Compositional Zero-Shot Learning for Attribute-Based Object Reference in Human-Robot Interaction

Peng Gao[1*], **Ahmed Jaafar**[1*], **Brian Reily**[2], **Christopher Reardon**[3], **and Hao Zhang**[1]
[1]University of Massachusetts Amherst, [2]DEVCOM Army Research Laboratory, [3]University of Denver
penggao.robotics@gmail.com, ajaafar@umass.edu, brian.j.reily.civ@army.mil
christopher.reardon@du.edu, hao.zhang@umass.edu
*Authors contributed equally to this paper

Abstract: Language-enabled robots have been widely studied over the past years to enable natural human-robot interaction and teaming in various real-world applications. Language-enabled robots must be able to comprehend referring expressions to identify a particular object from visual perception using a set of referring attributes extracted from natural language. However, visual observations of an object may not be available when it is referred to, and the number of objects and attributes may also be unbounded in open worlds. To address the challenges, we implement an attribute-based compositional zero-shot learning method that uses a list of attributes to perform referring expression comprehension in open worlds. We evaluate the approach on two datasets including the MIT-States and the Clothing 16K. The preliminary experimental results show that our implemented approach allows a robot to correctly identify the objects referred to by human commands.

Keywords: Object Reference, Zero-Shot Learning, Human-Robot Interaction

1 Introduction

Natural language-enabled robots have recently attracted considerable attention to enable intuitive, efficient, and transparent human-robot interaction and teaming [1, 2, 3], which has a wide variety of real-world applications throughout society, such as in elderly care, hospital assistance, education, inspection, and search and rescue [4, 5, 6]. Object reference, defined as the capability of identifying a particular object that a human teammate refers to, is essential for intelligent robots to appropriately communicate with humans [7]. In language-based communication, a language-enabled robot must comprehend a referring expression in order to recognize and localize a particular object from its visual perception using a set of referring attributes extracted from natural language by the human teammate.

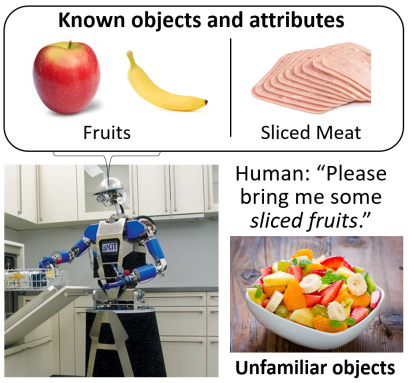


Figure 1: A motivating scenario of object reference for human-robot interaction.

While robot perception has shown promising performance for recognizing object categories, they are insufficient for referring expression comprehension to represent and identify object instances from language and vision. First, an object instance can be referred to through language before it is observed, and visual data may not be available for the object at the time it is referenced, meaning that the robot may have to visually identify an unfamiliar object that has not previously observed. Second, most object recognition and detection techniques focus on identifying the same categories of objects but

7th Conference on Robot Learning (CoRL 2023), Atlanta, USA.

are generally unable to discriminate between object instances, which are more typically referred to. Third, in an open world, even given a bounded number of object categories, the number of object instances and the number of object attributes can be unbounded. To enable referring expression comprehension and address the above challenges, we implement an attribute-based compositional zero-shot learning (CZSL) approach that composes a list of seen attributes and object labels to identify unseen object-attribute pairs from a robot's perception data in an open world.

2 Related Work

In this study, our primary focus is on compositional zero-shot learning (CZSL) [8]. CZSL involves working with a training dataset that comprises various combinations of states and objects. The primary objective in CZSL is to recognize and identify previously unseen combinations of these attributes and objects during testing. We generally divide the existing methods into two groups, including pure visual-based methods and language-prior methods.

For the first group of methods, several approaches have been proposed. Some methods tackle CZSL by separately learning classifiers for objects and states and then combining these classifiers to construct the final recognition model, such as using an SVM classifier trained for known combinations, and class weights for new combinations are inferred using a Bayesian framework [9]. LabelEmbed [10] introduces a transformation network built on top of pre-trained state and object classifiers. Another work [11] suggests encoding objects as vectors and states as linear operators that transform these vectors. Similarly, a recent work [12] enforces symmetrical representations of objects based on their state transformations. For the second group of methods, the mainstream is to minimize the loss between visual and linguistic learning space for unseen attribute-object pairs [13]. Some recent works use graph structure to leverage information transfer between seen to unseen pairs using graph convolutional networks [14] or graph attention networks [15] along with modular networks.

In this paper, we implement an attribute-based compositional zero-shot learning (CZSL) approach that composes a list of seen attributes and object labels to identify unseen object-attribute pairs from the robot's perception data. The approach allows robots to disentangle attributes and object appearances from observed compositions and predict unseen or unknown compositions in open-world scenarios.

3 Approach

3.1 Problem Formulation

We implement an approach for object reference in the context of human-robot interaction. In this scenario, we assume a shared environment containing a human and an assistant robot, such as a living room or a workspace. The primary objective is for the assistant robot to accurately identify the object to which the human is referring, based on the natural language instructions provided.

Specifically, let's consider a typical human referring expression like "black pen". This expression can be readily parsed by existing natural language models, such as BERT [16], into its attributes and class label. We denote the attribute "black" as w_{attr} and the class label "pen" as w_{obj}. Based upon the parsed keywords w_{attr}, w_{obj}, the robot then undertakes the task of identifying the referred object in the shared environment, drawing upon its own observation sequence, which we denote as $\mathcal{L} = \{\mathcal{I}_1, \ldots, \mathcal{I}_n\}$. Formally, we formulate the referring expression comprehension as follows:

$$s_i = cos\left(\phi(\mathcal{I}_i), \psi\left(w_{attr}, w_{obj}\right)\right) = \frac{\phi(\mathcal{I}_i)\psi(w_{attr}, w_{obj})}{||\phi(\mathcal{I}_i)||\ ||\psi(w_{attr}, w_{obj})||} \tag{1}$$

where ϕ denotes the network to encode visual features of $\mathcal{I}_i \in \mathcal{L}$, which is constructed based on a ResNet [17] or a Vision-Transformer [18] followed by an Average Pooling operation, then the final visual feature is computed through a linear layer. ψ denotes the network to encode word features of w_{attr}, w_{obj}, which is based on GLoVe [19] followed by a multi-layer perception (MLP). cos

denotes the cosine function that computes the similarity. s_i denotes the similarity between the query text command and the robot observation. Given the similarity, the object reference is formulated as a classification problem, which applies the SoftMax function to the similarity score and predicts which class of attribute-object pairs the observation belongs to, thus achieving object reference.

3.2 Attribute-Based CZSL

To make the classification generalizable to unseen attribute-object pairs, based on the existing work [20, 21], we use an attribute-based compositional zero-shot learning method to disentangle attributes of objects for unseen object-attribute pair prediction. Formally, the approach introduces two images $\mathcal{I}_{attr}$ and $\mathcal{I}_{obj}$ into the learning process. $\mathcal{I}_{attr}$ has the same attribute as $\mathcal{I}$ but with a different object. $\mathcal{I}_{obj}$ has the same object as $\mathcal{I}$ but with different attributes. For example, if $\mathcal{I}$ is a black pen, then $\mathcal{I}_{attr}$ can be a black fork with the same attribute "black" as $\mathcal{I}$ but with a different object class "fork", denoted as $w_{non-obj}$. $\mathcal{I}_{obj}$ can be a sliver pen with the same object class as $\mathcal{I}$ but with a different attribute "sliver", denoted as $w_{non-attr}$. Our goal is to make a robot not only recognize the black pen that is seen in training but also recognize the silver fork by composing the disentangled attributes and object classes.

Formally, first the correlation between $\mathcal{I}, \mathcal{I}_{attr}$ and between $\mathcal{I}, \mathcal{I}_{obj}$ needs to be computed to disentangle the correlated attribute and object class. The correlation matrix is defined as:

$$\mathbf{C}^{attr} = \frac{\phi(\mathcal{I}_i)\phi(\mathcal{I}_{attr})}{||\phi(\mathcal{I}_i)||_2 \, ||\phi(\mathcal{I}_{attr})||_2} \quad and \quad \mathbf{C}^{obj} = \frac{\phi(\mathcal{I}_i)\phi(\mathcal{I}_{obj})}{||\phi(\mathcal{I}_i)||_2 \, ||\phi(\mathcal{I}_{obj})||_2} \tag{2}$$

where ϕ denotes the visual feature encoder based on a ResNet or a Vision Transformer. $\mathbf{C}^{attr}$ denotes the correlation matrix between $\mathcal{I}$ and $\mathcal{I}_{attr}$ on the attribute w_{attr}, and $\mathbf{C}^{obj}$ denotes the correlation matrix between $\mathcal{I}$ and $\mathcal{I}_{obj}$ on the object class w_{obj}. Given the correlation matrix, row-wise and column-wise SoftMax operations are performed on it to get the mask for different attributes and object classes, which is defined as

$$\mathbf{A}_i = \frac{e^{\mathbf{C}^{attr}_{i,:}}}{\sum_{i=1}^{d} e^{s_{i,j}}} \quad and \quad \mathbf{A}^{attr}_j = \frac{e^{\mathbf{C}^{attr}_{:,j}}}{\sum_{i=1}^{d} e^{s_{i,j}}} \quad and \quad \mathbf{A}^{obj}_j = \frac{e^{\mathbf{C}^{obj}_{:,j}}}{\sum_{i=1}^{d} e^{s_{i,j}}} \tag{3}$$

where d denotes the dimensions of the feature channels , $\mathbf{A}_i$ and $\mathbf{A}^{attr}_j$ denotes the correlation masks between $\mathcal{I}$ and $\mathcal{I}_{attr}$ in the feature space. $\mathbf{A}^{obj}_j$ denotes the correlated region of $\mathcal{I}_{obj}$ on the object class in the feature space. To further disentangle the unseen attributes and object class for generalizable prediction, the following based on the negative correlation matrix is computed:

$$\mathbf{A}^{non-obj}_j = \frac{e^{-\mathbf{C}^{attr}_{:,j}}}{\sum_{i=1}^{d} e^{s_{i,j}}} \quad and \quad \mathbf{A}^{non-attr}_j = \frac{e^{-\mathbf{C}^{obj}_{:,j}}}{\sum_{i=1}^{d} e^{s_{i,j}}} \tag{4}$$

where $\mathbf{A}^{non-obj}_j$ denotes the non-correlation mask between $\mathcal{I}$ and $\mathcal{I}_{attr}$ about the object class $w_{non-obj}$. $\mathbf{A}^{non-attr}_j$ denotes the non-correlation mask between $\mathcal{I}$ and $\mathcal{I}_{obj}$ about the attribute $w_{non-attr}$. To get the final masks, the elements in these masks are added up along their feature channels, which is defined as $m_j = \sum_{i=1}^{d} \mathbf{A}_{ij}$, where $\mathbf{m} = \{m_j\}^l$ and $l = 49$ denotes the feature length in this paper. Similarly, $\mathbf{m}^{attr}$, $\mathbf{m}^{obj}$, $\mathbf{m}^{non-attr}$ and $\mathbf{m}^{non-obj}$ can be obtained based on $\mathbf{A}^{attr}, \mathbf{A}^{obj}, \mathbf{A}^{non-attr}, \mathbf{A}^{non-obj}$ respectively. Given the feature masks, the disentangled feature is computed as follows:

$$\mathbf{v}_{attr} = \mathbf{m} \cdot \phi(\mathcal{I}_{attr}) + \mathbf{m}^{attr} \cdot \phi(\mathcal{I}), \tag{5}$$

$$\mathbf{v}_{obj} = \mathbf{m} \cdot \phi(\mathcal{I}_{obj}) + s\mathbf{m}^{obj} \cdot \phi(\mathcal{I}) \tag{6}$$

$$\mathbf{v}_{non-attr} = \mathbf{m}^{non-attr} \cdot \phi(\mathcal{I}_{attr}), \tag{7}$$

$$\mathbf{v}_{non-obj} = \mathbf{m}_{non-obj} \cdot \phi(\mathcal{I}_{obj}) \tag{8}$$

where $\mathbf{v}_{obj}, \mathbf{v}_{attr}, \mathbf{v}_{non-obj}, \mathbf{v}_{non-attr}$ denote the disentangled features of $w_{obj}, w_{attr}, w_{non-obj}, w_{non-attr}$ respectively. Given the disentangled features, classification is performed given Eq. (1) by replacing $\phi(\mathcal{I})$ with the disentangled features to predict the object-attribute pair given human text command. Cross entropy loss is used to train the network, in which the loss is minimized between all the visual features defined in Eqs. (5-8) and their associated word embedding features.

4 Preliminary Experimental Results

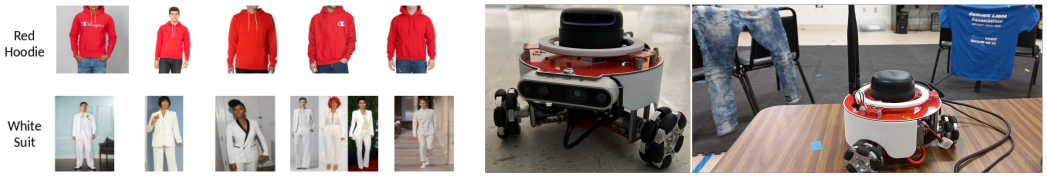
We use two existing datasets to evaluate the approach: MIT-States [22] and Clothing16K [23]. The MIT-States dataset contains object classes such as fish, rooms, etc. It also has attributes, such as mossy, dirty, etc. Clothing16K contains clothes, such as suits and pants, with different attributes, such as pink and black. In the training stage, all of the objects and attributes are seen. In the testing stage, the compositions of seen and unseen objects/attributes are used to evaluate the generalization of the approach. In addition, various scenarios are evaluated including seen object-attribute pairs (**Seen**), unseen object-attribute pairs (**Unseen**), unseen objects with seen attributes (**Object**), and seen objects with unseen attributes (**Attr**).

Table 1: Quantitative Results using the MIT-States Dataset based on the **AUC** (%) metric.

	Seen	**Unseen**	**Object**	**Attr**
Val @1	0.3368	0.2965	0.3683	0.3190
Val @2	0.4501	0.4382	0.4800	0.4531
Val @3	0.5206	0.5195	0.5508	0.5320
Test @1	0.316	0.2585	0.3336	0.2844
Test @2	0.4252	0.3811	0.4442	0.4035
Test @3	0.4987	0.4577	0.5160	0.4793

Area Under the Curve (**AUC**) is utilized to evaluate the accuracy of text-to-image retrieval. For each scenario, the top K ($@k$) candidates are retrieved as the results.

Table 1 presents the quantitative results using the MIT-States dataset. From these results, we observe that the approach can successfully retrieve the correct images given the human text commands. The performance of unseen pairs retrieval is slightly worse than the seen pairs, which indicates the generalization capability of the approach for object references in human-robot interaction.



(a) Evaluation on Clothing16K Dataset (b) Evaluation on Real Robot

Figure 2: Experiment setups. (a) Retrieval of unseen object-attribute pairs given text query command. (b) Case studies on object reference in human-robot interaction.

We further implement a method based on the recent work [21], it obtains the quantitative results using the Clothing16K dataset, as shown in Figure 2(a). Given the query text command, the approach can correctly retrieve the unseen object-attribute pairs. We deploy our implemented work on a physical robot, Triton, to conduct a case study. In this study, the Triton moves around its environment to identify the unseen attribute-object pair given via a human textual command. As shown in Figure 2(b), the human command is "blue shirt" which is unseen in the training stage. Eventually, the robot can correctly identify the "blue shirt" and move forward to it, given that some of the attribute-objects it was trained on are "red shirt" and "blue shorts".

5 Conclusion

In this paper, we implement a new attribute-based compositional zero-shot learning (CZSL) approach to enable referring expression comprehension for human-robot interaction. The approach composes a list of unknown attributes and object labels to identify unseen object-attribute pairs from the robot's perception data in an open world. Our preliminary experimental results have validated the effectiveness of the approach.

Our current implementation has some limitations. In the future, we will further study the following aspects, including 1) integrating natural language processing with the vision network to perform human-robot interaction tasks, 2) adding robot search and navigation modules to complete the loop, 3) doing a deeper analysis with more sophisticated robots in real-world experiments.

Acknowledgments

This research was partially supported by the ONR grant N00014-21-1-2418, the ARL A2I2 Program W911NF-23-2-0005, and the NSF CAREER Award IIS-2308492.

References

[1] R. Paul, J. Arkin, N. Roy, and T. M Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. 2016.

[2] P. Moolchandani, C. J. Hayes, and M. Marge. Evaluating robot behavior in response to natural language. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 197–198, 2018.

[3] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.

[4] N. Randall. A survey of robot-assisted language learning (rall). *ACM Transactions on Human-Robot Interaction (THRI)*, 9(1):1–36, 2019.

[5] C. E. Bartlett and N. J. Cooke. Human-robot teaming in urban search and rescue. In *Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 250–254, 2015.

[6] V. Raman, C. Lignos, C. Finucane, K. C. Lee, M. P. Marcus, and H. Kress-Gazit. Sorry dave, i'm afraid i can't do that: Explaining unachievable robot tasks using natural language. In *Robotics: Science and Systems*, 2013.

[7] P. Gao and H. Zhang. Bayesian deep graph matching for correspondence identification in collaborative perception. In *Robotics Science and Systems (RSS)*, 2021.

[8] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata. Open world compositional zero-shot learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[9] C.-Y. Chen and K. Grauman. Inferring analogous attributes. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[10] I. Misra, A. Gupta, and M. Hebert. From red wine to red tomato: Composition with context. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[11] T. Nagarajan and K. Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. In *The European Conference on Computer Vision*, 2018.

[12] Y.-L. Li, Y. Xu, X. Mao, and C. Lu. Symmetry and group in attribute-object compositions. 2020 ieee. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[13] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *The European Conference on Computer Vision*, 2012.

[14] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[15] Z. Xu, G. Wang, Y. Wong, and M. Kankanhalli. Relation-aware compositional zero-shot learning for attribute-object pair recognition. *arXiv*, 2021.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020.

[19] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *The Conference on Empirical Methods in Natural Language Processing*, 2014.

[20] N. Saini, K. Pham, and A. Shrivastava. Disentangling visual embeddings for attributes and objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[21] S. Hao, K. Han, and K.-Y. K. Wong. Learning attention as disentangler for compositional zero-shot learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[22] P. Isola, J. J. Lim, and E. H. Adelson. Discovering states and transformations in image collections. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[23] T. Zhang, K. Liang, R. Du, X. Sun, Z. Ma, and J. Guo. Learning invariant visual representations for compositional zero-shot learning. In *European Conference on Computer Vision*, 2022.