An Analysis of Active Learning Methods for Efficient Robotic Grounded Language Acquisition

Nisha Pillai **Francis Ferraro** **Cynthia Matuszek**
University of Maryland, Baltimore County, Baltimore, Maryland
npillai1@umbc.edu ferraro@umbc.edu cmat@umbc.edu

Abstract

In *grounded language acquisition*, language is combined with vision or sensor data to create a model of how it relates to the physical world. This approach often requires extensive natural language annotations, which can be difficult to obtain; however, *active learning* can lead to improvements in learning efficiently from smaller corpora. We conduct an exploration of active learning approaches applied to three grounded language problems of varying complexity. We demonstrate how different active learning methods can improve the efficiency of grounded language learning, and analyze how characteristics of the underlying task drives the best choice of approach.

1 Introduction

In grounded language, the semantics of language are given by how symbols connect to the underlying real world—the so-called "symbol grounding problem" (Harnad, 1990). This connection can be explored by using world sensors in conjunction with language learning: paired language and physical context are used to train a model of how linguistic constructs apply to the perceivable world.

While powerful, machine learning of grounded language often requires extensive annotation work. Meanwhile, it is desirable to have user-specific, customized agents, especially in the case of physically situated agents. Such agents must be able to handle the wide range of objects and situations in the world. Learning the meanings of language from unstructured communication with people is an attractive approach; however, individual users cannot provide large-scale language annotations for objects around themselves, meaning very efficient learning methods are required.

Active learning, in which a system queries for specific training data, has the potential to improve learning efficiency and reduce the number of labels needed in grounded language acquisition. However, active learning is not a magic bullet; when not carefully applied, it does not outperform sequential or random sampling baselines (Ramirez-Loaiza et al., 2017), meaning that careful selection of suitable approaches for a problem is required.

While active learning has been used for language grounding before (Kulick et al., 2013; Pillai et al., 2016), to the best of our knowledge, there has not previously been a principled exploration the efficacy of different approaches. In this paper, we test different active learning approaches on grounded language problems of varying difficulty, then use our experimental results to discuss how to use select active learning methods for grounded language acquisition in an informed way.

We focus on the problem of learning novel language about previously unseen object types and attributes. In this task, neither the language nor the perceptual targets are represented in the underly-

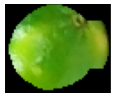
color		This is an **orange** object.
		This is a **purple** eggplant.
shape		This object is **half** of a yellow **cylinder** that has been divided across the diameter of its **base**.
		This looks like a green **upside down C shape**.
object type		This is an **Italian Eggplant**. It is firm and dark purple when ripe.
		This is a green **bulb** of some sort.

Table 1: Examples of images from the Kinect2 sensor used for this work, paired with descriptions provided by annotators. Although in practice all symbols are trained on all object/description pairs, words relevant to a particular attribute type (left column) are bolded to demonstrate the problem being addressed.

ing language model until they are learned from NL interactions (Matuszek et al., 2012). We use a dataset of objects paired with crowdsourced descriptions (see table 1), limiting training data to a single description of each object in order to mimic the limited training available from a human interlocutor. The task is then to find words 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.

Learning the connection between novel percepts and novel language has been explored before. Our primary contribution is a thorough analysis of active learning methods on grounded language problems of varying complexity, and a discussion of what characteristics of different problems make them suitable for different approaches. We find that choosing training data in a principled order makes it possible to learn successfully from many fewer descriptions in most cases, but also that the active learning methodology chosen must be sensitive to the nature of the specific learning problem.

This paper presents an exploration of a number of different active learning approaches on varied learning problems with minimal training data. Our focus is on developing guidelines by which active learning methods might be appropriately selected and applied in these very low resource settings. Although our contributions are primarily investigational rather than algorithmic, they are broadly applicable to grounded language understanding, an active research area in which questions of efficiency and data collection are widespread, and have the potential to support additional algorithmic developments in these areas.

2 Related Work

Active learning has been applied successfully to a number of human-robot interaction and robotics problems previously, providing performance improvements in areas as diverse as learning from demonstration (Cakmak et al., 2010), following directions (Hemachandra and Walter, 2015), and learning about object traits (Thomason et al., 2017). It can reduce the number of labels required for grounded language learning (Amershi et al., 2014; Pillai et al., 2016), but raises questions of what queries to ask, when (Cakmak and Thomaz, 2012; Tellex et al., 2013; Skočaj et al., 2016).

Active learning itself is a rich area of research (Settles, 2012). In this work, we draw on existing techniques, particularly pool-based learning (Zhang and Chaudhuri, 2014; Kontorovich et al., 2016) and uncertainty sampling (Lewis and Gale, 1994; Zhu et al., 2008; Yang et al., 2015). We take advantage of that body of research to select our set of experimental approaches, drawing from work on sample selection from probability distributions (Sarawagi and Bhamidipaty, 2002) and using Gaussian mixture models (Cohn et al., 1996) for sample selection (Khansari-Zadeh and Billard, 2011). We also rely on Determinantal Point Processes (DPPs) (Kulesza et al., 2012), which have proven effective in modeling diversity.

Existing work on grounded language learning has demonstrated success in a number of domains, for example learning to follow directions (Artzi and Zettlemoyer, 2013; Anderson et al., 2018) and understanding commands (Misra et al., 2016; Al-Omari et al., 2017; Chai et al., 2018). Parsing can be grounded in a robot's world and action models, taking into account perceptual and grounding uncertainty (Tellex et al., 2011; Walter et al., 2014; Matuszek, 2018) or natural language ambiguity (Chen and Mooney, 2011). The example problem space in this research requires neither language nor pre-existing models of the world to exist (Matuszek et al., 2012; Tucker et al., 2017), making the evaluation more broadly applicable.

Our work is related to that of Thomason et al., who incorporate "opportunistic" active learning in a robot that learns language in an unstructured environment (Thomason et al., 2017; Padmakumar et al., 2018). However, that work focuses on opportunistically querying for labels when 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 in most situations will be severely limited.

3 Approach

We focus on learning grounded language for three different types of characteristics: COLORs, such as red and yellow; SHAPEs, such as arch and cylinder; and OBJECT TYPEs such as eggplant and banana. Each of these characteristic types has traits that we wish to explore against different types of active learning approach (see table 2). In this dataset, COLORs are relatively easy to learn; SHAPEs, which depend in part on camera angle, are

more difficult; and OBJECT TYPEs are the finest grained grouping, having the highest perceptual feature dimensionality.

For the investigation presented here, we use pool-based active learning (Settles, 2012), in which training instances are chosen from a pre-existing pool of descriptions, rather than interactively seeking new descriptions from people for each experiment. Because our active learning focuses on choosing what object to obtain a description of, this is consistent with asking for a description of a selected object, but allows larger-scale and more replicable experiments.

As discussed in Sec. 3.1 we learn groundings by learning characteristic-specific classifiers for each trait. In order to learn these different classifiers, we we use TF-IDF to select and extract the most meaningful and relevant word types from the language corpus. We use RGB and RGB-D images of objects for learning the perceptual features. We extract color, shape, and object features from the images and train visual classifiers for every word associated with these features. We opt for the simplicity of the word-based approach for its past effectiveness in learning different categories of language concepts in association with vision.

Because our data is inherently noisy, we have found variations on Gaussian mixture models (GMMs) and determinantal point processes to be robust choices in our selection algorithms. GMMs accommodate mixed membership, and soft cluster assignments allow us to model uncertainty. As we focus on learning from limited data, we do not consider deep learning approaches, which generally operate best over large datasets.

Characteristic	Complexity	Traits	Size
Color	Simple	Visually **consistent** **Simple** language **Coarse**-grained	6
Shape	Intermediate	Visually **varied** **Complex** language **Medium** granularity	8
Object type	Complex	Visually **complex** **Moderate** language **Fine**-grained	18

Table 2: A summary of the characteristics of the learning problem, from simplest to most complex. COLORs are consistent across the surface of the object, can be represented with simple visual features, are described using consistent language, and are relatively coarse-grained. SHAPEs are a more difficult learning problem visually, and tend to be described inconsistently in language. OBJECT TYPEs are linguistically more consistent, but are the most difficult perceptual problem, in part due to the specificity of labels.

3.1 Data Corpus

Our dataset consists of 18 categories of objects, each of which has four instances associated with it (see fig. 1 for examples). Following Pillai and Matuszek (2018), we structure both our data and methods around how to associate descriptive words, such as *yellow* and *curved*, with characteristics of categories of objects, like `banana`. Each characteristic has different visual features, as perceived by a robot-mounted Kinect camera.

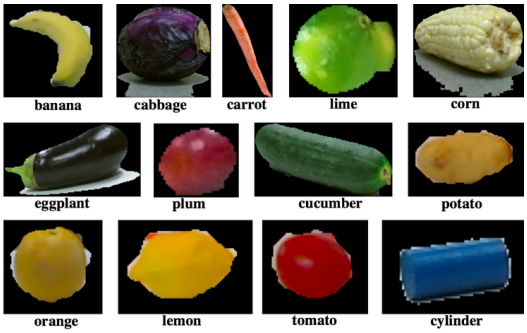


Figure 1: Sample RGB images in the dataset, as taken with a Kinect2 camera and shown to AMT annotators.

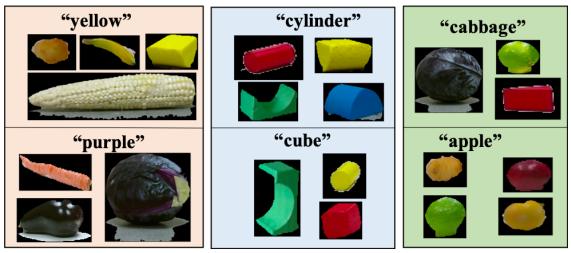


Figure 2: Samples of images and words used to describe them, grouped by characteristic (COLOR on the left, SHAPE in the middle, and OBJECT on the right). Each word was used by multiple annotators to describe one of the corresponding images. The shape descriptions "cylinder" and "cube" are especially noisy.

To provide physical world context into which to ground language, we take an image of each object in our dataset using a Kinect2 RGB-D camera mounted on a robot platform. From each image, we extract perceptual features η_{CHAR} for each different type of characteristic: average RGB values for color, HMP-extracted kernel descriptors (Bo et al., 2011; Lai et al., 2013) for shape, and a combination of the two for objects. We then learn to associate these perceptual inputs with descriptive words drawn from the descriptions.

In order to learn these associations, we acquire natural language descriptions of each object. We use the dataset of Pillai and Matuszek (2018), which contains approximately 6000 crowd-sourced descriptions of 72 objects; for each of the objects, we randomly select a single description of that object to create a training 2-tuple. We perform basic preprocessing to convert these descriptions into language *tokens*: we remove common stop words and lemmatize the remaining words. We then identify meaningful, relevant, representative words from the group of tokens by applying tf-idf, which selects important tokens such as 'banana' and 'yellow', while rejecting those such as 'object' and 'look' (Pillai and Matuszek, 2018).

Formally, given an instance x_i and a characteristic-specific perceptual representation $\eta_{\text{CHAR}}(x_i)$, we learn characteristic-specific probabilistic binary classifiers

$$p_{\text{CHAR}}(w_{value} \mid \eta_{\text{CHAR}}(x_i))$$

where $w_{value} \in \{0, 1\}$ represents the probability of x_i's characteristic CHAR being described as *value*. 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. For example, if x_i is a particular instance of a `banana`, $p_{\text{COLOR}}(w_{yellow} = 1 \mid \eta_{\text{COLOR}}(x_i))$ will be high, while both $p_{\text{COLOR}}(w_{red} = 1 \mid \eta_{\text{COLOR}}(x_i))$ and $p_{\text{COLOR}}(w_{arch} = 1 \mid \eta_{\text{COLOR}}(x_i))$ will be low (first, as bananas are not red, and second, as *arch* is not a proper color term). As we discuss in section 4, we use logistic regression for our basic classifier types p_{CHAR} and extract characteristic-specific features η_{CHAR}. We note that our primary aim is to study active learning methodologies for grounded language acquisition. Logistic regression's widespread familiarity and approachability allow us to focus our efforts and analysis.

3.2 Sampling Methods

As an active learning strategy, our models preferentially select the most informative and diverse objects for labeling from the pool of unlabeled objects. We utilize characteristics of probabilistic clustering, and point process modeling in particular, as active learning strategies. We employ two probabilistic clustering approaches—Gaussian mixture model (GMM) clustering and Determinantal Point Process (DPP) clustering—applied to visually grounded object features in order to select the most informative points from the pool of unlabeled objects.

We explore five active learning models using pool-based and uncertainty-based strategies:

1. **Pool-based** selection using a hard-assignment GMM density to select points closest to component centroids
2. **VL-GMM:** A Vision and Language joint pool based model to select informative and diverse data points using the language descriptions and visual features.
3. **DPP:** A DPP for diversity selection
4. **GMM-DPP:** A GMM-based structured DPP
5. **Uncertainty**-based selection that uses a hard-assignment GPP density to pick points with uncertain component affinity

We compare these variants of active learning strategies with two baselines of traditional sampling: instance-level random sampling of objects, and description-level random sampling across our three categories (color, shape, and object). Although initial experiments considered entropy-based sampling methods (as given by posterior entropy according to our GMM model), these approaches were found to perform substantially worse than those listed, and subsequent experiments accordingly did not include them. Overall, we select N samples from a pool of K uncertainty samples (Sarawagi and Bhamidipaty, 2002) using their probabilities as ranking criteria.

In our experiments, we select instances which are informative and diverse and query in batch mode (query for all K items at once) (Chattopadhyay et al., 2013). As described above, we draw from an existing pool of human-provided descriptions rather than explicitly seeking new labels via interaction, making broader and more repeatable experiments possible. In some of the pool-based active learning experiments, we cluster instances using their informativeness and use density measure as a ranking criterion.

3.2.1 Pool-Based Methods

Pool-based active learning methods are intended to pick the most representative and diverse data samples from a pool of data—in our case, object descriptions. We employ Gaussian mixture models and determinantal point processes as our selection approach to find the diversity and informativeness

from data samples in four variants of pool-based sampling approaches. For any GMM approach, we select the number of components K empirically. We fit the GMM with the standard expectation maximization (EM) algorithm: for a K-component GMM, we learn K-dimensional mixing weights $\pi_1, \ldots, \pi_K$, and K different means and covariances.

Max Log-Density-Based GMM Sampling. In this model, we use K-component GMM to first cluster unique image input features and rank in the order of informativeness from the unlabeled data pool for every visual characteristics. We rank data points by maximum multivariate posterior probability densities. Those with greater density are deemed to be more representative and potentially informative data samples. We calculate the log density of image features for K Gaussian components and select the data points which have the maximum density across mixture components.

k-DPP-Based Sampling. Determinantal Point Processes (DPPs) have proven effective in modeling diversity (Gong et al., 2014). We here use DPPs as a technique to find the most representative and diverse data points from the pool of data instances. Using a kernel function $K^{(0)}$, DPPs define a discrete probability distribution of all subsets of image data samples. If $\mathbf{X}$ is the random variable of selecting a subset X of a larger set $\mathcal{X}$, then:

$$P(\mathbf{X} = X) = \frac{\det(K_X^{(0)})}{\det(K_{\mathcal{X}}^{(0)} + I)}, \; X \subseteq \mathcal{X}$$

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 value of the kernel applied to items x_i and x_j. In this work, we use the RBF kernel, $K^{(0)}(x_i, x_j) = \exp\left(-h\|x_i - x_j\|_2^2\right)$, with h determined experimentally. Here, I represents the identity matrix.

GMM-Based Structured DPP Sampling. Following both Kulesza and Taskar (2010) and Affandi et al. (2014), 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 an image x according to the learned GMM—as the quality scores, and define a new kernel as:

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

Here, the marginal probability acts to modulate the diversity. It allows a separate model, with its own separate assumptions, to help designate what data is and is not diverse.

VL-GMM Sampling. This vision-language pool sampling method utilizes language informativeness together with visual features to choose sample points from the data pool. We use paragraph vectors (Le and Mikolov, 2014) to semantically represent a language description associated with the image data point in vector space. The combination of image features and description vectors are used in Gaussian mixture model-based pool sampling. We use K-component GMM to cluster our feature vectors and rank them according to their informativeness and diversity. We consider the features which are closest to the center of cluster points are the most informative data points and select them to learn our grounded language model.

3.2.2 Pool-Based Uncertainty Methods

Uncertainty sampling methods use posterior density or posterior probability entropy as a measure of the uncertainty in a model. We explore two variants of pool-based uncertainty sampling. The image data points that have the highest log densities are considered the most *informative* points, and the points which have the lowest densities (outliers) are the most *diverse* data points.

Max Log-Density-Based GMM Sampling. As before, for each data point x_i, we find the mixture component that gives the lowest log-density. With this approach, our aim is to select a combination of the most certain and uncertain data points will achieve the diversity in the dataset.

4 Experimental Results

The quality of grounded language acquisition is estimated by the predictive power of learned visual classifiers. Probabilistic active learning strategies were compared against a random sampling baseline. This baseline randomly picks images to train visual classifiers while the active learning approaches sample data points as described above. The baseline and our active learning methods all only observe a single text description for each image. The baseline is meant to mimic the performance of a robot asking random questions about objects in the environment. Visual classifiers are trained on the selected data, and their performance

on a test set is evaluated. A summary of the performance of all approaches is shown in fig. 3, and the area under the curve for each is shown in fig. 4.

As mentioned in Sec. 3.1, we use tf-idf to select tokens for which to train classifiers. Images which are described by these tokens are selected as positive instances. Similarity metrics are used to find negative examples for these language tokens. We combine the description of every object instance and represented them into vector space using Paragraph Vector (Mikolov et al., 2013; Le and Mikolov, 2014). Using cosine similarity, we selected the instances with the most dissimilar vectors as the negative examples in our evaluation. All results are averaged over 8–12 runs for each of object, shape, and color. We tested the learned classifiers on the images of positive and negative instances. We selected hyperparameters, such as the number of components of our GMM model empirically via cross-validation.

Color. COLOR is the simplest of the three categories of characteristics learned. This 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").

To train our color classifiers (eqn. 3.1), we extract RGB features of the segmented object; these define η_{COLOR} and were shared across all approaches. All active learning approaches tried outperformed a random baseline. Instance-based description labeling learns color tokens slowly; sequential ordering of this labeling does not find representative members of all color categories quickly. Random sampling based on the description is able to select informative instances from the pool.

Max posterior-based GMM pool sampling outperforms random sampling in learning groundings for color words. Whereas random sampling selects points or concepts that have already been learned, max posterior-based GMM can choose new, diverse points. This helps max posterior-based GMM perform 20% better and converge more quickly. Uncertainty-based max posterior GMM sampling also outperforms random sampling relatively quickly. On inspection, uncertainty-based sampling successfully picks objects showing a range of colors.

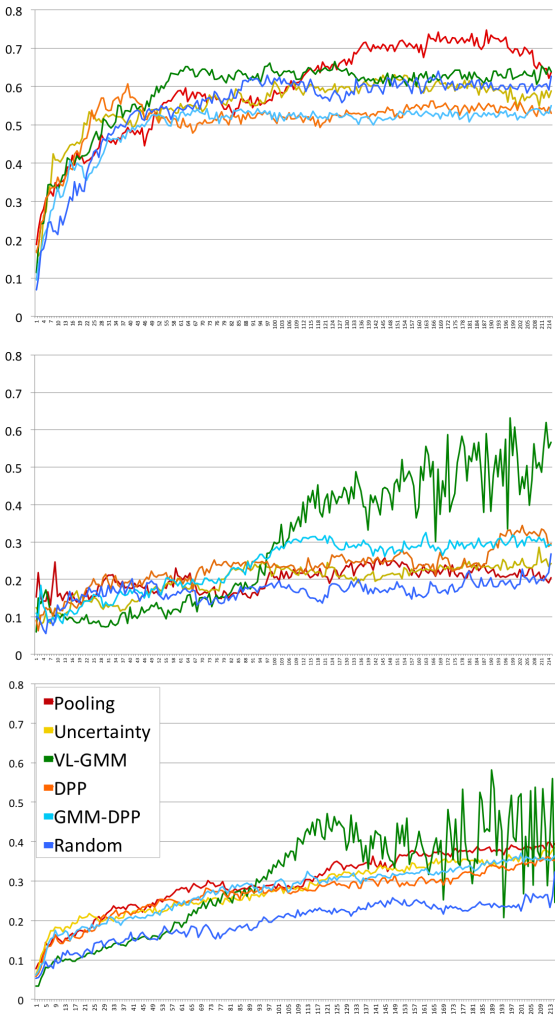


Figure 3: Performance of **visual classifiers** as learning progresses. F1-score is shown on the y axis, and number of samples seen is shown on the x axis. *Top:* color, *middle:* shape, *bottom:* object type. Dark blue lines show performance of the baseline. The GMM-based DPP approach and max posterior-based approaches learn faster than baseline. The VL-GMM approach is shows promising performance in the more complex shape and object classification problems.

4.1 Results and Per-Characteristic Analysis

Shape. The second category of results, SHAPE, is the most visually complex, but of intermediate linguistic difficulty. Learning shape classifiers is a comparatively complex problem, as the shape of an object varies with viewing angle. Users also tend not to explicitly specify objects' shapes; empirically, when asked to describe a lemon, most people say yellow, but relatively few say "round" or equivalent. Annotators also use a wider variety of words to describe shapes.

To train our shape classifiers (section 3.1), we

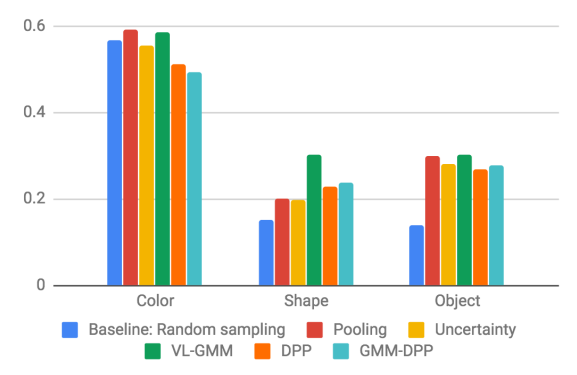


Figure 4: The area-under-curve for each method explored, grouped by each type of trait learned.

extract kernel descriptors of the segmented object (Bo et al., 2010); these define η_{SHAPE} and were shared across all approaches. The instance-based baseline is affected by the lack of shape tokens in the description, requiring nearly 30 descriptions to learn the first few shape words. The next baseline, description-based random sampling, performs better.

Max posterior-based GMM sampling shows a noticeable improvement in quality compared to random sampling, but noise in the descriptions causes inconsistencies in the learning curve. Uncertainty-based max posterior GMM sampling finds distinct shape words very fast compared to random sampling, picking diverse, representative points early in training. By contrast, neither DPP-based pool sampling nor entropy-based methods consistently outperform random sampling. All strategies initially improved slightly faster than random sampling. The VL-GMM approach has the strongest performance; this makes intuitive sense, as this method is using language as well as image characteristics to select training data, and as such has strictly more information.

Object Type. The most challenging grounding task considered in this work is OBJECT—object recognition, or learning what language describes membership in an object class. To train object classifiers (section 3.1), we extract both RGB and kernel descriptors (Bo et al., 2010); these define η_{OBJECT}, meaning that object recognition is treated in part as a superset of color and shape learning.

Performance is good for some but not all active learning methods, as shown in fig. 3. 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. With respect to baselines, 'instance-based labeling' learns more slowly than other approaches. Description-based random sampling improves on instance-based selection. In both cases, this performance can be explained by the sparser populations of different classes; random sampling can find diverse samples quickly, while instance-based learning covers less of the sample space initially.

Max posterior-based GMM pool sampling learns slightly faster initially compared to random sampling, again due to successfully selecting diverse samples. (This learner finds training samples from all object classes roughly 15% faster than random sampling.) Max posterior GMM uncertainty modeling selects a combination of data points which are not certain about their membership in any particular cluster, and data points with high membership certainty. This combination selects both diverse and informative samples, and, as expected, performs well.

There are two DPP variant active learning approaches in our research: k-DPP-based pool sampling, and GMM-based structured DPP pool sampling. Determinantal point processes (DPP) are designed to select diverse data from a pool of samples. In our experiments the model identified all objects except 'triangle' within 40 examples. The next variant of DPP, GMM-based structured DPP sampling, could identify all objects in that span.

5 Discussion

Broadly speaking, we find that one or more active learning methods exists that can improve on learning speed, overall performance, or both in all cases. The appropriate method depends on (at least) the complexity of the dataset in terms of perceptual complexity, complexity and coverage of language, and sparsity of objects in the data set. These results are summarized in fig. 5 and discussed further in this section.

5.1 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. This is consistent with the unsurprising result that diverse train-

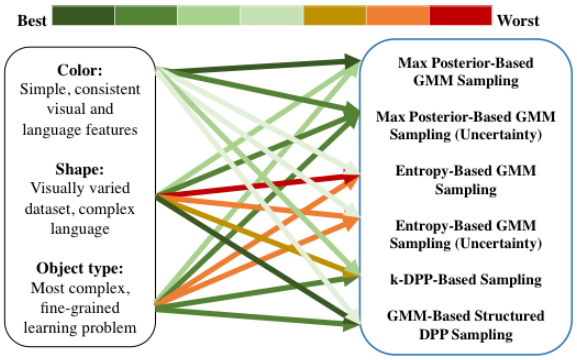


Figure 5: A graphical representation of what active learning approaches (right) performed best on grounding language describing attributes with varying levels of visual and linguistic complexity (right).

ing data improves learned groundings.

For visually distinct and linguistically complex datasets, the importance of having a wide variety of samples increases. DPPs (Kulesza and Taskar, 2011; Kulesza et al., 2012) are a class of repulsive processes suitable for increasing diversity (see section 3.2). 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 provide quite as strong performance for sparsely populated classes.

For cases in which neither visual percepts nor descriptive language vary widely, such as our COLOR attribute, different approaches may be appropriate. All active learning methods worked better than baselines in this set. In particular, max log-density-based GMM samplings approaches worked both well and quickly. Since the features are simple, the main consideration is to select representative data quickly, assuming that learning groundings (here, training visual classifiers, per Matuszek (2018)) will proceed quickly.

5.2 Method-Specific Findings

DPP variants of active learning methods are well suited selecting the most diverse points in early learning, which is suitable when data is less sparse (so that representativeness is less of a concern); for example, coverage of the "color" attribute space was attained significantly faster for these methods than random sampling.

Visually varied datasets require more examples of concepts overall to train classifiers, in addition to requiring diversity. k-DPP sampling provides diverse samples from the dataset, but is not alone sufficient for effective learning of visually complex concepts. However, GMM-based structured DPPs provide breadth as well as diversity, and perform well for complex data. This approach is somewhat weaker for simple data, which may be because the process of selecting of representative data adds unnecessary constraints.

Tuning with GMM parameters allows the DPP approach to choose both distinct and important points in the data set in very early stages. However, tuning the number of components is necessary to pick the most relevant data samples from the pool; a large number for components will result in underfitting, and a small number in overfitting. This tuning must be considered against the performance gain in complex datasets when a method is selected.

Gaussian mixture model clustering with simple features recovers a selection of data with meaningful, diverse representation of the dataset. This approach probabilistically clusters similar features in the same component when the features are visually simple; approximately 20 components provides an adequate visual representation. However, the GMM is unable to effectively find patterns in the dataset when feature dimensionality is higher. As pool-based selection of data points depends on the selection of Gaussian components, this approach performs poorly for visually complex problems. This echoes the performance reduction of max posterior-based GMM sampling in high dimensional, complex feature spaces.

In our experiments, information-gain based sampling methods did not consistently improve overall performance due to sharply peaked posterior probabilities. An investigation of this behavior is ongoing.

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 number of annotations necessary to ground language about object attributes, an active area of research in both NLP and robotics. We additionally provide suggestions for what approach may be suitable given the perceptual and linguistic complexity of a problem. We believe these guidelines will apply beyond attribute grounding problems, and intend to explore

this question in future work.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grants No. 1657469 and .

References

Raja Hafiz Affandi, Emily Fox, Ryan Adams, and Ben Taskar. 2014. Learning the parameters of determinantal point process kernels. In *International Conference on Machine Learning*, pages 1224–1232.

Muhannad Al-Omari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural language acquisition and grounding for embodied robotic systems. In *Proceedings of the 31st National Conference on Artificial Intelligence (AAAI)*, pages 4349–4356.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics (TACL)*, 1:49–62.

Liefeng Bo, Kevin Lai, Xiaofeng Ren, and Dieter Fox. 2011. Object recognition with hierarchical kernel descriptors. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Liefeng Bo, Xiaofeng Ren, and Dieter Fox. 2010. Kernel descriptors for visual recognition. In *Advances in neural information processing systems*, pages 244–252.

Maya Cakmak, Crystal Chao, and Andrea L Thomaz. 2010. Designing interactions for robot active learners. *Autonomous Mental Development, IEEE Transactions on*, 2(2):108–118.

Maya Cakmak and Andrea L Thomaz. 2012. Designing robot learners that ask good questions. In *Proceedings of the 7th annual ACM/IEEE international conference on Human-Robot Interaction*, pages 17–24. ACM.

Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9.

Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. 2013. Batch mode active sampling based on marginal probability distribution matching. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):13.

D.L. Chen and R. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI-2011)*, pages 859–865.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, pages 2069–2077.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Sachithra Hemachandra and Matthew R Walter. 2015. Information-theoretic dialog to improve spatial-semantic representations. In *Intelligent Robots and Systems (IROS)*. IEEE.

S Mohammad Khansari-Zadeh and Aude Billard. 2011. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 27(5):943–957.

Aryeh Kontorovich, Sivan Sabato, and Ruth Urner. 2016. Active nearest-neighbor learning in metric spaces. In *Advances in Neural Information Processing Systems*, pages 856–864.

Alex Kulesza and Ben Taskar. 2010. Structured determinantal point processes. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1171–1179. Curran Associates, Inc.

Alex Kulesza and Ben Taskar. 2011. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1193–1200.

Alex Kulesza, Ben Taskar, et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

Johannes Kulick, Marc Toussaint, Tobias Lang, and Manuel Lopes. 2013. Active learning for teaching a robot grounded relational symbols. In *IJCAI*.

Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. 2013. RGB-D object recognition: Features, algorithms, and a large scale benchmark. In Andrea Fossati, Juergen Gall, Helmut Grabner, Xiaofeng Ren, and Kurt Konolige, editors, *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. Springer.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML-14)*.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag.

Cynthia Matuszek. 2018. Grounded language learning: Where robotics and nlp meet. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proceedings of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*.

Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. 2016. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. In *International Journal of Robotics Research (IJRR)*.

Aishwarya Padmakumar, Peter Stone, and Raymond J. Mooney. 2018. Learning a policy for opportunistic active learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP-18)*, Brussels, Belgium.

Nisha Pillai, Karan K Budhraja, and Cynthia Matuszek. 2016. Improving grounded language acquisition efficiency using interactive labeling. In *Proc. of the R:SS 2016 Workshop on Model Learning for Human-Robot Communication*.

Nisha Pillai and Cynthia Matuszek. 2018. Unsupervised end-to-end data selection for grounded language learning. In *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*, New Orleans, USA.

Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. 2017. Active learning: an empirical study of common baselines. *Data mining and knowledge discovery*, 31(2):287–313.

Sunita Sarawagi and Anuradha Bhamidipaty. 2002. Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278. ACM.

Burr Settles. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Danijel Skočaj, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M Kruijff, Marc Hanheide, Nick Hawes, Jeremy L Wyatt, Thomas Keller, Kai Zhou, et al. 2016. An integrated system for interactive continuous learning of categorical knowledge. *Journal of Experimental & Theoretical Artificial Intelligence*.

S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S. Teller, and N. Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. 2013. Toward information theoretic human-robot dialog. *Robotics*, page 409.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Justin Hart, Peter Stone, and Raymond J Mooney. 2017. Opportunistic active learning for grounding natural language descriptions. In *Conference on Robot Learning*, pages 67–76.

Jesse Thomason, Jivko Sinapov, Raymond J Mooney, and Peter Stone. 2018. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*.

Mycal Tucker, Derya Aksaray, Rohan Paul, Gregory J Stein, and Nicholas Roy. 2017. Learning unknown groundings for natural language interaction with mobile robots. In *International Symposium on Robotics Research (ISRR)*.

Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. 2014. A framework for learning semantic maps from grounded natural language descriptions. *The International Journal of Robotics Research*, 33(9):1167–1190.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*.

Chicheng Zhang and Kamalika Chaudhuri. 2014. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1137–1144. Association for Computational Linguistics.