2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Learning Object Attributes with Category-Free Grounded Language from Deep Featurization

Luke E. Richards,[1,2] Kasra Darvish,[1] Cynthia Matuszek[1]

***Abstract*— While *grounded language learning,* or learning the meaning of language with respect to the physical world in which a robot operates, is a major area in human-robot interaction studies, most research occurs in closed worlds or domain-constrained settings. We present a system in which language is grounded in visual percepts without using categorical constraints by combining CNN-based visual featurization with natural language labels. We demonstrate results comparable to those achieved using handcrafted features for specific traits, a step towards moving language grounding into the space of fully open world recognition.**

I. Introduction

As robots become more capable and ubiquitous, they are increasingly moving into traditionally human-centric environments such as medical spaces, workplaces, and homes. In these environments, interacting with and learning from people is a key component of intuitive, natural human-robot interaction. Natural language is a powerful, widely understood way of conveying instructions and information. However, pre-defining language models for all possible tasks and objects in a dynamic human environment is infeasible, especially in *grounded language,* where unfamiliar language may refer to anything in a robot's perceptual world.

Grounded language learning and understanding has been an active research area in recent years. It is worth noting that 'grounded language' has different meanings in each of natural language, vision, and robotics research. In this paper, we refer to the understanding of human language in the context of the limited, noisy, and idiosyncratic perceptual data received by a physical agent. This task is qualitatively different from that of grounded language learned from large corpora of fixed images that incorporate only vision, or from learning language grounded in the virtual space of purely linguistic phenomena. In this work, models of the environment and language semantics are jointly learned in order to understand language in the context of a robot's sensed environment.

Significant progress has been demonstrated in robots that learn from language in tasks as diverse as understanding pick-and-place [1], kitchen [2] commands, and learning objects characteristics from unconstrained language [3] or via dialog [4]. Although there has been substantial effort on reducing and focusing the supervisory signal required for such learning, the majority of existing work still operates on pre-defined categories in which new concepts can occur—e.g., learning shapes or colors, learning the meaning of a pre-defined set of attributes such as weight, or learning to understand action commands in a context with a tightly constrained set of elements such as the kitchen domain.

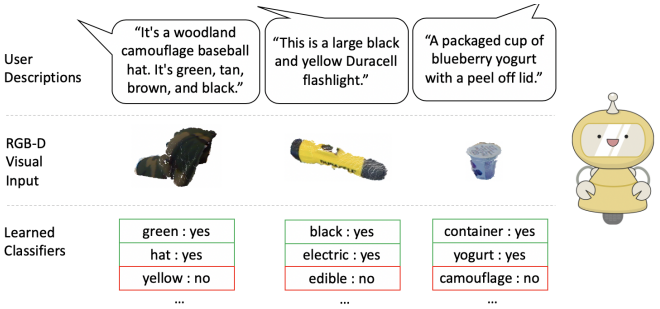

Fig. 1: Examples of object descriptions from Amazon Mechanical Turk workers above RGB point clouds. The robot then learns classifiers for each concept and whether they relate to the given visual input. Note the correct application of concepts not explicitly present in class labels, e.g., *electric*, which is learned solely from user descriptions.

This is a significant limitation. Adaptable robots should to be able to learn to understand previously unconsidered *categories* of things in the world, and not merely previously unseen members or combinations of those categories. We explore how such constraints may be relaxed, using convolutional neural network (CNN)-based visual feature understanding paired with a joint-probability grounded language model to learn language from user given descriptions to household objects without defining categories of language that can be learned. This work fits broadly in the category of open world recognition [5, 6], in which the set of objects present in the world is assumed to be unknown.

In this work, we demonstrate that it is possible to relax constraints that define the space of possible learned groundings in real, physical sensor data by using multimodal deep learning to understand language referring to an environment regardless of category. We refer to the method presented here as **category-free language learning** [7], in order to differentiate it both from approaches that learn into previously delineated spaces, and from fully open world learning.

Our primary contributions are twofold. First, we demonstrate the effectiveness of learning *concepts* from novel *language* descriptions. The system presented is able to learn descriptive concepts that are not explicit in the training data, but that human descriptions convey (e.g., the idea that an object is made of a particular substance, e.g., ceramic or aluminum). Second, we show that deep learning-driven

[1]University of Maryland, Baltimore County
[2]Booz Allen Hamilton
lurich1, kasra.darvish, cmat@umbc.edu

featurization is effective for perceptual open world language grounding, yielding results that are comparable to those obtained by learning language in specific, pre-existing attribute and object categories. In addition, we make the annotated UW-RGBD dataset available, providing a source for other efforts in this domain.[1]

II. Related Work

This work builds on substantial existing research on language grounding in robotics, deep learning, learning from scarce data, and open world learning. Here, we describe some of the most closely related efforts.

Open World Recognition and Zero-Shot Learning Our work is closely related to open world recognition and zero-shot learning. Open world recognition [6] and open set recognition [5] describe the concept of learning non-predefined concepts in an open world. In this work, we learn classifiers for novel classes not considered in the original vision model's categorizations, based on the natural language descriptions provided by human users. However, we differentiate our work from both. In most approaches to zero-shot learning, the mapping between raw-input and descriptions (semantic feature space) is learned, and then using a semantic knowledge base the unseen class can be labeled [8, 9]. In our work we assume no knowledge base except noisy natural language given by human users.

Deep Learning With recent advancements in multi-layer (deep) neural network in machine learning, it is a natural advancement to examine how these methods can apply to language grounding. Work in this space varies from multimodal approaches in which parallel images and language datasets are treated as an alignment problem [10], to image caption generation tasks [11], to question-answering given complex natural language queries and images [12]. In the specific space of deep learning for robotic language grounding, there auto-encoders have been used to learn the semantic meanings of words in a single-classifier approach by inputting multiple feature sets [7]. Our work extends [13], in which we outlined using deep learning vision models to obtain rich features for grounded language learning.

Few-Shot Learning Much of machine learning operates under the assumption that there will be copious data to learn from. However, this is rarely the case for robotics. Few-shot learning offers methods in which classes are learned with few examples [14]. In our work, we have instances where a single positive instance of a class is seen during training. This is particularly relevant in natural language descriptions, where a word or concept may only be mentioned once but still be critical for understanding. Training state-of-the-art object recognition models is not always plausible without large sets examples. Some methods try to utilize rich embedding spaces to curtail this issue [15, 16, 17, 18]. However, many of these works deal with RGB images, rather than RGB-D where rich natural language labels are less abundant.

[1] https://iral.cs.umbc.edu/datasets

RGB-D Object Recognition and Analysis Object recognition is a wide-ranging area of research. In our work we focus on a multimodal combination of RGB and depth images. RGB-D data is a typical perceptual source in robotics, and there has been substantial work demonstrating that depth significantly aids learning, especially in knowledge-constrained settings [19]. In this section we briefly describe the related work in vision using RGB-D data. Early methods of RGB-D object recognition focused on extracting features for separate categories such as gradient, color, and shape [20]. Extracting features with hierarchical matching pursuit (HMP) introduces an unsupervised feature extraction network, which allows learning models of high-level features through a layered approach to combining RGB and depth images [21]. This was early work in obtaining visual features that would be descriptive of the entire object.

Transfer learning combined with deep learning has been a major catalyst in the success of computer vision tasks by introducing *transferable layers* in vision models. These initial layers learn fundamental steps in visual perception and can be applied to other vision tasks. This concept has been particularly popular due to the success of the large-scale dataset ImageNet [22] being used to learn generalized concepts between vision tasks. While the ImageNet dataset [22] has furthered work in RGB object recognition, the RGB-D vision community is still exploring ways to transfer such transfer learned models to depth effectively.

Additionally, the concept of a neural network that can be trained on a low cost single graphics processing unit (GPU) that can be shared and trained, CaffeNet [23], led to even more success in the space. Eitel et al. [24] introduced a method that combines the benefits of transfer learned RGB models to both RGB and depth images for object recognition and the practicality of a lower compute neural network model. In our work, we use this joint transfer learned RGB and depth model paradigm to extract visual features (see III-B for details).

Grounded Language Learning Language is used to communicate about, refer to, and describe the physical world, so the use of robots in learning to comprehend and use natural language is intuitive. Grounded language learning is the concept of learning the groundings of language to perception [25, 26]. This has been an active area in the intersection of natural language processing and vision communities. Work such as image caption generation and recognition [27, 28] and text-to-image synthesis [29] showcase this overlapping interest.

When this interest moves into the physical world using robotic agents, the perceptual space that language can be grounded to increases. Language can be grounded to manipulation tasks [30, 31], navigation tasks [32, 33], and assistive robotics [34], among others. In all these tasks, there is a need to understand the referent language (nouns, adjectives, and more) that aligns with objects in physical spaces. Typically, the grounding of the rich language humans use for individual objects is abstracted in robotic learning environments to sparse labels such as color, shape, and object name.

Work in learning language models for color, shape, object, haptics, and sound with predefined unique feature channels [35, 36, 37] have resulted in successful groundings. However, our work explores using a set of general features to learn groundings outside of predefined feature channels. This allows our robotic learner to expand upon a single source of features derived through a neural network to ground language that ties to other visual phenomena such as texture, symbols, size, and complex color, to name a few. This work and previous work [7] classify this single classifier system as category-free learning. While previous work [7] has utilized methods to take category-based visual features and learn category-specific classifiers for each token, we use non-category-based visual features to learn a single classifier.

While datasets of rich natural language aligned with images [38, 39] are increasing, to our knowledge no such dataset exists for RGB images with depth (RGB-D) [40]. RGB-D data creates a new learning paradigm for grounded language learning that includes physical proxy. Our dataset is similar to [41] by aligning RGB-D sensor data with user assigned attributes; however, in that work the descriptions were limited to strict categories (color, shape, object, material). Our work uses the same sensor dataset, but we learn from full-sentence natural language rather than single word attributes in categories. Our hope is this work will assist in the effort to curate rich language datasets aligned with robotic data to benchmark further research within the community and encourage further complexity to be augmented into other grounded language tasks.

III. Approach

In this section, we introduce a novel dataset of language descriptions, collected to augment the widely used RGB-D objects dataset [42]. We describe how annotations were collected using crowdsourcing (see fig. 2 for details), then describe how we modify the robust RGB-D object identification approach of Eitel et al. [24] to extract visual features. Finally, we outline how we use a novel joint learning objective that combines both visual and language features to train word-as-classifier models for linguistic concepts (see fig. 5 for an overview). This system learns how a person might refers objects in at a semantic level, leaving low-level classification and action details to a hypothetical robot assistant.

A. Data Corpus

We extend the well-known UW RGB-D object recognition dataset [42, 43], which includes roughly 40,000 RGB-D images of 300 objects in 51 categories. This dataset includes point clouds as well as RGB and depth images and masks. We select five images of each object instance using stratified random sampling, giving us a sample of 1,500 RGB-D images with an unfixed collection of angles from each object. Images were then uploaded to Amazon Mechanical Turk, where workers gave short descriptions of each object as if they were speaking to another person (see fig. 2 for the instructions given to workers). In order to obtain diverse descriptions, we avoided using language that would 'prime'

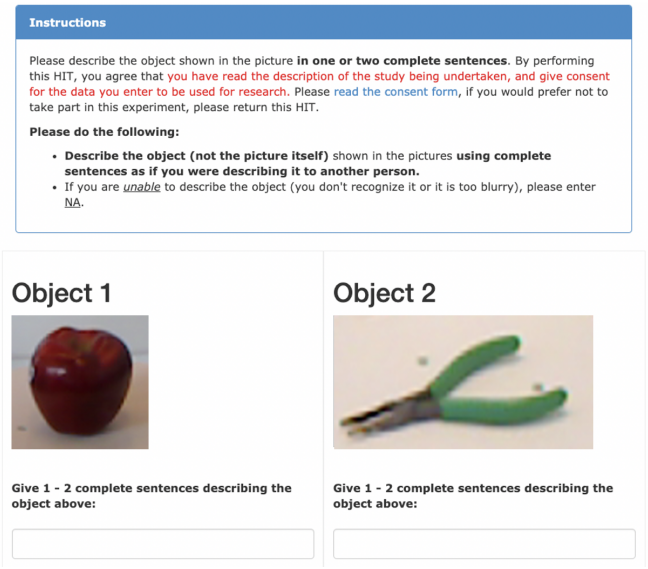

Fig. 2: Example of the Amazon Mechanical Turk Human Intelligence Task (HIT). We asked workers to provide complete sentences of the cropped RGB image, or "NA" if they were unable to recognize the cropped image. Each HIT contained images of 5 objects.

workers to describe objects in a particular way. Workers were only encouraged not to describe the picture itself (such as "the photo is blurry" or "the photo has a red cap").

114 unique workers participated in the study, an average of doing about 14 tasks (70 images) per worker. All workers were located in the United States. monitored results in real time, allowing for quality screening of work being submitted. This increasingly allowed workers to continue giving high-quality descriptions while discouraging subpar or non-conforming work. We found small amounts of noise in mistyping, spelling, and grammatical issues. While this creates a more challenging learning problem, it also adds noise that would be present in everyday human-robot language interactions.

We obtained a total of 8,186 raw object descriptions. While some workers provided incomplete sentences or described the photo rather than the object, the majority gave rich contextual language about the images. Descriptions which clearly did not follow instructions or in which the worker was unable to provide a description due to visual ambiguity were removed, yielding a total of 7,455 complete sentence descriptions of 300 objects in 51 categories, or almost 25 descriptions per object. There were 4 to 8 complete sentence description responses per image of the object instance, unique form of an object (see fig. 3 for a breakdown of responses for each object). These descriptions contained 59,998 total tokens, 23,366 tokens with stop words and punctuation removed, and 1,965 unique words. On average there were 7.32 tokens per description.

Nouns and adjectives are widely used by people in everyday task commands, and as such, are valuable in the language

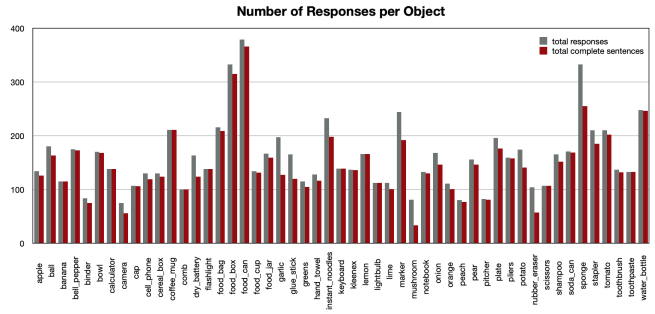

Fig. 3: A breakdown of the number of responses received for each of the objects versus the number that were complete sentence descriptions. While there were more instances in the dataset for some objects due to the distribution of the original dataset, this shows where Amazon Mechanical Turk workers faced ambiguity of what the image was showing and opted out of giving a description.

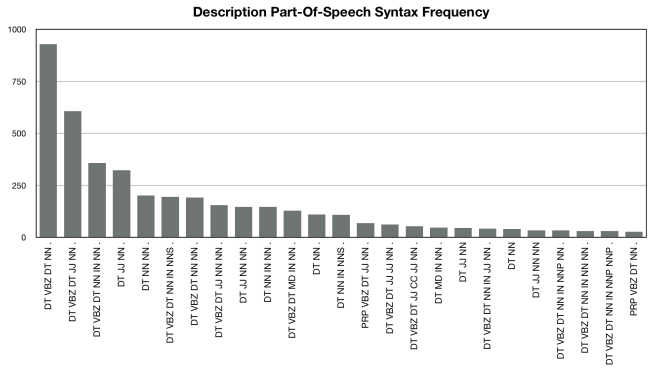

Fig. 4: The syntax of the descriptions was analyzed using Stanford Part-of-Speech tagging [44] and the format of the structures were then totaled. The most common structure was phrases like "This is a <noun>," followed by "This is a <adjective><noun>".

grounding task. In this corpus, 100% of the complete sentence descriptions have a noun present. 3,733 descriptions, or 45.6% of the dataset, also contain an adjective(s) according to the Stanford Part-of-Speech parser [44]. The frequency of description part-of-speech syntax is shown in fig. 4. The syntactic structures of the descriptions as well show variety and diversity as users were encouraged to describe the objects without any priming and as naturally as possible. While there a high percentage of workers who used short phrases, other users were more descriptive and thorough. This gives a diverse dataset of natural language for human-robot interaction in describing objects. To the best of our knowledge, this is the only openly available RGB-D dataset aligned with rich natural language.

B. Visual Features

Our approach to extracting visual features is drawn from the robust object recognition method of Eitel et al. [24]; however, in our results, we demonstrate that with minor modifications, this approach can be used to extract features suitable for understanding a user's high-level language in a variety of unspecified categories (see fig. 1). While more advanced and powerful vision models have been developed in recent years for RGB-D object recognition [45], many require large amounts of compute resources to train. We argue our methodology can be applied to these more advanced models with respective increase in performance. However, in this work, we explore the viability of using neural network features for robotic grounded language learning in the context of the word-as-classifiers, rather than the development and testing of advanced RGB-D vision models.

Broadly, artificial neural networks (ANNs) allow for high dimensional inputs to be condensed to meaningful representations of features in the data. Due to the nature of neural networks, the final layer offers high-level features for the objects. This is true of convolutional neural networks (CNNs), especially for extracting useful features in object recognition tasks [46]. While many employ a softmax function to perform classification, such as Eitel et al. [24], removing the softmax function exposes rich features that can be used directly for grounded language learning.

The network consists of two seven layer CNNs, one per sensor type (RGB and depth), that combine into two final fusion layers. The final layer of the network allows for 51 features to be extracted from the joint networks of the RGB and depth images. We extract these features for each of the 1,500 images sampled from the RGB-D object dataset paired with the natural language descriptions given from workers on Amazon Mechanical Turk. Once the visual features and language is paired, we start the grounded language model learning process.

C. Category-Free Joint Language Learning

In order to provide a meaningful evaluation of our work in the context of language grounding, we evaluate on an extension of the grounded language learning system of Pillai and Matuszek [3]. The basis of this work is a joint model combining perception and language models [36] to learn completely novel language groundings. For this model, groundings are learned solely from dataset. No prior representations of the objects or the language are required. In that comparison work, a words-as-classifiers approach is taken, meaning each token has a single binary classifier trained to predict whether an object is described by that word. In our infrastructure, *"red"-as-classifier* would classify a red apple and a green apple as positive and negative, while *"apple"-as-classifier* would classify both as positive.

The word-as-classifiers concept increases the number of concepts that can be learned. Meanwhile, a naïve end-to-end deep learning approach would require an extensive number of output classes—one for each word—as well as requiring retraining of the final layer. As well, with our negative example selection, there are cases in which an instance is neither a positive or negative example for a token (that is,

objects are not positively described as a concept are not necessarily counterexamples).

While previous work was constrained by the domains of language that could be learned (such as shape or object type), we seek to use multipurpose visual features from the method described in III-B to introduce end-user abstraction. Our system therefore learns from a single source of features rather than separate domain-specific sets such as those in previous work, which learned from hand-engineered average RGB channel values to, for example, encode color and depth kernel descriptors for learning shapes. Another change is learning language per *image*, rather than learning from a larger collection of descriptions of an object *instance*. This preserves visual differences in the object, such as orientation and appearance under different lighting, so the system has the opportunity to learn language relevant to only some images of a single object. (For example, from some angles the baseball cap in fig. 1 might look like a hemisphere, whereas in the image shown it does not.)

We aggregate all descriptions given by workers of each image, creating a more exhaustive 'descriptive document' for each image. These descriptive documents are processed to remove singletons and stop words. Visual features (extracted per image) are paired with documents aligned with that image. We consider an instance as a positive training example of the token if that token occurs more than once in the description document. Due to the non-exhaustive nature of natural language, determining which examples can be considered negative is an ongoing research area. We use the negative exemplar method of [3], which is unsupervised (in the sense negatives are not explicitly given. In this approach, a distributional semantic model is trained on language descriptions for each image, and then the cosine distance between two description document vectors is used to determine negative examples. Each token classifier is trained using logistic regression.

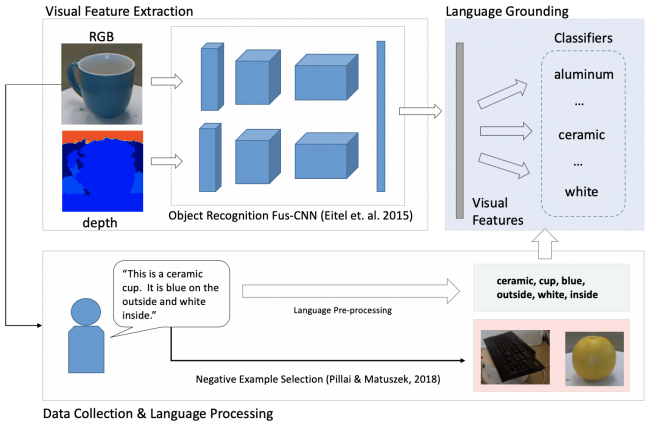

Fig. 5: Our proposed domain-free model using the visual features from object recognition system [24], creating word-as-classifier models. This method fuses two CNN architectures for RGB and depth images into fully connected fusion layers. We remove the softmax layer from their approach, exposing rich multimodal features for learning groundings.

IV. Experiment

We train our word-as-classifier models (see section III-A) using four-fold cross validation. We report the results of these tests averaged over 200 runs with random splits. First, we use embeddings extracted from the convolutional neural network (CNN) for both depth and RGB image inputs, and examine treating the layer before softmax as an embedding layer of 51-dimensional embedding. Previous work has shown value in various layers of a convolutional neural network providing embeddings that perform certain tasks better than later layer embeddings [47]. In using the multi-modal CNN as an embedding network, we examine using the second-to-last fully-connected layer as the embedding to examine whether previous layers with higher dimensionality exploit more generalized embeddings compared to the layer before the softmax for the language learning task.

For this penultimate layer, we run dimensionality reduction by using singular value decomposition (SVD) to reduce the embedding's dimensionality from 4,096 to 150 to create a computationally feasible learning space for the logistic regression model, which still preserves a higher dimensionality than the layer before the softmax. While in the era of end-to-end pipelines this reduction would be performed through training a final linear layer, our work utilizes a pre-trained network as a static visual feature extraction process. We report the results compared to features extracted from the final layer, which is 51-dimensional (see table I).

We examine our method in relation to a previous approach that uses feature extraction specifically oriented to learning color, shape, and object word-as-classifiers [3]. In that work, 'color' is featurized using k-means centroids, 'shape' features are extracted from depth kernel descriptors [48], and 'object' features are a vector in which color and shape features are concatenated. Each token then has three separate classifiers for each respective category. Additional classifiers are generated that encompass the hypothesis that a new token may be a synonym for a previously-encountered concept. This leads to a large set of classifiers that must be trained in tandem as new data is observed.

V. Results

Each classifier was trained on a median of two examples, and the maximum number of positive examples was 149, while the minimum was one (see fig. 6). This positions our work in the space of few-shot learning, as we attempt to classify labels with majority of the classifiers learning on less than two positive labels at training. For the classifiers trained with the category-free CNN features, the average F1-score is 0.68, with a standard deviation of 0.059. Meanwhile the best result from the previous method's work was the object-category classifier with an average F1-score of 0.668 with a standard deviation of 0.061. This shows that our method, while not drastically improving on the performance of the language grounding system, accomplished similar results

Features	F1-Score	Recall	Precision
CNN_51	0.6801	**0.6047**	0.8843
CNN_SVD_150	**0.6837**	0.6027	**0.8920**
Object	0.6675	0.6037	0.8602
Shape	0.6625	0.5990	0.8562
Color	0.4349	0.4722	0.5033

TABLE I: Average F1-Scores for classifiers trained on color (simple k-means), shape (depth kernel descriptors [48]), object (concatenation of shape and color features), or our proposed category-free CNN_51, and CNN_SVD_150

with a vastly less constrained set of possible learned features. This demonstrates that the shift to category-free, non-hand-engineered features matches or exceeds the performance of more specialized approaches. Additionally, for each token, there is only one classifier being trained rather than several, saving on compute time, as well as eliminating the need to re-engineer category features for varied datasets.

Qualitative analysis provides the following insights. We see drastic performance decreases from the color-category classifiers as compared to our category-free classifiers; the average F1-score is the lowest of all four types of classifiers, 0.435. This is likely due to the fact that datasets used in previous work [3] used very homogeneously colored objects. The complexity of household objects' color cannot be accurately featurized using k-means averaging of RGB channels. More specifically, we note that while CNN features do a better job at detecting colors even compared to color features, some of them were close. The 'yellow' token in this work has an average f1 score of 0.88, compared to 0.81 for previous work. In the hand-engineered feature space, even shape features do a better job of detecting colors in five cases: brown, white, black, pink, and blue. This reflects overfitting of shape features to specific cases in the household objects data. The tokens 'kettle,' 'mobile,' and 'pot' are some of the poorest-performing classifiers in this work, with the average f1 scores of 0.57, 0.55, and 0.57 respectively.

Some tokens in our dataset do not described grounded objects and instead represent more abstract concepts, such as "appears," "built," "image," and "called." These tokens are ambiguous and difficult. When removing token classifiers that were unsuccessful groundings results in an average F1 score of 0.7310. We report these scores help to examine failures and ambiguity in the methodology and system.

Finally, the result of comparison between 51-dimensional embedding (CNN_51) resulted from the layer before the softmax of CNN, and 150-dimensional embedding (CNN_SVD_150) resulted by applying SVD to reduce dimensionality on the second-to-last layer of CNN suggests that we gain 1% improvement in precision, and slightly better F1-score (0.6801 for CNN_51, and 0.6837 for CNN_SVD_150). This sheds light on the plausibility of earlier layers being relevant to grounded language learning as they may be more visually generalized beyond class-based classification. Both findings support that these features are broad enough to be used in the context of single source feature category-free grounded language learning.

Fig. 6: Distribution of the average number of positive examples each classifier is trained on. This showcases the nature of the problem with the majority of classifiers being learned on 2 positive examples or less.

VI. Conclusion

We present a grounded language learning system suitable for supporting user-specific, language-based human-robot interfaces. We employ the well-known object recognition system of Eitel et al. [24] to extract rich visual features for an intuitive, category-free joint model grounded language learning system, and introduce a dataset of natural language aligned with a popular real-world sensor dataset. A series of classifiers denoted by descriptions are trained and evaluated on a held-out data set. Our results support the theory that category-free language learning is both feasible and desirable. We outline that our method, relying on a single model's embedding output, performs just as well if not better than previous work that relied on hand-engineered feature representations from multiple models concatenated.

In future work, we intend to explore more sophisticated language models, using semantic parsing to further the information provided from the natural language descriptions. We plan to use the insights from this work in exploring multimodal object embeddings in pursuit of furthering work in category-free grounded language learning. We then plan to deploy this system on a mobile robot in a human-robot user study to collect more real world data to enrich the parallel dataset. Our hope is that this work will encourage the development of robust and effective robotic systems utilizing natural language.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. 1637614, 1657469, and 1637937.

References

[1] Muhannad Al-Omari, Paul Duckworth, David C Hogg, and Anthony G Cohn. Natural language acquisition

and grounding for embodied robotic systems. In *Proceedings of the 31st National Conference on Artificial Intelligence (AAAI)*, pages 4349–4356, 2017.

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

[3] Nisha Pillai and Cynthia Matuszek. Unsupervised selection of negative examples for grounded language learning. In *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*, New Orleans, USA, 2018.

[4] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond Mooney. Jointly improving parsing and perception for natural language commands through human-robot dialog. *Journal of Artificial Intelligence Research*, 67:327–374, 2020.

[5] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[6] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Computer Vision and Pattern Recognition*, pages 1893–1902, 2015.

[7] Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. Deep learning for category-free grounded language acquisition. In *Proc. of the NAACL Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics (NAACL-SpLU-RoboNLP)*, Minneapolis, MI, USA, June 2019.

[8] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

[9] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.

[10] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 35–44. ACM, 2018.

[11] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *Computer Vision and Pattern Recognition*, June 2015.

[12] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 361–369. Curran Associates, Inc., 2016.

[13] Luke E. Richards and Cynthia Matuszek. Learning to understand non-categorical physical language for human-robot interactions. In *Proceedings of the R:SS 2019 workshop on AI and Its Alternatives in Assistive and Collaborative Robotics (RSS: AI+ACR)*, Freiburg, Germnany, June 2019.

[14] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[15] Yanwei Fu and Leonid Sigal. Semi-supervised vocabulary-informed learning. In *Computer Vision and Pattern Recognition*, June 2016.

[16] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018.

[17] Jose L Part and Oliver Lemon. Incremental online learning of object classes using a combination of self-organizing incremental neural networks and deep convolutional neural networks. In *Workshop on Bio-inspired Social Robot Learning in Home Scenarios (IROS), Daejeon, Korea*, 2016.

[18] Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017.

[19] Xiaofeng Ren, Liefeng Bo, and Dieter Fox. Rgb-(d) scene labeling: Features and algorithms. In *IEEE Computer Vision and Pattern Recognition*, pages 2759–2766. IEEE, 2012.

[20] Liefeng Bo, Kevin Lai, Xiaofeng Ren, and Dieter Fox. Object recognition with hierarchical kernel descriptors. In *Computer Vision and Pattern Recognition*, 2011.

[21] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *Advances in neural information processing systems*, pages 2115–2123, 2011.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[23] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[24] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin A. Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust RGB-D object recognition. *International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, 2015.

[25] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[26] David L. Chen and Raymond J. Mooney. Learning

to sportscast: a test of grounded language acquisition. In *International Conference on Machine Learning (ICML)*, 2008.

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017.

[28] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015.

[29] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan R. Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015.

[30] Achyutha Bharath Rao, Krishna Krishnan, and Hongsheng He. Learning robotic grasping strategy based on natural-language object descriptions. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 882–887. IEEE, 2018.

[31] Lanbo She and Joyce Yue Chai. Interactive learning of grounded verb semantics towards human-robot communication. In *ACL*, 2017.

[32] Cynthia Matuszek, Evan Herbst, Luke S. Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *ISER*, 2012.

[33] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.

[34] Jake Brawer, Olivier Mangin, Alessandro Roncone, Sarah Widder, and Brian Scassellati. Situated human–robot collaboration: predicting intent from grounded natural language. *International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[35] Jesse David Thomason et al. *Continually improving grounded natural language understanding through human-robot dialog*. PhD thesis, University of Texas at Austin, 2018.

[36] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June 2012.

[37] Caroline Kery, Frank Ferraro, and Cynthia Matuszek. ¿es un plátano? exploring the application of a physically grounded language acquisition system to spanish. In *Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, 2019.

[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[39] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[40] Ziyun Cai, Jungong Han, Li Liu, and Ling Shao. Rgb-d datasets using microsoft kinect or similar sensors: a survey. *Multimedia Tools and Applications*, 76(3): 4313–4355, 2017.

[41] Yuyin Sun, Liefeng Bo, and Dieter Fox. Attribute based object identification. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2096–2103. IEEE, 2013.

[42] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *2011 IEEE international conference on robotics and automation*, pages 1817–1824. IEEE, 2011.

[43] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. RGB-D object recognition: Features, algorithms, and a large scale benchmark. In *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*, pages 167–192, 2013.

[44] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for computational Linguistics, 2003.

[45] Mingliang Gao, Jun Jiang, Guofeng Zou, Vijay John, and Zheng Liu. Rgb-d-based object recognition using multimodal convolutional neural networks: A survey. *IEEE Access*, 7:43110–43136, 2019.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60:84–90, 2012.

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[48] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Depth kernel descriptors for object recognition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 821–826. IEEE, 2011.