. A model learned in this manner should fit the labeled data better, as reflected in the supervised part of the loss function $\mathcal{L}_l$, and at the same time should obtain a good grasp of the data representations from unlabeled data in meaningful ways as reflected in the unsupervised part of the loss function $\mathcal{L}_u$. Note, that our focus is not optimizing the objectives $\mathcal{L}_l$ and $\mathcal{L}_u$ individually to improve the overall SSL performance, but rather illustrating how they can be integrated in an unified framework.

The design of the selection criterion plays a crucial role in integration of SSL and AL. To this end, our second novelty is that we propose a selection criterion $\mathcal{C}$ to better integrate AL selection mechanism in SSL training framework. It has been observed that predictions of deep neural networks are sensitive to small perturbations on the input data (Zheng et al., 2016; Azulay & Weiss, 2018). Recent successes in SSL (Athiwaratkun et al., 2019; Berthelot et al., 2019; Verma et al., 2019) 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 semi-supervised training, a common choice of loss is $\mathcal{L}_u(x, M) = D(P(\hat{Y} = \ell|x, M), P(\hat{Y} = \ell|\tilde{x}, M))$, where D is a distance function such as KL divergence (Xie et al., 2019), or L2 norm (Laine & Aila, 2017; Berthelot et al., 2019) and $\tilde{x}$ denotes a perturbation (augmentation) of the input x.

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	Our method	83.81	**87.57±0.31**	**89.20±0.51**	**90.23±0.49**

Table 1: Comparison with AL methods trained in supervised and semi-supervised setting on CIFAR-10. All methods start from 100 labeled samples (the third column). The following columns are results of different methods with the same selection batch size.

Our proposal is motivated by the following intuition. First, the unsupervised objective exploits unlabeled data by encouraging consistent predictions across slightly distorted version of each unlabeled sample. Labeling samples with low consistent predictions is valuable, since these samples are hard to be minimized using $\mathcal{L}_u$. Thus, they need human annotations to provide further useful supervision for model training. Second, the data that yields large model performance gain is not necessarily the data with most uncertainty score, since neural network prefers learning with a particular curriculum (Bengio et al., 2009). The most uncertain data could be too hard to learn, and including them in training would be misleading. Thus, we argue that the samples that can be recognized to some extent but not consistently, should benefit more for learning compared to the most uncertain ones.

Based on these insights, we propose a simple metric $\mathcal{C}$ that aims to measure the inconsistency across a certain number of meaningful perturbations. There are various ways to quantify consistency. Due to its empirically-observed superior performance, we choose the metric $\mathcal{C}(B, M) = \sum_{x \in B} \mathcal{E}(x, M)$, where

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

J is the number of response classes and we consider N perturbed samples of the original input data x, $\{\tilde{x}_1, ..., \tilde{x}_N\}$, which can be obtained by standard augmentation operations (e.g. random crops and horizontal flips for image data). Our method prefers data samples with high $\mathcal{C}$ values, which may possess varying level of difficulty for the model to classify.

3.2 Comparisons with Baselines

The practical performance of our method is demonstrated on two commonly used datasets: CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) on the image classification task. Both datasets have 60K images in total, of which 10K images are split out for testing. CIFAR-10 consists of 10 classes and CIFAR-100 has 100 fine-grained classes. Different variants of SSL methods encourage consistency loss in different ways. In our implementation, we adopt the recently-proposed state-of-the-art method, Mixmatch (Berthelot et al., 2019), which proposes a specific loss term to encourage consistency. Following (Berthelot et al., 2019), we use Wide ResNet-28 (Oliver et al., 2018) with 32 filters as the base model and keep the default hyper-parameters for different settings in (Berthelot et al., 2019). In each cycle, M_t is initialized with M_{t-1}. We select $K = 0.5 \cdot |L_0|$ samples for labeling by default. 50 augmentations of each sample are used in our experiments, but we observe that 5 augmentations can produce comparable results. For a fair comparison, different selection methods start from the same initial model (M_0) and the reported results are over 5 trials.

We consider three representative selection methods for comparison:

- *Uniform* indicates random selection. This method shows the performance when no AL is involved. It should be compared to verify the effectiveness of AL.
- *Entropy* is widely considered as an uncertainty-based baseline in previous methods (Sener & Savarese, 2018; Yoo & Kweon, 2019a). It selects uncertain samples that has maximum entropy of its predicted class probabilities.
- *k-center* (Sener & Savarese, 2018) is a state-of-the-art diversity based AL method. It selects representative samples by maximizing the distance between a selected sample and its nearest neighbor in the labeled pool. The feature from the last fully connected layer of the target model is used to calculate distance between samples.

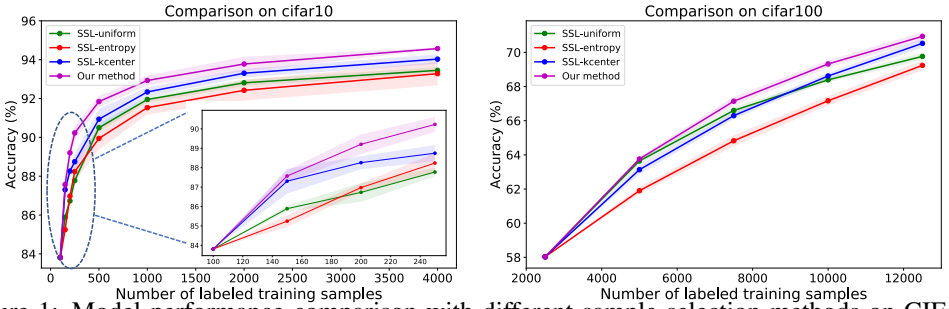

Figure 1: Model performance comparison with different sample selection methods on CIFAR-10 and CIFAR-100. Solid lines are average results over 5 trials. Shadows represent standard deviation.

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
Our method	**90.23±0.39**	**91.84±0.29**	**92.93±0.26**	**93.78±0.38**	**94.57±0.06**

Table 2: Comparison of different sampling methods in the SSL setting on CIFAR-10. Note that all the methods are under the SSL setting and start from 100 labeled samples. When the number of labeled samples reaches 250, AL batch size K is set to be $|L_t|$.

As shown in Tab. 1, our method significantly outperforms the comparison methods which only learn from labeled data at each cycle, as expected. When only 100 (0.2%) samples are labeled, our method outperforms *kcenter* by 39.24% accuracy. Next, we focus on comparing different methods in our used SSL framework. Figure 1 shows the effectiveness of our consistency-based selection in SSL setting by comparing with the baselines, when they are integrated into SSL. 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*, the state-of-the-art method, 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. When having access to all the labels (50K) for the whole training set, a fully-supervised model achieves an accuracy of 95.83% (Berthelot et al., 2019). Our method has comparable performance (1.26% lower accuracy) with *only* 4*K labeled samples*. CIFAR-100 is a more challenging dataset as it has 10$\times$ more categories to classify into. On CIFAR-100, we observe a consistent outperformance of our method at all AL cycles.

3.3 Analyses of consistency-based selection

To explain its superior, we analyze the samples selected by our method from several perspectives, which are known to be important for AL.

Uncertainty and overconfident mis-classification: Uncertainty-based AL methods query the data samples that present high entropy for the predicted probability, which are 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 (Guo et al., 2017; Lakshminarayanan et al., 2017). Such overconfident mis-classification cases would not be distinguished by the *entropy*-based AL metrics and result in suboptimal selection. Figure 2 (left) demonstrates that our *consistency*-based selection is indeed far superior in detecting high-confident mis-classification cases than *entropy*. We use entropy to measure the uncertainty of the selected samples by different methods in Figure 2 (middle). It compares different approaches and show that *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 should contribute to the performance gap with *entropy*. Figure 2 (right) exemplifies some selected samples that mis-classified with high confidence.

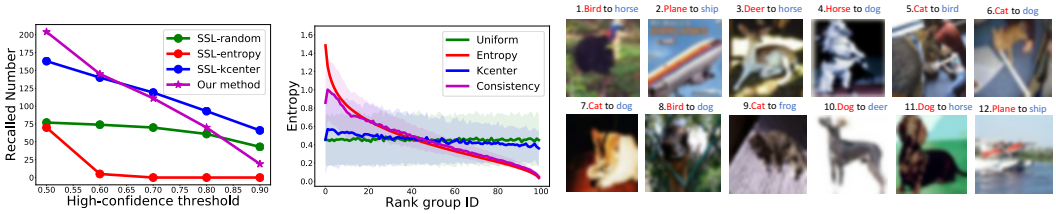

Figure 2: Left: Recalled number of overconfident mis-classified samples in top 1% samples ranked by different methods. Overconfident samples are defined as those has the highest class probability larger than a high-confidence threshold. Middle: the average entropy of unlabeled samples ranked by different selection metrics. The ranked samples are divided into 100 groups for computing average entropy. Right: Examples of overconfident mis-classified samples selected by *consistency*, which yield low entropy yet high inconsistency. Bird to horse indicates bird is mis-classified as horse.

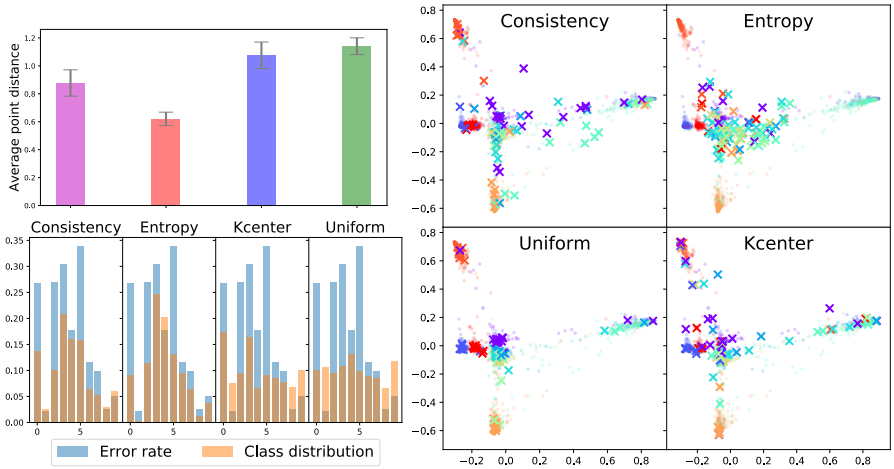

Figure 3: Average distance between samples (top-left): the average pair-wise L_2 distance of top 1% unlabeled samples ranked by different selection metrics. Per-class error rate vs. sample class distribution (bottom-left) are shown. Diversity visualization (right): Dots and crosses mean unlabeled (un-selected) samples and the selected samples (top 100), respectively. Each color represent a ground truth class.

Diversity: Diversity is proposed to be a key factor for AL besides uncertainty (Yang et al., 2015). *k-center* is a state-of-the-art AL method based on diversity (it prefers to select data points that span over the whole input space). Towards this end, Figure 3 visualizes the sample diversity selected by different methods. We use principal components analysis to reduce dimension of embedded samples to a two-dimensional space. *uniform* chooses samples equally-likely from the unlabeled pool. Samples selected by *entropy* are clustered at certain regions. On the other hand, *consistency* selects data samples as diverse as those selected by *k-center*. Average distances between top 1% points selected by different methods are shown in Figure 3 (top-left). We can see that *entropy* chooses samples have small average distances, while our *consistency* has much larger distance compared.

Class distribution complies with classification error: Figure 3 (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 *k-center*.

As above, we show that the consistency measure in SSL results in a well-balanced selection in consideration of these three aspects.

4 When can we start learning-based AL selection?

4.1 Cold-start failure

When the size of the labeled data is too small, the learned decision boundaries could be skewed and AL selection based on the model outputs could be biased. To better illustrate the problem,

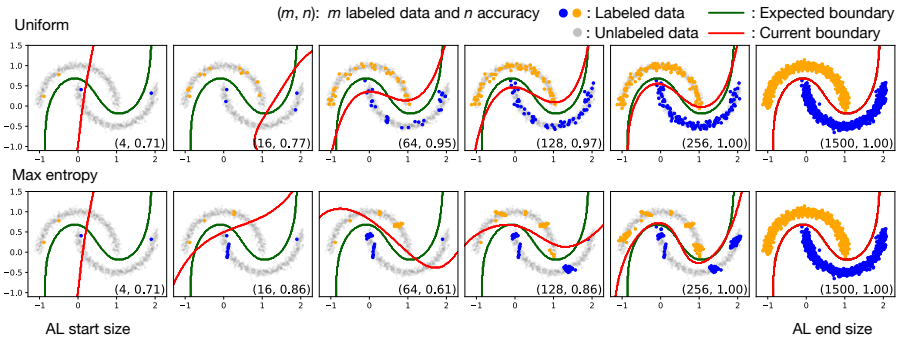

Figure 4: Illustration of the cold-start issue for uncertainty-based AL. The start size and end size has the same labeled data. When the learned decision boundary is far away from the expected boundary (e.g. the second and third columns), the selected samples by uncertainty-based AL is biased.

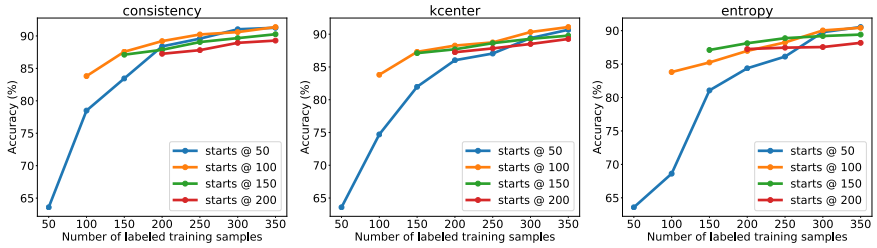

Figure 5: Performance comparison of different sampling methods trained with SSL framework on CIFAR-10 when AL starts from different number of labeled samples.

Figure 4 shows the toy two-moons dataset using a simple support vector machine (in supervised setting with the RBF kernel) to learn the decision boundary (Oliver et al., 2018). 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 the *max entropy* concentrate around the vicinity of a poorly learned boundary. In another example, we study the effects of cold start using deep neural networks on CIFAR-10, shown in Figure 5. Using uniform sampling to select different starting sizes, AL methods achieve different predictive accuracy. 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. Naively increasing the start size for uniform sampling is not practically desirable, because it would not utilize learning-based selection optimally. Given a limited budget, such sub-optimal utilization might result in worse performance. For example, our method starting from $K_0 = 100$ labeled samples has better performance than starting from 150 or 200, since we can 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 learned model can still be skewed at extreme early AL stages. These observations motivate us to propose a systematic way of inferring a proper starting size. We analyze this problem and propose an exploratory approach to assist in determining the start size in practice.

4.2 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 data set have labels, 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 function for any classifier $p(\hat{Y}|X)$, which we call the *AL target loss*:

$$\mathcal{L}_l \left[L_T, p(\hat{Y}|X) \right] = -\frac{1}{|L_T|} \sum_{(x,y)\in L_T} \log p(\hat{Y} = y|X = x). \tag{2}$$

Note that the ultimate goal of an AL method can be viewed as minimizing the AL target loss with a small subset of the entire training set L_T (Zhu et al., 2003). In any intermediate AL step, we expect the loss on the current labeled subset to mimic the target loss. If cold start occurs, the model

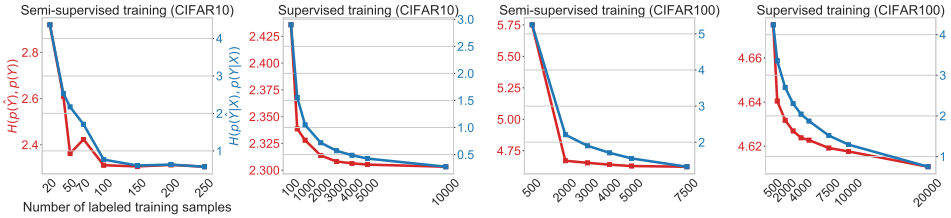

Figure 6: Empirical risk (i.e. the target loss) on the whole training data (in blue) and cross-entropy between $p(\hat{Y})$ and $p(Y)$ show strong correlations in both semi-supervised and supervised settings.

does a poor job in approximating and minimizing equation 2. The quality of the samples selected in the subsequent AL cycles would be consequently poor. Therefore, it is crucial to understand the performance of the currently-learned model in minimizing the criterion in equation 2. However, since the labeled data set L_t at cycle t is a strict subset of the total training set L_T, it is impossible to simply plug the most recently learned model $\hat{Y}$ in equation 2 for direct calculation.

The basic premise of our approximation to the target loss stems from the following proposition, which gives upper and lower bounds on the expected loss, to which the target loss approximates:

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

$$H\left[p(Y), p(\hat{Y})\right] - H[p(X)] \leq R_H\left[p(\hat{Y}|X)\right] = \mathrm{E}_X\left\{H\left[p(Y|X), p(\hat{Y}|X)\right]\right\}$$
$$\leq H\left[p(Y), p(\hat{Y})\right] - H[p(X)] - \log \hat{Z}, \tag{3}$$

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 expected cross-entropy loss 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 $R_H[p(\hat{Y}|X)]$ for different number of samples. Unlike the intractable target loss on the whole training set, $H[p(Y), p(\hat{Y})]$ does not need all data to be labeled. Actually, to compute $H[p(Y), p(\hat{Y})]$, we just need to specify a distribution of Y, which could be assumed from prior knowledge or estimated using all of the labels in the starting cycle.

In Figure 6, we observe a strong correlation between the target loss and $H[p(Y), p(\hat{Y})]$, where Y is assumed to follow a uniform distribution. We see how the tractable quantity $H[p(Y), p(\hat{Y})]$ can be used to show the trend when the actual target is minimized. Particularly, in SSL setting, a practitioner may set the starting set size of 100 or 150 labeled samples on CIFAR-10, as the value of $H[p(Y), p(\hat{Y})]$ essentially ceases to decrease, which coincide with the oracle stopping points if we were given the access to the target loss. In contrast, a start size of 50 has much higher $H[p(Y), p(\hat{Y})]$, which leads to less favorable performance. A similar pattern in the supervised learning setting is shown in Figure 6. Exploratory analysis of $H[p(Y), p(\hat{Y})]$ could thus provide a qualitative evaluation of any intermidiate AL model in minimizing the loss objective on the entire yet unavailable training set.

5 Conclusion and Future Work

In this paper, we propose a simple pool-based AL selection metric to select data for labeling by leveraging unsupervised information of unlabeled data during training. Our experiments shows that our method outperforms previous state-of-the-art AL methods under the SSL setting. Our proposed metric implicitly balance uncertainty and diversity when making selection. In addition, we study and address a very practically valuable yet challenging question — "When can we start learning-based AL selection?". We present a measure to assist in determining proper start size. The design of our method focuses on the principles of consistency in SSL. For alternative SSL methods based on other principles, it is necessary to revisit AL selection with respect to their training objectives, which will be considered in future work. Experimental analysis demonstrates that the proposed measure correlates well with the AL target loss (i.e. the ultimate the supervised loss on all labeled data). In practice, it can be tracked to examine the model without requesting a large validation set.

REFERENCES

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. *International Conference on Learning Representations*, 2019.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pp. 35–50. Springer, 2007.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

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*, pp. 9368–9377, 2018.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48. ACM, 2009.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 59–66, 2003.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

Corinna Cortes, Giulia DeSalvo, Mehryar Mohri, and Ningshan Zhang. Agnostic active learning without constraints. In *International Conference on Machine Learning*, 2019a.

Corinna Cortes, Giulia DeSalvo, Mehryar Mohri, Ningshan Zhang, and Claudio Gentile. Active learning with disagreement graphs. In *International Conference on Machine Learning*, pp. 1379–1387, 2019b.

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 208–215. ACM, 2008.

Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *Advances in neural information processing systems*, pp. 353–360, 2008.

Thomas Drugman, Janne Pylkkönen, and Reinhard Kneser. Active and semi-supervised learning in ASR: benefits on the acoustic and language models. *arXiv:1903.02852*, 2019.

Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and S Shankar Sasrty. A convex optimization framework for active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 209–216, 2013.

Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision*, pp. 562–577. Springer, 2014.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. *International Conference in Machine Learning*, 2019.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

Yuhong Guo. Active instance sampling via matrix partition. In *Advances in Neural Information Processing Systems*, pp. 802–810, 2010.

Mahmudul Hasan and Amit K Roy-Chowdhury. Context aware active learning of activity recognition models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4543–4551, 2015.

Neil Houlsby, José Miguel Hernández-Lobato, and Zoubin Ghahramani. Cold-start active learning with robust ordinal matrix factorization. In *International Conference on Machine Learning*, pp. 766–774, 2014.

Juan Eugenio Iglesias, Ender Konukoglu, Albert Montillo, Zhuowen Tu, and Antonio Criminisi. Combining generative and discriminative models for semantic segmentation of ct scans via active learning. In *Biennial International Conference on Information Processing in Medical Imaging*, pp. 25–36. Springer, 2011.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2372–2379. IEEE, 2009.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *Advances in Neural Information Processing Systems*, pp. 4225–4235, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *International Conference on Learning Representations*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *Neural Information Processing Systems*, 2018.

David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pp. 148–156. Elsevier, 1994.

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR94*, pp. 3–12. Springer, 1994.

Oisin Mac Aodha, Neill DF Campbell, Jan Kautz, and Gabriel J Brostow. Hierarchical subquery evaluation for active learning on a graph. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 564–571, 2014.

Andrew Kachites McCallumzy and Kamal Nigamy. Employing em and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning (ICML)*, pp. 359–367. Citeseer, 1998.

Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 79. ACM, 2004.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 3235–3246, 2018.

Phill Kyu Rhee, Enkhbayar Erdenee, Shin Dong Kyun, Minhaz Uddin Ahmed, and Songguo Jin. Active and semi-supervised learning for object detection with imperfect data. *Cognitive Systems Research*, 45:109–123, 2017.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pp. 413–424. Springer, 2006.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pp. 441–448, 2001.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *International Conference on Learning Representations*, 2018.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pp. 1289–1296, 2008.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294. ACM, 1992.

Katrin Tomanek and Udo Hahn. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1039–1047. Association for Computational Linguistics, 2009.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *International Conference on Learning Representations*, 2019.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *International Joint Conferences on Artifical Intelligence*, 2019.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 93–102, 2019a.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *2019 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2019b.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 4480–4488, 2016.

Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, volume 3, 2003.

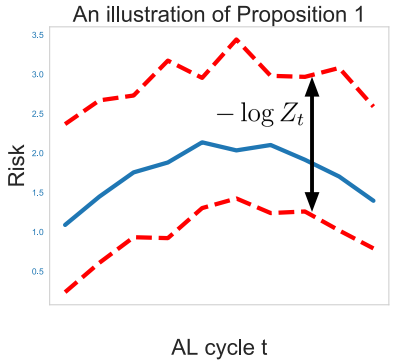

Figure A1: An illustration of Proposition 1: the blue curve represents the (expected) cross-entropy, and the two red curves are the lower and upper bounds. The value $-\log \hat{Z}_t$ characterizes the range of the bounds.

APPENDIX

A PROOF OF PROPOSITION 1

Proof. Denote $\mathcal{X}$ as the feature space and $\{1, \ldots, J\}$ as the label space. Note that by Baye's formula and the law of total probability, we have

$$
\begin{aligned}
&R_H[p(\hat{Y}|X)] = \mathrm{E}_X\left\{H\left[p(Y|X), p(\hat{Y}|X)\right]\right\} \\
=&-\sum_{x\in\mathcal{X}}\sum_{y=1}^{J} p(Y=y|X=x)\log p(\hat{Y}=y|X=x)p(X=x) \\
=&-\sum_{y=1}^{J}\sum_{x\in\mathcal{X}} p(X=x,Y=y)\log\left[\frac{p(\hat{Y}=y)p(X=x|\hat{Y}=y)}{p(X=x)}\right] \\
=&-\sum_{y=1}^{J}\sum_{x\in\mathcal{X}} p(X=x,Y=y)\log p(\hat{Y}=y) - \sum_{y=1}^{J}\sum_{x\in\mathcal{X}} p(X=x,Y=y)\log\left[\frac{p(X=x|\hat{Y}=y)}{p(X=x)}\right] \\
=&-\sum_{y=1}^{J} p(Y=y)\log p(\hat{Y}=y) - \sum_{x\in\mathcal{X}}\sum_{y=1}^{J} p(X=x,Y=y)\log\left[p(X=x|\hat{Y}=y)\right] \\
&+\sum_{x\in\mathcal{X}}\sum_{y=1}^{J} p(X=x,Y=y)\log\left[p(X=x)\right] \\
=&H\left[p(Y), p(\hat{Y})\right] + \sum_{x\in\mathcal{X}} p(X=x)\log\left[p(X=x)\right] - \sum_{x\in\mathcal{X}}\sum_{y=1}^{J} p(X=x,Y=y)\log\left[p(X=x|\hat{Y}=y)\right] \\
=&H\left[p(Y), p(\hat{Y})\right] - H\left[p(X)\right] - \sum_{x\in\mathcal{X}}\sum_{y=1}^{J} p(X=x,Y=y)\log\left[p(X=x|\hat{Y}=y)\right]. \qquad (4)
\end{aligned}
$$

We first give a lower bound. Note that $p(X=x|\hat{Y}=y) \leq 1$ for any $(x,y) \in \mathcal{X}\times[J]$, so equation 4 implies that

$$
\mathrm{E}_X\left\{H\left[p(Y|X), p(\hat{Y}|X)\right]\right\} \geq H\left[p(Y), p(\hat{Y})\right] - H\left[p(X)\right].
$$

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
Our method		76.63±0.17	79.39±0.31	80.99±0.39	82.75±0.26

Table 3: Comparison of different sampling methods in the supervised setting on CIFAR-10. All methods start from 1000 labeled samples.

To prove the upper bound, denote $\min_{(x,y)\in\mathcal{X}\times[J]} p(X=x|\hat{Y}=y)=\hat{Z}\in(0,1)$ where $(x,y)\in\mathcal{X}\times[J]$. Then from equation 4

$$\begin{aligned}\mathrm{E}_X\left\{H\left[p(Y|X),p(\hat{Y}|X)\right]\right\}\leq& H\left[p(Y),p(\hat{Y})\right]-H\left[p(X)\right]-\log\hat{Z}\sum_{x\in\mathcal{X}}\sum_{y=1}^{J}p(X=x,Y=y)\\ =& H\left[p(Y),p(\hat{Y})\right]-H\left[p(X)\right]-\log\hat{Z}.\end{aligned}$$

□

B More Discussion

B.1 Consistency-based AL in supervised learning

We also curious about how well our method performs under supervised learning using only labeled samples. Following Yoo & Kweon (2019b), we start with 1000 labeled samples on CIFAR-10. After 4 AL cycles ($B=500$, totaling 3000 labels), *uniform*, *k-center*, *entropy* and our method (*consistency*) achieves accuracy of 80.81%, 81.70%, 82.67% and 82.75%, respectively. It shows that *consistency* still works even if only trained using labeled samples. However, the improvement of *consistency* compared with other compared methods (especially *entropy*) is marginal.

B.2 Out-of-distribution and challenging samples

In real-world scenarios, it is very likely that not all unlabeled data are irrelevant to the task. Therefore, if a sample remains high uncertainty given arbitrary perturbations, it is probably a out-of-distribution example (Lee et al., 2018). In addition, selecting the hardest samples are not preferred, because it could be "over-challenging" for current model as suggested by the study of curriculum learning (Bengio et al., 2009). It can be easily inferred that our proposed selection can avoid such cases (see equation 1). More exploration of active learning with out-of-distribution samples is left for future work.