Contrastive Coding for Active Learning under Class Distribution Mismatch

Pan Du[1,2], Suyun Zhao[1,2,*], Hui Chen[2,†], Shuwen Chai[2,†], Hong Chen[1,2], Cuiping Li[1,2]
Key Lab of Data Engineering and Knowledge Engineering of MOE Renmin University of China[1]
Renmin University of China, Beijing, China[2]
{dupan, zhaosuyun, chenhui1025, chaishuwen, chong, licuiping}@ruc.edu.cn

Abstract

Active learning (AL) is successful based on the assumption that labeled and unlabeled data are obtained from the same class distribution. However, its performance deteriorates under class distribution mismatch, wherein the unlabeled data contain many samples out of the class distribution of labeled data. To effectively handle the problems under class distribution mismatch, we propose a contrastive coding based AL framework named CCAL. Unlike the existing AL methods that focus on selecting the most informative samples for annotating, CCAL extracts both semantic and distinctive features by contrastive learning and combines them in a query strategy to choose the most informative unlabeled samples with matched categories. Theoretically, we prove that the AL error of CCAL has a tight upper bound. Experimentally, we evaluate its performance on CIFAR10, CIFAR100, and an artificial cross-dataset that consists of five datasets; consequently, CCAL achieves state-of-the-art performance by a large margin with remarkably lower annotation cost. To the best of our knowledge, CCAL is the first work related to AL for class distribution mismatch.

1. Introduction

Deep Learning, which largely depends on sufficient labeled data, has achieved unprecedented breakthroughs in supervised learning [23]. Nevertheless, it is impractical to obtain abundant labeled data because labeling requires enormous human and financial costs [36].

Active learning (AL) selects the most informative samples to query their labels, delivering a competitive target model while saving annotation costs relative to the supervised learning [36]. In traditional AL methods, it's generally assumed that labeled and unlabeled data are drawn from the same class distribution, i.e., the categories of unlabeled data are the same as those of labeled ones. Unfortunately, this assumption cannot be maintained in many real-world scenarios since that unlabeled data always contain lots of samples out of the class distribution of labeled data, i.e., some categories of unlabeled data are not presented in labeled ones. For example, when crawling a large set of images (as shown in Figure 1) [49] by keyword filtering ("dog", "cat") from the Internet for binary image classification, many unlabeled images not belonging to target classes (such as "deer", "horse", "airplane", "ship", "car", "flower") are collected. The same problem has already been found in medical diagnoses containing unseen lesions [11, 48] and the house annotation of remote sensing images containing numerous natural sceneries. These scenarios have been formalized as the learning framework, called class distribution mismatch [11] [7].

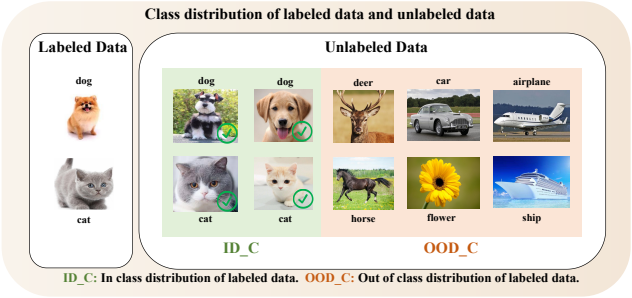


Figure 1: An instance of class distribution mismatch. Unlabeled data contains some samples that are out of the class distribution of labeled data.

Under class distribution mismatch, an AL algorithm will suffer a sharp drop in performance if only focusing on querying the "most informative" samples. One main reason for this phenomenon is that a large set of samples with mismatched categories will be queried, which are invalid for the target model, thereby wasting the annotation budget. Thus, it is essential to reduce the cost for invalid queries while improving the information of samples queried (valid queries) in class distribution mismatch. Heuristically, we introduce both invalid query error and valid query error to combat the problem, as described in Eq.1. Specifically, an invalid query error is attributed to those queried samples invalid for improving the target model, and a valid query error is due to

*Corresponding Author
†Equal Contribution

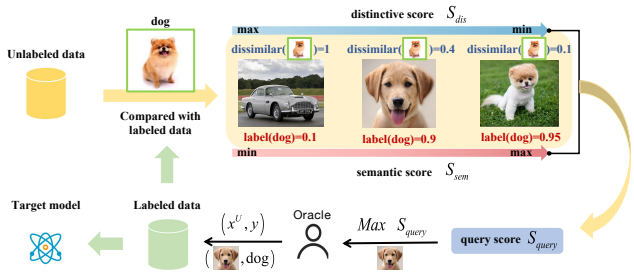


Figure 2: CCAL is combining the semantic score S_{sem} and distinctive score S_{dis} to select samples for annotating.

the less representativeness of queried data to the entire data. For example, in Figure 1, the "horse" and "flower" images may increase the invalid query error, while the "cat" image in the left bottom corner of unlabeled data may lead to the increment of valid query error.

Inspired by the theoretical analysis presented in Subsection 3.5, we propose a contrastive coding based AL framework, which extracts semantic and distinctive features by contrastive learning, named CCAL, as shown in Figure 2. For one thing, semantic features, which function as category-level features, can be exploited to filter invalid samples with mismatched categories, thereby reducing the invalid query error in CCAL. For another, distinctive features, which describe the features at the individual level, can be used to select the most representative and informative ones to extend the decision boundary, bring the lessen of valid query error. In conjunction with semantics and distinctiveness, CCAL selects the most informative samples with matched categories, excelling on tasks under class distribution mismatch. Our theoretical study shows that the AL error of CCAL has a tight upper bound. Experimental results on different datasets further validate that CCAL outperforms the compared methods.

1.1. Contributions

The major contributions of this study are:

1) Proposing an AL framework, called CCAL, which combines semantics with distinctiveness as a new AL criterion and selects the most informative unlabeled samples with matched categories to query under class distribution mismatch. This is the first work of AL related to class distribution mismatch to the best of our knowledge.

2) Learning semantic and distinctive features, which describe the samples in category-level and individual-level, respectively, and conductive to keep away from the unlabeled samples with mismatched categories and select the most informative ones with matched categories from unlabeled data.

3) Dividing the AL error into invalid query error and valid query error, and proving that the AL error of CCAL has a tight upper bound under class distribution mismatch.

The remainder of this paper is organized as follows. In Section 2, we review some related work. In Section 3, the proposed approach CCAL and theoretical studies are introduced. Section 4 presents the experiments, followed by the conclusion in Section 5.

2. Related Work

Active learning: AL reduces the labeling cost by actively selecting the most valuable data to query their labels [36]. Existing AL methods can be roughly divided into pool-based methods and generation-based methods.

Most pool-based methods evaluate a sample in terms of its informativeness, representativeness, or both [40]. The first considering informativeness, contain uncertainty [45, 41, 44], query by committee [29, 37], etc. A naive way of defining uncertainty involves using the posterior probability of a sample predicted by a model [24, 18, 33]. Entropy [28] is one of the most usual methods. Recently, Sinha et al. [38] propose to assess the uncertainty based on whether unlabeled samples share the same distribution as labeled ones. Yoo et al. [50] take the uncertainty into account by estimating the loss of the sample. The second one is on representativeness, which focuses on diversity [9] and density [31], etc. Coreset [35] is a classic diversity-based method that minimizes the Euclidean distance between the sampled points and remaining points in the feature space. It has been verified that it is better to consider both informativeness and representativeness instead of either [46, 15]. T et al. [1] measure uncertainty with regard to the gradient magnitude concerning parameters in the final (output) layer and collect a batch of examples in which gradients span a diverse set of directions to capture the diversity.

Generation-based methods attempt to generate informative samples to reduce the annotation budget. GAAL [53] intends to generate samples at the target model's decision boundary; it introduces the generative adversarial network into AL for the first time. BGADL [42] combines AL and data augmentation [43] to continuously generate informative samples to accelerate the learning of the training model.

However, the AL methods mentioned above are based on the assumption that labeled and unlabeled data originate from the same class distribution. Hence, their performance undergoes a sharp deterioration under class distribution mismatch.

Contrastive learning: Contrastive learning is an efficient tool in learning representations, which yields a specific feature space that benefits the downstream tasks. A practical way of implementing contrastive learning involves the construction of positive and negative pairs in the training

stage and embedding the anchor close to the positive samples while pushing it away from negative ones [17].

Contrastive learning has achieved significant attention in recent years owing to its success in self-supervised learning [22]. Transformation is vital in contrastive learning, and several studies [6, 30, 5, 39] focus on designing varied augmentations to yield useful representations. SimCLR [5] is a self-paced approach that considers the potential value and easiness of an instance simultaneously. CSI [39] contrasts the sample with distributionally-shifted augmentations of itself. Moreover, various methods attempt to learn invariant feature by combining contrastive learning with clustering, such as [25] and [4]. Beyond this, to ensure the effect of contrastive learning, Dosovitskiy et al. [8] propose a memory bank mechanism to store the representations computed in the training process. MoCo [12] treats contrastive learning as a dictionary lookup problem, wherein a dynamic dictionary is designed with a queue and considered the recent representations to be more important. Khosla et al. [19] integrate class information with contrastive learning, wherein it is considered that the samples obtained from the same categories are positive pairs and those obtained from different classes are negative pairs.

Semi-supervised learning: Semi-supervised learning (SSL) aims to solve the problem of insufficient labeled data. Different from AL, it utilizes unlabeled data to improve the target model's performance, thus reducing the demand for more labeled data. Based on the scenario of class distribution mismatch, Guo et al. propose a deep SSL framework, DS^{3}L [11], which uses unlabeled data selectively to ensure that the accuracy of supervised learning is not compromised. UASD [7] combines self-distillation and out-of-distribution filtering, which produces soft targets to avoid catastrophic error propagation.

3. Proposed Method: CCAL

Let labeled data as $\mathcal{D}_L = (X^L, Y^L) = \{(x_i^L, y_i^L)\}_{i=1}^{n_L}$ and unlabeled data as $\mathcal{D}_U = X^U = \{x_j^U\}_{j=1}^{n_U}$, wherein the labeled samples are $i.i.d.$ ones over space $\mathcal{D}$, i.e., $\mathcal{D}_L \sim \mathcal{D}$, and $n^L \ll n^U$. Each labeled sample belongs to one of the K known categories in label space $\mathcal{Y}$, and $\mathcal{Y} = \{y_i\}_{i=1}^{K}$. However, an unlabeled one's category may be excluded in $\mathcal{Y}$ under class distribution mismatch. Let $X^{ID\text{-}C}$ and $X^{OOD\text{-}C}$ denote the samples in and out of the class distribution of the labeled data, respectively. Then, the entire data can be redefined as the combination of $X^{ID\text{-}C}$ and $X^{OOD\text{-}C}$. Specifically, $X^L \subseteq X^{ID\text{-}C}$, $X^U \cap X^{ID\text{-}C} \neq \emptyset$, and $X^{OOD\text{-}C} \subseteq X^U$. Suppose that X_{query} is the query set composed of all the queried samples in AL cycles, and then it may contain the samples from $X^{ID\text{-}C}$ and $X^{OOD\text{-}C}$. Accordingly, X_{query} is the set of $X_{query}^{ID\text{-}C}$ and $X_{query}^{OOD\text{-}C}$, which indicates the queried samples in and out of the class distribution of labeled data.

3.1. AL Error Analysis

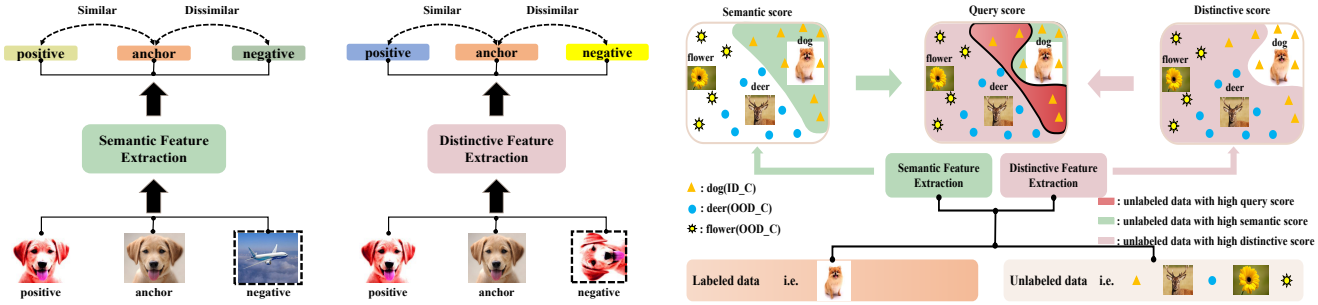
In AL, population risk is jointly controlled by the generalization gap, training error, and AL error. We formulate the population risk as Eq.1. In Eq.1, the generalization gap is the gap between the population risk and generalization loss in $X^{ID\text{-}C}$; the training error is the average empirical loss over X^{tr}, where $X^{tr} = X^L \cup X_{query}^{ID\text{-}C}$ denotes the samples trained for building target model in AL; the AL error consists of valid query error and invalid query error. Invalid query error is due to those queried samples, $X_{query}^{OOD\text{-}C}$, which are invalid for improving the target model; it is measured by the average empirical loss over X^{re}, where X^{re} denotes the samples belonging to $\left(X^{ID\text{-}C} - X_{query}^{ID\text{-}C}\right)$, but replaced by $X_{query}^{OOD\text{-}C}$ in the query process. The valid query error is measured by the difference between the average empirical loss over X^{tr} and the average empirical loss over $X^{ID\backslash re} = X^{ID\text{-}C} - X^{re}$. When those queried samples, i.e., $X_{query}^{ID\text{-}C}$, are more informative, the smaller both valid query error and invalid query error are. Denote that in Eq.1, $p = |X^{ID\text{-}C}|$, $q = |X^{tr}|$, $d = |X^{re}|$.

$$
\begin{aligned}
&E_{(x,y)\sim D}[l(x,y;\mathbf{w})] \\
&\leq \underbrace{\left|E_{(x,y)\sim D}[l(x,y;\mathbf{w})] - \frac{1}{p}\sum_{i=1}^{p} l(x_i^{ID\text{-}C}, y_i^{ID\text{-}C};\mathbf{w})\right|}_{\text{generalization gap}} + \underbrace{\left|\frac{1}{q}\sum_{i=1}^{q} l(x_i^{tr}, y_i^{tr};\mathbf{w})\right|}_{\text{training error}} \\
&+ \underbrace{\left|\frac{1}{p}\sum_{i=1}^{p-d} l(x_i^{ID\backslash re}, y_i^{ID\backslash re};\mathbf{w}) - \frac{1}{q}\sum_{i=1}^{q} l(x_i^{tr}, y_i^{tr};\mathbf{w})\right|}_{\text{valid query error}} + \underbrace{\left|\frac{1}{p}\sum_{i=1}^{d} l(x_i^{re}, y_i^{re};\mathbf{w})\right|}_{\text{invalid query error}} \\
&\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{AL error (CCAL error)}}
\end{aligned}
\tag{1}
$$

Traditional strategies risk querying many samples of unseen classes in terms of class distribution mismatch, resulting in a drastic waste of annotation budget and a higher AL error. For the sake of higher generalization ability and less annotation budget, we propose a scheme combining the semantic and distinctive features, named CCAL, to minimize the AL error, especially the invalid query error. Followed by, we exploit contrastive learning to extract semantic and distinctive features, forming a combined query strategy. Finally, we present a theorem to analyze the upper bound of the AL error, which theoretically shows the effectiveness of CCAL.

3.2. Learning Semantic Features

Under class distribution mismatch, one drawback of the existing AL methods is that the drastic increase in the number of samples in unseen classes reduces the informativeness and representativeness of selected data. Thus, filtering unlabeled samples with mismatched categories poses a significance. Heuristically, one feasible and effective means to handle this tricky problem is to distinguish $X^{ID\text{-}C}$ and



(a) Contrast of semantics. (b) Contrast of distinctiveness. (c) Joint query strategy.

Figure 3: A effective CCAL framework that extracts semantic and distinctive features by contrastive learning and combines the semantic score S_{sem} and distinctive score S_{dis} to select samples for labeling.

X^{OOD_C} in semantics. Contrastive learning is such a discriminative approach, which does not pay attention to the detailed information but learns invariant semantic properties of the samples [22]. Based on the Instance Discrimination task theory introduced by Wu et al. [47], the semantics of the samples are learned via extracting the semantic feature by contrastive learning in CCAL. Specifically, the positive sample is the one that applies several transformations at random, such as Random crop, Horizontal flip, etc., and the negative one is randomly sampled from the remainder unlabeled data, as shown in Figure 3(a).

Let $B = \{x_i\}_{i=1}^{a}$ denotes a batch, and $\hat{x}_i^+$ indicates the sample x_i augmentated by the random and independent transformation mentioned as above. Then, the positive pairs can be represented as $(\hat{x}_i, \hat{x}_i^+)$ and the negative pairs as $(\hat{x}_i, \hat{B}_{-i})$. The loss of the semantic contrasting is formulated as Eq.2.

$$L_s(B) = \frac{1}{|\hat{B}|} \sum_{i=1}^{|\hat{B}|} L_{con}\left(\hat{x}_i, \hat{x}_i^+, \hat{B}_{-i}\right), \tag{2}$$

where $\hat{B} = \{\hat{x}_i\}_{i=1}^{a} \cup \{\hat{x}_i^+\}_{i=1}^{a}$, $\tilde{B}_{-i} = \{\hat{x}_j\}_{j\neq i} \cup \{\hat{x}_j^+\}_{j\neq i}$, $x_i \in X^L \cup X^U$, and L_{con} is a contrastive learning loss [5].

Then, the semantic feature, $z_s(\cdot)$, can be learned by Eq.2, basing on the semantic score defined as Eq.3.

$$S_{sem}\left(x_i^U\right) = \sigma\left[\max_j cos\left(z_s(x_j^L), z_s(x_i^U)\right)\right], \tag{3}$$

where $\sigma[\cdot]$ is a Min-Max normalization operator [16].

Eq.3 achieves the maximum semantic similarity degree of an unlabeled sample to the labeled data. The larger the value of the semantic score, the higher the probability of the unlabeled one belonging to the known categories, and the lower the invalid query error is. Thus, the semantic score is available to design the query strategy.

3.3. Learning Distinctive Features

The samples in and out of class distribution can be roughly distinguished by semantic feature learning, considerably reducing the invalid query error. However, if only the semantic score is used as a query measure, many uninformative samples within the class distribution may be queried. Consequently, such samples will not significantly improve the target model's performance since they are similar or even identical to labeled data. It is, therefore, necessary to extract the distinctive features of labeled and unlabeled data and filter those uninformative samples in query processing.

Prompted by [39], the rotation transform [10], shifts the sample input distribution but retains invariant semantics. We consider the rotated sample the negative one, and the transformations in semantic feature learning augment the positive one. This allows the feature extractor to concentrate on the distinctiveness of the features.

Let $x_{i,k}$ denotes the sample x_i rotated by k degrees. Then, the positive pairs can be denoted as $(\hat{x}_{i,k}, \hat{x}_{i,k}^+)$, and the negative ones as $(\hat{x}_{i,k}, \hat{B}_{-i,k}^R)$. The distinctive contrasting loss is formulated as Eq.4.

$$L_d(B;R) = \frac{1}{|\hat{B}^R|}\frac{1}{|R|}\sum_{i=1}^{|\hat{B}^R|}\sum_{k\in R}\Big[L_{con}(\hat{x}_{i,k}, \hat{x}_{i,k}^+, \hat{B}_{-i,k}^R) + log\, p(k|\hat{x}_{i,k})\Big] \tag{4}$$

where $\hat{B}_{-i,k}^R$ includes all the augmentated samples except the $x_{i,k}$ and $x_{i,k}^+$.

Then, the distinctive features learned by Eq.4, $z_d(\cdot)$, are exploited to measure the distinctiveness of unlabeled to the labeled samples with the same semantics. Each unlabeled sample is annotated with a pseudo semantic label, which is the same as the label of its closest labeled one. Given an unlabeled sample x_i^U, let $x_{i,st}^L$ and $x_{i,nd}^L$ denote its closest and second closest labeled samples with the same semantic label, then the distinctiveness of the unlabeled sample can be measured by Eq.5, where $\sigma[\cdot]$ is a normalization operator as same as Eq.3.

$$S_{dis}(x_i^U) = 1 - \sigma\Big[cos\big(z_d(x_i^U), z_d(x_{i,st}^L)\big) - cos\big(z_d(x_i^U), z_d(x_{i,nd}^L)\big) + cos\big(z_d(x_{i,st}^L), z_d(x_{i,nd}^L)\big)\Big]. \tag{5}$$

The first two-term in $\sigma[\cdot]$ of Eq.5 measures the difference of x_i^U to labeled samples. The larger the difference, the greater the similarity of x_i^U with the labeled one, followed

by a smaller distinctiveness. The third term, to some extent, measures the information of x_i^U. A larger one makes the information of x_i^U smaller. To conclude, the larger the value of S_{dis}, the higher the distinctiveness of sample x_i^U, and the lower the valid query error. This means that the distinctive score can be used in the final query strategy.

3.4. Joint Query Strategy

As shown in Figure 3(c), if only semantic features are used, numerous redundant unlabeled samples within the class distribution may be queried. For example, in the green area of the top-left graph in Figure 3(c), those dots near the image "dog" are less informative to the target model. However, simply using distinctive features may lead to many invalid queries. For example, in the pink area of the top-right graph in Figure 3(c), those dots with the label "flower" are invalid for improving the target model. Thus, the valid queries should contain those samples within the class distribution but with distinctive features. Consequently, it is paramount to design a combined strategy with contrastive coding on semantic and distinctive features.

A naive way to balance the semantic and distinctive scores is to bring them to the same range and assign them different weights, such as $\beta S_{sem} + S_{dis}$. However, this idea is not suitable for class distribution mismatch. It is unreasonable to simultaneously enlarge or narrow all the samples' semantic scores by β. As the samples with lower semantic scores are more likely to be out of class distribution, their semantic scores should be narrowed in joint query scores. On the contrary, the samples with higher semantic scores should be enlarged. Therefore, we enlarge the semantic score from [0,1] to [-1,1] by the mapping $tanh[x] = \frac{1-e^{-x}}{1+e^{-x}}$, and then we have the final strategy S_{query} as Eq.6.

$$S_{query}(x_i^U) = tanh\left[\psi\left(S_{sem}(x_i^U)\right)\right] + S_{dis}(x_i^U) \quad (6)$$

where $\psi\left(S_{sem}(\cdot)\right) = k \times (S_{sem}(\cdot) - t)$, the nonlinear function, $tanh[\cdot]$, selectively narrows the semantic scores of samples by threshold t, and k is introduced to control the slop of $tanh[\cdot]$. The larger the value of S_{query}, the lower the AL error. The detail of algorithm for CCAL query is shown in Algorithm 1.

3.5. Theoretical Studies

In this Subsection, we discuss and analyze the upper boundary of the AL error and then verify the effectiveness of CCAL. As shown in Eq.1, population risk is jointly controlled by the generalization gap, training error, and AL error. Empirically, it has been widely observed that the training error can be reduced to near zero in Convolutional Neural Networks (CNNs). Theoretically, it has been proved that the generalization gap of CNNs can be bounded [26]. Hence, the essential part for AL under class distribution mismatch is the AL error, also called CCAL error. Further, we analyze the upper boundary of CCAL error in Theorem 1.

Algorithm 1: Joint Query Strategy in CCAL

Input: Labeled data (X^L, Y^L), Unlabeled data X^U, Budget: b, Number of categories in labeled data: K, semantic encoder: θ_s, distinctive encoder: θ_d

Output: $(X^L, Y^L), X^U$

```
Calculate the semantic features of X^L and X^U using θ_s: z_s(X^L) = θ_s(X^L), z_s(X^U) = θ_s(X^U);
Calculate the distinctive features of X^L and X^U using θ_d: z_d(X^L) = θ_d(X^L), z_d(X^U) = θ_d(X^U);
for e in AL cycles do
    X_query_e = ∅;
    Calculate S_sem(X^U) using Eq.3 and obtain pseudo-semantic set X^{U_l}, l ∈ Y^L. Where X^{U_l} is composed of the unlabeled samples with the same pseudo-semantic label l;
    for l in Y^L do
        Calculate S_dis(X^{U_l}) using Eq.5;
        Calculate S_query(X^{U_l}) using Eq.6;
        Select the unlabeled samples with max_{b/K} S_query(X^{U_l}) and then add to X_query_e;
    end
    Given the labels Y_query_e of X_query_e by oracle;
    (X^L, Y^L) ← (X^L, Y^L) ∪ (X^{ID_C}_{query_e}, Y^{ID_C}_{query_e});
    X^U ← X^U − X_query_e;
end
Return (X^L, Y^L), X^U;
```

Theorem 1 *Given p i.i.d. samples drawn from $\mathcal{D}$ as $\{x_i, y_i\}_{i=1}^p$, and set of points X^{tr}, X^{re} with size of q, d respectively. If the loss function $l(\cdot, y; \mathbf{w})$ is λ^l − Lipschitz continuous for all y, $\mathbf{w}$ and bounded by T, regression function is λ^μ − Lipschitz continuous, training error $l(\mathbf{x_j}, \mathbf{y_j}; \mathbf{w}) = \mathbf{0}$, $j \in \{1, 2, ..., q\}$, and CCAL strategy can maximize the lower boundary α of information measure S_{dis} (S_{dis} is further defined in Appendix.); with probability of at least $1 - \gamma$,*

$$\left\{\left|\frac{1}{p}\sum_{i=1}^{p-d} l(x_i^{ID\backslash re}, y_i^{ID\backslash re}; \mathbf{w}) - \frac{1}{q}\sum_{j=1}^{q} l(x_j^{tr}, y_j^{tr}; \mathbf{w})\right| + \left|\frac{1}{p}\sum_{i=1}^{d} l(x_i^{re}, y_i^{re}; \mathbf{w})\right|\right\} \quad (7)$$
$$\leq \sqrt{6 - 2\alpha}(\lambda^l + \lambda^\mu TK) + \sqrt{\frac{T^2 log(1/\gamma)}{2p}} + \frac{dT}{p}.$$

According to Theorem 1, the smaller d and the larger α, followed by the fewer the invalid queried samples become and the more informative the valid queried samples are. This results in a tighter upper boundary of CCAL error. CCAL then obtains a more satisfied target model. The detailed proof for Theorem 1 was provided in Appendix.

4. Experiments

In the following, we evaluate the CCAL's performance on classification tasks in Subsection 4.1 and perform ablation study and sensitivity analysis in Subsections 4.2 & 4.3, respectively. Besides, CCAL is compared with two state-of-the-art SSL methods in Subsection 4.4. The code is available at https://github.com/RUC-DWBI-ML/CCAL.

Datasets. There are two benchmark datasets used, i.e., CIFAR10 [21] and CIFAR100 [21]. CIFAR10 contains 50000 training images and 10000 test images from 10 categories with a size of $32 \times 32 \times 3$. The size of CIFAR100 is the same as that of CIFAR10; however, CIFAR100 contains 100 categories, which contain 500 training images and 100 test images, each, with a size of $32 \times 32 \times 3$. Moreover, an artificial cross-dataset consist of five datasets as CIFAR10, CIFAR100, Flowers [32], Places-365 [52], and Food-101 [3], is utilized.

In all experiments, the labeled data is initialized by randomly sampling 8% samples from $X^{ID\text{-}C}$, and 1500 samples are selected to request the labels from oracle at each cycle of AL. To evaluate CCAL on datasets with different mismatch ratios, we use Eq.8 to construct artificial unlabeled data. The mismatch ratio is set as 20%, 40%, 60%, and 80%, respectively, in the subsequent experiments.

$$ratio(mismatch) = \frac{|X_U^{OOD\text{-}C}|}{|X_U^{ID\text{-}C}| + |X_U^{OOD\text{-}C}|} \tag{8}$$

where $|X_U^{ID\text{-}C}|$ and $|X_U^{OOD\text{-}C}|$ denotes the number of unlabeled samples in $X^{ID\text{-}C}$ and $X^{OOD\text{-}C}$, respectively.

Baselines. CCAL is compared with the following six state-of-the-art AL algorithms which can be divided into the following five categories. (1) Random sampling: Randomly select samples from unlabeled pool for labeling. (2) Uncertainty: Entropy [28] and VAAL [38]. (3) Diversity: Coreset [35]. (4) Uncertainty and Diversity: BADGE [1]. (5) Generation: BGADL [42].

Implementation details. For the target classification model training, ResNet18 [13] is adopted as the target task module cooperates with the data augmentation of random horizontal flip. In each AL cycle, training continues 100 epochs, and Adam is adopted as the optimizer [20] with an equal learning rate of 5e-4 and batch size of 32. In Subsection 4.4, the accuracy is the result over 1 run due to the stability of the adopted classifier, Wide ResNet-28-2 [51], in comparison. The reported accuracy of these compared AL methods is the average of the results over 5 runs in the remaining parts, and the shaded area in the reported results represents the standard deviation of the five runs. Unless otherwise specified, all parameters in semantic and diatinctive feature learning are the same as [5] and [39], respectively. Note that except the sensitivity analysis in Subsection 4.3, k and t are set as 100 and 0.9, respectively.

4.1. Evaluation on image classification benchmarks

In CIFAR10, we perform a binary classification task on airplane and automobile. The remainder 8 categories ("bird," "cat," "dog," "deer," "frog," "horse," "ship," and "trunk") are seen as unknown classes not presented in labeled data. In CIFAR100, a 20-class classification task is conducted on 4 superclasses of large carnivores, large omnivores and herbivores, medium-sized mammals, and small mammals, each of which contains 5 categories. The remainder 80 categories are unknown classes. In the cross-dataset, the classification task is conducted on the 6 animal categories from CIFAR10, while 668 categories originating from the four external datasets are regarded as unknown.

Evaluation results on CIFAR10, CIFAR100, and cross-dataset. The results of CIFAR10, CIFAR100, and cross-datset are shown in Figure 4 & 5 & 7, respectively. Then, we have the following five observations. 1) Unlike the existing AL methods whose performance deteriorates as the ratio of class mismatch increases, CCAL achieves satisfactory performance even on 80% ratio of class mismatch. More specifically, in CIFAR10 and CIFAR100, CCAL achieve a supervised accuracy by just annotating 16.3% and 26.09% of the unlabeled data under 80% class mismatch, respectively. This shows the design of joint semantic and distinctive feature extraction is effective and it makes CCAL query more valid and informative samples. 2) As shown in Figure 4 & 5 & 7, with the mismatch ratio increasing, the accuracies of the existing AL methods, such as VAAL and Coreset, approximate to random sampling. This indicates that the query measure of the existing AL methods work less effectively or even ineffectively under class distribution mismatch. 3) With the increment of the mismatch ratio, the performance of the existing AL methods decreases more obviously compared with CIFAR10, as shown in Figure 4 & 5. For example, CCAL outperforms random sampling by a margin of 2.44% and 7.65% in CIFAR10 and CIFAR100 respectively, when mismatch ratio is 80% and the number of AL cycle is 5. This suggests the existing AL methods are more sensitive when unlabeled data contain numerous mismatch categories. 4) As shown in Figure 7, CCAL performs remarkably well in comparison to the AL method. This represents CCAL success even the class distribution across muti-datasets. 5) It is worth noting that almost all the compared methods can achieve a super-

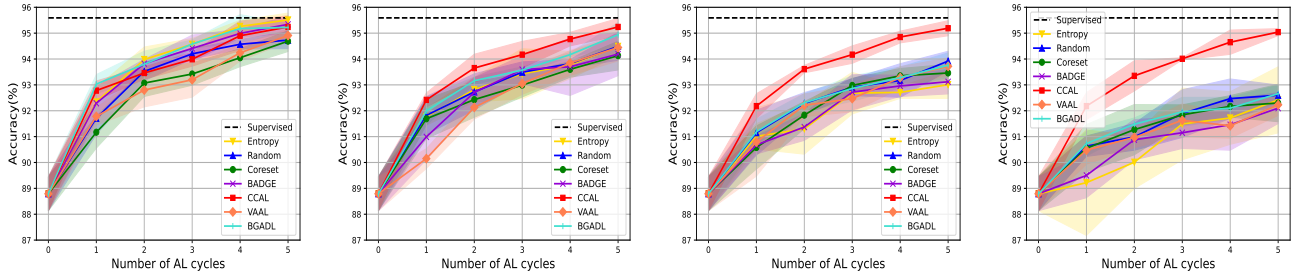


Figure 4: Classification accuracy of CCAL and compared AL algorithms on CIFAR10 under different mismatches.

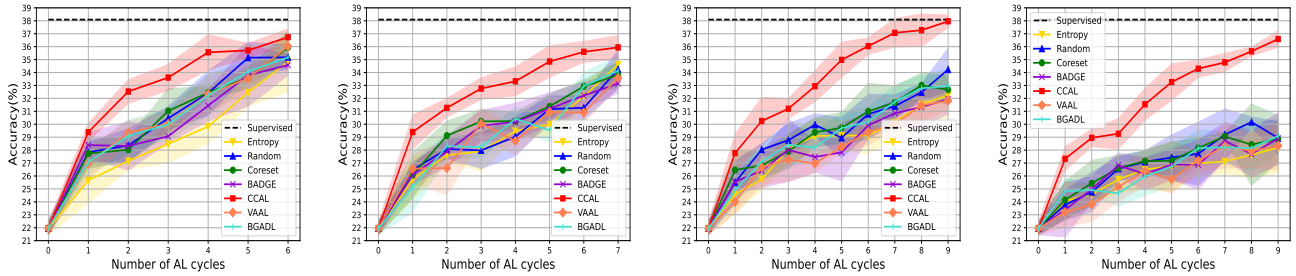


Figure 5: Classification accuracy of CCAL and compared AL algorithms on CIFAR100 under different mismatches.

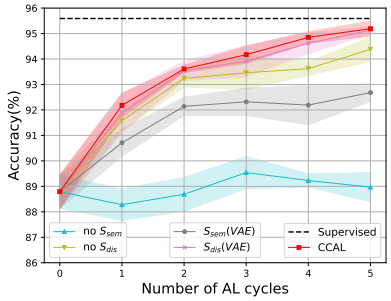


Figure 6: Ablation study to analyze the influence of each part of CCAL.

vised accuracy after 5 AL cycles when the mismatch ratio is 20%, due to the problem of class distribution mismatch degenerates into the problem with noise.

4.2. Ablation study

The ablation study is conducted on CIFAR10 with the mismatch ratio is 60% to demonstrate the effectiveness of CCAL. The experimental results are shown in Figure 6.

• **no $\mathbf{S_{sem}}$:** It denotes that semantic score S_{sem} is not used. The accuracy of "no S_{sem}" rapidly deteriorates and oscillates compared with the CCAL. This demonstrates that semantic score S_{sem} plays an essential role in our query strategy. Without S_{sem}, CCAL no long works well.

• **no $\mathbf{S_{dis}}$:** It indicates that distinctive score S_{dis} is not used. When removing the distinctive score S_{dis}, we observe the accuracy of "no S_{dis}" decreases compared to the CCAL. This shows that filtering those uninformative samples within class distribution does work in CCAL.

• **$\mathbf{S_{sem}}$(VAE):** It states that VAE learns the semantic feature. To assess the role of contrastive learning, we replace contrastive learning with VAE. CCAL exceeds S_{sem}(VAE) by a margin of 3.74% when the number of AL cycles is 5. This suggests that contrastive learning is a great technique for learning semantic features.

• **$\mathbf{S_{dis}}$(VAE):** It points out that VAE learns the distinctive feature. As shown in Figure 6, S_{dis}(VAE) exhibits comparable performance compare with CCAL. Hence, the distinctive score is appropriate for the feature VAE has learned.

4.3. Sensitivity analysis

The sensitivity of parameters k and t is analyzed on CIFAR10 with a mismatch ratio of 60%. Figure 8 shows the accuracy and the distribution of $tanh[\psi(S_{sem})]$ with different parameters.

• **Evaluate the effect of t:** The experiments with t equal to 0.7, 0.8, 0.9 while k fixed as 100 are performed to evaluate the sensitivity of t. As shown in Figure 8(a), with t increasing, the samples with a score from 0.9 to 1 in $X_U^{OOD\_C}$ become less numerous, and CCAL accuracy improves. Thus in our experiments, t is set as 0.9.

• **Evaluate the effect of k:** The experiments with k equal to 10, 70, 130 while t fixed as 0.9 are done to assess the sensitivity of k. As illustrated in Figure 8(b), the samples with a [-0.9,0.9] score decreasing with the increment of k. Moreover, the accuracy of k equal to 70 has a barely noticeable difference compared to 130. But the accuracy will decrease when k equal to 10 since $tanh[\psi(S_{sem})]$ cannot reach 1. Therefore, CCAL's performance is not sensitive to k, and it would be even better when $k > 10$.

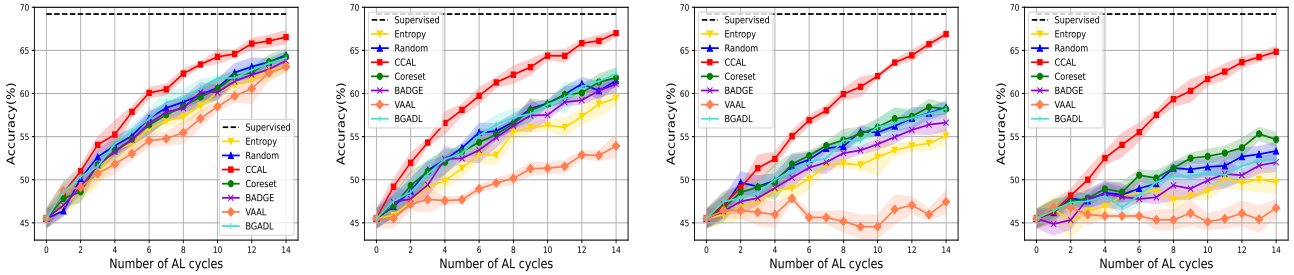
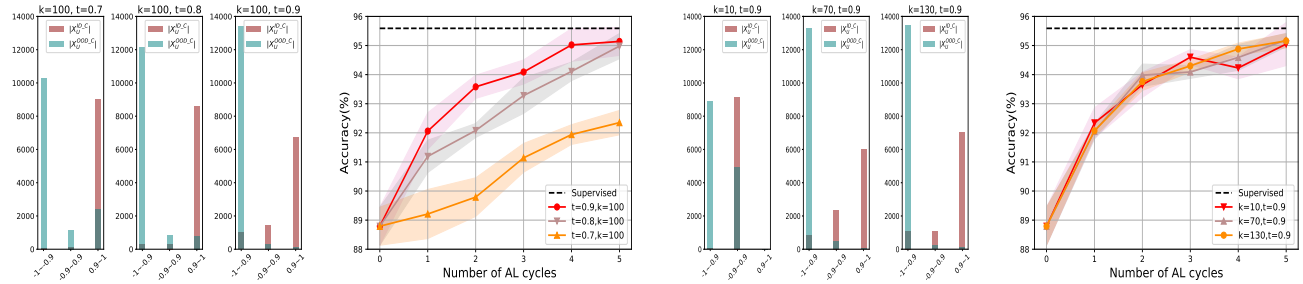


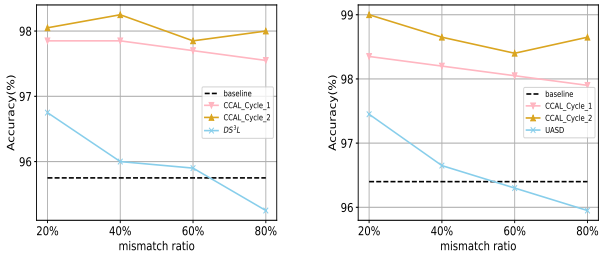
Figure 7: Classification accuracy of CCAL and compared AL algorithms on cross-dataset under different mismatches.



(a) Analyze the influence of the change of t in S_{query}.

(b) Analyze the influence of the change of k in S_{query}.

Figure 8: Analyze the sensitivity of parameters k and t in S_{query} on the experimental results.



(a) DS^{3}L and CCAL.

(b) UASD and CCAL.

Figure 9: The classification accuracy of CCAL and SSL methods DS^{3}L and USAD which based on the problem of class distribution mismatch.

4.4. Comparison of CCAL and SSL

In this Subsection, CCAL is compared to only two SSL methods focusing on class distribution mismatch, DS^{3}L and UASD, on CIFAR10 under different mismatch ratios. Except for the mismatch data construction, all the experiment settings are followed with DS^{3}L [11] and UASD [7], respectively. In Figure 9, "baseline" represents training the target model with 800 initialized labeled samples. "CCAL_Cycle_1" and "CCAL_Cycle_2" indicate that the number of AL cycles is 1 and 2, respectively. CCAL accuracy is remarkably superior to DS^{3}L and UASD after just one AL cycle. Therefore, in practical applications, people can choose a suitable method depending on their needs: budget saving or high accuracy.

5. Conclusion

This paper finds that AL error is composed of valid query error and invalid query error under class distribution mismatch. Following the theoretical, a contrastive coding based AL framework CCAL is proposed, which learns semantic and distinctive features by contrastive coding and joins them in query strategy. Unlike the existing AL algorithms, CCAL can effectively keep away from the unlabeled samples with mismatched categories on one hand, and on the other hand, it can seek the most informative samples with matched categories from the unlabeled pool. Experimental results on two benchmark datasets and an artificial cross-dataset demonstrate that CCAL achieves state-of-the-art performance with much lower annotation costs by a large margin. In the future, we expect to extend our algorithm to the text classification task under class distribution mismatch by designing an AL architecture with specific contrastive coding.

6. Acknowledgement

This work is supported by the National Key Research & Development Plan of China (2018YFB1004401), National Natural Science Foundation of China (No. 61732006, 61772537, 61772536, 62072460, 62076245), and Beijing Natural Science Foundation (4212022).

References

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learn-

ing by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2019. 2, 6

[2] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 6

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4, 6

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 3

[7] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3569–3576, 2020. 1, 3, 8

[8] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. Citeseer, 2014. 3

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017. 2

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 4

[11] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, pages 3897–3906. PMLR, 2020. 1, 3, 8

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[15] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23:892–900, 2010. 2

[16] Anil Jain, Karthik Nandakumar, and Arun Ross. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005. 4

[17] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021. 3

[18] 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*, pages 2372–2379. IEEE, 2009. 2

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[22] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 2020. 3, 4

[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[24] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994. 2

[25] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 3

[26] Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *arXiv preprint arXiv:1806.05159*, 2018. 5

[27] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. *arXiv preprint arXiv:1410.1141*, 2014.

[28] Wenjie Luo, Alex Schwing, and Raquel Urtasun. Latent structured active learning. *Advances in Neural Information Processing Systems*, 26:728–736, 2013. 2, 6

[29] Andrew Kachites McCallumzy and Kamal Nigamy. Employing em and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning (ICML)*, pages 359–367. Citeseer, 1998. 2

[30] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 3

[31] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004. 2

[32] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006. 6

[33] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer, 2006. 2

[34] Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3839–3848, 2018.

[35] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 2, 6

[36] Burr Settles. Active learning literature survey. 2009. 1, 2

[37] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992. 2

[38] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019. 2, 6

[39] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *34th Conference on Neural Information Processing Systems (NeurIPS) 2020*. Neural Information Processing Systems, 2020. 3, 4, 6

[40] Ying-Peng Tang and Sheng-Jun Huang. Self-paced active learning: Query the right thing at the right time. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5117–5124, 2019. 2

[41] 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. 2

[42] Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. Bayesian generative active deep learning. In *International Conference on Machine Learning*, pages 6295–6304. PMLR, 2019. 2, 6

[43] Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. A bayesian data augmentation approach for learning deep models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2794–2803, 2017. 2

[44] Sudheendra Vijayanarasimhan and Kristen Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International journal of computer vision*, 108(1):97–114, 2014. 2

[45] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. 2

[46] Zheng Wang and Jieping Ye. Querying discriminative and representative samples for batch mode active learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(3):1–23, 2015. 2

[47] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 4

[48] Haiqin Yang, Kaizhu Huang, Irwin King, and Michael R Lyu. Maximum margin semi-supervised learning with irrelevant data. *Neural Networks*, 70:90–102, 2015. 1

[49] Haiqin Yang, Shenghuo Zhu, Irwin King, and Michael R Lyu. Can irrelevant data help semi-supervised learning, why and how? In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 937–946, 2011. 1

[50] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 93–102, 2019. 2

[51] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 6

[52] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6

[53] Jia-Jie Zhu and José Bento. Generative adversarial active learning. *arXiv preprint arXiv:1702.07956*, 2017. 2