Sampling Approach Matters: Active Learning for Robotic Language Acquisition

Nisha Pillai
Univ. of Maryland, Baltimore County

Edward Raff
Booz Allen Hamilton
Univ. of Maryland, Baltimore County

Francis Ferraro
Univ. of Maryland, Baltimore County

Cynthia Matuszek
Univ. of Maryland, Baltimore County

***Abstract*—Ordering the selection of training data using *active learning* can lead to improvements in learning efficiently from smaller corpora. We present an exploration of active learning approaches applied to three grounded language problems of varying complexity in order to analyze what methods are suitable for improving data efficiency in learning. We present a method for analyzing the complexity of data in this joint problem space, and report on how characteristics of the underlying task, along with design decisions such as feature selection and classification model, drive the results. We observe that representativeness, along with diversity, is crucial in selecting data samples.**

I. Introduction

In grounded language theory, the semantics of language are given by how symbols connect to the underlying real world—the so-called "symbol grounding problem" [24]. For example, we want a robotic system that sees an eggplant (a set of visual percepts from the real world) to ground the recognition object to a canonical symbol for 'eggplant.' When a user asks "Please grab me the eggplant," the robot should ground the natural language *word* "eggplant" to the same *symbol* that denotes the relevant visual percepts. Once both language and vision successfully ground to the same symbol, it becomes feasible for the robot to complete the task. We learn this connection by using physical sensors in conjunction with language learning: paired language and perceptual data are used to train a joint model of how linguistic constructs apply to the perceivable world.

Machine learning of grounded language often demands large-scale natural language annotations of things in the world, which can be expensive and impractical to obtain. It is not feasible to build a dataset that encompasses every object and possible linguistic description. Novel environments will require symbol grounding to occur in real time, based on inputs from a human interactor. Learning the meanings of language from unstructured communication with people is an attractive approach, but requires fast, accurate learning of new concepts, as people are unlikely to spend hours manually annotating even a few hundred samples, let alone the thousands or millions commonly required for machine learning.

In this work we study *active learning*, in which a system deliberately seeks information that will lead to improved understanding with less data, to minimize the number of samples/human interactions required. The field of active learning typically assumes that a pool of unlabeled samples is available, and the model can request specific example(s) that it would like to obtain a label for. By having the model select the most informative data points for labeling, the number of samples that need to be labeled is reduced. This maps to the goal of human-robot learning with minimum training data provided by the human. Furthermore, active learning can be part of a pipeline with other few-shot learning methods [19].

However, active learning is not a magic bullet. When not carefully applied, it does not outperform sequential or random sampling baselines [56]. Thoughtful selection of suitable approaches for problems is required. While active learning has been used for language grounding [48], [52], to the best of our knowledge, we present the **first broad exploration of the best methods for active learning for grounding vision-language pairs**. In this paper, our focus is on developing guidelines by which active learning methods might be appropriately selected and applied to vision-language grounding problems. We test different active learning approaches on grounded language problems of varying linguistic and sensory complexity, and use our results to drive a discussion of how to select active learning methods for different grounded language data acquisition problems in an informed way.

We consider the grounded language task of learning novel language about previously unseen object types and characteristics. Our emphasis is on **determining what methods can reduce the amount of training data** needed to achieve performance consistent with human evaluation. Primarily, we address five relevant questions concerning characteristic-based grounded language learning: (1) How much do active learning techniques help when learning with limited data? (2) Do different active

learning techniques, e.g., pool-based vs. uncertainty-based approaches, lead to noticeable differences in performance? (3) Are the methods robust across both neural and non-neural features and classifiers? (4) How important are the characteristics of the dataset? and (5) How much does incorporating some seed language affect the performance? We make conclusions with respect to these questions in §IV. In addition to addressing the above research questions, we verify how generalizable these learning techniques are beyond characteristic-based grounding.

We find that a right ordering of training data makes it possible to learn successfully from significantly fewer descriptions in most cases, but also that the active learning methodology chosen is specific to the nature of the learning problem. Our main contribution is a **principled analysis of using active learning methods as unsupervised data sampling techniques** in language grounding with a discussion of what aspects of those problems are relevant to approach selection. While our contributions are primarily analytic rather than algorithmic, we argue they address a critical need in language understanding, a research area in which questions of efficiency and data collection are widespread.

II. Related Work

Grounded language learning has been successful in learning to follow directions [2], [3], generating referring expressions [64], visual storytelling [28], video grounding [61] and understanding commands [11], among others. Parsing can be grounded in a robot's world and action models, taking into account perceptual and grounding uncertainty [68], [71]. or language ambiguity [12], [14]. The problem space considered in this paper assumes that there are no pre-existing models of language or objects in the world—an agent is learning from novel language about previously unseen objects [45], [70], making the evaluation more broadly applicable.

Active learning has been applied successfully to a number of problems [7], [32], [63], providing performance improvements in areas as diverse as learning from demonstration [8], [9], following directions [25], and learning about object characteristics [67]. A well-chosen active learning approach can reduce the number of labels required for grounded language learning [43], [52], but raises questions of what queries to ask and when to ask them [10], [41], [66].

Advances in active learning techniques have improved the ability to find the most useful data points. Unsupervised learning techniques, such as subspace clustering, have been shown to find influential points from a cluster [50]. A hybrid method that connects active learning and data programming [47] has shown improvements in the reduction of noisy data in large scale workspaces [15]. Similar to our work, active learning approaches [23], [62], [76] have been effective while training biased and highly varied datasets. Similarly, researchers have put effort into utilizing different active learning methods depends on the complexity of the problem [20]. Also, traditional active learning methods have helped to improve performances in other tasks, such as data fault or fake news detection [4], [26]. Though we consider efficiency over time complexity, researchers have studied methods that are time efficient, especially in large scale applications [27]. Similar to our research, [16] also compares two traditional active learning algorithms for selecting important points from a pool of training data. But we also consider distinct machine learning approaches with small scale and large scale datasets in our comparisons. Various Bayesian techniques have been used in selecting diverse points as the most influential [55] is widely popular, and we use different variants of DPP to select distinct data points as our active learning technique in batch sample selection.

In this work, our goal is to perform a principled exploration of selecting what data to query for labeling [44], using informativeness and uncertainty metrics [7], [72] in grounded language problems of varying complexity. We draw on existing techniques, particularly pool-based learning [34], [65], [77], uncertainty sampling [33], [42], and probabilistic sample selection [59]. We take advantage of that body of research to select our set of experimental approaches, which include sample selection via Gaussian mixture models [17], [31] and Determinantal Point Processes (DPPs) [38], which have proven effective in modeling diversity [21], [74]. Using supervised learners as the active learning techniques [5], [65] are not suitable for our current study since we concentrate on building a language model without prior knowledge [35].

Our work is most closely related to that of Thomason et al. [69], who incorporate 'opportunistic' active learning in a system that learns language in an unstructured environment [48], [67]. However, that work focuses on opportunistically querying for labels whenever annotators are present; this work, in contrast, is focused on exploring the best way of selecting good choices from a large range of possible queries, reflecting the assumption that opportunities to query users will often be severely limited.

III. Approach

For different active learning methods, we learn associations between RGB-D images (color+Depth) of objects in a dataset and the language that describes them. The task is then to find concepts that have a grounded meaning, create lexical terms in an underlying formal meaning representation, and learn visual classifiers that correctly identify things that are referred to in later language interpretation tasks.

At a high level, we ground language by learning characteristic-specific classifiers such as color, shape, and

object for a concept. Consistent with previous work [54], the different types of concepts are obtained from human-provided descriptions of selected objects. In this approach, each concept is associated with a (learned) classifier, and all selected objects described by that concept are used as training data for that classifier. We rely on existing datasets and classification approaches for the actual grounding. We note that the evaluations done in this work are intended to *compare* the success of different active learning approaches for the same problem.

We limit training data to a single description of each object to mimic the limited training available from human interactions. In order to perform replicable experiments, we use active learning approaches in which objects (and associated training and evaluation information, such as descriptions and identified concepts) are drawn from a pre-existing pool of data, rather than obtained *de novo* through human interaction. In our primary experiments we vary the active learning approach used to select new descriptions of objects to add to the training pool. We additionally experiment with different features, and classification techniques. Because our problem focuses on choosing objects to obtain labels for, this is consistent with the task of asking a person for a description of a particular object, but allows us to perform larger-scale and more replicable experiments.

Our goal is to explore the data selection decisions in limited settings to improve performance at the early stages. It is not to improve absolute learning performance; using a novel or complex approach runs the risk of introducing poorly understood confounding factors.

A. Data: Corpora, Concepts, and Features

We use two existing datasets for learning from descriptions: the UMBC dataset [51], which contains 72 objects (see Fig. 1), and the UW RGBD+ dataset [39], which contains 300 objects. Each object instance has multiple associated language descriptions. We follow existing literature [29], [30], [57] on learning to understand language referring to different types of characteristics: COLOR, SHAPE, and OBJECT TYPE. The corpora consist of Kinect2 depth images of objects, paired with human descriptions. The UMBC object dataset contains 8 color, 9 shape, and 18 object characteristics, while the UW RGBD+ dataset includes 14 color, 13 shape, and 51 object characteristics. Shape concepts are reported only in 9.5% descriptions of all 300 object UW RGBD+ dataset annotations. In the UMBC dataset, 53%, 14%, and 73% of annotations reported color, shape, and object concepts.

Type	Image	English annotation
color		This is an **orange** object.
shape		This looks like a green **upside down C shape**.
object type		This is an **Italian Eggplant**. It is firm and dark purple when ripe.

Fig. 1 RGB-D sensor data and descriptions [54]. Each concept was used by multiple annotators to describe each of the corresponding images, showing the noise and variability of human descriptions.

Language. The UMBC dataset contains approximately 430 natural language object descriptions, while the UW RGBD+ dataset contains 1500 descriptions; all were obtained via Amazon Mechanical Turk. For each image of an object, a single description is randomly selected to pair with an image it describes; this is intended to replicate the limited labels available from human interaction. We opt for the simplicity of learning meanings for individual concepts, based on past effectiveness of this approach [46], [54]. Following this previous work, we first convert descriptions into language *concepts*, removing common stop words and lemmatizing the remainder. We then identify meaningful, relevant, and representative concepts by applying tf-idf [58], which yields concepts such as 'banana' and 'yellow', while rejecting those such as 'object' and 'look' (See section IV for details).

Sensor Data. Physical context for language grounding is provided by depth and color images of each object, taken with an RGB-D camera mounted on a robot platform (Fig. 1). From each RGB-D image, we extract perceptual features η_{TRAIT} for each different type of characteristic. We use two different kinds of visual features for learning. In the first, a kernel descriptor-based approach, we use the average RGB values for color, HMP-extracted kernel descriptors [39] for shape, and a combination of the two for objects. In our second approach, we use a convolutional neural network, Neural Architecture Search Network (NASNetLarge) [78], with pretrained ImageNet [18] weights for extracting a 1024 dimension feature vector.

Across all combinations we found that COLOR is relatively easy to learn; SHAPE, which depends in part on camera angle and is less likely to be mentioned, is more difficult; and OBJECT TYPE is the finest grained, with the highest visual complexity.

B. Learning Concept Classifiers

The task we use to test approaches is to learn associations between perceptual inputs and descriptive concepts. Once perceptual features are extracted from the images, a visual classifier for each characteristic is learned. These classifiers are trained using every image that has been described with a concept *and* selected by an active learning method.

Given an instance x_i and a characteristic-specific perceptual representation $\eta_{\text{TRAIT}}(x_i)$, we learn characteristic-

specific probabilistic binary classifiers for each concept, $p_{\text{TRAIT}}(w_{concept} \mid \eta_{\text{TRAIT}}(x_i))$, where $w_{concept} \in \{0, 1\}$ represents the probability of x_i's characteristic TRAIT being described as *concept*. Note that this problem is two-fold: we must learn how to both describe objects properly, and how to *avoid* characterizing objects in a way that does not make sense. We use logistic regression (LR) as our primary classifier type p_{TRAIT} (see §V-D for the impact of this decision) and extract characteristic-specific features η_{TRAIT}.

C. Core Sampling Methods

Intuitively, we want our algorithms to select preferentially the most informative and diverse objects for labeling from the pool of unlabeled objects. Driven by both long-standing and recent findings in active learning [17], [21], [60], [73], we use probabilistic clustering—and point process modeling in particular—as active learning strategies. Because our data is inherently noisy, we found in our early experiments that variations on Gaussian mixture models (GMMs) and determinantal point processes (DPPs) were robust selection algorithms. GMMs accommodate mixed membership, and soft cluster assignments allow us to model uncertainty. We select parametric methods in our learning techniques as they are statistically stable [49] compared to nonparametric models. We therefore focus on GMM- and DPP-based approaches, applied to visually grounded object features, in order to select the most informative points from a set of unlabeled instances.

As we focus on learning from limited data, we do not consider deep learning approaches, which generally operate best over large datasets. Across all of our experiments, we examine five different active learning models: four pool-based methods (GMM Max Log Density Based, VL-GMM,[1] and DPP), and one uncertainty-based (GMM Log Density) method. We introduce a structured DPP (GMM-DPP)-based active learning technique, a novel approach for the grounded language problem. We compare these variants of active learning strategies with a random sampling baseline across our three characteristics (color, shape, and object). Although initial experiments considered entropy-based sampling methods (computed by our GMM's posterior entropy), these were found to perform substantially worse than those listed, and subsequent experiments did not include them. For all GMM approaches, we select the number of components C empirically using four-fold cross-validation. In GMM based methods, we compared the test performance with the number of components ranging from 5 to 35 and received the best results with 15 components. In GMM-based pool sampling experiments, we cluster instances using their informativeness and rank the instances according to their learned conditional densities.

Our methods select instances which are informative and diverse by querying from all N items at once. This is also called querying in "batch mode" and has been applied successfully in the past [13], [59]. We draw from an existing pool of human-provided descriptions, rather than explicitly seeking new labels via interaction, to enable broader and more repeatable experiments.

Max Log-Density-Based GMM Sampling: This model uses a C-component GMM to cluster unique image features and rank them according to their maximum multivariate densities from the unlabeled data pool. Those with greater density are selected as they are potentially more informative. We used 15 Gaussian components (selected empirically as a hyperparameter), initialized the mixing weights and Gaussian parameters using k-means, and fit the GMM with the standard expectation maximization algorithm to learn the parameters.

DPP Sampling: DPPs have proven effective in modeling diversity [22]. We use DPPs as a technique to find the most representative and diverse data points from the pool of data instances. This method uses the pool of all unlabeled image samples to find the most diverse data points by using a radial basis function (RBF) kernel with carefully selected parameters. In our setting, DPPs define a discrete probability distribution of all subsets of image data samples. If $\mathbf{X}$ is the random variable of selecting a subset of images X from a larger set $\mathcal{X}$, then $P(\mathbf{X} = X) = \det(K_X^{(0)})/\det(K_{\mathcal{X}}^{(0)} + I)$, where I represents the identity matrix. Applied to all pairwise elements of X, the *kernel* $K_X^{(0)}$ is a positive semi-definite matrix, where the (i, j) element of the matrix is the value of the kernel applied to items x_i and x_j. We use the RBF kernel, $K^{(0)}(x_i, x_j) = \exp\left(-h\|x_i - x_j\|_2^2\right)$, by cross-validating with $h \in \{100, 25, 4\}$.

GMM-DPP: We combine the DPP kernel with the GMM marginal probability derived from the image samples to rank input samples based on diversity. Following Kulesza and Taskar [36] and Affandi et al. [1], we combine a DPP Kernel $K^{(0)}(x_i, x_j)$ defined on images x_i and x_j with individual "quality" scores for each of the images. We use $P_{\text{GMM}}(x)$—the marginal probability of image x according to the GMM—as the quality scores, and define a new kernel as:

$$K^{(1)}(x_i, x_j) = P_{\text{GMM}}(x_i) K^{(0)}(x_i, x_j) P_{\text{GMM}}(x_j)$$

The marginal probability modulates the diversity of the data. It allows a separate model, with its own assumptions, to help designate what data is and is not diverse. To the best of our knowledge, this is a novel kernel for grounded language learning. Similar to the GMM-based sampling approach, we used 15 Gaussian components in

[1] VL-GMM is included to show the difference between vision-only *vs.* vision-language clustering-based learning, and so does not occur in other reported results.

Sampling	Color	Shape	Object
Baseline: Random	0.75	0.19	0.49
Max-Log-Density-Based GMM Pool	0.82	0.25	0.62
DPP Sampling	0.8	0.22	0.59
GMM - DPP Sampling	0.78	0.27	0.58

Table I AUC summaries for each method's F_1 performance, grouped by the characteristic learned. All AL techniques performed better in characteristic grounding by selecting significant points from the pool.

the GMMs and we initialized the mixing weights and Gaussian parameters using k-means.

IV. Experimental Setup

We estimate the quality of grounded language acquisition by the predictive power of learned concept classifiers against the test objects. In Table I we calculate area-under-the-curve (AUC) from the F_1-score performance of concept classifiers. Our baseline randomly picks images to train visual classifiers while the active learning approaches sample data points as described above. This is meant to mimic the performance of a robot asking random questions about objects in the environment.

The baseline and our active learning methods all only observe concept words from a single text description for each image. Images which are described by these words are selected as positive instances. Similarity metrics are used to find negative examples for these words [53]. All results are averaged over 4–12 runs for each of object, shape, and color. We selected hyperparameters, such as the number of components of our GMM model empirically via cross-validation. We also selected the query size for each experiment empirically.

V. Results and Per-Characteristic Analysis

The overall performance of each approach on the language learning task is shown in Table I, divided into the three characteristic learning problems addressed, namely color, shape, and object.

A. In-Depth Analysis of Active Learning Performances

The effect of active learning techniques in grounded characteristics learning is measured by comparing thre three pool-based active learning techniques described previously, with the random sampling baseline for color, shape, and object characteristics (Table I). Below we will give an analysis of the results with respect to Color, Shape, and Object grounding.

Color. COLOR is the simplest of the three categories of characteristics learned. This observation is, in part, a result of the dataset, in which objects are primarily all of one color; it is also a simpler vision problem overall. Similarly, there is little variation in the color descriptions. Most annotators used simple color names (e.g., "red") rather than the full range of available English terms (e.g., "crimson"). Noisy annotations such as a carrot being described as "purple" and "rose" make the learning problem difficult. To train our color classifiers, we extract RGB features of the segmented object; these define η_{COLOR} and were shared across all approaches.

All active learning techniques outperformed a random baseline in learning groundings for color concepts. When neither visual percepts nor descriptive language varies widely, the primary consideration is to choose representative data quickly. DPP-based sampling methods, which by design select diverse points, also learned effective classifiers with limited data.

Shape. The second category of results, SHAPE, is the most visually complex, and of extreme linguistic difficulty due to limited annotations. Learning shape classifiers is a comparatively complex problem, as the shape of an object varies with viewing angle. A wider variety of words is used to describe shapes but unlike describing colors, users tend not to explicitly specify objects' shapes, e.g., when asked to describe a lemon, most people say yellow, but relatively few say "round". To train shape classifiers, we extract kernel descriptors of the segmented object; these define η_{SHAPE} and were shared across all approaches.

The random sampling baseline is affected by the lack of shape tokens in the description, requiring nearly 30 descriptions to learn the first few shape words. GMM-based DPP showed a noticeable improvement in speed of learning, and also, on inspection, found distinct shape words faster than random sampling. All active learning approaches that found diverse points at earlier stages also outperformed the random baseline.

Object Type. The next challenging grounding task considered in this work is OBJECT—learning language that describes membership in an object class, i.e., object recognition. To train object classifiers, we extract both RGB and kernel descriptors [6]; these define η_{OBJECT}, meaning that object recognition is treated approximately as a superset of color and shape learning.

Performance of Max-Log-Density-Based GMM Pool sampling approach is significantly better than the random baseline. We believe we observed this result because the number of classes is larger (and membership is therefore sparser) than for color and shape characteristics, reflecting the complexity of 'real world' sensor data. This sparsity makes careful selection of samples particularly critical.

B. Pool Vs. Uncertainty-Based Active Learning Methods

Uncertainty sampling methods use learned probability models to measure the uncertainty in unlabeled data points. *Log-Density-Based GMM Uncertainty Sampling:* uses a learned GMM to pick outliers. We select these

Sampling	Color	Shape	Object
Baseline: Random	0.75	0.19	0.49
Max-Log-Density-Based GMM Pool	0.82	0.25	0.62
Log Density Based GMM Uncertainty	0.83	0.23	0.44

Table II AUC summaries of F_1 performance for Pool and Uncertainty sampling performance, grouped by the characteristic learned. Uncertainty sampling (which depends on the feature variability) does not perform well in object grounding, which has a noisy, highly varied data pool.

Sampling	CNN	KernelDesc
Baseline: Random	0.53	0.49
Max-Log-Density-Based GMM Pool	0.66	0.62
DPP Sampling	0.55	0.59
GMM - DPP Sampling	0.49	0.58

Table III AUC summary results for each visual feature's F_1 performance for "object" characteristics. DPP, and GMM pool are able to consistently outperform the baseline with both types of visual features (non-neural kernel descriptors and CNN features).

by finding the images that have the *lowest* log-density from any GMM component. We aim to select the most uncertain data points in order to get a diverse dataset.

Max log-density-based GMM pool based sampling (Table II) chooses representative data points from the unlabeled pool of objects, whereas uncertainty sampling selects the diverse points by considering outliers as the useful points. The selection depends on the variability of the features. For learning color and shape concepts, both pool- and uncertainty-based sampling performed better than the baseline. But while learning object types, the uncertainty sampling could not get required concepts from the most varied visual set and limited annotation dataset.

We hypothesize that the deterioration of uncertainty pooling on the Object task relates to the nature of information's utility in an active learning context. As more information and descriptors become available in the Object scenario, it becomes easier for outliers to occur: points with unusual shapes and color combinations that are not well described will increase model uncertainty. Obtaining a label for an outlier may have limited utility for future data due to the precise nature of being an outlier: its behavior is inconsistent with the rest of the data. This may make uncertainty based approaches less attractive as more complex grounded language datasets become available, or indicate a need in refinement to uncertainty based approaches.

C. The Impact of Visual Features

Convolutional Neural Network (CNN) features have been shown to be effective in learning characteristic types [75]. In this section, we examine how robust our active learning methods are across both neural and non-neural features. In contrast to the "kernel descriptors" (the RGB and HMP features used in the previous section), we extracted 1024 dimension features from the Neural Architecture Search Network (NASNetLarge), which is pre-trained on ImageNet. We refer to the NASNetLarge features as the "CNN" features.

Table III shows that, similar to grounded learning with kernel descriptors, most of the active learning techniques are able to outperform the random baseline on CNN features. DPP and Max log-density-based GMM pool active learning techniques are able to pick diverse and representative points at earlier stages. The characteristic learning example above shows that active learning is effective in selecting meaningful and diverse points faster irrespective of the underlying visual features. These results also show that in a low-data regime, using a CNN over Kernel descriptors without considering the specific method of active learning used can lead to inferior results. Using CNN features with both DPP sampling approaches yields lower AUC than Kernel Descriptors. While the Max-Log-Density approach dominates in this setting, it shows why the study of the impact of features in combination with active learning is necessary.

D. Analysis with Different Classifiers

In this section, we revisit our choice to use a logistic regression classifier for $p_{\text{TRAIT}}(w_{concept} \mid \eta_{\text{TRAIT}}(x_i))$, and we examine how robust our active learning methods are across different classifiers. We consider a support vector machine (SVM) and a multilayer perception (MLP). The SVM is a well-known linear model that finds the maximum-margin hyperplane, which distinctly classifies the data samples. An MLP is a feed-forward artificial neural network that uses nonlinear activation functions. Both have been widely used for classification.

Sampling	LR	SVM	MLP
Baseline: Random	0.75	0.72	0.62
Max-Log-Density-Based GMM Pool	0.82	0.66	0.54
DPP Sampling	0.8	0.66	0.6
GMM - DPP Sampling	0.78	0.63	0.5

Table IV AUC summary results for each classifier's F_1 performance for "color" characteristics. Logistic regression is effectively able to classify the types with diverse and meaningful points.

In this experiment, we examined the "color" characteristic learned with the three classifiers (LR, SVM, and MLP). In Table IV we see that, across active learning methods, logistic regression classifiers are able to classify better than the random sampling baseline. In contrast, neither the SVM nor the MLP resulted in effective classification models when paired with active learning approaches. These results suggest that complex classification methods may not yield improved performance,

and show the need to consider the selection of active sampling methods and downstream classifiers jointly.

E. Analysis with Different Datasets

In this section we examine if our techniques are effective for a large dataset that is visually and linguistically noisy and diverse. In addition to the limited features dataset, we tested our active learning techniques over a 300 object UW RGDB+ multi-colored dataset (Table V) for just "color" characteristics due to space constraints. It contained 51 objects and 1500 annotations (Sec. III-A). In the UW RGBD+ dataset, not every description contains color information. Additionally, the words used to describe the color concepts are inconsistent. Since the dataset contains fewer monochromatic objects, the visual variation is also high, making the vision-language grounding a challenging task. Even in these experiments, most of the learning techniques which selected diverse and representative points were able to perform better than a random baseline. DPP fails to rank in order of importance when the linguistic and visual data is inconsistent. These results indicate that our active learning techniques are generalizable and equally beneficial to datasets on different scales.

Sampling	**UMBC**	**UW RGBD+**
Baseline: Random	0.75	0.53
Max-Log-Density-Based GMM Pool	0.82	0.58
DPP Sampling	0.8	0.51
GMM - DPP Sampling	0.78	0.64

Table V AUC summary results for each dataset's F_1 performance for COLOR. GMM pool and GMM-DPP are able to consistently outperform baseline even with a multi-colored UW RGBD+ dataset.

F. The Impact of Seed Language

So far, our methods have selected images without considering the concepts of the objects represented; in this section, we revisit that restriction and examine whether active learning methods can benefit from considering both the image and language description together. To do this we define a joint vision-language pool-based model that uses a combination of language informativeness and visual features to choose sample points from the data pool. We call this method *VL-GMM Sampling*. We use paragraph vectors [40] to semantically represent a language description associated with the image data point in vector space. We use C-component GMMs to cluster our feature vectors—combined image features and paragraph vectors—and rank them. We consider the features which are closest to the center of cluster points to be the most informative data points and select them for training.

Sampling	**Color**	**Shape**	**Object**
Baseline: Random	0.75	0.19	0.49
Max-Log-Density-Based GMM Pool	0.82	0.25	0.62
VL-GMM Sampling	0.8	0.22	0.57

Table VI AUC summaries for each method's F_1 performance, grouped by the characteristic learned. Both AL techniques performed better in characteristic grounding by selecting significant points from the pool.

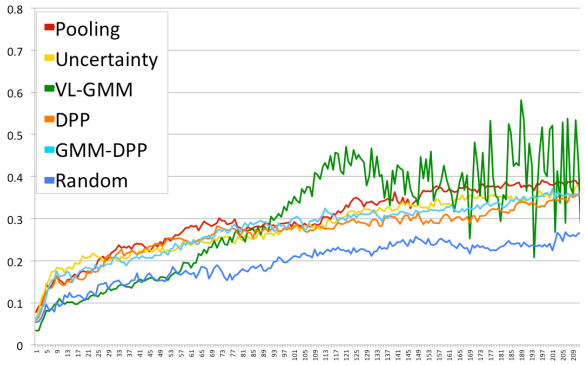

Fig. 2 Performance of **visual classifiers** for Object type as learning progresses with varying data size. Two hundred sixteen distinct object images and their annotations are used in training. F_1-score is shown on the y axis, and number of data samples seen is shown on the x axis. The VL-GMM approach shows promising performance in the more complex shape and objects classification problems. Still, the addition of noisy, highly varied descriptions in training affects the consistency in classification.

VL-GMM sampling (Table VI) outperformed a random baseline in learning groundings for color, shape, and object concepts, selecting the most diverse and informative data points at the earlier stages. VL-GMM consistently exhibited better performance; this makes intuitive sense, as this method uses language as well as image characteristics to select training data, and as such, has more information. While learning object types, VL-GMM selects only informative points at the initial stages, and initial performance is comparable to the baseline. After 50 data samples, it found diverse and representative data samples and ultimately outperformed all other sampling strategies.

G. Performance with Varying Data Size

In this experiment (Fig. 2), we try to mimic real-world human-robot learning that uses noisy, inconsistent, and limited data resources. For training, we used 216 distinct depth images and each image's description for training. We used the remaining 72 images for testing. The descriptions are highly noisy and varied. Most of them do not provide shape or object information. Our

objective here is to understand how our active learning methods perform across varying amounts of available training data. Due to space constraints, and to examine our methods under the "harder" setting where concepts are not frequently described, we show the results for "object" classification in Fig. 2. With highly varied and noisy features, all active learning algorithms could select diverse and important points from the pool using image features and perform better than the baseline for shape and object type words. However, linguistic variability within the description caused VL-GMM's performance to oscillate as it uses language while training. The results show that Max-Log-Density-Based pool sampling is consistently effective in all cases. This experiment also suggests that active learning algorithms that select informative and diverse points increase language acquisition quality especially when the training data is diverse and noisy.

VI. Analysis of Results

The main conclusion we draw from our results is that the selection of the appropriate active learning method depends on the difficulty of the problem in terms of perceptual complexity, complexity and coverage of the language, and sparsity of objects in each class. However, we find that one or more active learning methods exist that can improve learning speed, overall performance, or both in all cases. We overall found GMMs to be a reliable choice for enhancing overall performance. These results are discussed further in this section.

A. Method-Specific Findings

GMM clustering with image features recovers a selection of data with both informative and diverse representation. This approach probabilistically clusters similar features in the same component. The uncertainty-based GMM is unable to effectively find patterns faster at initial stages in the dataset, when object classes are scattered in the visual space. Uncertainty sampling depends on the feature variability for finding uncertain points in GMM clustering, and the sampling selects the noisy outliers when the variability is greater. This finding echoes the performance reduction of uncertainty-based sampling in object feature spaces compared to pool-based approaches.

DPP variants of active learning methods with careful parameter tuning are well suited for selecting the most diverse points in the early stages of learning, which is appropriate when highly varied perceptual features make sample diversity important. Coverage of the more complex SHAPE and OBJECT attributes were attained significantly faster through these methods than random sampling. Visually varied datasets require more examples of concepts overall, in addition to requiring diversity. k-DPP sampling provides diverse samples from the dataset, and is proven sufficient for faster convergence of characteristic concepts. The DPP-based method is able to find the diverse data samples at the initial stages and provide faster convergence to the classification tasks with kernel descriptors as well as CNN features. However, *representativeness*, along with *diversity*, ensure consistent improvement. DPP sampling does not ensure representativeness and is not effective in the case of multi-colored, confusing samples. GMM-based structured DPPs provide breadth as well as diversity and perform well for simple and complex kernel descriptors data. However, this approach is weaker for CNN-based object classification, which may be because the process of selecting representative data adds unnecessary constraints.

While the requirement for language in selecting data samples would be a limitation for large datasets, we found that sampling methods that could consistently augment the visual features with a small amount of language yielded improved grounded language systems.

Time Considerations. For a dataset with N number of training data with D dimension, the determinantal point process computation requires $O(Nk + k^2)$ [37] if the eigen decomposition of the positive semi-definite kernel $K^{(0)}$ is available. And, eigen decomposition approximately takes $O(D^3)$. Here k denotes the size of subsets considered in DPP sampling. Similarly, the Gaussian Mixture model requires $O(D^3)$ to calculate weights that involve finding inverse and determinants. Since we calculate weights for every component and every data point, the overall time complexity of Max-Log-Density based approaches would be $O(C{*}N{*}D^3)$, where N is the number of data points, and C is the number of components. Structured DPP calculation involves GMM and DPP, so it requires $O(((N + k) * k + C) * N * D^3)$ operations in total. After comparing the time complexities, Max-Log-Density based pool sampling seems suitable for large scale datasets.

B. General Considerations

In all but the most trivial cases, random sampling from a dataset outperforms a sequential baseline. Since describing objects in order is a normal human behavior, this suggests that, lacking any other change, having an agent ask widely ranging questions in varying order may improve learning efficiency compared to passive learning.

For cases in which neither visual percepts nor descriptive language varies widely, such as COLOR, all active learning techniques are appropriate. We show that careful selection of *informative* points is most critical under these circumstances: since the features are simple, the main consideration is to select representative data quickly, assuming that learning groundings (here, training visual classifiers) will proceed quickly.

For visually differentiated, linguistically complex datasets, the importance of having a wide *variety* of samples increases. DPPs [38] are a class of 'repulsive' processes designed for increasing diversity (see the discussion of k-DPP, above). Tuning with GMM parameters allows the DPP method to choose distinct, representative, and salient points in the data set in very early learning. Uncertainty-based max posterior GMM sampling performs well on complex data but does not perform as strongly for sparsely populated features.

We have shown that active learning techniques with carefully selected points reduce the amount of training data needed (see Table I). We see that when dealing with more complex datasets, choosing diverse and meaningful points increases performance compared to choosing outliers. Our experiments have also shown that active learning helps set the right order of data points that can improve learning efficiency for both neural and non-neural visual features, and the addition of language features is not necessary for pool-based learning techniques to reduce the label cost.

To summarize, GMM pool sampling, which decides certainty based on the density of the clustered data points, is the most reliable active learning choice for simple, complex, noisy, multi-colored, and highly varied datasets. It is consistently able to outperform random selection at least with 5% increased predictive power. GMM uncertainty sampling is not a reliable choice in case of visual data with extremely noisy outliers. Logistic regression is the most robust classification model in modeling diverse limited data compared to SVM and MLP. DPP based and GMM based pool sampling produces good results in the case of neural and non-neural visual features. We observe that feature variability affects the selection techniques than the characteristics of the dataset. We believe that the vision-and-language sampling method considers the complexity and variance in visual features as well as the language features, and it aids in selecting the most diverse samples.

VII. Conclusion

In this work, we present a thorough exploration of different active learning approaches to grounding unconstrained natural language in real-world sensor data. We demonstrate that active learning has the potential to reduce the amount of data necessary to ground language about objects, an active area of research in both NLP and robotics as well as machine learning from sparse data generally. We additionally provide suggestions for what approach may be suitable given the perceptual and linguistic complexity of a problem. Given our analysis of the causes of performance for different algorithms and cases, we believe these results will prove to generalize beyond the relatively simple data seen here, making it possible for these guidelines to apply to more complicated language grounding tasks in future.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grants No. IIS-1940931 and IIS-2024878.

References

[1] R. H. Affandi, E. Fox, R. Adams, and B. Taskar, "Learning the parameters of determinantal point process kernels," in ICML, 2014.

[2] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in CVPR, 2018.

[3] Y. Artzi and L. Zettlemoyer, "Weakly supervised learning of semantic parsers for mapping instructions to actions." TACL, 2013.

[4] S. D. Bhattacharjee, A. Talukder, and B. V. Balantrapu, "Active learning based news veracity detection with feature weighting and deep-shallow fusion," in IEEE Big Data, 2017.

[5] M. Bloodgood, "Support vector machine active learning algorithms with query-by-committee versus closest-to-hyperplane selection," CoRR, 2018.

[6] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in NeurIPS, 2010.

[7] K. Bullard, Y. Schroecker, and S. Chernova, "Active learning within constrained environments through imitation of an expert questioner," in IJCAI, 2019.

[8] K. Bullard, A. L. Thomaz, and S. Chernova, "Towards intelligent arbitration of diverse active learning queries," in IROS, 2018.

[9] M. Cakmak, C. Chao, and A. L. Thomaz, "Designing interactions for robot active learners," TAMD, 2010.

[10] M. Cakmak and A. L. Thomaz, "Designing robot learners that ask good questions," in HRI, 2012.

[11] J. Y. Chai, Q. Gao, L. She, S. Yang, S. Saba-Sadiya, and G. Xu, "Language to action: Towards interactive task learning with physical agents." in IJCAI, 2018.

[12] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov, "Gated-attention architectures for task-oriented language grounding," in AAAI, 2018.

[13] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Batch mode active sampling based on marginal probability distribution matching," TKDD, 2013.

[14] D. Chen and R. Mooney, "Learning to interpret natural language navigation instructions from observations," in AAAI, 2011.

[15] X. Chen and B. Wujek, "Autodal: Distributed active learning with automatic hyperparameter selection." in AAAI, 2020.

[16] R. Chhatwal, N. Huber-Fliflet, R. Keeling, J. Zhang, and H. Zhao, "Empirical evaluations of active learning strategies in legal document review," in IEEE Big Data, 2017.

[17] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," JAIR, 1996.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in CVPR, 2009.

[19] E. Gavves, T. Mensink, T. Tommasi, C. G. M. Snoek, and T. Tuytelaars, "Active transfer learning with zero-shot priors: Reusing past datasets for future tasks," in ICCV, 2015.

[20] Y. Geifman and R. El-Yaniv, "Deep active learning with a neural architecture search," in NeurIPS, 2019.

[21] J. A. Gillenwater, A. Kulesza, S. Vassilvitskii, and Z. E. Mariet, "Maximizing induced cardinality under a determinantal point process," in NeurIPS, 2018.

[22] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in NeurIPS, 2014.

[23] D. Gudovskiy, A. Hodgkinson, T. Yamaguchi, and S. Tsukizawa, "Deep active learning for biased datasets via fisher kernel self-supervision," in CVPR, 2020.

[24] S. Harnad, "The symbol grounding problem," Physica D: Nonlinear Phenomena, 1990.

[25] S. Hemachandra and M. R. Walter, "Information-theoretic dialog to improve spatial-semantic representations," in IROS, 2015.

[26] H. Homayouni, S. Ghosh, I. Ray, and M. G. Kahn, "An interactive data quality test approach for constraint discovery and fault detection," in IEEE Big Data, 2019.

[27] E.-C. Huang, H.-K. Pao, and Y.-J. Lee, "Big active learning," in IEEE Big Data, 2017.

[28] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, M. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell, "Visual Storytelling," in NAACL, 2016.

[29] C. Kery, "Esta es una naranja atractiva: Adventures in adapting an english language grounding system to non-english data," Ph.D. dissertation, University of Maryland, Baltimore County, 2019.

[30] C. Kery, N. Pillai, C. Matuszek, and F. Ferraro, "Building language-agnostic grounded language learning systems," in Ro-Man, 2019.

[31] S. M. Khansari-Zadeh and A. Billard, "Learning stable nonlinear dynamical systems with gaussian mixture models," T-RO, 2011.

[32] W. B. Knox, P. Stone, and C. Breazeal, "Training a robot via human feedback: A case study," in ICSR, 2013.

[33] Q. Kong, B. Tong, M. Klinkigt, Y. Watanabe, N. Akira, and T. Murakami, "Active generative adversarial network for image classification," in AAAI, 2019.

[34] A. Kontorovich, S. Sabato, and R. Urner, "Active nearest-neighbor learning in metric spaces," in NeurIPS, 2016.

[35] B. Krawczyk and A. Cano, "Adaptive ensemble active learning for drifting data stream mining," in IJCAI, 2019.

[36] A. Kulesza and Taskar, "Structured determinantal point processes," in NeurIPS, 2010.

[37] A. Kulesza and B. Taskar, "k-dpps: Fixed-size determinantal point processes," in ICML, 2011.

[38] A. Kulesza, B. Taskar et al., "Determinantal point processes for machine learning," Foundations and Trends® in Machine Learning, 2012.

[39] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in ICRA, 2011.

[40] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in ICML, 2014.

[41] Y. Lewenberg, Y. Bachrach, U. Paquet, and J. Rosenschein, "Knowing what to ask: A bayesian active learning approach to the surveying problem," in AAAI, 2017.

[42] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in SIGIR, 1994.

[43] C. Liang, J. Ye, S. Wang, B. Pursel, and C. L. Giles, "Investigating active learning for concept prerequisite learning," in AAAI, 2018.

[44] Z.-Y. Liu and S.-J. Huang, "Active sampling for open-set classification without initial annotation," in AAAI, 2019.

[45] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A Joint Model of Language and Perception for Grounded Attribute Learning," in ICML, 2012.

[46] H. Mei, M. Bansal, and M. R. Walter, "Listen, attend, and walk: Neural mapping of navigational instructions to action sequences," in AAAI, 2016.

[47] M. Nashaat, A. Ghosh, J. Miller, S. Quader, C. Marston, and J.-F. Puget, "Hybridization of active learning and data programming for labeling large industrial datasets," in IEEE Big Data, 2018.

[48] A. Padmakumar, P. Stone, and R. J. Mooney, "Learning a policy for opportunistic active learning," in EMNLP, 2018.

[49] B. Pan, H. Dong, W. Chen, and C. Xu, "Semiparametric clustering: A robust alternative to parametric clustering," TNNLS, 2019.

[50] H. Peng and N. G. Pavlidis, "Subspace clustering with active learning," in IEEE Big Data, 2019.

[51] N. Pillai and C. Matuszek, "Identifying negative exemplars in grounded language data sets," RSS Workshop on Spatial-Semantic Representations in Robotics, 2017.

[52] N. Pillai, K. K. Budhraja, and C. Matuszek, "Improving grounded language acquisition efficiency using interactive labeling," in R:SS 2016 Workshop MLHRC, 2016.

[53] N. Pillai, F. Ferraro, and C. Matuszek, "Optimal semantic distance for negative example selection in grounded language acquisition," RSS Workshop on Models and Representations for Natural Human-Robot Communication, 2018.

[54] N. Pillai and C. Matuszek, "Unsupervised end-to-end data selection for grounded language learning," in AAAI, 2018.

[55] R. Pinsler, J. Gordon, E. Nalisnick, and J. M. Hernández-Lobato, "Bayesian batch active learning as sparse subset approximation," in NeurIPS, 2019.

[56] M. E. Ramirez-Loaiza, M. Sharma, G. Kumar, and M. Bilgic, "Active learning: an empirical study of common baselines," Data mining and knowledge discovery, 2017.

[57] L. E. Richards and C. Matuszek, "Learning to understand non-categorical physical language for human-robot interactions," in R:SS 2019 workshop on AI+ACR, 2019.

[58] G. Salton and M. J. McGill, Introduction to modern information retrieval. McGraw-Hill Book Company, 1983.

[59] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning," in KDD, 2002.

[60] B. Settles, "Active learning," Synthesis Lectures on Artificial Intelligence and Machine Learning, 2012.

[61] J. Shi, J. Xu, B. Gong, and C. Xu, "Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses," in CVPR, 2019.

[62] W. Shi and Q. Yu, "Integrating bayesian and discriminative sparse kernel machines for multi-class active learning," in NeurIPS, 2019.

[63] E. S. Short, A. Allevato, and A. L. Thomaz, "Sail: simulation-informed active in-the-wild learning," in HRI, 2019.

[64] M. Shridhar and D. Hsu, "Interactive visual grounding of referring expressions for human-robot interaction," in RSS, 2018.

[65] Y.-P. Tang and S.-J. Huang, "Self-paced active learning: Query the right thing at the right time," in AAAI, 2019.

[66] S. Tellex, P. Thaker, R. Deits, D. Simeonov, T. Kollar, and N. Roy, "Toward information theoretic human-robot dialog," RSS, 2012.

[67] J. Thomason, A. Padmakumar, J. Sinapov, J. Hart, P. Stone, and R. J. Mooney, "Opportunistic active learning for grounding natural language descriptions," in CoRL, 2017.

[68] J. Thomason, A. Padmakumar, J. Sinapov, N. Walker, Y. Jiang, H. Yedidsion, J. Hart, P. Stone, and R. J. Mooney, "Improving grounded natural language understanding through human-robot dialog," in ICRA, 2019.

[69] J. Thomason, J. Sinapov, R. J. Mooney, and P. Stone, "Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions," in AAAI, 2018.

[70] M. Tucker, D. Aksaray, R. Paul, G. J. Stein, and N. Roy, "Learning unknown groundings for natural language interaction with mobile robots," in ISRR, 2017.

[71] M. R. Walter, S. Hemachandra, B. Homberg, S. Tellex, and S. Teller, "A framework for learning semantic maps from grounded natural language descriptions," IJRR, 2014.

[72] H. Wang, R. Zhou, and Y.-D. Shen, "Bounding uncertainty for active batch selection," in AAAI, 2019.

[73] Z. Wang, C. Da Cunha, M. Ritou, and B. Furet, "Comparison of k-means and gmm methods for contextual clustering in hsm," Procedia Manufacturing, 2019.

[74] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu, "Tagging like humans: Diverse and distinct image annotation," in CVPR, 2018.

[75] W. Xu, H. Wang, F. Qi, and C. Lu, "Explicit shape encoding for real-time instance segmentation," in ICCV, 2019.

[76] S. Yan, K. Chaudhuri, and T. Javidi, "The label complexity of active learning from observational data," in NeurIPS, 2019.

[77] C. Zhang and K. Chaudhuri, "Beyond disagreement-based agnostic active learning," in NeurIPS, 2014.

[78] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in CVPR, 2018.