Pillai, Nisha, and Matuszek, Cynthia. 2018. Unsupervised Selection of Negative Examples for Grounded Language Learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. February 2018, New Orleans.

Unsupervised Selection of Negative Examples for Grounded Language Learning

Nisha Pillai, Cynthia Matuszek
npillai1 | cmat @ umbc.edu
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, Maryland

Abstract

There has been substantial work in recent years on *grounded language acquisition,* in which a model is learned that relates linguistic constructs to the perceivable world. While powerful, this approach is frequently hindered by ambiguities and omissions found in natural language. One such omission is the lack of negative descriptions of objects. We describe an unsupervised system that learns visual classifiers associated with words, using semantic similarity to automatically choose negative examples from a corpus of perceptual and linguistic data. We evaluate the effectiveness of each stage as well as the system's performance on the overall learning task.

Introduction

Semantic representations of real-world environments are a powerful tool for supporting user interaction and action planning. Our goal is to obtain such representations from conversation with users, allowing physically situated agents to learn appropriate world models "on the fly" for a wide range of situations. Learning these models from natural language provides a framework for learning such semantics at the right granularity in an intuitive, natural way.

One promising approach is to treat learning language about percepts as a *joint modeling problem* (Matuszek* et al. 2013; Pillai, Budhraja, and Matuszek 2016), in which descriptive language paired with sensor and actuator data is used to jointly train visual classifiers in conjunction with language models. In this approach, descriptions are treated as labels for visual percepts, making it possible to learn novel language describing entirely novel visual concepts.

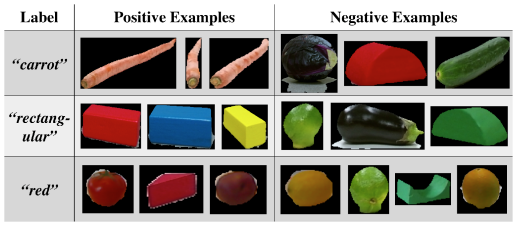

Figure 1: Automatically selected terms and training data for grounded language learning.

However, building semantic models from natural language is challenging. People's use of language is frequently not a good match for statistical learning systems. For example, descriptions of physical things rarely contain negative data: It is unusual for people to provide negative examples without prompting. (Objects are rarely described as "not yellow.") A lack of a positive label does not imply a negative grounding; something described as "an apple" is not a good negative grounding for a "red" classifier. This problem has an effect on parser learning, (Hastings and Lytinen 1994), lexical acquisition (Roy 2002), and human grammar acquisition (Bowerman 1988; Lasnik 1989).

In this paper, we use statistical language processing tools to address two outstanding problems

in grounded language learning. First, we automatically select terms to consider as candidate labels for visual classifiers; second, we use document similarity metrics to select appropriate negative examples from a corpus of training data (see Figure 1). We evaluate our approach on a new data set of objects and descriptions, and our initial results support the idea that purely linguistic tools can be used to overcome weaknesses in corpora of perceptual training data.

Related Work

Much of the work on learning to understand grounded language relies in some part on algorithms that use negative labels as part of learning. The most straightforward approach is to explicitly collect negative labels (Tellex et al. 2013; Dindo and Zambuto 2010), possibly through crowdsourcing (Tellex et al. 2014; Knepper et al. 2015) or gameplaying (Thomason 2016). However, this may not be applicable to all mechanisms for gathering language. Another possibility is to associate randomly chosen groundings with terms that are not used to describe those images (Silberer, Ferrari, and Lapata 2016; Chrupala, Gelderloos, and Alishahi 2017). Because language is not exhaustive, this approach is noisy and may require manual cleanup (Tellex et al. 2011).

Another practical technique is to design language collection trials that either use objects that have no shared visual characteristics (Matuszek* et al. 2013), or explicitly design trials that exhibit negative characteristics (Schenck and Fox 2017). Our work is most similar to the fully unsupervised label identification of Roy (2002), but uses document similarity metrics, rather than term clustering.

In order to choose appropriate language terms for which to train classifiers, we rely on the well-known tf-idf algorithm, which can be used to determine the descriptive power of terms (Salton and McGill 1983), their relevance to particular documents (Zobel and Moffat 1998), or as a document similarity metric (Salton and Buckley 1988). Our selection of negative labels uses the Paragraph Vector algorithm, which learns representations of features from varying length documents (Mikolov et al. 2013a; 2013b). We employ the Distributed Memory Model of Paragraph Vectors (PV-DM) for this work (Le and Mikolov 2014).

Notable research exists in generating descriptions from images or videos (Yu and Ballard 2004; BenAbdallah et al. 2010; Kojima, Tamura, and Fukunaga 2002a; 2002b; Chen and Lawrence Zitnick 2015); for this work we used Amazon Mechanical Turk to obtain descriptions. This work is similar to zero-shot learning for visual classifiers (Elhoseiny, Saleh, and Elgammal 2013), but we use color/depth images to learn classifiers, rather than purely textual descriptions. Like Berg, Berg, and Shih (2010) and Farhadi et al. (2009), our focus is learning classifiers for object attributes; however, we learn the fixed attributes color, shape, and object while they infer higher level attributes.

Linguistic indexing (Li and Wang 2003; 2005) is a related area, but here we intend to learn one attribute/word association. Visual Question Answering (VQA) (Antol et al. 2015) learns image attributes and produces answers to open-ended questions, while we limit ourselves to learning attributes. Previous language representations have used vector models and multimodal topic models for image retrieval (Socher et al. 2014; Lienhart, Romberg, and Hörster 2009), whereas we use a vector model of language to measure the similarity between descriptions of images. We use a simple bag-of-words model, unlike work on generating advanced sentences to describe images by predicting the most likely nouns, verbs, scenes and prepositions (Yang et al. 2011).

Background

TF-IDF: In order to select relevant terms to learn the meanings of, we use tf-idf, for *term frequency-inverse document frequency*, a well-studied metric reflecting how important a word is to a document in a corpus. The tf-idf value *increases* proportionally to the number of times a term appears in the document, which reflects the term's relevance to that document, and *decreases* with the number of documents containing that term, reflecting its discriminative power.

In this work, $tf(t, d)$ is a raw count of the number of times a term t appears in a document d. Inverse document frequency is the inverse logarithmic fraction of the number of documents that contain the term from the set of all documents, D. This gives the tf-idf value of t for a particular descrip-

tive document d:

$$tf\text{-}idf(t,d,D) = tf(t,d) \cdot \log \frac{N}{|\{d \in D : t \in d\}|}$$

Where $tf(t,d)$ is the number of times a term t appears in document d, N is the size of the set of documents $N = |D|$, and $|\{d \in D : t \in d\}|$ is the number of documents in which the term t appears.

Paragraph Vector: In order to find negative examples for terms selected by tf-idf, we use a similarity metric to maximize the semantic distance between object descriptions. Paragraph Vector is an unsupervised learning algorithm that maps documents into a fixed-length feature vector that is robust against varying document sizes (Le and Mikolov 2014). A neural network with one hidden layer is used to derive the error gradients from the loss function, which is calculated using the probability of words in a visual context given the input terms. We use that model to measure dissimilarity between descriptions. In the Paragraph Vector model, paragraphs and words in these paragraphs are mapped to vectors P and W respectively. We calculate the non-normalized log-probability vector of P:

$$y = b + Uh$$

Here y_i is the non-normalized log-probability of a word in the vector. U and b are softmax parameters, and h is a vector formed by a concatenation of word vectors W and paragraph vector P. Prediction of the 'next word' in the context or 'topic' of the paragraph is achieved using a softmax classifier. A fixed length sliding window is applied to choose contexts. Here, w_1,w_2, ..., w_T denote the sequence of words being trained on:

$$p(w_t|w_{t-k}, ...w_{t+k}) = \frac{e^{y_{wt}}}{\sum_i e^{y_i}}$$

The average log probability is then maximized:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k}, ...w_{t+k})$$

Training is performed using gradient descent with backpropagation. The output is a fixed length dense vector, as in a bag of words model, but retains the predictive power of a more semantically informed model. The trained paragraph vector represents the "topic" of a document, and has shown good performance for predicting other terms that may be found in that document. Paragraph Vector maps every document to a point in fixed-dimensional space irrespective of their varying description size; empirically, 2000 dimensions gives sufficient representative power.

Approach

We build on previous work that treats the grounding problem as one in which words are associated with classifiers, jointly training classifiers and descriptive language to develop semantic understanding of the visual characteristics of objects (Matuszek* et al. 2013; Pillai, Budhraja, and Matuszek 2016). We use a two-step approach: first, choosing relevant terms for which to train visual classifiers; second, using semantic dissimilarity between descriptions of objects to find negative examples of that term.

Specifically, we treat all of the descriptions of a particular object, concatenated, as a "document" associated with that object. We use tf-idf to find the most discriminative terms for a particular document, and use all objects people described using that term as positive examples for a classifier. We choose negative examples by learning a paragraph vector for each document, and using cosine similarity to find the most distant paragraph vectors.

Data Corpus

Our data set contains 72 objects, divided into 18 classes. (Classes included both food objects, such as 'banana,' 'cabbage,' and 'carrot,' and children's blocks in various shapes, such as cylinders and cuboids.) We took 3-4 RGB-D images of each object from a variety of angles (see Figure 2).

To obtain descriptive language, the RGB images were posted on Amazon Mechanical Turk, and users provided short descriptions. A total of 3055 descriptions were collected, an average of 42 per object. All descriptions of a single object are concatenated into a "document" describing that object. Documents range from 200–450 words, and our corpus contains 19,947 unique words. A short list of stop words is stripped from the documents, and the remaining words are lemmatized as "terms."

Selecting Relevant Terms

In order to select words to learn,we employ tf-idf to find discriminative terms from the set of descrip-

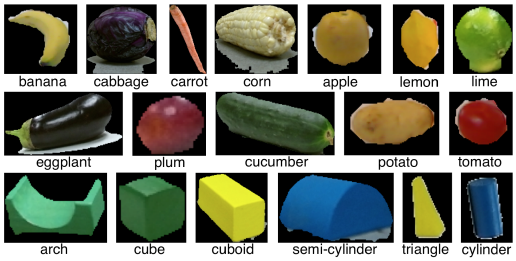

Figure 2: Sample RGB images for each class in the dataset, as taken with a Kinect2 camera and presented to Mechanical Turk annotators.

tive documents and pass it through an activation function to learn how important the term is to that document. This function is currently thresholding; in future, we plan to experiment with more sophisticated context-aware functions. Important terms are then used as labels for visual classifiers (see Figure 3 for examples). Varying this threshold affects the precision of this selection process (see Experimental Results).

For each term, all images that have been described using that term become positive examples for training a classifier. From the original 19,947 words used to describe 72 objects, 230 words were selected as tokens for classifier training. This process successfully screens out words that are used frequently when people are asked to describe objects, but that have poor discriminative or semantic power (such as 'picture', 'look', or 'image').

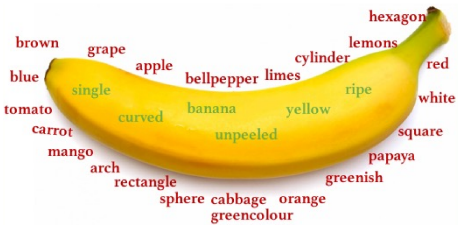

Figure 3: Selected and discarded terms after tf-idf. Terms above the threshold (green) will name a classifier that uses this object as a training example; terms below the threshold (red) will not.

Finding Negative Examples for Concepts

We are building a world model in which both the words being used and the concepts they are describing are initially unknown. Once a set of images has been selected as positive training examples, the next step is to find dissimilar objects in the corpus to serve as negative examples. This presents a bootstrapping problem: counterexamples are critical to efficient learning of word meanings (Elkan 2001) for a new term, but no classifier has yet been trained to automatically select negative examples. However, we expect that the descriptions of similar objects will be semantically similar.

A Paragraph Vector model is used to find the semantic distance between descriptions in vector space, which can then be treated as reflective of dissimilarity between objects in the world. All descriptions of each object are concatenated into an unordered "document," from which a PV is generated. The cosine similarity of these PVs then serves as a distance metric (Figure 4). From a matrix of all cosine similarities, we choose the objects with the most semantically dissimilar descriptions as negative training data. Our experimental results validate this approach.

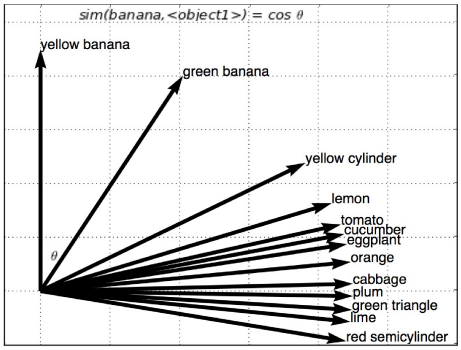

Figure 4: Cosine similarity of the Paragraph Vectors of descriptive documents for a single banana in our dataset vs. selected other objects. Each PV represents an individual object in the dataset.

Classifier Learning

We first select terms for which to create classifiers, as described above. For the perceptual learning problem, we use RGB and RGB-D images of

objects. We extract RGB features from the color channel and use kernel descriptors (Bo et al. 2011; Lai et al. 2013) to extract shape and object features from the depth channel. Kernel descriptors model size, 3D shape, and depth edge from the depth channel, and experiments show that it significantly enhances the quality of object classification results.

To test the effectiveness of our approach, we use three different types of classifiers: color, shape, and type of object. The first two are suitable for the current problem and have been used in previous work on this topic, (Pillai, Budhraja, and Matuszek 2016) while object type classifiers demonstrate the possibility of learning more complex concepts. Because an unsupervised learner has no way of knowing which of these categories a word actually refers to, it is necessary to train multiple classifiers for each term, one of each type (Matuszek* et al. 2013). All objects that were described with that term are used as positive examples. Training is performed using logistic regression.

Experimental Results

In this section, we present experiments testing each stage of the learning pipeline: selecting semantically meaningful words, finding negative training data, and the quality of the final trained classifiers.

Selecting Terms

To evaluate our approach to finding semantically meaningful words, we compared the results to ground truth provided by human annotators. All unique words in the data set were given to two annotators to categorize as 'Visually meaningful' or 'Not meaningful.'[1] Figure 5 shows precision and recall as the tf-idf threshold used for term selection is adjusted. Our method gives promising results in determining the significance of words for which to learn visual groundings.

Discussion: As presented, this method selects preferentially for precision, i.e., reliably returns semantically meaningful terms at the cost of thoroughness. This is appropriate; as classifiers trained on visually uninformative words will show poor predictive power and can be screened later, the purpose of the term selection is to focus the learning effort on the most promising terms.

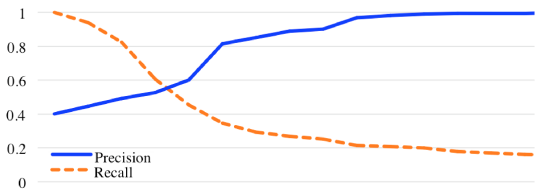

Figure 5: Precision (blue) and recall (orange) of term selection as the tf-idf threshold is varied.

Negative Example Selection

One of our primary contributions is a distance metric for perceptual training data based entirely on paired, novel language. Using the Paragraph Vector model addresses a major failing in the simpler bag-of-words model: it considers the ordering and semantics of words, but still allows vector-space-based comparisons. We treat the cosine distance between the Paragraph Vectors as an implicit distance in the grounding space (see Figure 4). Images of the most distant objects can then be used as negative samples for training the visual classifier (see Figure 1 for examples).

As the "similarity" of objects is highly contextual, ground truth for this distance metric is not clearly defined. We approximate ground truth by using the Amazon Mechanical Turk (AMT) infrastructure to ask people for evaluations of object similarity. Because asking for a complete ordering of objects in the dataset is impractical, we tested a subset of cases, asking five annotators to decide which of two objects was most similar to another. We presented 360 comparisons of the 72 objects in our dataset to five different evaluators for a total of 1800 comparisons. A simple majority of annotators agree with our similarity metric in 84% of cases. Figure 6 shows examples of the results.

Discussion: Our paragraph vector model is generally able to select good negative samples from the corpus, according to comparison with human evaluators. Visual classifiers trained using these negative samples outperform baseline classifiers trained using random sampling from the dataset. A more

[1]For ease of annotation, the choices 'Hard to say' and 'Not a word' were provided, but were selected too infrequently to affect results.

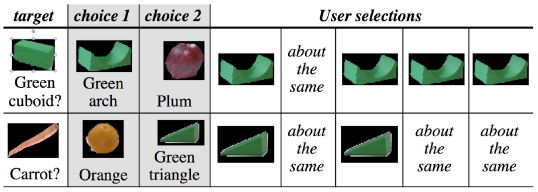

Figure 6: Examples of AMT similarity results. Five participants select which of two choices was more similar to a target object. In the first row, most users selected the green arch; the second row shows a less clear case.

complex evaluation of similarity with better defined parameters might be appropriate in the future; for example, some users never considered color when designating similarity, while others clearly based their decisions on whether something was food or not. These are informed and reasonable aspects of similarity, but did not always align with the visual classifier training problem.

End-to-End Quality of Trained Classifiers

The quality of the grounded language model—the learned model of the relationship between language and percepts—is a product of the association between language tokens and the trained visual classifiers. Ideally, attribute descriptions should be associated primarily with a single classifier with good predictive power.

As a baseline, we compared classification accuracy of the end-to-end system described in this paper with a model that chooses random negative samples and all non-overlapping samples from the data set. We used the same dataset for evaluating our method, random selection, and all other samples method. Our evaluation was conducted on our corpus of images and descriptions. Cross-validation was used for testing. As described above, we trained color, shape and object classifiers for all selected terms.

Color: Our color classification results show good results on color labels (see Figure 7). There is some overfitting resulting from the relatively small set of objects. For example, objects were frequently described as being on a white background, leading to conflation in the "white.' The "orange" and "red" classifiers overlap, in part because users described both tomatoes and carrots using both terms; in addition, polysemy had a negative impact, as the term "orange" can refer to the color or object.

One possible solution to the need for extensive annotation is using efficient active learning techniques. Previous grounded language acquisition experiments that exercise active learning techniques (Pillai, Budhraja, and Matuszek 2016) have shown promising outcomes in reducing annotation efforts without compromising classification accuracy.

Shape: Training shape classifiers on small RGB-D images is significantly more difficult than color, in part because the shape of an object from different angles can vary considerably. While still performing well, the quality of the results is somewhat less. A few sources of complication included the tendency of annotators not to describe the shape of common objects; cucumbers were frequently referred to as green, but never as cylindrical. In addition, certain terms, such as rectangular, were overused. Figure 8 shows the results of some selected shape classifiers.

Object class: Object classifiers, which are intended to determine the class an object belongs

Color classifier denoted by "term"	Ground truth: yellow	red	green	white	orange
"yellow"	0.93	0.20	0.37	0.05	0.02
"building"	0.09	0.11	0.00	0.00	0.17
"red"	0.00	0.89	0.05	0.16	0.35
"green"	0.27	0.00	0.89	0.02	0.00
"tomato"	0.24	0.94	0.00	0.00	0.00
"white"	0.06	0.68	0.55	0.85	0.73
"orange"	0.50	0.93	0.21	0.26	0.66

Figure 7: Performance of color classifiers for words (y-axis) versus ground truth (x-axis). Only a small subset of representative classifiers are shown, since one is created for each keyword in the corpus. This confusion matrix show the confidence of trained classifiers when run against objects of each type; for example, the trained model for the word "yellow" classifies the first object as positive with 93% confidence, but is only 20% confident that the second object matches. Classifiers associated with color words have strong predictive power, as does the color classifier associated with the token "tomato." The visually uninformative word "building," by contrast, is not strongly associated with a classifier.

Shape classifier denoted by "term"	Ground truth: cube	cylinder	sphere	arch	triangle
"cylinder"	0.32	0.87	0.06	0.29	0.29
"rectangular"	0.82	0.43	0.51	0.78	0.30
"circle"	0.25	0.25	0.75	0.26	0.21
"archshaped"	0.29	0.27	0.12	0.82	0.33
"triangle"	0.54	0.60	0.52	0.31	0.82

Figure 8: Performance of selected shape classifiers (x-axis) against objects (y-axis). The confusion between rectangles and arches is a product of the data, as the blocks usually described as arch-shaped have a rectangular top. This confusion matrix show the confidence of trained classifiers when run against sample objects of each type.

to, are trained using a combination of color and shape features. While our object classification has good results on our data set, this is partly due to the strong influence of color in classification; both the toys and the food objects in our data set tended to be primarily a single strong color.

Object classifier denoted by "term"	Ground Truth: corn	semi-cylinder	banana	eggplant	tomato
"corn"	0.92	0.01	0.77	0.04	0.00
"building"	0.08	0.61	0.30	0.02	0.03
"banana"	0.00	0.15	1.00	0.00	0.04
"tomato"	0.00	0.00	0.05	0.00	0.94
"wedge"	0.49	0.30	0.00	0.43	0.00
"eggplant"	0.26	0.24	0.01	0.84	0.11

Figure 9: Performance of selected object classifiers (x-axis) against objects (y-axis). This confusion matrix show the confidence of trained classifiers when run against sample objects of each type.

Overall: Our system convincingly outperforms two baseline models, one that randomly selects objects to serve as negative examples, and one using all other objects as negative examples (see Figure 10), demonstrating improvement in the state of the art on unsupervised grounded language acquisition. A classifier trained with all other samples as negative data performs well, while random sampling performs almost as well in most cases but represents a fair comparison in terms of training time and resource.

The overall goal of this work is to allow agents to improve their ability to learn semantic representations of their perceived environments, using natural language as the training signal. While not a complete metric, one way of considering whether this work makes progress towards that goal is to verify that the most obvious terms for the intended ground truth have been identified as having important semantic relevance, and how accurately the classifiers associated with those terms perform on the complete dataset. By this metric, we find that all of our ground truth labels have been discovered; classifier performance is shown in Figure 11.

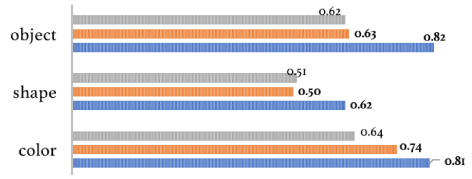

Figure 10: Average performance of color, shape, and object classifiers. Negative data is selected randomly (red), by using all non-overlapping objects (gray), and using our dissimilarity measure. Using meaningful negative examples improves performance in every category.

Conclusion and Future Work

While a number of different approaches have explored how to acquire semantic representations of perceptual data, the need for automated selection of learning targets and, especially, negative natural language exemplars recurs throughout the literature. Our results demonstrate that statistical tools from natural language can be applied to corpora of mixed language and perceptual data, automatically identifying terms that should be considered as candidates for learning groundings and selecting negative examples automatically for training classifiers. This reduces the need for human supervision, allowing language-learning agents to learn end-to-end in an unsupervised fashion, from collecting data to fully trained grounded language models.

An evaluation of our process for finding meaningful words and selecting negative examples suggests that these approaches are effective. These results illustrate the performance and effectiveness of the classification model by comparing it with two baselines, either randomly selecting negative samples or using all non-positive examples as negatives. In future, our intention is to extend this

work to a more varied set of objects, additional kinds of classifiers, and complex visual classification tasks, as well as to apply the identification of negative grounding examples to ongoing work on grounded language acquisition for robotics.

We use the word-as-classifier approach because, while is a simplification of the language problem, it is an applicable starting point for the robotic language understanding task applied to noisy perceptual data. This language model is preliminary, and we intend to extend this to more semantically driven and context-sensitive model in future. We also hope to extend this research to a conversational agent. In a conversation-based interaction, the system will have the opportunity to ask for negative examples explicitly, which we hope will improve results. The approach in this paper would then be useful to reduce the number of (possibly repetitive) questions and to enhance the quality of the dialogue.

blue:	0.995	arch:	0.532
green:	0.947	cube:	0.590
orange:	0.720	cylinder:	0.725
purple:	0.499	rectangle:	0.621
red:	0.844	triangle:	0.649
white:	0.772		
yellow:	0.918		
banana:	0.942	lemon:	0.777
cabbage:	0.879	lime:	0.936
carrot:	0.887	orange:	0.921
corn:	0.922	potato:	0.715
cucumber:	0.615	tomato:	0.926
eggplant:	0.646		

Figure 11: Average cross-validation performance of classifiers for words. In general, color classifiers (top left) perform best; the outlier, purple, reflects the color differences between the objects described as purple (typically eggplants, red cabbage, and plums). Classifiers for object types (bottom left and right) perform well in general. Shape classifiers (top right) perform worst, stemming from the fact that people do not provide a shape description as often as the other classes.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 1657469.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

BenAbdallah, J.; Caicedo, J. C.; Gonzalez, F. A.; and Nasraoui, O. 2010. Multimodal image annotation using non-negative matrix factorization. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE.

Berg, T.; Berg, A.; and Shih, J. 2010. Automatic attribute discovery and characterization from noisy web data. *Computer Vision–ECCV 2010*.

Bo, L.; Lai, K.; Ren, X.; and Fox, D. 2011. Object recognition with hierarchical kernel descriptors. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Bowerman, M. 1988. The'no negative evidence'problem: How do children avoid constructing an overly general grammar? In *Explaining language universals*.

Chen, X., and Lawrence Zitnick, C. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chrupala, G.; Gelderloos, L.; and Alishahi, A. 2017. Representations of language in a model of visually grounded speech signal. In *Association for Computational Linguistics*.

Dindo, H., and Zambuto, D. 2010. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *The IEEE International Conference on Computer Vision (ICCV)*.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*. Lawrence Erlbaum Associates Ltd.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *Computer Vision and Pattern Recognition*. IEEE.

Hastings, P. M., and Lytinen, S. L. 1994. The ups and downs of lexical acquisition. In *Twelfth AAAI National Conference on Artificial Intelligence*. AAAI Press.

Knepper, R. A.; Tellex, S.; Li, A.; Roy, N.; and Rus, D. 2015. Recovering from failure by asking for help. *Autonomous Robots*.

Kojima, A.; Tamura, T.; and Fukunaga, K. 2002a. Natural language description of human activities from video

images based on concept hierarchy of actions. *International Journal of Computer Vision*.

Kojima, A.; Tamura, T.; and Fukunaga, K. 2002b. Textual description of human activities by tracking head and hand motions. In *Pattern Recognition*. IEEE.

Lai, K.; Bo, L.; Ren, X.; and Fox, D. 2013. RGB-D object recognition: Features, algorithms, and a large scale benchmark. In Fossati, A.; Gall, J.; Grabner, H.; Ren, X.; and Konolige, K., eds., *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*.

Lasnik, H. 1989. On certain substitutes for negative data. In *Learnability and linguistic theory*.

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

Li, J., and Wang, J. Z. 2003. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on pattern analysis and machine intelligence*.

Li, J., and Wang, J. Z. 2005. Alip: the automatic linguistic indexing of pictures system. In *Computer Vision and Pattern Recognition*. IEEE.

Lienhart, R.; Romberg, S.; and Hörster, E. 2009. Multilayer plsa for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM.

Matuszek*, C.; FitzGerald*, N.; Zettlemoyer, L.; Bo, L.; and Fox, D. 2013. A Joint Model of Language and Perception for Grounded Attribute Learning. In *29th International Conference on Machine Learning (ICML)*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. In *ICLR Workshop*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Pillai, N.; Budhraja, K. K.; and Matuszek, C. 2016. Improving grounded language acquisition efficiency using interactive labeling. In *Robotics: Science and Systems workshop on Model Learning for Human-Robot Communication*.

Roy, D. K. 2002. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*.

Salton, G., and McGill, M. J. 1983. *Introduction to modern information retrieval*. New York: McGraw - Hill Book Company.

Schenck, C., and Fox, D. 2017. Towards learning to perceive and reason about liquids. In *Proceedings of the International Symposium on Experimental Robotics (ISER)*.

Silberer, C.; Ferrari, V.; and Lapata, M. 2016. Visually grounded meaning representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Socher, R.; Karpathy, A.; Le, Q.; Manning, C.; and Ng, A. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*.

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

Tellex, S.; Thaker, P.; Joseph, J.; and Roy, N. 2013. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning*.

Tellex, S.; Knepper, R. A.; Li, A.; Rus, D.; and Roy, N. 2014. Asking for help using inverse semantics. In *Robotics: Science and systems*.

Thomason, J. 2016. 1. continuously improving natural language understanding for robotic systems through semantic parsing, dialog, and multi-modal perception.

Yang, Y.; Teo, C. L.; Daumé, III, H.; and Aloimonos, Y. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yu, C., and Ballard, D. H. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)*.

Zobel, J., and Moffat, A. 1998. Exploring the similarity space. In *ACM SIGIR Forum*. ACM.