Multi-level Adaptive Active Learning for Scene Classification

Xin Li and Yuhong Guo

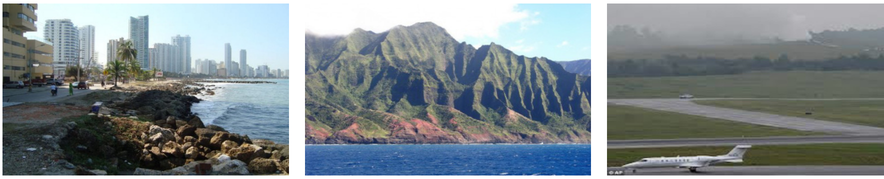
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122, USA
{xinli, yuhong}@temple.edu

Abstract. Semantic scene classification is a challenging problem in computer vision. In this paper, we present a novel multi-level active learning approach to reduce the human annotation effort for training robust scene classification models. Different from most existing active learning methods that can only query labels for selected instances at the target categorization level, i.e., the scene class level, our approach establishes a semantic framework that predicts scene labels based on a latent object-based semantic representation of images, and is capable to query labels at two different levels, the target scene class level (abstractive high level) and the latent object class level (semantic middle level). Specifically, we develop an adaptive active learning strategy to perform multi-level label query, which maintains the default label query at the target scene class level, but switches to the latent object class level whenever an "unexpected" target class label is returned by the labeler. We conduct experiments on two standard scene classification datasets to investigate the efficacy of the proposed approach. Our empirical results show the proposed adaptive multi-level active learning approach can outperform both baseline active learning methods and a state-of-the-art multi-level active learning method.

Keywords: Active Learning, Scene Classification.

1 Introduction

Scene classification remains one of the most challenging problems in computer vision field. Different from the classification tasks in other fields such as NLP, where the meanings of features (e.g., words) are perceivable by human beings, the low-level features of an image are primarily built on some signal responses or statistic information of mathematical transformations. Though these low-level features are useful and powerful as proved by numerous works for decades, the *semantic gap* between the semantically non-meaningful low-level features and the high-level abstractive scene labels becomes a bottleneck for further improving scene classification performance. Recent advances on scene classification [24, 19] and other related tasks such as semantic segmentation [29, 3, 12] and object



(a) coast/city (b) mountain/coast (c) field/airport

(d) theater/auditorium (e) airport/mall (f) terminal/auditorium

Fig. 1: Examples of ambiguous scene categories. *(a)-(c)* are confusing examples of outdoor scenes and *(d)-(e)* are examples of indoor scenes.

detection/recognition [32, 11, 5] have demonstrated the importance of exploiting semantic information and extracting high-level scene label structures, where a scene label (e.g., coast) can be viewed as a semantic concept comprising of a set of important high level visual objects (e.g., sky, sand and sea). The work in [14] particularly demonstrated the strength of predicting scene labels based on the high-level object-based representations of images. However, this work requires supervised training of object detectors, which can significantly increase the demand for human annotation effort. Moreover, to produce a good scene classification model, a sufficient amount of target scene labels need to be acquired as well, which induces expensive human annotation cost. In this work, we address the important problem of reducing human annotation effort for learning scene classification models.

Active learning is a well studied technique for reducing the cost of manual annotations by performing selective instance sampling. In contrast to "passive" learning where the learner uses randomly generated labeled instances, "active" learners iteratively select the most informative instances to label in an interactive learning process [25]. Traditional active learners query labels for the selected instance at the target prediction label level, which however is not the best strategy in many cases of scene classification tasks. Scene labels are highly abstractive and semantic labels. Without accurately identifying their high level object-based semantic representations, some scene labels can be very difficult to be distinguished from each other even by a human labeler in many scenarios. For example, it is hard to tell for a human labeler whether the image in Figure 1(b) is indeed a mountain scene or a coast scene; similarly, it is hard to tell whether the image in Figure 1(e) is the seating area of a mall or an airport terminal. From Figure 1 we can see that such ambiguities exist not only among outdoor scenes but also in indoor scenes. However, the objects contained in these images are much more easier to be identified by a human labeler. The object level labels may successfully infer the scene labels based on the object-based statistical semantic scene structure induced from the labeled data.

Based on these observations, in this paper we develop a novel multi-level adaptive active learning approach to reduce the annotation effort of learning accurate scene classification models. This approach is based on a latent object-based hierarchical scene classification model, which involves both scene classifier and object classifiers. It selects both instance and label types to query, aiming to reduce the overall prediction uncertainty of the multi-class scene classification model over all labeled and unlabeled instances. By default, it performs label query at the target scene class level and selects instance based on a maximum conditional mutual information criterion. But whenever an "unexpected" target scene label is returned by the labeler in a given iteration, it will switch to perform label query at the latent object class level in the next iteration for once. After querying for a scene label, only the scene classifier will be updated. But if an object label is queried, both object and scene classifiers will be updated. We conduct experiments on two standard scene classification datasets to investigate the efficacy of the proposed approach. Our empirical results show the proposed adaptive multi-level active learning approach can outperform a few baseline active learning methods and a state-of-the-art multi-level active learning method.

2 Related Work

In this section, we present a brief review over the related scene classification and active learning works developed in computer vision field.

Scene classification has long gained its popularity in the literature. Previous works on scene classification can be categorized into two main groups: data representation centered methods and classification model centered methods. In the first group, mid-level representations built from low-level features such as SIFT [17] or HOG [2] features have been exploited for scene classification. For example, [4] introduces a bag-of-words (BoW) model based on low-level features to represent a natural scene image. [13] proposes a spatial pyramid matching model to further improve the BoW model by taking the spatial relationship between the visual words into account. [33] proposes a novel holistic image descriptor for scene classification. More recent efforts have centered on representing a scene with semantically meaningful information rather than statistic information of low-level hand-designed features. [24] proposes an image representation based on discriminative scene regions detected using a latent SVM model. [14] proposes an object-centered approach called object bank, where each image is represented as the response map to a large number of pre-trained generic object detectors. Our classification model shares similarity with this work on using the presence of objects as attributes for scene classification. However, the object bank method requires supervised training of a large number of object detectors which is extremely expensive in terms of annotation cost, while the object classifiers in our model are learned on the fly in a semi-supervise manner and require very limited annotations. Moreover, the object detectors of the object bank model take the whole image as input, while our object classifiers pursue patch-based training. Another work [22] also proposes an attribute based scene representation which contains binary attributes to describe the intra- and inter-class scene variations. But similar to the object bank method, their attribute learning is quite expensive and they predict the presence of attributes using the sliding window technique which further increases the computational cost.

For methods centered on classification model development, we would like to mention a few works with widely used techniques [19, 23, 20]. In [19], a deformable part-based model (DPM) has been applied to address scene categorization. [23] proposes a prototype based model for indoor scenes that captures the characteristic arrangements of scene components. [20] proposes a latent structural SVM for the reconfigurable version of a spatial bag of words model. These methods also demonstrate the usefulness of exploiting mid-level representations for scene classification. Nevertheless, all these methods are passive learning methods and require a large number of labeled instances for training.

Active learning methods have been widely used in computer vision field to reduce human labeling efforts in image and video annotation [10, 34], retrieval [31], recognition [7–9] and segmentation [29]. These active learning methods iteratively select the most informative instance to annotate according to a given instance selection criterion. Recently, some researchers have observed that exploiting single criterion for instance selection lacks the capacity of handling different active learning scenarios, and an adaptive active learning strategy that integrates strengths of different instance selection criteria has been proposed in [15]. Nevertheless, all these active learning methods are limited to querying labels in the target prediction label space, and lack sufficient capacity of handling the highly semantic scene classification problems and exploiting advanced scene classification models, especially when the scene images are ambiguous to categorize as demonstrated in Figure 1. Our proposed active learning approach will address the limitation of these current methods by exploiting a latent object-based scene classification model and performing multi-level adaptive label querying at both the scene class level and the object class level.

There are a number of existing active learning methods that query the labelers for information beyond the target image labels. For example, [18] considers attributed based prediction models and asks users for inputs on the attribute level to improve the class predictions, while assuming fixed attribute configurations for each give image class label. [30] treats the overall object classification problem as a multi-instance learning problem and considers the same type of labels at two levels, instance level (segments) and bag level (images). These works [18, 30] nevertheless are still limited to exploiting the same type of standard queries, while another few works [1, 21, 27, 11] have exploited semantic or multiple types of queries. [1, 21] introduces a new interactive learning paradigm that allows the supervisor to additionally convey useful domain knowledge using *relative* attributes. [27] presents an active learning framework to simultaneously learn appearance and contextual models for scene understanding. It explores three different types of questions: regional labeling questions, linguistic questions and contextual questions. However, it does not handle scene classification problems but evaluate the approach regarding the region labels. [11] presents an active learning approach that selects image annotation requests among both object category labels and the object-based attribute labels. It shares similarity with our proposed approach in querying at multi-levels of label spaces, but it treats image labels and attribute labels in the same way and involves expensive computations. Nevertheless, these active learning works tackle object recognition problems using pre-fixed selection criteria. Our proposed approach on the other hand uses an adaptive multi-level active learning strategy to optimize a latent object-based hierarchical scene classification model.

3 Proposed Method

In this section, we first establish the hierarchical semantic scene classification model based on latent object level representations in Section 3.1 and then present our multi-level adaptive active learning method in Section 3.2.

3.1 Hierarchical Scene Classification Model

Learning mid-level representations that capture semantic meanings has been shown to be incredibly useful for computer vision tasks such as scene classification and object recognition. In this work, we treat object category values as high level scene attributes, and use a hierarchical model for scene classification that has a mid-level object representation layer. The work flow of our approach has four stages: Firstly, we preprocess each image into a bag of patches and a bag of low-level feature vectors can be produced from the patches. For the sake of computational efficiency, we only used aligned non-overlapping patches. We expect each patch presents information at the local object level. Secondly, we perform unsupervised clustering over the patches using a clustering method K-Medoids and then assign an object *class name* to each patch cluster by querying the object level labels for the center patch in each cluster. Thirdly, we train a set of binary object classifiers based on these named clusters of patches using the one-vs-all scheme. Then for each image, its mid-level object-based representation can be obtained by applying these object classifiers over its patches. That is, each image will be represented as a binary indicator vector, where each entry of the vector indicates the presence or absence of the corresponding object category in the image. Figure 2 presents examples of this mid-level object-based representation of images. Finally, a multi-class scene classifier is trained based on the mid-level representation of labeled images. To further improve the scene classifier, we have also considered using hybrid features to train the scene classifier. That is, we train the scene classifier based on both the mid-level representation features and the low-level features of the labeled images. This turns out to be more robust for scene classification than using the mid-level representation alone. More details will be discussed in the experimental section.

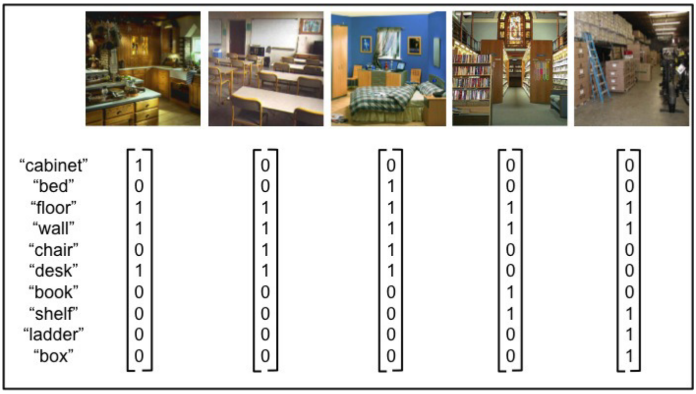


Fig. 2: Examples of the mid-level semantic representation employed in our scene classification model. Each 1 value indicates the presence of an object and each 0 value indicates the absence of an object in a given image.

Our system uses logistic regression as the classification model at both object and scene levels. Given the patch *labels* produced by clustering, for each object class, we have a set of binary labeled patches $\{(\tilde{\mathbf{x}}_i, \tilde{z}_i)\}_{i=1}^{N_o}$ with $\tilde{z}_i \in \{+1, -1\}$. We then train a probabilistic binary logistic regression classifier for each object class to optimize a ℓ_2-norm regularized log-likelihood function

$$\min_{\mathbf{u}} \quad -C \sum_{i=1}^{N_o} \log P(\tilde{z}_i|\tilde{\mathbf{x}}_i) + \frac{1}{2}\mathbf{u}^T\mathbf{u} \tag{1}$$

where

$$P(\tilde{z}_i|\tilde{\mathbf{x}}_i) = \frac{1}{1 + \exp(-\tilde{z}_i \tilde{\mathbf{x}}_i^T \mathbf{u})} \tag{2}$$

For scene classification, given the labeled data $\mathcal{L} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{z}_i$ is the mid-level indicator representation vector for the i-th image $\mathcal{I}_i$, and y_i is its scene class label, we train a multinomial logistic regression model as the scene classifier. Specifically, we perform training by minimizing a ℓ_2-norm regularized negative log-likelihood function

$$\min_{\mathbf{w}} \quad -C \sum_{i=1}^{N} \log P(y_i|\mathbf{z}_i) + \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{3}$$

where

$$P(y_i = c|\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i^T \mathbf{w}_c)}{\sum_{c'} \exp(\mathbf{z}_i^T \mathbf{w}_{c'})} \tag{4}$$

The minimization problems in both (1) and (3) above are convex optimization problems, and we employ the trust region newton method developed in [16] to perform training.

We can see that our hierarchical scene classification model has similar capacity with the object bank method regarding exploiting the object-level representations of images. For object-based representation models, one needs to determine what object classes and how many of them should be used in the model. The object bank model chooses object classes based on some statistic information drew from several public datasets and their object detectors are trained on several large datasets with a large amount of object labels as well. However, our model only requires object labels for a relatively very small number of representative patches produced by K-Medoids clustering method to automatically determine the object classes and numbers involved in our target dataset. In detail, for each cluster center patch, we will seek an object label from a human labeler through a crowd-sourcing system and take it as the class label for the whole cluster of patches. However, due to the preferences of different labelers, the labels can be provided at different granularity levels, e.g., "kid" vs "sitting kid". Moreover, typos may exist in the given labels, e.g., "grooound" vs "ground". We thus apply some word processing technique [28] on the collected object labels. When the given label is a phrase, we will not process it as a new category if one of its component words is already a category keyword. Hence "sitting kid" will not be taken as a category if "kid" is already one. After object labels being purified, we merge the clusters with the same object labels and produce the final object classes and number for the given data. In our experiments, the numbers of object classes resulted range from 20 to 50, which fits into the principle of Zipf's Law and implies that a small proportion of object classes account for the majority of object occurrences.

3.2 Multi-level Adaptive Active Learning

Let $\mathbf{z}_i$ denote the mid-level feature vector for image $\mathcal{I}_i$, $Y = \{1 \dots K_y\}$ denote the scene class label space, $\mathcal{L} = \{(\mathbf{z}_1, y_1), \dots, (\mathbf{z}_N, y_N)\}$ denote the set of labeled instances, and $\mathcal{U}$ denote the large pool of unlabeled instances. After initializing our training model based on the small number of labeled instances, we perform multi-level active learning in an iterative fashion, which involves two types of iterations, *scene level iterations* and *object level iterations*. In a *scene level iteration*, it selects the most informative unlabeled instance to label at the scene class level, while in an *object level iteration*, it selects the most informative unlabeled instance to label at the object class level. An adaptive strategy is used to perform switch between these two types of iterations.

Scene level iteration. In such an iteration, we select the most informative unlabeled instance to label based on a well-motivated utility measure, named maximum conditional mutual information (MCMI), which maximizes the amount of information we gain from querying the selected instance:

$$\mathbf{z}^* = \arg\max_{\mathbf{z} \in \mathcal{U}} \left(H(\mathcal{L}) - H(\mathcal{L} \cup (\mathbf{z}, y)) \right) \tag{5}$$

where the data set entropy is defined as

$$H(\mathcal{L}) = -\sum_{i=1}^{|\mathcal{L}\cup\mathcal{U}|}\sum_{l=1}^{|Y|} P_{\mathcal{L}}(y_i = l|\mathbf{z}_i)\log P_{\mathcal{L}}(y_i = l|\mathbf{z}_i) \tag{6}$$

which measures the total entropy of all labeled and unlabeled instances. $P_{\mathcal{L}}(y|\mathbf{z})$ denotes the probability estimate produced by the classification model that is trained on the labeled data $\mathcal{L}$. Note the first entropy term $H(\mathcal{L})$ remains to be a constant for all candidate instances and can be dropped from the instance selection criterion, which leads to the selection criterion below:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}\in\mathcal{U}} H(\mathcal{L}\cup(\mathbf{z}, y)) \tag{7}$$

Though Equation (7) provides a principled instance selection criterion, it is impossible to compute given the true label y is unknown for the unlabeled query instance $\mathbf{z}$. We hence adopt the "optimistic" strategy proposed in [6] to pursue an alternative optimistic selection criterion below:

$$(\mathbf{z}^*, l^*) = \arg\min_{\mathbf{z}\in\mathcal{U}}\min_{l\in\mathcal{Y}} \ H(\mathcal{L}\cup(\mathbf{z}, l)) \tag{8}$$

which selects the candidate instance $\mathbf{z}^*$ and its a label option l^* that leads to the smallest total prediction uncertainty over all instances. Once the true label y^* of the select instance $\mathbf{z}^*$ being queried, we added ($\mathbf{z}^*$,y^*) into the labeled set $\mathcal{L}$ and retrain the scene classifier. This optimistic selection strategy however requires retraining the scene classifier for $O(|\mathcal{U}|\times|\mathcal{Y}|)$ times to make the instance selection decision: For each of the $|\mathcal{U}|$ unlabeled instances, one scene classifier needs to be trained for each of its $|\mathcal{Y}|$ candidate labels. The computational cost can be prohibitive on large datasets. To compensate this drawback, one standard way is to use random sub-sampling to select a subset of instances and label classes to reduce the candidate set in Equation (8).

Object level iteration. Querying labels at the object class level raises more questions. First, what, image vs patch, should be presented to the human labeler? What information should we query? A naive idea is to present a patch to the human labeler and query the object class label of the patch. However, it will be very difficult to select the right patch that contains a perceivable and discriminative object. Hence, instead of presenting patches to the annotators, we present a whole image to the labeler and ask whether the image contains a particular set of selected objects. Such specific questions will be easy to answer and will not lead to any ambiguities.

Next, we need to decide which image and what objects to query. We employ a most uncertainty strategy and select the most uncertain image (with the maximum entropy) to query under the current scene classification model:

$$\mathbf{z}^* = \arg\max_{\mathbf{z}\in\mathcal{U}} \ -\sum_{l=1}^{|Y|} P_{\mathcal{L}}(y = l|\mathbf{z})\log P_{\mathcal{L}}(y = l|\mathbf{z}) \tag{9}$$

For the selected image $\mathbf{z}^*$, we then select the top M most important objects regarding the most confident scene label $\hat{l}^*$ of $\mathbf{z}^*$ under the current scene classifier to query (We used $M = 5$ in our experiments later). Specifically, $\hat{l}^*$ will be determined as $\hat{l}^* = \arg\max_l P_{\mathcal{L}}(l|\mathbf{z}^*)$. Then we choose M objects that correspond to the largest M entries of the weight parameter vector $|\mathbf{w}_{\hat{l}^*}|$ under the current multi-class scene classifier. Our query questions submitted to the annotators will be in a very specific form: "Does object o_i appear in this image?" We will ask M such questions, one for each selected object.

The last challenge in the object level iteration is on updating the scene classification model after the selected object labels being queried. If the answer for a question is "No", we simply re-label all patches of the selected image as negative samples for that object class, and retrain the particular object classifier if needed. On the other hand, if the answer for a question is "Yes", it means at least one patch in this image should have a positive label for the particular object class. We hence assign the object label to the most confident patch within the selected image under the current particular object classifier. Then we will refine our previous unsupervised patch clustering results by taking the newly gathered patches into account. Our clustering refine scheme is very simple. Given the previous clustering result with K clusters, we set the new labeled patch as a new cluster center and perform K-Medoids updates with $K+1$ clusters. Note two of these $K+1$ clusters share the same object label and we will merge them after the end of the clustering process. Finally, all object classifiers will be updated based on the new clustering results. Consequently, the mid-level representations of each labeled image changes as well, and the scene classifier needs to be updated with the new mid-level features.

Adaptive active learning strategy. The last question one needs to answer to produce an active learning algorithm is how do we decide which type of iterations to pursue. We employ an adaptive strategy to make this decision: By default, we will perform active learning with scene level iterations, as most traditional active learners pursued. In each such iteration, an instance $\mathbf{z}^*$ and its optimistic l^* will be selected, and its true label y^* will be queried. However, once we found the true label y^* is different from the optimistic guess l^*, which means the strategy in the scene level iteration has been misled under the current scene classifier, we will then switch to the object level iteration in the next iteration to gather more information to strengthen the scene classification model from its foundation. We will switch back to the traditional scene label iteration after that. The overall multi-level adaptive active learning algorithm is summarized in *Algorithm 1*.

4 Experimental Results

We investigate the performance of the proposed active learning approach for scene classification on two standard challenging datasets, Natural Scene dataset and MIT Indoor Scene dataset. Natural scene dataset is a subset of the LabelMe dataset, which contains 8 scene categories (coast, forest, highway, inside city,

Algorithm 1 Multi-level Adaptive Active Learning

Input: Labeled set $\mathcal{L}$, unlabeled set $\mathcal{U}$, and record set $\mathcal{V} = \emptyset$;
  M: number of objects to query on each image,
  K: number of patch clusters.
Procedure:
Apply K-Medoids clustering on patches $\{\tilde{\mathbf{x}}_i \in \mathcal{L}\}$.
Query object labels for each cluster center patch.
Merge clusters with the same object labels.
Train object classifiers based on the clusters.
Obtain mid-level representation for each image $\mathbf{z} \in \mathcal{L} \cup \mathcal{U}$.
Train a scene classifier on $\mathcal{L}$.
Set itype = 1. %scene level=1, object level = 0
repeat
  if itype == 1 **then**
    Select $(\mathbf{z}^*, l^*)$ from the unlabeled set $\mathcal{U}$ based on Equation (8) and purchase its true label y^*.
    Drop $\mathbf{z}^*$ from $\mathcal{U}$ and add $(\mathbf{z}^*, y^*)$ into $\mathcal{L}$.
    Retrain the scene classifier on the updated $\mathcal{L}$.
    if $y^* \neq l^*$ **then**
      Set itype =0.
    end if
  else
    Select $\mathbf{z}^* \in \mathcal{U} \setminus \mathcal{V}$ according to Equation (9).
    Predict most confident scene label $\hat{l}^*$ for $\mathbf{z}^*$.
    Query the top M most important objects based on the absolute weight values $|\mathbf{w}_{\hat{l}^*}|$ for scene class $\hat{l}^*$.
    Update the clustering result if necessary.
    Update object classifiers.
    Add $\mathbf{z}^*$ into $\mathcal{V}$.
    Update the mid-level representation for all images.
    Update scene classifier on $\mathcal{L}$.
    Set itype =1.
  end if
until run out of money or achieve the aim

mountain, open country, street, and tall building) and each category has more than 250 images. We randomly selected 100 images from each category and pooled them together into a training set and used the rest as the test set. We further randomly selected 5 images per category (40 in total) as the initial labeled set. MIT indoor scene dataset contains 67 indoor categories and a total of $15,620$ images. The number of images varies across categories, but there are at least 100 images per category. We randomly selected 50 images per category to form the training set and the rest are used for testing. Within the training set, 2 images are randomly selected from each category as labeled instances and the rest images are pooled together as unlabeled instances.

The natural scene dataset has object level annotations available to use and the MIT indoor scene dataset also has object level annotations for a proportion of its images. We thus simulated the human annotators' answers based on these available object level annotations for our multi-level active learning. For the MIT indoor scene dataset, we further preprocessed it by discarding the categories that contain less than 50 annotated images (at the object level). After this preprocessing, only 15 categories were left. We produced all non-overlapping patches in size of 16×16 pixels that cover each image. We used the 128-dimension SIFT feature as the low-level features in our experiments.

In our experiments, we compared the proposed *Multi-Level Adaptive active learning (MLA)* method to three baselines: (1) *Single-Level Active learning (SLA)* method, which is a variant of MLA that only queries the scene labels; (2) *Single-Level Random sampling (SLR)* method, which randomly selects an image from the unlabeled pool in each iteration and queries its scene label; and (3) *Multi-Level Random sampling (MLR)* method, which randomly selects an image from the unlabeled pool in each iteration and then randomly chooses to query its object labels or scene label with equal probability. Moreover, we have also compared to the method, Active Learning with Object and Attribute annotations (*ALOA*), developed in [11]. This *ALOA* method is the state-of-the-art active learner that utilizes both attribute and image labels. We used $K = 200$ (for the K-Medoids clustering) and $M = 5$ for the proposed and the baseline methods. For the trade-off parameters C in Eq.(1) and Eq. (3), we set C as 10 for the object classifiers and 0.1 for the scene classifier, aiming to avoid overfitting for the scene classifier with limited labeled data at the scene level. Starting from the initial randomly selected labeled data, we ran each active learning method for 100 iterations, and recorded their performance in each iteration. We repeated each experiment 5 times and reported the average results and standard deviations.

Figure 3 presents the comparison results in terms of scene classification accuracy on the MIT Indoor scene dataset and the Natural scene dataset. For the proposed approach *MLA* and the baselines *SLA, MLR, SLR*, we experimented two different ways of learning scene classifiers. [1] A straightforward way is to learn the scene classifier based on the mid-level semantic representation produced by the object classifiers. Alternatively, we have also investigated learning the scene classifier based on *hybrid* features by augmenting the mid-level representation with the low-level SIFT features. Such a mechanism was shown to be effective in [26]. Specifically, we built a 500-words codebook with K-Means clustering over the SIFT features and represented each image as a 500-long vector with vector quantization. This low-level representation together with the mid-level representation form the hybrid features of images for scene classification. The comparison results based only on mid-level representation are reported on the left column of Figure 3 for the two datasets respectively; and the comparison results based on the hybrid features are reported on the right column of Figure 3. We can see in terms of scene classification accuracy, our proposed method *MLA* beats all other comparison methods, especially the baselines, across most of the comparison range, except at the very beginning. At the beginning of the active learning process, *ALOA* produces the best performance with very few labeled images. Given

[1] The *ALOA* from [11] works in a different mechanism with a latent SVM classifier.

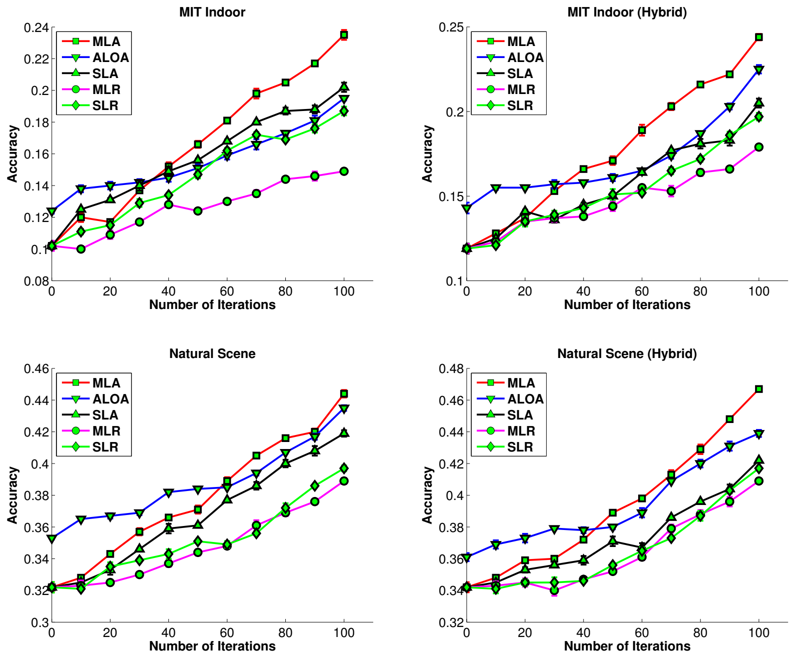


Fig. 3: The average and standard deviation results in terms of scene classification accuracy on both MIT Indoor scene dataset and Natural Scene dataset.

that *ALOA* [11] uses the state-of-the-art latent SVM classifier, and our approach uses a simple logistic regression model, this seems reasonable. But the gap between *ALOA* and the proposed *MLA* quickly degrades with the active learning process; after a set of iterations, *MLA* significantly outperforms *ALOA*. This demonstrates that our proposed multi-level adaptive active learning strategy is much more effective and it is able to collect most useful label information that makes a simple logistic regression classifier to outperform the state-of-the-art latent SVM classifier. Among the three baseline methods, *SLA* always performs the best. On MIT-Indoor dataset, it even outperforms *ALOA* when only semantic representation is used. This suggests the MCMI instance selection strategy we employed in the scene level iterations is very effective. On the other hand, the random sampling methods *MLR* and *SLR* produce very poor performance. Another interesting observation is that at the start of active learning, though we only have very few labeled instance available for each category, the accuracy of our latent object-based hierarchical scene classification model already reaches around 12% on 15-category MIT indoor scene subset and reaches around 34% on Natural scene dataset. This demonstrates the mid-level representation is very descriptive and useful for abstractive scene classification. By comparing the two versions of results across columns, we can see that with hybrid features, the proposed *MLA* produces slightly better results, which suggests that low-level features and mid-level representation features can complement each other.

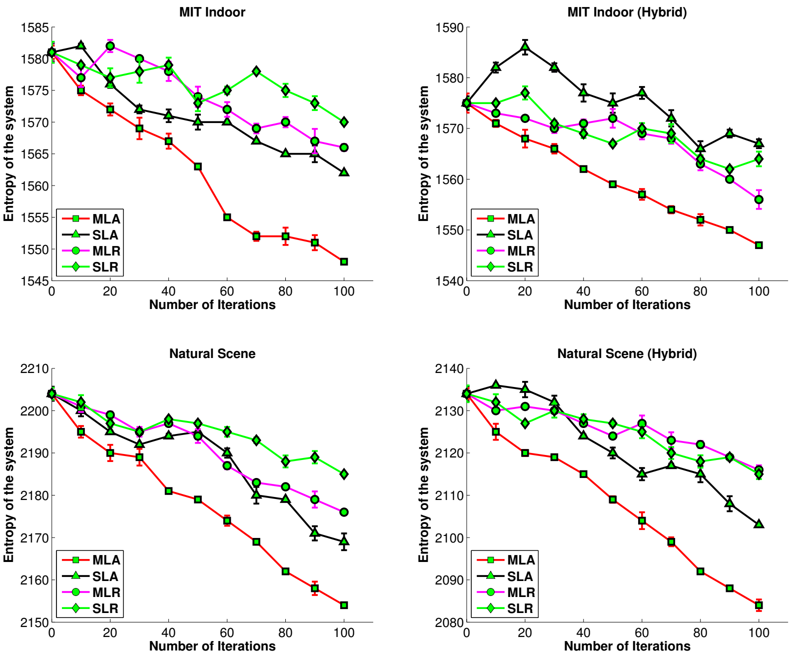


Fig. 4: The entropy reduction results on both MIT Indoor Scene dataset and Natural Scene dataset.

In addition to scene classification accuracy, we have also measured the performance of the comparison methods in terms of system entropy (i.e., data set entropy). We recorded the reduction of the system entropy with the increasing number of labeled instances. The *ALOA* method from [11] uses a Latent SVM model, the system entropy of which is contributed by both the image classifier and the model's inner attribute classifiers. However, the entropies of all other methods are only associated with the target image label predictions, which makes the computed entropy of *ALOA* and others not comparable. Therefore, we only consider the other four methods in this experimental setting. The results are reported in Figure 4. It is easy to see that the proposed *MLA* method reduces the entropy much quickly than other baselines, which verifies the effectiveness of our proposed adaptive active learning strategy. The curve of *MLA* is monotone decreasing, indicating that every query is helpful in terms of entropy reduction. The curves of the other baselines nevertheless have fluctuations. Among them, *SLA* is almost always the runner-up except on the MIT indoor dataset with hybrid features. By comparing the two versions of results across columns, we can see the system entropy with hybrid features is relatively lower than its counterpart with mid-level semantic representation alone, which again suggests that the low-level features can provide augmenting information for the mid-level semantic representations.

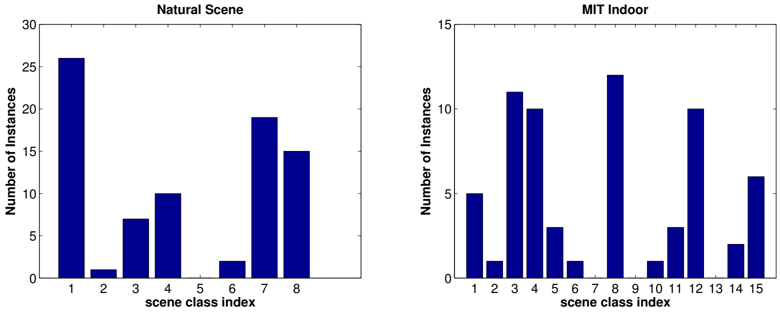


Fig. 5: Distribution of queried instances in scene label space for the proposed approach on MIT Indoor and Natural Scene datasets.

Finally, we collected the number of queries in each scene category on the two datasets for the proposed approach and presented the results in Figure 5. We can see, obviously the instances are not selected according to a uniform distribution across categories. The total numbers of scene level label queries among the 100 iterations are 65 and 80 on the MIT Indoor and Natural scene datasets respectively. The remaining querying effort is on the object-level annotations. On the MIT indoor dataset, the ratio between the numbers of queries on scene labels and object annotations is about $2:1$. In contrast, this ratio is $4:1$ on the Natural scene dataset. This observation indicates that our model can adaptively switch query levels based on the complexity of the data. When the object layout is easy, it will put more effort on querying scene labels; when the scene becomes complicated and ambiguous, it will ask more questions about object annotations.

5 Conclusions

In this paper, we developed a novel multi-level active learning approach to reduce the human annotation effort for training semantic scene classification models. Our idea was motivated by the facts that latent object-based semantic representations of images are very useful for scene classification, and the scene labels are difficult to distinguish from each other in many scenarios. We hence built a semantic framework that learns scene classifiers based on latent object-based semantic representations of images, and then proposed to perform active learning with two different types of iterations, the scene level iteration (abstractive high level) and the latent object level iteration (semantic middle level). We employed an adaptive strategy to automatically perform switching between these two types active learning iterations. We conducted experiments on two standard scene classification datasets, the MIT Indoor scene dataset and the Natural Scene dataset, to investigate the efficacy of the proposed approach. Our empirical results showed the proposed adaptive multi-level active learning approach can outperform both traditional baseline single level active learning methods and the state-of-the-art multi-level active learning method.

References

1. Biswas, A., Parikh, D.: Simultaneous active learning of classifiers & attributes via relative feedback. In: Proceedings of CVPR (2013)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of CVPR (2005)
3. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Scene parsing with multiscale feature learning, purity trees,and optimal covers. CoRR abs/1202.2160 (2012)
4. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of CVPR (2005)
5. Gould, S., Gao, T., Koller, D.: Region-based segmentation and object detection. In: Proceedings of NIPS (2009)
6. Guo, Y., Greiner, R.: Optimistic active learning using mutual information. In: Proceedings of IJCAI (2007)
7. Jain, P., Kapoor, A.: Active learning for large multi-class problems. In: Proceedings of CVPR (2009)
8. Joshi, A., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: Proceedings of CVPR (2009)
9. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with gaussian processes for object categorization. In: Proceedings of ICCV (2007)
10. Kapoor, A., Hua, G., Akbarzadeh, A., Baker, S.: Which faces to tag: Adding prior constraints into active learning. In: Proceedings of ICCV (2009)
11. Kovashka, A., Vijayanarasimhan, S., Grauman, K.: Actively selecting annotations among objects and attributes. In: Proceedings of ICCV (2011)
12. Kumar, M., Koller, D.: Efficiently selecting regions for scene understanding. In: Proceedings of CVPR (2010)
13. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of CVPR (2006)
14. Li, L., Su, H., Xing, E., Fei-Fei, L.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Proceedings of NIPS (2010)
15. Li, X., Guo, Y.: Adaptive active learning for image classification. In: Proceedings of CVPR (2013)
16. Lin, C., Weng, R., Keerthi, S.: Trust region newton method for logistic regression. J. Mach. Learn. Res. 9 (Jun 2008)
17. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2) (Nov 2004)
18. Mensink, T., Verbeek, J., Csurka, G.: Learning structured prediction models for interactive image labeling. In: Proceedings of CVPR (2011)
19. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: Proceedings of ICCV (2011)
20. Parizi, S., Oberlin, J., Felzenszwalb, P.: Reconfigurable models for scene recognition. In: Proceedings of CVPR (2012)
21. Parkash, A., Parikh, D.: Attributes for classifier feedback. In: Proceedings of ECCV (2012)
22. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proceeding of CVPR (2012)
23. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Proceedings of CVPR (2009)

24. Sadeghi, F., Tappen, M.: Latent pyramidal regions for recognizing scenes. In: Proceedings of ECCV (2012)
25. Settles, B.: Active Learning. Synthesis digital library of engineering and computer science, Morgan & Claypool (2011)
26. Sharmanska, V., Quadrianto, N., Lampert, C.: Augmented attribute representations. In: Proceedings of ECCV (2012)
27. Siddiquie, B., Gupta, A.: Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In: Proceedings of CVPR (2010)
28. Sparck Jones, K., Willett, P. (eds.): Readings in Information Retrieval. Morgan Kaufmann Publishers Inc. (1997)
29. Vezhnevets, A., Buhmann, J., Ferrari, V.: Active learning for semantic segmentation with expected change. In: Proceedings of CVPR (2012)
30. Vijayanarasimhan, S., Grauman, K.: Multi-level active prediction of useful image annotations for recognition. In: Proceedings of NIPS (2008)
31. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: Training object detectors with crawled data and crowds. In: Proceedings of CVPR (2011)
32. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: Proceedings of ECCV (2010)
33. Wu, J., Rehg, J.: CENTRIST: A Visual Descriptor for Scene Categorization. IEEE Transactions on PAMI 33 (2011)
34. Yan, R., Yang, L., Hauptmann, A.: Automatically labeling video data using multiclass active learning. In: Proceedings of ICCV (2003)