Grounding Language in Exploratory Behaviors and Multi-Modal Perception

Zachary Osman
Tufts University
zachary.osman@tufts.edu

Abstract

Grounded language is the task of learning the meaning of natural language terms based on sensory data. Often grounded language solely leverages visual data to identify properties of objects. However, by relying only on visual data, there are features of objects that a robot would not be able to identify. For example, if a robot wanted to learn the meaning of the word "soft" based on sensory data, visual data alone would likely not be sufficient for the robot to develop an accurate understanding of the word. In order to identify non-visual properties of objects, the robot must leverage other types of sensory data, as well as a variety of behaviors.

Dataset

The dataset of objects used consisted of 100 samples that were divided into categories such as bottles, cans, cups, and balls. This dataset had previously been used for category classification, but never for language learning [Tatiya and Sinapov, 2019]. Each of these objects were assigned various descriptive words such as "hard", or "empty". In total, there were 68 different descriptive words for the 100 objects.

For each of these objects, the robot collected sensorimotor data in order to train classifiers. The robot performed various behaviors on each object and collected data from several modalities. The behaviors the robot performed were *lift, crush, grasp, hold, look, drop, poke, push, shake,* and *tap.* The modalities that the robot used were *audio, vibration, flow, haptics, SURF,* and *finger position.* Since not all of these actions and modalities were compatible, the result was 48 different behavior-modality datasets. For each object in each of these data sets, 5 trials were recorded, so for each behavior-modality combination there were 500 samples.

Experimental Setup

Let W be the set of descriptive words, C be the set of behavior-modality contexts, and O be the set of objects explored. Since some words only applied to a very small subset of objects, a threshold of 5 applicable objects was set and words that did not meet this threshold were removed from W. The resulting W contained 47 words. Using a 5-fold train-test split, O was divided into a test set, O_{test} and a training set, O_{train}, for each individual classifier.

An instance of an SVM classifier $X_i = (W_i, C_i)$ was created for each combination of words $W_i \in W$ and behavior-modality contexts $C_i \in C$. Since most words had far more negative instances where the word did not apply to an object than positive instances, the classifier would be trained to simply always predict that the word does not apply to achieve better accuracy. In order to prevent this from happening, the positive samples had to be up-sampled so that the number of negative and positive examples of a word applying were equal. For each trained X_i, the accuracy, precision, recall, F1-Score, and Cohen's Kappa were calculated based on a confusion matrix generated from predicting on O_{test} resulting in stats for each W_i-C_i combination. The classifiers for each of the train-test splits were saved, resulting in a total of about 13000 classifiers.

Once these key statistics were recorded, combinations of C_i could be made to produce more accurate identification of object properties. In order to combine different X_i, a weighted sum based on the classifier's Cohen's Kappa scores was used to calculate a class distribution probability estimate. The number of X_i used was varied as well to observe the change in Cohen's Kappa as the number of combined classifiers was increased.

Results and Analysis

For each X_i, the resulting accuracy and Cohen's Kappa were on average very low. This was

not unexpected though since for just a single behavior-modality it would make sense for the classifier to have a difficult time identifying if the word applied or not. Certain X_i had relatively higher Cohen's Kappa scores and accuracy, for example the "tap-audio" behavior-modality was relatively accurate for the "hard" property of objects, which is to be expected since tapping a hard object would produce distinctive audio data.

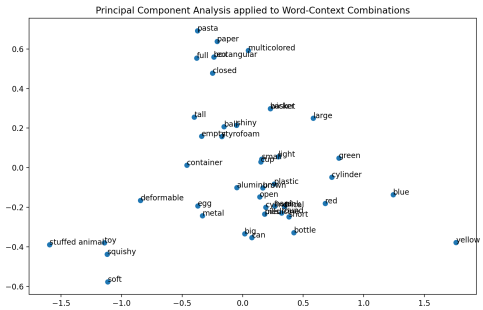

Figure 1: Principal Component Analysis (PCA) for each word-context combination

Fig. 2 shows the Principal Component Analysis (PCA) using the Cohen's Kappa for each word-context combination with each data point representing a word. The PCA is useful for seeing which words are "similar" to one another in that similar contexts are used to effectively identify them.

Combining every classifier X_i for a given word W_i, the results showed that prediction accuracy and Cohen's Kappa increase over any X_i. This can be shown in the heatmap in Fig. 3 which shows the Cohen's Kappa for a small sample of contexts and words and then the Kappa score for the combination of all contexts. These results show that the Kappa for the combination of all contexts is significantly higher than using almost any individual context alone.

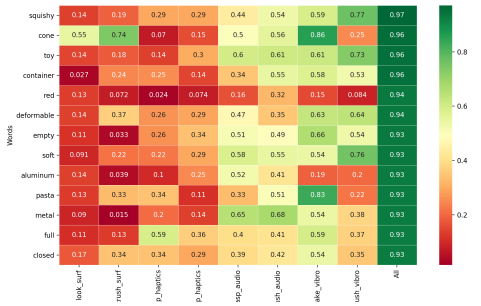

Figure 2: Heatmap of Kappa Scores for word-behavior-modality combinations

Using all contexts, it was shown that the Cohen's Kappa and accuracy can be improved significantly. However, this process of using every context takes a lot of time for the robot to perform. To test whether a high Cohen's Kappa could still be achieved while trying fewer contexts on average, a random sampling of 5 trials for every train-test split on 1, 2, 5, 10, 20, 30, 40, and all of the contexts was taken. The results for the word "empty" are shown in Fig. 4 and indicate that the Cohen's Kappa increases significantly at a low number of contexts, but begins leveling off when around 10 are used.

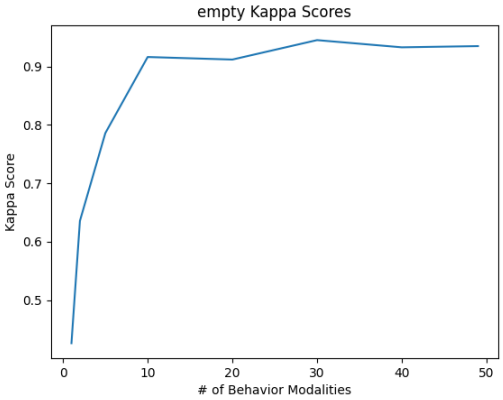

Figure 3: Average Kappa score for different number of behavior modalities used

Conclusion and Future Work

The results indicate that a multimodal approach can effectively identify object properties that can't be picked up by visual data alone. Using a combination of sensorimotor features also allows for more accurate identification than using any one sensorimotor feature alone. In future work, a deep neural network could be implemented instead of using individual SVMs for each sensorimotor feature. Another extension could be to try and transfer the knowledge learned by one robot to another in order to save time. One approach to knowledge transfer is the Kernel Manifold Alignment in which all the feature spaces for the different robots are combined into a common space for training [Tatiya *et al.*, 2020].

References

Tatiya G., and Sinapov J. (2019) **Deep Multi-Sensory Object Category Recognition Using Interactive Behavioral Exploration** In

proceedings of the IEEE International Conference on Robotics and Automation (ICRA)
Tatiya, G., Shukla, Y., Edegware, M., and Sinapov, J. (2020) **Haptic Knowledge Transfer Between Heterogeneous Robots using Kernel Manifold Alignment** In *proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*