Online Adaptation for Joint Scene and Object Classification

Jawadul H. Bappy, Sujoy Paul and Amit K. Roy-Chowdhury

Dept. of ECE, University of California, Riverside, CA 92521
{mbappy,supaul,amitrc}@ece.ucr.edu

Abstract. Recent efforts in computer vision consider joint scene and object classification by exploiting mutual relationships (often termed as context) between them to achieve higher accuracy. On the other hand, there is also a lot of interest in online adaptation of recognition models as new data becomes available. In this paper, we address the problem of how models for joint scene and object classification can be learned online. A major motivation for this approach is to exploit the hierarchical relationships between scenes and objects, represented as a graphical model, in an active learning framework. To select the samples on the graph, which need to be labeled by a human, we use an information theoretic approach that reduces the joint entropy of scene and object variables. This leads to a significant reduction in the amount of manual labeling effort for similar or better performance when compared with a model trained with the full dataset. This is demonstrated through rigorous experimentation on three datasets.

Keywords: Scene Classification, Object detection, Active learning

1 Introduction

Scene classification and object detection are two challenging problems in computer vision due to high intra-class variance, illumination changes, background clutter and occlusion. Most existing methods assume that data will be labeled and available beforehand in order to train the classification models. It becomes infeasible and unrealistic to know all the labels beforehand with the huge corpus of visual data being generated on a daily basis. Moreover, adaptability of the models to the incoming data is crucial too for long-term performance guarantees. Currently, the big datasets (e.g. ImageNet [1], SUN [2]) are prepared with intensive human labeling, which is difficult to scale up as more and more new images are generated. So, we want to pose a question, '*Are all the samples equally important to manually label and learn a model?*'. We address this question in the context of joint scene and object classification.

Active learning [3] has been widely used to choose a subset of most informative samples that can achieve similar or better performance than all the data being manually labeled. In order to identify the informative samples, most active learning techniques choose the samples about which the classifier is most uncertain.

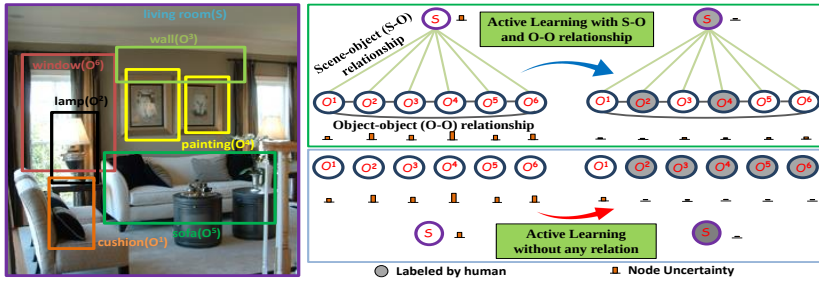


Fig. 1. This figure presents the motivation of incorporating relationship among scene and object samples within an image. Here, scene (S) and objects ($O^1, O^2, \ldots, O^6$) are predicted by our initial classifier and detectors with some uncertainty. We formulate a graph exploiting scene-object (S-O) and object-object (O-O) relationships. As shown in the figure, even though $\{S, O^2, O^3, O^4, O^5, O^6\}$ nodes have high uncertainty, manually labeling only 3 of them is good enough to reduce the uncertainty of all the nodes if S-O and O-O relationships are considered. So, the manual labeling cost can be significantly reduced by our proposed approach.

Expected change in gradients [3], information gain [4], expected prediction loss [5] are some approaches used in the literature to obtain the samples for query. These approaches consider the individual samples to be independent. However, there are various tasks, such as document classification [6] and activity recognition [7], where interrelationships between samples exist. In such cases, it will be advantageous to exploit these relationships to reduce the number of samples to be manually labeled. Some active learning frameworks consider this idea and exploit different contextual relations such as link information [8], social relationships [9], spatial information [10], feature similarity [11], spatio-temporal relationships [12].

We leverage upon active learning for identifying the samples to label in the problem of joint scene and object recognition. Similar to the applications mentioned above, exploiting mutual relationships between scene and objects can yield better performance [13] than if no relationships are considered. For example, it is unlikely to find a 'cow' in a 'bedroom', but, the probability of finding 'bed' and 'lamp' in the same scene may be high. Thus gaining information about a scene can help in enhanced prediction on objects and vice versa. Previously, research in [14–17] has shown how to exploit the scene-object relationships to yield better classification performance. However, these methods require data to be manually labeled and available before learning. Although there exist some works involving active learning in scene and object classification [4, 5, 18], they do not exploit the scene-object(S-O) and object-object(O-O) inter-relationships. This is critical because of the hierarchical nature of the relationships between objects and scenes. This relationship can be represented as a graphical model with the samples on the graph, which need to be labeled by a human, chosen using a suitable criterion. The labeling effort can be significantly reduced in this process - labeling a scene node in the graph can possibly resolve ambiguities for multiple object classes. This motivation is portrayed in Fig. 1.

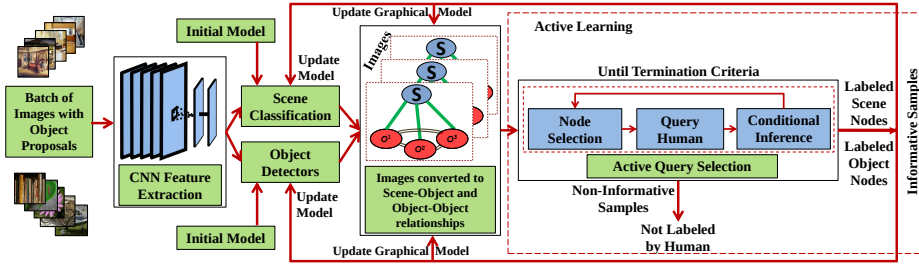


Fig. 2. This figure presents a pictorial representation of the proposed framework. At first, initial classification models and relationship model are learned from a small set of labeled images. Thereafter, as images are available in batches, scene & object classification models provide prediction scores of scene and objects. With these scores and the relationship model, the images are represented as graphs with scene and object nodes. Then, the active learning module is invoked which efficiently chooses the most informative scene or object nodes to query the human. Finally, the labels provided by the human are used to update the classification & relationship models.

Motivated by the above, we propose a novel active learning framework which exploits the S-O and O-O relationships to jointly learn scene and object classification models. Using mutual relationships between scene and objects, we can leverage upon the fact that manual labeling of one reduces the uncertainty of the other, and thus reduces labeling cost. This is achieved using an information theoretic approach that reduces the joint entropy of a graph. As presented in the figure, exploiting relationships between scene and objects can lead to lesser human labeling effort, compared to when relationships are not considered.

Framework Overview. The flow of the proposed algorithm is presented in Fig. 2. We perform two tasks simultaneously:

1. Selection of an image that contains the most informative samples (scene,objects)
2. Given an image, a sample (i.e., a node in the graph representing that image) is chosen in a way that reduces the uncertainty on other samples.

Our framework is divided into two phases. At first, we learn the initial classification models as well as the S-O and O-O relationship model with small amount of labeled data. In the second phase, with incoming unlabeled data, we first classify the unlabeled scene and object samples using the current models. Then, we represent each incoming image as a graph, where scene classification probabilities and object detection scores are utilized to represent the scene and object node potentials. S-O and O-O relations delineate the edge potentials. We compute the marginal probabilities of node variables from the inference on the graphs.

Thereafter, we formulate an information-theoretic approach for selecting the most informative samples. Joint entropy of a graph is computed from the joint distribution of scene and objects that represents the total uncertainty of an image. For a batch of data, our framework chooses the most informative samples based on some uncertainty measures (discussed in Sec. 3) that lead to the maximum decrease in the joint entropy of the graph after labeling. After receiving the label of a node from the human, we infer on the graph conditioned upon the known

label. Due to this inference, the other unlabeled nodes gain information from the node labeled by human, which leads to a significant reduction in uncertainties of other nodes. The labels obtained in this process are used to update the scene and object classification models as well as the S-O and O-O relationships.

Main Contributions. Our main contributions are as follows.

• In computer vision, most of the existing active learning methods involve learning a classification model of one type of variable, e.g., scene, objects, activity, text, etc. On the other hand, the proposed active learning framework learns scene and object classification models *simultaneously*.

• In the proposed active learning framework, both scene and object classification models take advantage of the interdependence between them in order to select the most informative samples with the least manual labeling cost. To the best of our knowledge, any previous work using *active learning to classify scene and objects together* is unknown.

• Leveraging upon the inter-relationships between scene and objects, we propose a new information-theoretic sample selection strategy along with inference on a graph based on the intuition that learning a sample reduces the uncertainties of other samples. Moreover, our framework facilitates continuous and incremental learning of the classification models as well as the S-O and O-O relationship models, thus dynamically adapting to the changes in incoming data.

1.1 Related Works

Scene and Object Recognition. Many of the scene classification methods use low dimensional features such as color and texture [19], GIST [20], SIFT descriptor [21] and deep feature [22]. In object detection, current state-of-the-art methods are R-CNN [23], SPP-net [24] and fast R-CNN [25]. Another promising approach in recognition tasks has been to exploit the relationships between objects in a scene using a graphical model [26], [27], [13]. A Conditional Random Field (CRF) for integrating the scene and object classification for video sequences was proposed in [14]. A model for joint image segmentation, object and scene class inference was proposed in [13]. In [15], the spatial relationships between the objects within an image were exploited to compute the scene similarity score, based on which the indoor scene categories were predicted. In [16], a CRF model was constructed based on scene, object and the textual data associated with the images on the web, to label the scenes and localize objects within the image. In [28], a projection was formulated from images to a space spanned by object banks, based on which, the image was classified into different categories. In [17], a framework was developed for multiple object classification within an image, where a conditional tree model was learned based on the co-occurrences of objects.

Active Learning. Although the above mentioned works exploit the contextual relationships, they assume that all the data are labeled and available beforehand, which is not feasible and involves huge labeling cost. Active learning has been widely used to reduce the effort of manual labeling in different computer vision tasks including scene classification [4], video segmentation [29], object detection

[30], activity recognition [12], tracking [31]. A generalized active learning framework for computer vision problems such as person detection, face recognition and scene classification was proposed in [32]. They used the two concepts of uncertainty and sample diversity to choose the samples for manual labeling. Some of the common techniques to measure uncertainty for selecting the informative data points are presented in [33]. Active learning has been separately used for scene or object classification [4, 18, 30, 34], but not in their joint classification.

In [18], a framework for actively learning scene classification model was proposed, where the authors incorporated two strategies - Best vs. Second Best (BvSB) and K-centroid to select the informative subset of images. A framework based on information density measure and uncertainty measure to obtain the best subset of images for querying the human was proposed in [5]. Although their algorithm can be applied separately for both scene and object classification, they do not exploit the relationships between scene and objects. An active learning framework for object categories was proposed in [35] which considers the case where the labeler itself is uncertain about labeling an image.

In [4], the authors present an active learning framework for scene classification. In their hierarchal model, they focus on querying at the scene level, and whenever unexpected class labels are returned by the human, queries are made at the object level. Thus in their method, there exists a flow of information from the object level to the scene level. However, in our method, there is a flow of information from scene to object level and vice versa, in a collaborative manner, which paves the path for a joint scene-object classification framework.

2 Joint Scene and Object Model

In this section, we discuss how we represent an image in a graphical model with scene and object as hidden variables.

***A.* Scene Classification Method.** In order to represent scenes, we extract features using Convolution Neural Networks (CNN). Given an image, we get a feature vector f from the $fc7$ layer of a CNN architecture, where $f \in \Re^{4096\times 1}$. We train a linear multi-class Support Vector Machine (SVM) [36] to compute the probability of n^{th} class, $p(S = s_n|f^j)$, where f^j implies the feature vector corresponding to sample j. We denote the learned model for scene classification as $\mathcal{P}_s$. Given an image, $\Phi_S \in \Re^N$ represents the classification score. N is the total number of scene categories considered in the experiment.

***B.* Object Detection Method.** We use R-CNN presented in [23] to detect the objects in an image. In R-CNN, we extract features from deep network for each object proposal. Then, we train a binary SVM classifier for each object category to get the probability of appearance of an object. After classifying the region we form a vector that represents the confidence scores of the binary classifiers for each category. Thus, for each p^{th} region we get Φ_{O^p} that represents the detection score vector. Finally, we use bounding box regression method [37] for better object localization. We denote the learned model for scene classification as $\mathcal{P}_o$.

***C.* Graphical Model Representation.** In this model, two levels of nodes are

used - one represents scene υ_s and other set of nodes implies detected objects υ_o. υ_o is generally represented by $\upsilon_o = \{\upsilon_{o^1}, \upsilon_{o^2}, ..\upsilon_{o^D}\}$, where D is the number of bounding boxes appearing in an image. The link between them is depicted by edges. The joint distribution of υ_s and υ_o over the CRF can be written as

$$P(\upsilon_s, \upsilon_o) = \frac{1}{Z}\, \Psi_\xi(\upsilon_s, \upsilon_o) \prod_{\substack{i,j \in D \\ i \neq j}} \Psi_\xi(\upsilon_{o^i}, \upsilon_{o^j}) \prod_{w \in \{\upsilon_s, \upsilon_o\}} \Psi_v(w) \tag{1}$$

where, Z is normalizing constant. $\Psi_v(.)$ and $\Psi_\xi(.)$ denote node and edge potentials.

Node Potentials. Given an image, the scene classifier ($\mathcal{P}_s$) produces a vector that contains the probabilities of all the scene labels. From these probabilities we compute scene node potential $\Psi_v(\upsilon_s)$ as presented in Eqn. 2. Similarly, given an image, the object detection scores are used to model the object node potentials $\Psi_v(\upsilon_o)$ as shown in Eqn. 3.

$$\Psi_v(\upsilon_s) = \sum_{n \in N} \mathcal{I}(S_n) \beta_n^T\, \Phi_S \tag{2}$$

$$\Psi_v(\upsilon_o) = \sum_{p \in D} \sum_{m \in M} \mathcal{I}(O_m^p) \Omega_m^T\, \Phi_{O^p} \tag{3}$$

Here, Φ_S is a vector of the probability of the scene labels obtained from multi-class SVM classifier. β_n is the feature weight vector corresponding to scene label S_n and $\mathcal{I}(.)$ is the indicator function, i.e., $\mathcal{I}(S_n) = 1$ when $S = S_n$, otherwise 0. Ω_m is the weight corresponding to the detection score of the object O_m. Φ_{O^p} is the score vector of detecting all the objects in the p^{th} bounding box. M is the number of object Classes.

Edge Potentials. We use two type of relationships, S-O and O-O. We use co-occurrence frequencies to represent edge potential. The probability of the presence of an object in a particular scene is determined by the co-occurrence statistics. For instance, in a context of *'highway'* scene, the probability of appearance of *'car'* will be higher than *'table'* or *'chair'*. In Eqn.4, $\Psi_\xi(\upsilon_s, \upsilon_o)$ represents the relationship between S and O. Similarly, $\Psi_\xi(\upsilon_{o^i}, \upsilon_{o^j})$ models the O-O relations.

$$\Psi_\xi(\upsilon_s, \upsilon_o) = \sum_{p \in D} \sum_{n \in N} \sum_{m \in M} \mathcal{I}(S_n) \mathcal{I}(O_m^p) \Phi_\xi(S_n, O_m) \tag{4}$$

$$\Psi_\xi(\upsilon_{o^i}, \upsilon_{o^j}) = \sum_{m' \in M} \sum_{m \in M} \mathcal{I}(O_{m'}^i) \mathcal{I}(O_m^j)\, \Phi_\xi(O_{m'}, O_m) \tag{5}$$

$\Phi_\xi(S_n, O_m)$ represents the co-occurrence statistics between scene and objects. Larger value implies higher probability of co-occurrence of S_n and O_m. Here, $\Phi_\xi(O^i, O^j)$ is the co-occurrence [38] between the detected objects O^i and O^j. It encodes the information about how often two objects can co-occur in a scene.

Parameter Learning. The initial model parameters of the CRF model are learned from a set of annotated images, object detectors and scene classifier. Given the ground truth object bounding boxes, we use object detectors to obtain detection scores for the corresponding bounding box region. Similarly,

we get the classification score from the annotated scene label. Thus, we can easily apply maximum likelihood estimation approach to learn all the parameters $\{\beta, \Omega, \Phi_\xi(S_n, O_m), \Phi_\xi(O_{m'}, O_m)\}$ in the model.

Inference of Scene and Object Labels. To compute the marginal distributions of the node and edge, we use Loopy Belief Propagation (LBP) algorithm [39], as our graph contains cycles. LBP is not guaranteed to converge to the true marginal, but has good approximation of the marginal distributions.

3 Active Learning Framework

In the previous section, we represent an image as a graph containing υ_s and υ_o nodes. If we select a node from a graph, such that querying it will minimize the joint entropy of the graph maximally, then it means that the classifier will be able to gain maximum amount of information by labeling that node.

Formulation of Joint Entropy. Consider a fully connected graph $G = (V, E)$, where V and E are the set of nodes and edges respectively. It may be noted that $V = \{S, O^1, O^2, \ldots, O^D\}$. Let $\mu_i(\upsilon_i)$ and $\mu_{ij}(\upsilon_i, \upsilon_j)$ be the marginal probabilities of the node and edge of the graph. Let υ_i and υ_j represent the random variables for nodes $i, j \in V$. In our joint scene and object classification, $i \in \{S, O^1, O^2, \ldots, O^D\}$ as discussed in Sec. 2. The node entropy $H(\upsilon_i)$ and mutual information $I(\upsilon_i, \upsilon_j)$ between a pair of nodes are defined as,

$$H(\upsilon_i) = \mathbb{E}[-\log_2 \mu_i(\upsilon_i)] \qquad I(\upsilon_i, \upsilon_j) = \mathbb{E}[\log_2 \frac{\mu_{ij}(\upsilon_i, \upsilon_j)}{\mu_i(\upsilon_i)\mu_t(\upsilon_j)}] \tag{6}$$

Considering Q nodes in the graph, its joint entropy can be expressed as,

$$\begin{aligned} H(V) &= H(\upsilon_1) + \sum_{i=2}^{Q} H(\upsilon_i|\upsilon_1, \ldots, \upsilon_{i-1}) \\ &= H(\upsilon_1) + \sum_{i=2}^{Q} \Big[H(\upsilon_i) - I(\upsilon_1, \ldots, \upsilon_{i-1}; \upsilon_i)\Big] \end{aligned} \tag{7}$$

using $I(\upsilon_1, \ldots, \upsilon_{i-1}; \upsilon_i) = H(\upsilon_i) - H(\upsilon_i|\upsilon_1, \ldots, \upsilon_{i-1})$. Again, using the chain rule, $I(\upsilon_1, \ldots, \upsilon_{i-1}; \upsilon_i) = \sum_{j=1}^{i-1} I(\upsilon_j; \upsilon_i|\upsilon_1, \ldots, \upsilon_{j-1})$, Eqn. 7 becomes

$$H(V) = \sum_{i=1}^{Q} H(\upsilon_i) - \sum_{i=2}^{Q} \sum_{j=1}^{i-1} I(\upsilon_j; \upsilon_i|\upsilon_1, \ldots, \upsilon_{j-1}) \tag{8}$$

It becomes computationally expensive to compute the conditional mutual information, as the number of node increases [40]. As we consider only pair-wise interactions between S-O and O-O, we approximate the conditional mutual information $I(\upsilon_j; \upsilon_i|\upsilon_1, \ldots, \upsilon_{j-1}) \approx I(\upsilon_j; \upsilon_i)$. Thus, the joint entropy of the graph can be approximated as,

$$H(V) \approx \sum_{i=1}^{Q} H(\upsilon_i) - \sum_{i=2}^{Q} \sum_{j=1}^{i-1} I(\upsilon_j; \upsilon_i) = \sum_{i \in V} H(\upsilon_i) - \sum_{(i,j) \in E} I(\upsilon_i; \upsilon_j) \tag{9}$$

This expression is actually exact for a tree, but approximate for a graph having cycles. The approximation leads to the expression of joint entropy in Eqn. 9, which is similar to the joint entropy expression in Bethe method [40].

Informative node selection. In our problem, an image is represented by a graph having several nodes with two types of hidden variables υ_s and υ_o. So, we require not only to find the most informative image but also need to choose the node to be manually labeled. If we manually label a node, then we assume that there is no uncertainty involved in that node. Thus, after labeling a node υ_i with the label l, the node entropy becomes zero, i,e. $H(\upsilon_i = l) = 0$.

Let $H^p(V)$ be the the joint entropy of image p which can be computed using Eqn. 9. We query the node such that $H^p(V)$ is maximally reduced after labeling the node and inferring the graph conditioned on the new label. Then, after labeling υ_i, we find the optimal node q of image p to be queried as [1],

$$q^* = \arg\max_q \ \left[H^p(\upsilon_q) - \frac{1}{2}\sum_{j\in\mathcal{N}(q)} I^p(\upsilon_q, \upsilon_j)\right] \tag{10}$$

where $\mathcal{N}(q)$ represents the neighbor nodes of q. For simplicity, let us define the uncertainty associated with node q of image p as $J_q^p = H^p(\upsilon_q) - \frac{1}{2}\sum_{j\in\mathcal{N}(q)} I^p(\upsilon_q, \upsilon_j)$ where the joint entropy for an image p is $H^p(V) = \sum_{q=1}^{n} J_q^p$ from Eqns. 9 and 10. From Eqn. 10, we choose the node to query, which has the maximum uncertainty considering not only the node entropy but also the mutual information between the nodes. Next, we explain how to choose a set of nodes from a batch of images.

Simultaneous Image and Node Selection. We query the nodes of image p only if its joint entropy $H^p(V) \geq \delta$, where δ is a threshold. Since we have the information about all the node uncertainties of all images, we can perform multiple queries across multiple different images such that the learner can learn faster and more efficiently. In this paper, we consider that there is no relation between the images, thus the conditional inference on one image is independent of the other images. Thus, graphs of different images can be conditionally inferred in a parallel manner.

Let, a vector, $J^p = [J_1^p, J_2^p, \ldots, J_Q^p]^T$ contain the uncertainty associated with Q (dependent on the image) nodes for an image p. Consider another vector, $\hat{J} = [J^1\ J^2 \ldots J^P]^T$ which is obtained after concatenating all the vectors J^p for P images, whose joint entropy is higher than threshold δ. We sort the vector $\hat{J}$ in descending order to obtain a new vector $\hat{J}_s$. Then, we perform multiple queries based on $\hat{J}_s$, which contain uncertainty of nodes from multiple images of a batch. For each image, we choose the node appearing first in $\hat{J}_s$ for labeling. We perform conditional inference with the new labels in a parallel manner over all the images. The $\hat{J}_s$ vector is again obtained using the updated uncertainties of the nodes and the process is repeated until $H^p(V) \leq \delta, \forall p$. It may be noted that P decreases or at least remains same in succeeding iterations, because nodes belonging to images attaining joint entropy less than δ are not queried and thus not included in $\hat{J}_s$. Inference reduces the uncertainty on other nodes of the same image.

[1] See derivation in supplementary

As uncertainty of nodes decreases, joint entropy is also reduced. Consider a matrix S having dimension $N_n \times 2$, where N_n is the total number of nodes of all images in the batch. The first and second columns of S contain the node index of a graph (image) and the image index respectively. The order in which the elements of S are populated is the same as that of $\hat{J}_s$. We refrain from choosing more than one node per image in each iteration because labeling one node can help the other nodes attain a better decision after inference. The set of nodes $\mathcal{M}$, chosen for labeling in each iteration can be expressed as,

$$\mathcal{M}^* = \underset{\substack{\mathcal{M} \\ s.t. |\mathcal{M}|=P \\ S^{i,2} \neq S^{j,2}, i,j \in \mathcal{M}}}{\arg\max} \sum_{k \in \mathcal{M}} \left[\hat{J}_s\right]_k \tag{11}$$

where $\left[\hat{J}_s\right]_k$ denote the k^{th} element of $\hat{J}_s$ and $S^{i,m}$ denote the i^{th} row and m^{th} column of S, where $m \in \{1, 2\}$. All the steps of active learning are shown in Algorithm 1. The first column of S is used to identify which node of an image should be labeled. To summarize Eqn. 11, the optimal set $\mathcal{M}$ can be obtained by choosing one node which has the highest entropy from each image.

Classifier Update. To classify scene and objects, we use a linear support vector machine (SVM) classifier. The probability of predicted label can be defined as $\hat{y} = w^T f(x) + b$, where $f(x)$ is the feature of scene or objects and w, b are parameters that determine the hyperplane between two classes. We use soft margins formulation presented in [36] to find the solution of w, b. The solution can be found by optimizing, $\frac{1}{2}w^2 + C\sum_1^n \epsilon_i$ subject to $y_i(w^T f(x_i) + b) \geq (1 - \epsilon_i)$ and $\epsilon_i \geq 0$ for all i samples, where ϵ_i is the slack variable.

Edge Weight Update. We update the co-occurrence statistics with new manually labeled data as presented in Eqns. 4 and 5. lets denote them by $\Phi'_\xi(S_n, O_m)$ and $\Phi'_\xi(O_{m'}, O_m)$. The updated co-occurrence matrix will be $[\Phi_\xi(S_n, O_m)]_{t+1} \leftarrow [\Phi_\xi(S_n, O_m)]_t + \Phi'_\xi(S_n, O_m)$ and $[\Phi_\xi(O_{m'}, O_m)]_{t+1} \leftarrow [\Phi_\xi(O_{m'}, O_m)]_t + \Phi'_\xi(O_{m'}, O_m)$, where the subscript $t+1$ indicates the edge potentials after t updates.

4 Experiments

In this section, we provide experimental analysis of our active learning framework for joint scene and object recognition models on three challenging datasets. For convenience, we will use terms 'inter-relationship' and 'contextual relationship' to denote scene-object and object-object relationship.

Datasets. In our experiments, we use SUN [41], MIT-67 Indoor [42] and MSRC [43] datasets in order to analyze scene classification and object recognition performance and compare our results. These datasets are appropriate as they provide rich source of contextual information between scene and objects. In SUN dataset, we choose 125 scene classes and 80 object categories to evaluate scene classification and object detection performance, as those contain annotation for both scene and objects. MIT-67 indoor [42] dataset consists of 67 indoor scene categories with large varieties of object categories. For MSRC [43] dataset, we

Algorithm 1: Online Learning for Scene and Object Sample Selection

INPUTS. 1. Learned scene, object and relation models after processing images in Batch$_{K-1}$: $\{\mathcal{P}_s, \mathcal{P}_o, \Phi_\xi(S_n, O_m)$ & $\Phi_\xi(O_{m'}, O_m)\}$
2. Unlabeled Batch$_K$: $\mathcal{U}$
OUTPUTS. Learned Models after processing images in Batch$_K$: $\{\mathcal{P}_s, \mathcal{P}_o, \Phi'_\xi(S_n, O_m)$ & $\Phi'_\xi(O_{m'}, O_m)$
Initialize: $L_s = \{\}$ (Empty set)
Step 1: Compute $H(v_i)$ and $I(v_i, v_j)$ using Eqn. 6
Step 2a: Compute vector $J^p = [J_1^p, J_2^p, \ldots J_Q^p]$ containing the node uncertainties involving entropy and mutual information, for all images.
Step 2b: Obtain vector $\hat{J}$ by concatenating the vectors J^p, $\forall p$, s.t. $H^p(V) \geq \delta$
Step 2c: $\hat{J}_s \leftarrow sort(\hat{J})$ in descending order
Step 2d: Obtain a vector S storing the image id in the sequence as in $\hat{J}_s$
if $length(\hat{J}_s) \neq 0$ **then**
  Step 3a: Select nodes for manual labeling to form a set $\mathcal{M}$ using Eqn. 11
  Step 3b: Query the nodes in $\mathcal{M}$ to the human
  Step 4: $L_s = L_s \cup \mathcal{M}$ (Labels provided by human)
  Step 5: Infer on the graphs conditioned on the labels provided by human
  Step 6: Update $\hat{J}_s, S$ using Steps 1 & 2a-d
else
  Step 7: Update models $\{\mathcal{P}_s, \mathcal{P}_o, \Phi_\xi(S_n, O_m)$ & $\Phi_\xi(O_{m'}, O_m)\}$ with L_s

evaluate our results comparing with the ground truth which is available in [13].
Experimental Setup. We use a publicly available software- '*UGM Toolbox*' [44] to infer the node and edge belief in image graphs. We use pre-trained model '*VGG net*' [22] which is trained on 'places-205' dataset to extract the scene features from CNN. For object recognition, we use the model as presented in [25].

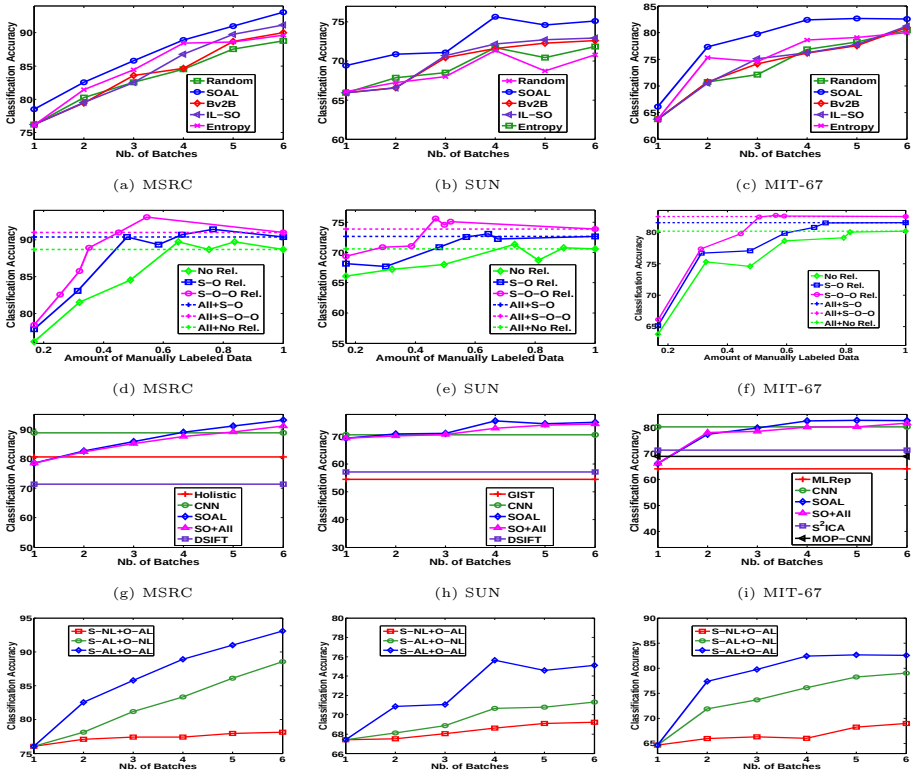
In our online learning process, we perform 5 fold-cross validation, where one fold is used as testing set and the rest are used as training set. We divide the training set into 6 batches. We assume that human-labeled samples are available in the first batch and we use it to obtain the initial S and O classification models and the S-O and O-O relations. It might be possible that we do not have all the classes for scene and objects in the first batch. So, new classes are learned incrementally as batches of data come in. Now, with current batch of data we apply our framework to choose the most informative samples to label and then, update the classification and relationship models with newly labeled data. Finally, we compute our recognition results on the test set with each updated models.
Evaluation Criterion. In order to train the object detectors, we first choose positive and negative examples. We apply standard hard negative mining [37] method to train the binary SVM. We calculate the average precision (AP) of each category by comparing with the ground truth. Precision depends on both correct labeling and localization (overlap between object detection box and ground truth box). Let the computed bounding box of an object be O_b and the ground truth box be G_b, then the overlap ratio, $OR = \frac{O_b \cap G_b}{O_b \cup G_b}$. $OR \geq 0.5$ is considered as correct localization of an object. Before presenting our results, we define all the abbreviations that will be used hereafter

◇ **SOAL:** proposed scene-object active learning (SOAL) as discussed in Sec. 3.
◇ **Bv2B:** Best vs Second Best active learning strategy proposed in [18].
◇ **IL-SO:** Incremental learning (IL) approach presented in [45] is implemented for scene and object (SO) classification.
◇ **No Rel:** No relation is considered between scene and objects.
◇ **S-O Rel:** Only S-O relations are considered but not O-O relations.
◇ **S-O-O Rel:** Both S-O and O-O relationships are considered.
◇ **All+S-O:** All samples with S-O relations are considered.
◇ **All+S-O-O:** All samples with both S-O and O-O relations are considered.
◇ **All+No Rel:** All samples without any relation are considered.
◇ **SO+All:** All samples in batch are considered for scene and object classification with S-O-O relationship.
◇ **NL, AL:** NL implies no human in the loop, i.e., we do not invoke any human to learn labels. AL denotes active learning. For example, S-NL+O-AL means scene nodes are not queried but object nodes are queried..

Experimental Analysis. We perform the following set of experiments - 1. Comparison with other active learning methods, 2. Comparison of the baselines with different S-O and O-O relations, 3. Comparison against other scene and object recognition methods, and 4. Recognition performance of scene and object models while labeling either scene or object.

Comparison with Other Active Learning Methods. In Figs. 3(a,b,c) and 4(a,b,c), we compare our active learning framework with some existing active learning approaches- Bv2B [18], Random Selection, Entropy [46] and IL-SO [45]. In the case of random selection, we pick the samples with uniform distribution. For Bv2B,Entropy and IL-SO, we implement the methods to select the informative samples for scene and objects. The feature extraction stages are the same as ours. We observe that our approach outperforms other methods by a large margin in selecting the most informative samples in both scene and object recognition.

Is Contextual Information Useful in Selecting the Most Informative Samples? We conduct an experiment that implements our proposed active learning strategy by exploiting different set of relations of scene (S) and objects (O). Figs.3(d,e,f) and 4(d,e,f) show the plots for S and O respectively on three datasets. It is noticed that the highest accuracy is yielded by S-O-O Rel (proposed), followed by S-O Rel and No-Rel in scene classification as well as in object recognition. This brings out the advantage of exploiting both S-O and O-O relations in actively choosing the samples for manual labeling. Moreover, the manual labeling cost is significantly reduced when we consider more relations. It may also be noted that our proposed framework achieves similar or even better performance by only choosing a smaller subset of training data than building a model with full training set for both scene and objects. For scenes, this subset is **35**%, **30**% and **42**% of whole training set on MSRC, SUN and MIT datasets respectively. Similarly, for objects, we require only **39**%, **61**%, **60**% of whole training set to be manually labeled on these three datasets.



(a) MSRC (b) SUN (c) MIT-67

(d) MSRC (e) SUN (f) MIT-67

(g) MSRC (h) SUN (i) MIT-67

(j) MSRC (k) SUN (l) MIT-67

Fig. 3. In this figure, we present the scene classification performance for three datasets-MSRC [43], SUN [2] and MIT-67 Indoor [42] (left to right). Plots (a,b,c) present the comparison of SOAL (proposed) against other state-of-the-art active learning methods. Plots (d, e, f) demonstrate comparison with different contextual relations. Plots (g,h,i) demonstrate the comparison of other scene classification methods. Plots (j,k,l) show the classification performance by utilizing our active learning framework either on scene or objects and both. Please see the experimental section for details. Best viewable in color.

Comparison against other Scene and Object Classification Methods. We also compare our S and O classification performance with other state-of-the-art S and O recognition methods. For scene, we choose Holistic [13], CNN [22], DSIFT [21], MLRep[47], S^2ICA [48] and MOP-CNN [49]. Similarly, we compare against Holistic [13], R-CNN [23], DPM [37] for object detection performance. Holistic approach exploits interrelationship among S and O using graphical model. We also compare with SO-All. From Figs. 3(g,h,i) and 4(g,h,i), we can see that our proposed framework outperforms the other state-of-the-art methods.

How does scene and object sample selection affect classification score of each other? We perform an experiment to observe how S and O recognition performs, when we implement active sample selection of either scene or object nodes and exploit S-O and O-O relationships to improve the decisions

of the other type of nodes. The results are shown in Figs. 3(j,k,l) and 4(j,k,l). Let us consider the first scenario (S-NL+O-AL) where we perform AL on the O nodes but use relationships to update the classification probabilities of the S node. We use the first batch to learn the S and O models, but thereafter query to label only object nodes and not scene nodes.

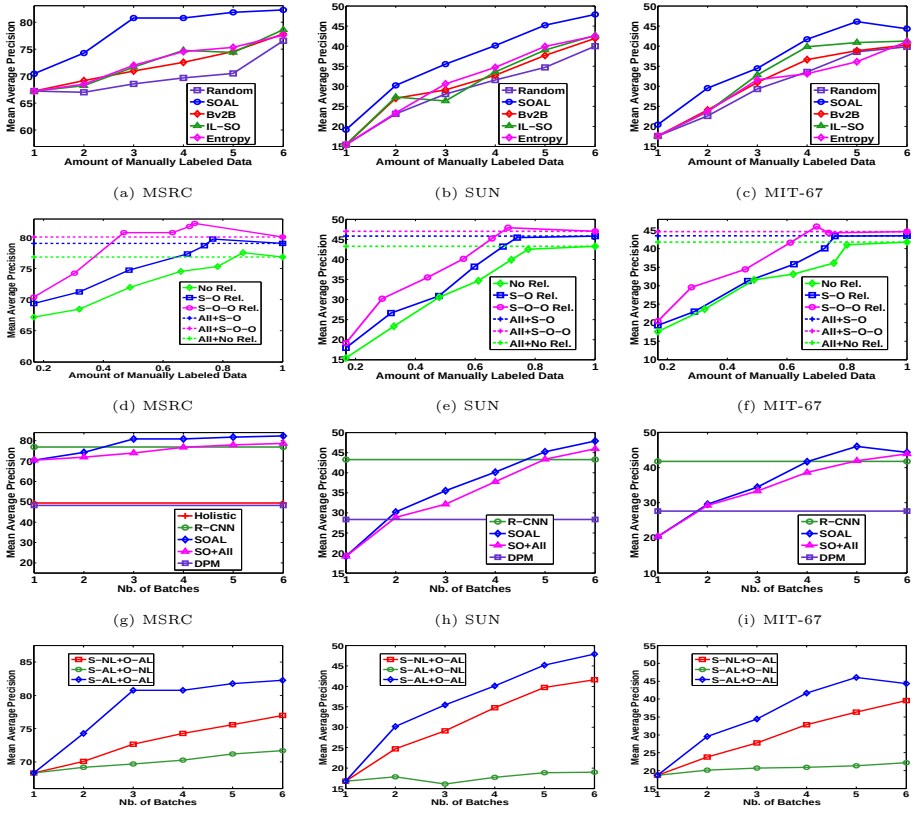


Fig. 4. In this figure, we show the object detection performances on MSRC [43], SUN [2] and MIT-67 Indoor [42] (left to right). Plots (a, b, c) present the comparison of SOAL with other state-of-the-art active learning methods. Plots (d, e, f) demonstrate comparison with different graphical relations. Plots (g, h, i) present the comparison of other object detection methods. Plots (j,k,l) show the detection performance by implementing our active learning framework either on scene or objects and both. Please see the experimental section for details. Best viewable in color.

The relationship models are updated based on the confidence of scene classifier and manual labeling of the objects obtained from a human annotator. With each update on context model, scene classification accuracy goes up even though the scene classification model is not updated. Similarly, the second scenario involves manual labeling of only S nodes but not O nodes. In this scenario, we do not

consider O-O relationships. We can not rely on confidence of object classifiers to model O-O relations as it might provide wrong prediction of object labels. However, involvement of human in both scene and objects makes the sample selection even more efficient and outperforms all the scenarios mentioned above. As shown in Figs. 3(j,k,l) and 4(j,k,l), S-AL+O-AL achieves better performance than S-AL+O-NL by approximately **4-5**% and **4.5-5.5**% in both scene and objects on three datasets.

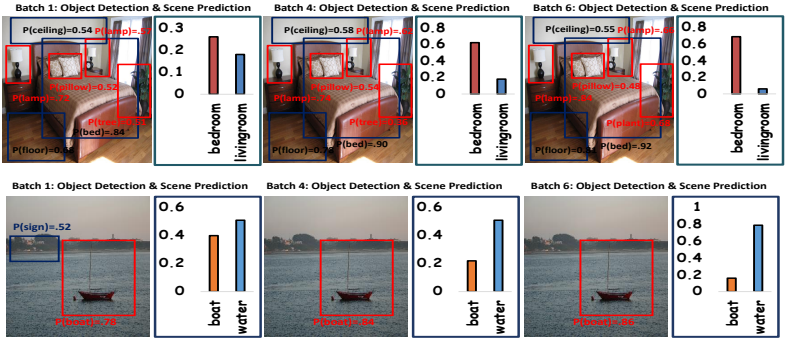


Fig. 5. Scene prediction and object detection performance on test image with updated model learned from the data of 1^{st}, 4^{th} and 6^{th} batch.

Some Examples of Active Learning (AL) Performance. We provide some examples of scene prediction and object detections as shown in Fig. 5. Here, scene prediction and detections are changing as models are learned over samples from each batch. Scene and object models are updated continuously with upcoming batch of data using our AL approach. With each improved model from the batch of data, classifiers become more confident in predicting scene and object labels on test image. More such examples are provided in the supplementary material.

5 Conclusions

In this paper, we propose a novel active learning framework for joint scene and object classification exploiting the interrelationship between them. We exploit the scene-object and object-object interdependencies in order to select the most informative samples to develop better classification models for scenes and objects. Our approach significantly reduces the human effort in labeling samples. We show in the experimental section that with only a small subset of the full training set we achieve better or similar performance compared with using full training set.

Acknowledgment: The work was partially supported by NSF grant IIS-1316934 and US Office of Naval Research contract N00014-15-C-5113 through Mayachitra, Inc.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)

2. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR. (2010)
3. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **6**(1) (2012) 1–114
4. Li, X., Guo, Y.: Multi-level adaptive active learning for scene classification. In: ECCV. (2014)
5. Li, X., Guo, Y.: Adaptive active learning for image classification. In: CVPR. (2013)
6. Moraes, R., Valiati, J.F., Neto, W.P.G.: Document-level sentiment classification: An empirical comparison between svm and ann. Expert Systems with Applications **40**(2) (2013) 621–633
7. Zhang, Y., Liu, X., Chang, M.C., Ge, W., Chen, T.: Spatio-temporal phrases for activity recognition. In: ECCV. (2012)
8. Shi, L., Zhao, Y., Tang, J.: Batch mode active learning for networked data. ACM Transactions on Intelligent Systems and Technology (TIST) **3**(2) (2012) 33
9. Hu, X., Tang, J., Gao, H., Liu, H.: Actnet: Active learning for networked texts in microblogging. In: SDM, SIAM (2013) 306–314
10. Li, J., Bioucas-Dias, J.M., Plaza, A.: Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning. Geoscience and Remote Sensing, IEEE Transactions on **51**(2) (2013) 844–856
11. Mac Aodha, O., Campbell, N., Kautz, J., Brostow, G.: Hierarchical subquery evaluation for active learning on a graph. In: CVPR. (2014)
12. Hasan, M., Roy-Chowdhury, A.K.: Context aware active learning of activity recognition models. In: ICCV. (2015)
13. Yao, J., Fidler, S., Urtasun, R.: Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In: CVPR. (2012)
14. Wojek, C., Schiele, B.: A dynamic conditional random field model for joint labeling of object and scene classes. In: ECCV. (2008)
15. Alberti, M., Folkesson, J., Jensfelt, P.: Relational approaches for joint object classification and scene similarity measurement in indoor environments. In: AAAI 2014 Spring Symposia: Qualitative Representations for Robots. (2014)
16. Wang, B., Lin, D., Xiong, H., Zheng, Y.: Joint inference of objects and scenes with efficient learning of text-object-scene relations. Multimedia, IEEE Transactions on **PP**(99) (2016) 1–1
17. Nimmagadda, T., Anandkumar, A.: Multi-object classification and unsupervised scene understanding using deep learning features and latent tree probabilistic models. arXiv preprint arXiv:1505.00308 (2015)
18. Li, X., Guo, R., Cheng, J.: Incorporating incremental and active learning for scene classification. In: ICMLA. (2012)
19. Yue, J., Li, Z., Liu, L., Fu, Z.: Content-based image retrieval using color and texture fused features. Mathematical and Computer Modelling **54**(3) (2011) 1121–1127
20. Li, Z., Itti, L.: Saliency and gist features for target detection in satellite images. TIP **20**(7) (2011) 2017–2029
21. Liu, C., Yuen, J., Torralba, A.: Dense scene alignment using sift flow for object recognition. In: CVPR. (2009)
22. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS. (2014) 487–495
23. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014)
24. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: ECCV 2014. (2014) 346–361

25. Girshick, R.: Fast r-cnn. In: ICCV. (2015)
26. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR. (2011) 3273–3280
27. Zhu, Y., Nayak, N., Roy-Chowdhury, A.: Context-aware activity modeling using hierarchical conditional random fields. PAMI **37**(7) (2015) 1360–1372
28. Zhang, L., Zhen, X., Shao, L.: Learning object-to-class kernels for scene classification. TIP **23**(8) (2014) 3241–3253
29. Fathi, A., Balcan, M.F., Ren, X., Rehg, J.M.: Combining self training and active learning for video segmentation. In: BMVC. Volume 29. (2011) 78–1
30. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: Training object detectors with crawled data and crowds. IJCV **108**(1-2) (2014) 97–114
31. Vondrick, C., Ramanan, D.: Video annotation and tracking with active learning. In: NIPS. (2011)
32. Elhamifar, E., Sapiro, G., Yang, A., Sasrty, S.: A convex optimization framework for active learning. In: ICCV. (2013)
33. Settles, B.: Active learning literature survey. University of Wisconsin, Madison **52**(55-66) (2010)
34. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with gaussian processes for object categorization. In: ICCV. (2007)
35. Kading, C., Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Active learning and discovery of object categories in the presence of unnameable instances. In: CVPR. (2015)
36. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(3) (2011) 27
37. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI **32**(9) (2010) 1627–1645
38. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV. (2007)
39. Li, Y., Nevatia, R.: Key object driven multi-category object recognition, localization and tracking using spatio-temporal context. In: ECCV. (2008)
40. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Constructing free-energy approximations and generalized belief propagation algorithms. Information Theory, IEEE Transactions on **51**(7) (2005) 2282–2312
41. Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S.: Exploiting hierarchical context on a large database of object categories. In: CVPR. (2010)
42. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR. (2009)
43. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: BMVC. (2007)
44. Schmidt, M.: Ugm: A matlab toolbox for probabilistic undirected graphical models (2010)
45. Hasan, M., Roy-Chowdhury, A.: Incremental activity modeling and recognition in streaming videos. In: CVPR. (2014)
46. Druck, G., Settles, B., McCallum, A.: Active learning by labeling features. In: EMNLP. (2009)
47. Doersch, C., Gupta, A., Efros, A.A.: Mid-level visual element discovery as discriminative mode seeking. In: NIPS. (2013)
48. Hayat, M., Khan, S.H., Bennamoun, M., An, S.: A spatial layout and scale invariant feature representation for indoor scene classification. arXiv preprint arXiv:1506.05532 (2015)
49. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: ECCV. (2014)