Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Class Active Learning

Behjat Siddiquie
Department of Computer Science
University of Maryland, College Park
behjat@cs.umd.edu

Abhinav Gupta
Robotics Institute
Carnegie Mellon University
abhinavg@cs.cmu.edu

Abstract

We present an active learning framework to simultaneously learn appearance and contextual models for scene understanding tasks (multi-class classification). Existing multi-class active learning approaches have focused on utilizing classification uncertainty of regions to select the most ambiguous region for labeling. These approaches, however, ignore the contextual interactions between different regions of the image and the fact that knowing the label for one region provides information about the labels of other regions. For example, the knowledge of a region being sea is informative about regions satisfying the "on" relationship with respect to it, since they are highly likely to be boats. We explicitly model the contextual interactions between regions and select the question which leads to the maximum reduction in the combined entropy of all the regions in the image (image entropy). We also introduce a new methodology of posing labeling questions, mimicking the way humans actively learn about their environment. In these questions, we utilize the regions linked to a concept with high confidence as anchors, to pose questions about the uncertain regions. For example, if we can recognize water in an image then we can use the region associated with water as an anchor to pose questions such as "what is above water?". Our active learning framework also introduces questions which help in actively learning contextual concepts. For example, our approach asks the annotator: "What is the relationship between boat and water?" and utilizes the answer to reduce the image entropies throughout the training dataset and obtain more relevant training examples for appearance models.

1. Introduction

Object recognition is one of the most challenging problems in computer vision. The performance of most recognition approaches, generally, depends upon the diversity and quantity of examples in the training dataset. There have been recent efforts aimed at gathering large training [1, 3]. However, these approaches have sought to obtain annotations for all the images in the dataset without prioritizing them on the basis of diversity. Such an approach leads to sub-optimal performance under finite/limited resources (manpower).

Due to the difficulty in obtaining a large amount of human labeling, many recent efforts have employed an active learning framework to choose regions to be labeled by human annotators. These approaches utilize the uncertainty in classification, asking humans to label examples which are hard to classify using the classifiers learned from previously labeled data. However, most of the work in active learning for visual recognition has focused on obtaining labeling for binary classification problems, especially where objects occur in isolation (such as the CALTECH-256 dataset [4]). In the case of multi-class classification, these approaches seek to obtain the labels of high entropy regions.

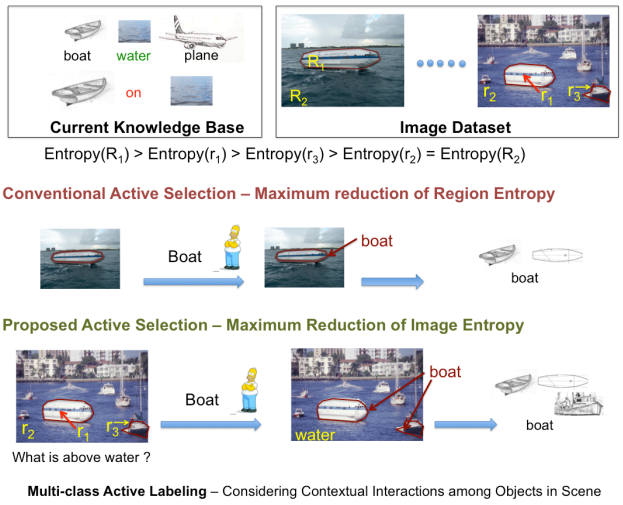


Figure 1. **Region Entropy vs. Image Entropy:** If we utilize region entropy only, region R_1 is selected for labeling since it has higher entropy than all other regions. Therefore, obtaining label of R_1 would lead to maximum reduction of entropy. On the other hand, if we consider image entropy and model the information yield due to contextual interactions, region r_1 is selected over R_1 since the label for r_1 would also provide information about other uncertain regions, such as r_3.

We present a new framework for active selection of questions that simultaneously learns appearance and contextual models for scene understanding (multi-class classification) tasks. Our framework is based on active learning from natural images containing multiple objects. Traditionally, active learning approaches select questions which solicit the labels of uncertain regions. In contrast, we model contextual in-

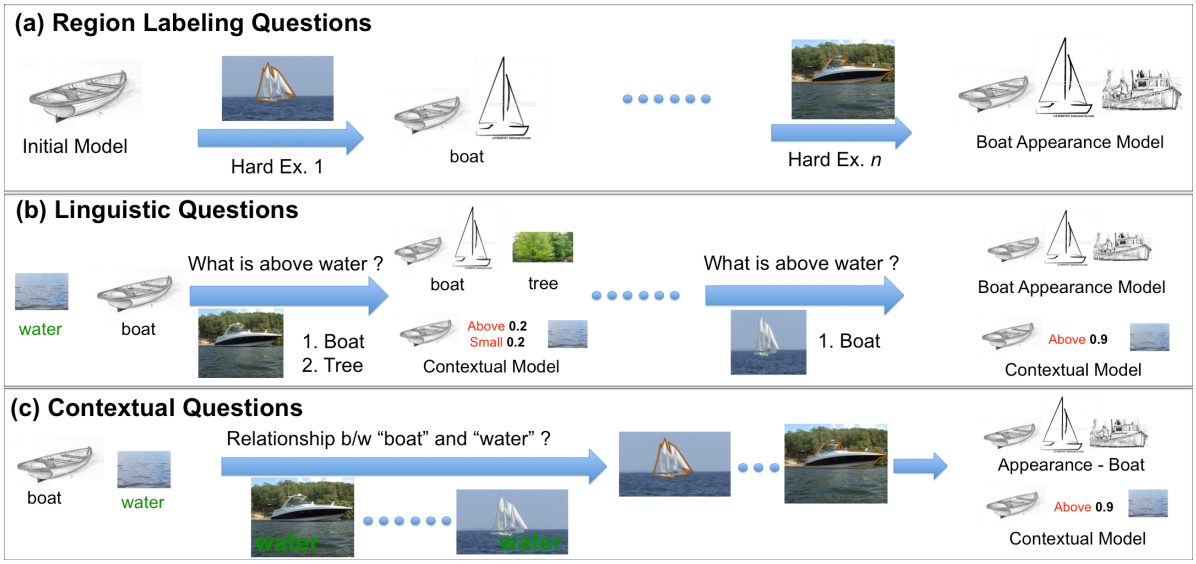


Figure 2. **Types of Questions:** Region labeling questions are the conventional questions utilized by active learning approaches. Here at each iteration the system asks the annotator to annotate the most uncertain region. Linguistic questions use the high confidence labels in the image to pose questions about uncertain regions. For example, since water is easy to recognize, the region associated with it is used to ask "what is above water". Contextual questions are the questions about contextual interactions between pairs of objects in the world. For example, the system asks "what is relationship between boat and water". Contextual questions can be utilized to reduce the entropy of the all the training images since concepts can help in dis-ambiguating other uncertain regions.

teractions between image regions and solicit labels for those regions that yield significant reduction in the combined entropy of all the regions in the image (image entropy). Therefore, our criteria selects regions which are likely to yield information about the other confusing regions in the image as well. For example, consider the scenario shown in figure 1. Traditional active learning approaches would select region R_1 to be labeled, since it is the most uncertain region. In contrast, our approach would evaluate the importance of each label not only based on the local region entropy, but also on how much new information that labeled region would provide about other uncertain regions in the image. Therefore, our approach selects r_1 since knowledge of r_1 label (boat) would yield information that would help reduce entropy of other regions, such as r_3.

One issue with using muti-object images for learning is localization of the objects of interest. Current active learning approaches handle this by either asking annotators to provide the boundaries or prompting labels on segmentations / super-pixels [22]. While such conventional labeling questions can be included in our active learning criteria, we also introduce linguistic questions which utilize additional constructs (such as prepositions or adjectives) in language for handling localization. In linguistic questions, the regions that can be linked to a concept with high confidence are used as anchors to ask questions about unknown regions in the scene. For example, in figure 2(b), the water region (easy to recognize) can be utilized as an anchor to ask questions such as " what is on the water?". Visual attributes of regions can also be used for anchoring, and lead to questions such as "What is the white region in the image?". These linguistic questions mimic the way humans solicit information to actively learn about their environment. These questions are also vital for obtaining labels when conventional labeling interfaces (mouse and screen) are not available [1].

The contributions of the paper are three-fold: (1) We introduce a new criteria for active selection of labeling questions based on reduction in the joint entropy of all the regions in the image (image entropy). By considering image entropy as opposed to the entropy of individual regions, we generate labeling questions which yield information about the region not only whose label is solicited, but other regions in the image as well. Experiments indicate that this criteria outperforms two baseline approaches by a wide margin. (2) We introduce linguistic questions in the active learning framework. In such questions, high confidence regions in the scene are used as anchors to pose questions about high entropy regions. (3) Finally, we introduce a new active learning framework which not only prompts for labels of regions but also poses questions about contextual concepts. For example, as shown in figure 2, our approach asks the annotator: " What is the relationship between boat and water? ". By learning contextual concepts directly from the annotator, we achieve reduction in global entropy over the entire dataset. This leads to faster learning of appearance models, as the concept can be applied throughout the training dataset to obtain new training examples (fig. 2).

2. Related Work

There has been recent interest in utilizing humans as resources for gathering visual recognition datasets[**?**, 1, 3,

[1] A typical example of this is an interaction between a robot and a human where robot asks questions to actively learn about the environment.

6, 7]. Some research has focused on generating human-friendly interfaces for labeling [1] or keeping human interest level high by formulating the labeling task as a game [7]. However, in most of these approaches the selection of regions/images to be labeled is mostly random. In machine learning, active learning approaches [18, 19, 11, 10] are used to rank unlabeled points based on classification uncertainty- difficult examples are chosen for labeling. Criteria for selection include heuristics based on the version space of SVMs [10], disagreement among classifiers [11] and expected informativeness [13, 12].

Early work on active learning in computer vision focused on obtaining binary labels of isolated objects. In multi-class scenarios, these approaches[14, 15, 16] extend the framework by utilizing multiple binary 1-vs-all classifiers. These approaches have two drawbacks: (1) They cannot compare the uncertainty in prediction of an example for two different binary subproblems, and hence cannot identify the classes that require more training data. (2) They assume localized object windows are available in the training dataset. These methods are appropriate for prioritizing labeling of isolated object datasets like CALTECH-256 [4], but would fail for obtaining annotations where multiple objects occur in the same image.

More recent approaches attempted to overcome these two problems. Jain et. al [21] presented an approach for multi-class active annotation utilizing a probabilistic variant of K-Nearest Neighbors. However, they still utilize active learning for selection of images with isolated objects. Settles et. al [20] present an active learning formulation of multiple-instance learning, where localization of positive examples is not required. In a recent paper, Vijayanarasimhan et. al [22] present an active learning formulation where multiple type questions can be used - one type of question solicits location information by labeling of super-pixels. However, they consider only binary classification problems and not contextual interactions. Our work is also related to [23] which exploits the same-class and different-class membership relations between multiple data-points for active learning. This framework [23], however, cannot be extended easily to include spatial interactions (such as above, below) and other relationships (such as bigger, brighter) between data-points.

3. Problem Formulation

3.1. Contextual Object Recognition Model

Our contextual object recognition model is based on the generative model used by Gupta and Davis [2]. In this approach, the authors represent contextual relationships between objects using constructs in language such as prepositions and comparative adjectives. Object appearance models are based on features of a region and relationship models are based on differential features.

We briefly describe the generative model (see figure 3(a)) and refer the readers to the paper[2] for details: Each image is segmented into regions and each region is assumed to be associated to a noun node. Every pair of noun nodes is connected by a relationship edge. The relationship edge provides the constraints on the type of relationships that can exist between the nouns (based on priors learned from data – for example, sun should occur above water). Relationship edges also draw their likelihood from the differential features extracted from the pair of regions. For an image I, let I_j be the region appearance features for the jth region of the image, R_j, and I_{jk} be the differential features computed between regions R_j and R_k. Then, the joint probability $P(n_1, n_2..|I)$ can be written as:

$$\begin{aligned} & = P(n_1, n_2..|I_1, I_2..I_{12}..C_A, C_R) \quad (1) \\ & \propto \prod_i P(I_i|n_i, C_A) \prod_{(j,k)} \sum_{r_{jk}} P(I_{jk}|r_{jk}, C_R) P(r_{jk}|n_j, n_k) \end{aligned}$$

where n_j represents the noun associated with region R_j, r_{jk} is the relationship between regions R_j, R_k while C_A and C_R represent the parameters learned for noun and relationship models respectively.

The inference equation above consists of three terms: the first term is the noun likelihood term, which reflects how well the appearance of the regions matches the appearance of the noun-classes. The second term is a relationship likelihood term which indicates how well differential features match with relationship word models and the third term is the prior which restricts the possible relationships between pairs of noun-classes. Inference over this network is conducted using belief propagation.

3.2. Active Learning

During active learning we pose one of the three types of questions to the user, and utilize the user's answer to update the existing object recognition model. Our objective at each stage, is to select the question, whose answer will lead to the maximum improvement in the current recognition model. The three types of questions are:

- **Regional Labeling Question:** This is the type of question used in traditional active learning methods for building visual classifiers. The user is simply asked to provide the label of a selected region in an image[Figure 2(a)].
- **Linguistic Question:** Motivated by the way humans actively learn about new objects using additional linguistic constructs, we propose a new type of active learning question. In this question, regions linked to "certain" concepts are used as anchors in the image to pose questions about other regions. For example, in the scenario shown in figure 2(b), a user is asked a question such as "what is above the water?", and is required to list the objects in the image which satisfy the question. The user simply answers "boat" and "tree" and does not specify which regions correspond to which objects in the answer.
- **Contextual Question:** The user is asked to provide the possible relationships between a pair of object classes, n_i and n_j. For each possible relationship the user also specifies whether the objects are positively or negatively related with respect to the relationship.

Compared to previous active learning methods [21, 23], which proceed by determining the best region to label next, our task is much more complex. We must identify both the type of question to ask and select the most (potentially) informative question from the set of possible questions of that type. The size of the set of possible questions, especially the linguistic questions, is much more larger than in traditional active learning methods.

Many active learning approaches use uncertainty/entropy as the criterion to choose the region to label. The region with the highest entropy is chosen based on the assumption that fixing its label would lead to maximum reduction in the overall entropy of the system. These approaches, however, ignore the interactions between different regions in the image and the information a label provides about other regions in the image. In contrast, we consider contextual interactions and formulate the selection based on likely reduction of image entropy (entropy based on all the regions of the image). For computational reasons, we ignore the effect of fixing the label of a region in an image on the other unlabeled images. Some approaches [22] choose questions whose answers(labels) are expected to minimize the uncertainty over the entire unlabeled dataset. However, during each round of active learning, they require evaluating the uncertainty on the entire unlabeled dataset for each possible answer of every question. This is impractical in the case of large multi-class problems, more so in our case where the number of possible questions is much higher than in traditional active learning methods. In the following section, we describe the information-theoretic measure, based on Shannon entropy, to quantify information gain for a question.

3.2.1 Entropy of the system

Our training set consists of a set of images $\mathcal{I}$, of which a small subset $\mathcal{I}_L$ is completely labeled, while the remaining, much larger, subset $\mathcal{I}_U$, is unlabeled. We use $\mathcal{I}_L$ to learn the initial contextual object recognition model and then employ our active learning framework to ask the user conventional and linguistic questions about images from the unlabeled subset $\mathcal{I}_U$ along with contextual questions, while attempting to minimize the total entropy on $\mathcal{I}_U$(defined below).

Equation 1, gives the probabilities of all possible class label assignments to the different regions of an image, while taking into account the contextual relations between them. We can use these probabilities to compute the joint entropy of an image:

$$H(I) = \sum_{(n_1,n_2..)\in N} -P(n_1, n_2..|I)\log(P(n_1, n_2..|I)) \quad (2)$$

Directly computing the joint entropy is impractical due to its computational complexity, hence we need to approximate it. An obvious approximation is the the first order entropy, which is the sum of the entropies of each region considered individually:

$$H_{fo}(I) = \sum_{I_j\in I}\sum_{n_j\in N} -P(n_j|I_j)\log(P(n_j|I_j)) \quad (3)$$

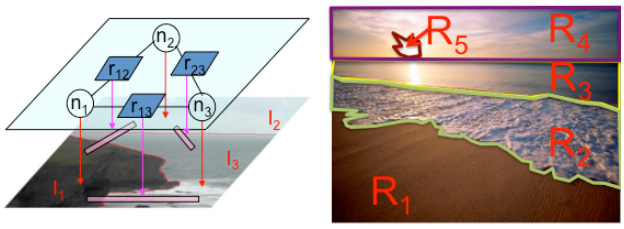
However, this completely ignores the contextual uncertainty of the system. Hence we use the second order approximation of the joint entropy, which is a special case of the Bethe entropy approximation [5], defined as:

$$H_{so}(I) = \sum_{(I_j,I_k)}\sum_{(n_j,n_k)\in N} -P(n_j, n_k|I_j, I_k, I_{jk}) \quad (4)$$
$$\log(P(n_j, n_k|I_j, I_k, I_{jk})) - (m-1)H_{fo}(I)$$

where m is the number of regions in the image I and $P(n_j, n_k|I_j, I_k, I_{jk})$ denotes the pairwise probability of regions R_j and R_k, which can be computed from Eqn. 1 assuming that the image contains only regions R_j and R_k. The total entropy of the system $H_{so}(\mathcal{I}_U)$, is then defined as the sum of the entropies of all the images, as they are independent of each other.

$$H_{so}(\mathcal{I}_U) = \sum_{I^i\in\mathcal{I}_U} H_{so}(I^i) \quad (5)$$

Based on this entropy measure, we define the importance of a question as the reduction in the system entropy resulting from knowing the answer to that question. Therefore, we compute the expected entropy reduction for each question and choose the one leading to the maximum expected reduction in entropy irrespective of its type. We now describe the method for computing the expected entropy reduction for each type of question and the procedure for updating the current appearance and context models based on the answer to each question.



(a) Graphical Model (b) Linguistic Question: Example

Figure 3. (a) The graphical model used in [2]. (b) Linguistic Questions : An example of how certainty of some regions can be used to pose questions.

3.2.2 Region Labeling Questions

In region labeling questions, an annotator is prompted for the label of region R_j in image I. The expected reduction in the entropy of the image can be written as the reduction in entropy given that region R_j has the label c (and marginalizing over c). The reduction in entropy based on labeling the region R_j in image I is thus:

$$\Delta H_{so}(I, R_j) = \sum_{c \in \mathcal{C}} P(I_j|c, C_A)(H_{so}(I) - H_{so}(I|n_j = c))$$

where $H_{so}(I|n_j = c)$ denotes the entropy of the image, given that region R_j belongs to class c. After being labeled, the new class likelihood of region R_j is simply:

$$P(I_j|n_j) = \begin{cases} 1 & \text{if } n_j = c \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Substituting the new likelihood $P(I_j|n_j)$, in (4), we obtain $H_{so}(I|n_j = c)$. Intuitively, it can be seen that in (4) $P(n_j, n_k|I_j, I_k, I_{jk}) = 0 \quad \forall n_j \neq c$ thereby decreasing the number of possible states of the image, leading to a reduction in its entropy. As the other images are independent of image I, $\Delta H_{so}(I, R_j)$ is also the total reduction in the system entropy. When the user provides the label(c) of region R_j, the corresponding features (I_j) are added to the training set and the appearance model of the class c is updated. Relationship priors are also updated based on the labels obtained.

3.2.3 Linguistic Questions

Linguistic questions utilize the high-confidence regions in images and additional constructs (such as prepositions and comparative adjectives) in the language to ask labeling questions. For example, consider the image shown in figure 3(b). If one can recognize with certainty that region R_3 is water, then using this region as an anchor, questions such as "what is above water ?" or "what is brighter than water?" can be posed.

We need to estimate the expected change in entropy for questions of the form: "What objects obey relationship r_k with respect to object A_c ?"(Expressed as $q = (r_k, A_c)$). The answer given by the user to this question is the list of classes $\mathcal{C}_q$ that satisfy the relationships. Let the regions that satisfy the relationship r_k w.r.t object class A_c in the image be represented by $\mathcal{R}_q$ (For example in fig.3(b), if $q = (above, water)$ then $\mathcal{R}_q = \{R_4, R_5\}$ since region R_3 is water). The entropy of the system is reduced since regions ($\mathcal{R}_q$) have a higher likelihood of belonging to the classes listed in $\mathcal{C}_q$. The new joint probability of the of the image is given by

$$P(n_1, n_2....|I, \mathcal{C}_q) = \sum_{\mathcal{R}_q} P(n_1, n_2....|I, n_{\mathcal{R}_q} \in \mathcal{C}_q) P(\mathcal{R}_q|I) \quad (7)$$

To compute $P(n_1, n_2....|I, n_{\mathcal{R}_q} \in \mathcal{C}_q)$, we modify the likelihood of the regions $R \in \mathcal{R}_q$ and recompute $P(n_1, n_2....|I)$ using equation 1. The new likelihoods are given by

$$P(I_j|n_j, C_A) = \begin{cases} 0 & \text{if } c \notin \mathcal{C}_q; \\ \frac{P(I_j|n_j, C_A)}{\sum_{c \in \mathcal{C}_q} P(I_j|n_j = c, C_A)} & \text{if } c \in \mathcal{C}_q \end{cases} \quad (8)$$

We also need to compute $P(\mathcal{R}_q|I)$. The set of regions that satisfy relationship r_k with anchor concept A_c in the image depends on the location of the anchor region R_{A_c} and the regions which satisfy relation r_k with the anchor region. Therefore, we can write it as:

$$P(\mathcal{R}_q|I) = \sum_{R_{A_c}} P(\mathcal{R}_q|r_k, R_{A_c}) P(I_{R_{A_c}}|A_c, C_A) \quad (9)$$

The new pairwise probabilities, $P(n_j, n_k|I_j, I_k, I_{jk}, \mathcal{C}_q)$ can be similarly computed. For a given answer $\mathcal{C}_q$, the entropy reduction is computed as:

$$\Delta H_{so}(q, \mathcal{C}_q) = H_{so}(I) - H_{so}(I|n_{\mathcal{R}_q} \in \mathcal{C}_q) \quad (10)$$

where $H_{so}(I|n_{\mathcal{R}_q} \in \mathcal{C}_q)$ denotes the new entropy of the image, which can be computed by substituting the new pairwise probabilities and the new likelihoods.

The entropy reduction computed above depends on the answer, $\mathcal{C}_q$, to the question. However, at the time of selection the answer is not known. One could compute the entropy reduction for all possible sets of classes which could be the answer to the question and compute the expected entropy reduction as:

$$\Delta H_{so}(q) = \sum_{\mathcal{C}_q \in Pr(C)} P(\mathcal{C}_q|I) \Delta H_{so}(q, \mathcal{C}_q) \quad (11)$$

where $Pr(C)$ is the power set consisting of all possible combinations of classes. This clearly is prohibitively expensive due to the large number of possible answers. Therefore, we employ importance sampling, where $\mathcal{C}_q$ is sampled based on the joint probability distribution computed from the current model.

The user answers a linguistic question by providing the list of class-labels $\mathcal{C}_q$ corresponding to the set of relevant regions. We can then compute the set of revised class probabilities for possible relevant regions, and then infer the classes of the regions. On obtaining the class assignments of the regions, we update the the appearance models of the corresponding classes by adding the regions to the training set. We also update the relationship priors $P(r_k|n_i, n_j)$ for the object pairs from the regions $\mathcal{R}_q$ and any other previously labeled regions in the image. Thus, linguistic questions, help in improving both the visual as well as the contextual components of our object recognition model.

3.2.4 Contextual Questions

In contextual questions, the annotator is asked for the relationships between a pair of object classes n_i and n_j, and he provides a list of possible relationships and whether these relationships occur "always" or "never". For example, if an annotator is asked : " List Relationship between sky and sea " then he can answer: "sky always occurs above sea and sky never occurs below sea".

For an object-object-relationship triplet the expected reduction in entropy can be obtained as:

$$\Delta H_{so}(r_k, n_i, n_j) = \max \begin{cases} H_{so}(\mathcal{I}_U) - H_{so,high_{ijk}}(\mathcal{I}_U) \\ H_{so}(\mathcal{I}_U) - H_{so,low_{ijk}}(\mathcal{I}_U) \\ 0 \end{cases}$$

where $H_{so}(\mathcal{I}_U)$ denotes the entropy of the system according to the current model, given by Eqn. 5. $H_{so,high_{ijk}}(\mathcal{I}_U)$ denotes the system entropy under the assumption that the relation r_k positively holds between the object pair (n_i, n_j), which can be estimated by computing the system entropy with a modified contextual model where the relationship prior $P(r_k|n_i, n_j)$ is set to high. Similarly $H_{so,low_{ijk}}(\mathcal{I}_U)$ is the system entropy assuming that the relation r_k negatively holds between (n_i, n_j), and is obtained by computing the system entropy with $P(r_k|n_i, n_j)$ set to low. Here the assumption is that, if the current relationship priors do not accurately model a strong relationship(or the lack of it) between a pair of object classes, then correcting the relationship priors should result in a large reduction in the system entropy. Additionally, the entropy reduction will be relatively larger in the case of highly co-occurring object pairs, thereby favoring contextual questions on highly co-occurring pairs whose relationship priors are inaccurate. There can exist more than one strong relationship between an object pair, and representing each of them in the contextual model is important. Hence, we define the total expected entropy reduction of an object-pair as the sum of the entropy reductions due to all the individual relationships:

$$\Delta H_{so}(n_i, n_j) = \sum_{r_k \in Rel} \Delta H_{so}(r_k, n_i, n_j) \qquad (12)$$

Computing the entropy reduction, for all pairs of object classes over the entire unlabeled dataset is, again, computationally expensive. To reduce the computational cost, we compute $\Delta H_{so}(n_i, n_j)$ only from images in which the object pair (n_i, n_j) is expected to have a high joint likelihood. The joint likelihood in each image is determined from the current recognition model. The complexity can be further reduced by restricting the entropy reduction computation to only highly co-occurring object-class pairs.

On obtaining the relationship labeling for the pair (n_i, n_j), the model is updated by setting the the positive relationship priors $P(r_{jk1}|n_i, n_j)..P(r_{jkc}|n_i, n_j)$, to a high value and the negative relationship priors to a low value.

4. Experimental Results

Implementation : Our appearance likelihoods are based on the approach in [21], which is a probabilistic variant of the K-nearest neighbor classifier, to model the likelihood of nouns. The relationship likelihood is modeled using a decision stump similar to [2]. Region and differential features used in the paper are the same as those used in [2]. The region features consist of color(mean rgb, mean hsv, hue and saturation histograms), texture(texture response and texture histograms) and location/shape(mean x-y locations, area), while the differential features are extracted from pairs of regions - for example, difference in brightness of two regions. Our relationship vocabulary consists of *above, below, left, right, more blue, more green, brighter*, which is a subset of the vocabulary used in [2], containing the most relevant contextual relationships. For segmentation, we use the SWA algorithm [24] and perform stability analysis for estimating the stable segmentation level [25]. In all the experiments, the role of annotator is played by an Oracle which utilizes ground truth to obtain the answer to the questions.

We now present experimental results to demonstrate the effectiveness of our active learning framework. We present a detailed experimental analysis of our approach on the MSRC dataset, along with additional results on the recently introduced Stanford dataset [8]. For evaluation, we compare our active learning framework to simple random sampling of questions and a state-of-the-art active learning method introduced in [21]. Both these baselines utilize only region labeling questions.

4.1. MSRC Dataset

We first show the performance of our approach on the standard MSRC dataset which consists of 532 images containing objects from 21 different categories. We use the standard training and test splits [9], consisting of 276 training images and 256 test images [2].

Ground Truth Segmentations: We first evaluate the performance of our approach under perfect segmentation by utilizing the ground-truth segmentations provided with the dataset. By isolating the errors due to segmentation, we can better understand the behavior of our active learning framework. A set of 34 fully-annotated images is chosen from the training set, such that each object category has at least 2 instances, is used for building the initial model. Active learning is then used to improve the model by asking the Oracle the three types of questions and using the response for updating the current model. Figure 4 shows the accuracy(region-level labels) of the different methods as a function of the number of questions answered, starting from the initial model.

It is clear from Figure 4 that our combined active learning framework is significantly better than the other methods. After 40 questions, our combined method has at-least a 14% improvement over all the other methods. As seen in the figure, utilizing a framework with different types of questions allows selection of the question-type which maximizes the entropy reduction. Therefore, initially our system asks contextual questions, since they reduce the entropy the fastest. This is generally followed by region labeling questions, which help in improving the appearance models. Once we have reasonably good appearance and context models, our system is able to find anchors to pose linguistic questions. The figure also shows the importance of utilizing image entropy over region entropy (Compare Region labeling curve to [21]). Utilizing the region labeling questions alone, our criteria outperforms both the selection criteria

[2]The generative model used in the paper yields 72% recognition rate when trained using the perfect segmentations and the entire training set. This rate is comparable to state of the art approaches

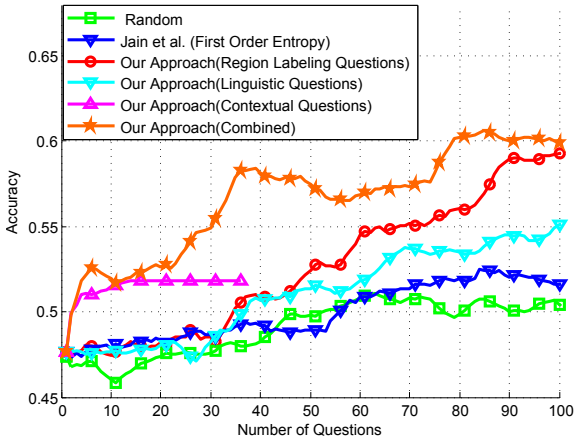


Figure 4. Performance on MSRC dataset when we utilize the ground truth segmentations of the images.

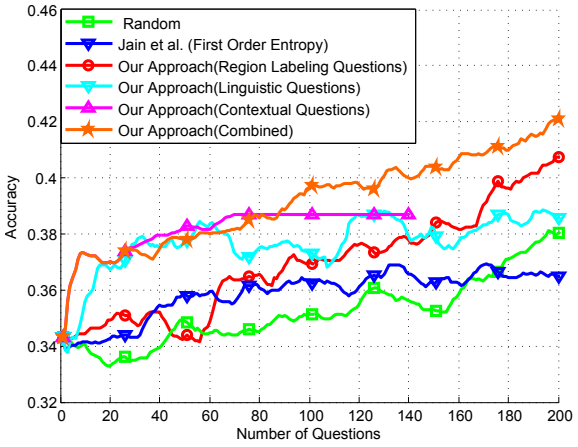


Figure 6. Performance on MSRC dataset using imperfect segmentations.

proposed in [21] and random selection, both of which are based only on region labeling questions. Note that, after a point none of the remaining contextual questions reduce the system entropy, hence the corresponding curve terminates earlier. Another interesting observation is that, as the number of unlabeled regions decreases the performance gain decreases (due to non-availability of informative questions).

Figure 5(a) shows some qualitative examples of questions asked by our active learning framework. It can be seen how our system utilizes high confidence regions associated with grass, sky, ground to pose questions about other regions. Contextual questions asked by the system are also very important and relevant for recognition. Figure 5(b) shows some qualitative examples of improvement in selection by our framework. For example, [20] often selects regions from images where an object(face) occurs in isolation, based on the classification uncertainty of the region, for learning the appearance model. In contrast, our system selects regions(face) from images where other related regions(body) are also present, as fixing the label of those regions also provides information about the other labels. Another example of better selection is that while [21] selects regions such as sky to be labeled (in case of high uncertainty), our approach prefers to ask question or solicit labels about other regions in the image such as house. Fixing the house label also provides information about the region above. Since only tree or sky can occur above a house, the likelihood of confusing those regions with other objects decreases. Whereas fixing the sky label provides very less information about other regions in the image, as most objects generally occur below the sky.

Imperfect Segmentations: In this case we use a set of 50 fully-annotated images for the initial training and active learning is performed as described above. However, here the regions correspond to segments that are automatically generated by the segmentation algorithm and this directly influences the region labeling and the linguistic questions that are selected. The evaluation of the test images is also performed based on the automatically generated segments. Figure 6 shows the accuracy of each method versus the number of questions answered. Here, again, it is clear that our method performs better than the other approaches. In case of imperfect segmentation the rate of increase of performance is slower. This is because ground-truth labels are provided only when the overlap between the segmentation and ground truth region is high; otherwise the Oracle does not provide any answer to the question. In our experiments we found that approximately half of the regions were left unlabeled by the Oracle due to this reason. Furthermore, in case of imperfect segmentation the performance of linguistic questions saturates earlier. This is partly because of the poor performance of linguistic questions at the later stages when only the images without candidate anchor regions remain (poorly segmented images).

4.2. Stanford Dataset

We also evaluate our approach on the Stanford dataset [8], which has been compiled from several already existing datasets and has accurate annotations collected using Amazon Mechanical Turk. It consists of 715 images, consisting of objects from 8 different categories. The images are randomly divided into a training set containing 415 images and a test set consisting of the remaining 300 images. A set of 8 images chosen from the training set, is used for building the initial model and active learning is employed for incrementally improving it. We consider only the top five regions(by area) in each image for both training as well as evaluation purposes. Figure 7 shows the accuracy(region-level labels) versus the number of questions, for each of the different methods. This dataset has 8 classes and therefore the initial context priors are very similar to final context priors and therefore contextual questions are not very helpful. However, due to good initial recognition rate our system finds anchors for linguistic questions more frequently and therefore linguistic questions outperform region labeling questions

Conclusion: We have presented an active learning framework that utilizes contextual interactions between regions in an image for selecting the regions to be labeled. Our criteria prefers regions which have high entropy and provide information about other regions in the image

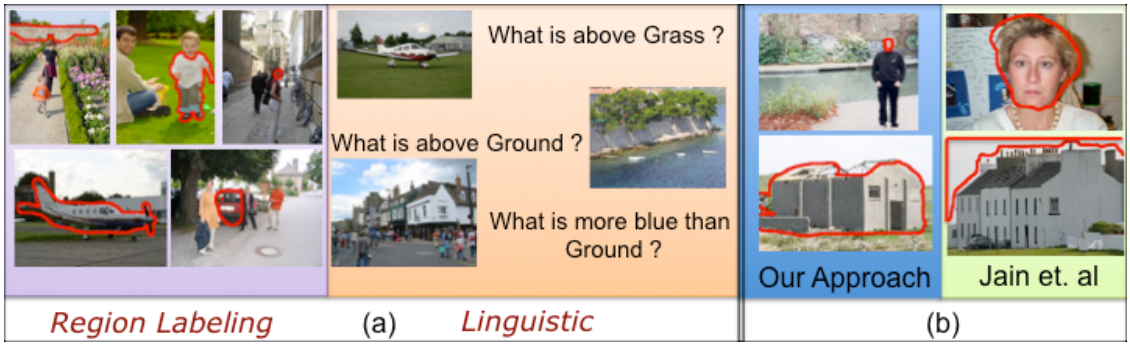


Figure 5. (a) A few examples of region labeling and linguistic questions posed by our framework in MSRC dataset with ground truth segmentations. Contextual questions posed by the system include: (1) What is relationship between grass and cow ? (2) What is relationship between sky and grass ? (3) What is relationship between tree and grass ? (b) Qualitative improvement in selection of questions.

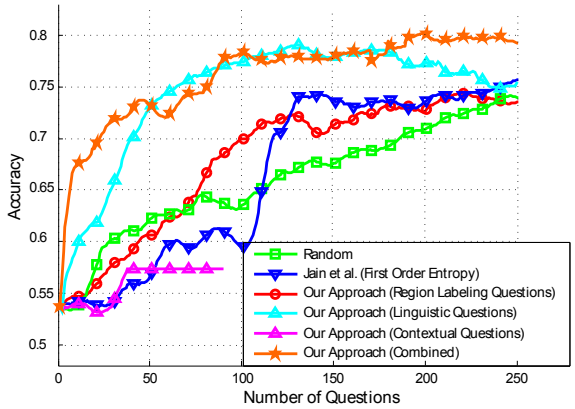


Figure 7. Performance of our system on Stanford dataset using ground truth segmentations.

through contextual interactions. We present linguistic questions which utilize high confidence regions as anchors and additional constructs in language (prepositions, comparative adjectives) to pose questions about uncertain regions. In future, we plan to explore the usage of language ontologies for linguistic question based active learning and how to extend it for videos using temporal prepositions [17].

Acknowledgements: This research was partially supported by ONR grant N00014-09-10044 and NSF award IIS-0905402.

References

[1] B. C. Russell and A. Torralba and K. P. Murphy and W. T. Freeman, LabelMe: a Database and Web-based Tool for Image Annotation, In IJCV 2008.

[2] A. Gupta and L. S. Davis, Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers, In ECCV 2008.

[3] J. Deng and W. Dong and R. Socher and L.-J. Li and K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, In CVPR 2008.

[4] Griffin, G. Holub, AD. Perona, P. The Caltech-256, Caltech Technical Report.

[5] J. Yedidia, W. Freeman and Y. Weiss, Constructing free-energy approximations and generalized belief propagation algorithms, IEEE Transactions on Information Theory, 2005.

[6] A. Sorokin and D. Forsyth, Utility data annotation with Amazon Mechanical Turk, Work. on Internet Vision, 2008.

[7] L. Ahn and L. Dabbish, Labeling Images with a Computer Game, In ACM CHI 2004.

[8] S. Gould, R. Fulton, D. Koller, Decomposing a Scene into Geometric and Semantically Consistent Regions, ICCV'09.

[9] J. Shotton and J. Winn and C. Rother and A. Criminisi, TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context, In IJCV 2007.

[10] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In ICML, 2000.

[11] Y. Freund, H. S. Seung, E. Shamir, N. Tishby. Selective sampling using the query by committee algorithm. ML, 1997.

[12] D. MacKay. Information-based objective functions for active data selection. Neural Computation, 4(4), 1992

[13] N. Lawrence, M. Seeger, R. Herbrich. Fast sparse Gaussian Process method: Informative vector machines. NIPS, 2002

[14] X. Li, L. Wang, and E. Sung. Multilabel svm active learning for image classification. In ICIP, 2004.

[15] R. Yan, J. Yang, and A. Hauptmann, Automatically labeling video data using multi-class active learning. In ICCV, 2003.

[16] A. Holub, P. Perona, and M. Burl. Entropy-based active learning for object recognition. In CVPR workshop on Online Learning for Classification, 2008.

[17] A. Gupta, P. Srinivasan, J. Shi and L. S. Davis. Understanding Videos, Constructing Plots: Learning a Visually Grounded Storyline Model from Annotated Videos, In CVPR 2009.

[18] Lindenbaum, M., Markovitch, S., Rusakov, D.: Selective Sampling for Nearest Neighbor Classifiers. ML (2004)

[19] Kapoor, A., Horvitz, E., Basu, S.: Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning. In: IJCAI. (2007)

[20] Settles, B., Craven, M., Ray, S.: Multiple-Instance Active Learning. In: NIPS. (2008)

[21] P. Jain and A. Kapoor, Active Learning for Large Multi-class Problems, In CVPR 2009.

[22] S. Vijayanarasimhan and K. Grauman, Multi-Level Active Prediction of Useful Image Annotations for Recognition, NIPS 2008

[23] A. Kapoor, G. Hua, A. Akbarzadeh and S. Baker, Which Faces to Tag: Adding Prior Constraints into Active Learning, In ICCV 2009

[24] M. Galun, E. Sharon, R. Basri and A. Brandt, Texture segmentation by multiscale aggregation of filter responses and shape elements, ICCV, 2003.

[25] A. Rabinovich and T. Lange and J. Buhmann and S. Belongie, Model Order Selection and Cue Combination for Image Segmentation, In CVPR 2006