Feature Based Disambiguation Dialog Model for Object Retrieval Task

Shivam Gandhi

May 15 2019

1 Abstract

Object retrieval tasks are important both in industry, where robots are expected to perform them in a variety of situations, but also in academia, where they provide an arena to develop improved dialog models and graspers. Past attempts at robot grasping have utilized dialog models that enumerate the set of objects in front of the robot to the human and ask questions based on the enumeration ("is this the object you want?") but fail to disambiguate the objects based on their features. We are training a robot to ask questions when it's confused about an object retrieval task that a human has assigned it such that it completes disambiguation based on features of the objects in front of it. The human asks the robot for a specific item, and then the robot will either retrieve the object if there is no ambiguity as to which object, or ask which instance of the object the human wants. This is done by turning the human utterance into a caption and scoring how good of a caption the utterance is for each object in front of the robot, and returning the object with the highest score. Then the robot can ask questions based on the features of the objects if no object in front of it passes a threshold to qualify for retrieval. Once the full project is implemented, the robot will be able to ask questions about which object the user wants based on features of the objects such as color, shape, etc.

2 Introduction

Retrieval tasks are one of the major tasks we would like implemented on robots. However, the tasks is difficult for a variety of reasons. First, deciding which object to retrieve is challenging when outside of a laboratory situation as multiple objects may fulfill the requirements of what the human would like. There can also be context that the robot is missing which a human accidentally implies. Aside from the object identification and clarification task, grasping objects to be retrieved also is challenging.

Previous methods implement a dialog model to allow the robot to disambiguate which object is to be retrieved. However, there is not yet a dialog model

that does disambiguation based on object features. Other robot object retrieval models are constrained by environment, or can only retrieve certain types of objects.

Our approach is to implement an image segmentation model to identify the different images in front of the robot and then use an image captioning model to decide which object best fits the description the human gave. We modify the captioning model such that, given an image and a caption on input, returns a score of how good a fit the caption is to the image. We score each of the objects in front of the robot in this manner, and if none of them pass a certain threshold that indicates that the robot can be confident that the specific object is the one the human wants, the robot begins clarification of which object the human wants with a dialog model.

We evaluate the success of this project through user tests with humans. The primary score of interest is a score of how accurate the robot is in retrieving the correct object.

3 Related Work

One current method of task disambiguation is detailed in [1] using a FETCH POMDP method. In this paper, the robot is tasked with retrieving an object in front of it. However, there is ambiguity, so the robot is able to point to each object and ask if the object it's pointing to is the one the human wants. It does this for each object it is deciding between. FETCH POMDP advanced the state of the art since it allowed robots to disambiguate via interaction with a human. However, it can be improved upon in various ways. FETCH does not allow for enough flexibility in tasks, as it much enumerate every single object it has ambiguity between. It also breaks down in situations where it cannot point to each object that it is deciding between.

Another technique is detailed in [2] in which the user asks the robot for an object and the robot asks for clarification if there is ambiguity in which object it should return. The primary issue with the technique in this paper is that the robot is only able to ask for clarification by asking "which one?" when faced with ambiguity in which object it should return. Furthermore, the user needs to clarify which object it should return by describing the object's position relative to the other objects around it. This lack of flexibility in clarification means that the robot is only able to operate on tasks where the human has full knowledge of where the objects are positioned relative to each other. Another downfall of this technique was that bounding boxes need to be given to the robot. While this is not a significant issue, we must recognize that bounding boxes cannot always be given.

4 Technical Approach

First, we implemented an image segmentation model, MaskRCNN trained on MSCoco, onto the robot. This segmentation model segments out the individual objects in front of the robot. We call the set of segmentations $S = \{\phi\}$. We call the caption C. We would like to solve the problem

$$\phi_R = \underset{\phi \in S}{\operatorname{argmax}} \Phi(\phi, C) \tag{1}$$

where Φ is a scoring function that takes in a segmentation and a caption as an argument. We create Φ by reconstructing the architecture of an existing image captioning model that has been trained on MSCoco. We then fine tune this model on the set of objects that the robot will be trying to choose between.

It is important to note that we desire a threshold of confidence before completing the retrieval. Define the threshold as η, and the second highest scored object as ϕ_2. We desire that

$$\Phi(\phi_R, C) - \Phi(\phi_2, C) > \eta \tag{2}$$

If this condition is not met, then we have our original captioning model caption the segmentations that the robot is trying to decide between. The dialog model then asks which object the human would like, using the captions in the dialog model. We then select the specific object that the human clarified, or if the robot is unable to understand what the human is saying, ask another question.

5 Evaluation

The project is not yet in a state to be evaluated based on object retrieval. We plan to evaluate the robot based on its accuracy in returning the correct object to the human as well as the clarity of the questions it asks and the number of questions it needs to ask. The first metric will be a simple percentage that quantifies what percentage of tasks the robot is successful at.

6 Conclusion

Based on initial tests using the segmentation and image captioning models, we are very confident that this task can be achieved. The next step in the project is to implement the scoring architecture from the captioning architecture and fine tune it on the objects we will be testing it. Since this model is not yet implemented, it is too early to make conclusions about whether the technique detailed in this report will be successful or not.

References

[1] David Whitney, Eric Rosen, James MacGlashan, Lawson L.S. Wong, Stefanie Tellex. Reducing Errors in Object-Fetching Interactions through Social Feedback. In *IEEE International Conference on Robotics and Automation (ICRA), 2017* DOI: 10.1109/ICRA.2017.7989121.

[2] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, Jethro Tan Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In 2018 IEEE International Conference on Robotics and Automation (ICRA).