Karl Dill

EMSE-6301 Natural Language Processing

Paper:

<https://aclanthology.org/2022.acl-long.373/>

Paper Title:

HOLM: Hallucinating Objects with Language Models for Referring Expression Recognition in Partially-Observed Scenes

Authors:

Volkan Cirik – 1013 Citations – Carnegie Mellon University

Louis-Philippe Morency – 29022 Citations – Carnegie Mellon University

Taylor Berg-Kirkpatrick – 4941 Citations – University of California San Diego

Problem:

The paper addresses the problem of finding a target location in a partially observed 360-degree scene by using natural language instructions. The paper proposes a novel method called HOLM (**H**allucinating **O**bjects with **L**anguage **M**odels ), which uses pre-trained language models to hallucinate objects that are not visible in the current field of view, based on their co-occurrence with other objects in language.

Prior Work:

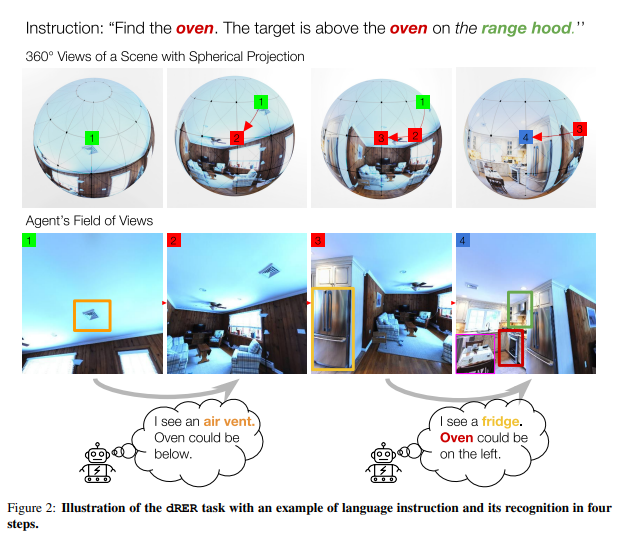
This study relies on the Dynamic Referring Expression Recognition (dRER) task, where the goal is to find a target location by dynamically adjusting the field of view (FoV) in partially observed 360 scenes.

The dRER task requires a system to track the referent of an expression across different frames or turns of a dynamic environment and resolve any ambiguities or uncertainties that may arise. For example, if the expression is "the ball", the system needs to determine which ball is being referred to and whether it is the same ball throughout the video or dialogue. The dRER task is challenging because it involves complex linguistic phenomena such as anaphora, deixis, ellipsis, coreference, and pragmatics.

Summary:

The paper argues that the HOLM approach can help the system reason about the spatial relationships of objects and guide its actions to adjust the field of view until it finds the target location. The paper evaluates HOLM on two datasets for dynamic referring expression recognition (dRER) and shows that it outperforms the state-of-the-art methods on both indoor and outdoor settings. HOLM is a novel approach to dReER where an agent has to find a target location in a partially-observed 360 scene by adjusting its field of view (FoV) based on natural language instructions. The main challenge of this task is that the target object may not be visible in the initial FoV, and the agent has to infer its existence and location from the linguistic and visual context.

Example:



HOLMleverages large pre-trained language models (LMs) to hallucinate (AKA “dream” or “imagine”) objects that are likely to co-occur with the target object in the environment. For example, if the instruction is "find the chair next to the table", and the table is not visible, HOLM can use an LM to generate possible objects that are associated with tables, such as lamps, books, or cups. These hallucinated objects are then used as additional visual features to guide the agent's FoV adjustment.

HOLM consists of three main components: a hallucination module, a recognition module, and a navigation module. The hallucination module uses an LM to generate object candidates for each word in the instruction, and ranks them based on their affinity scores with the target object. The affinity scores are computed from the co-occurrence statistics of objects in a large-scale image captioning dataset (Visual Genome). The recognition module uses a transformer-based model to encode the instruction and the FoV, and predicts a bounding box for the target object. The navigation module uses a reinforcement learning algorithm to adjust the FoV based on the recognition module's output and a reward function.

The paper evaluates HOLM on two datasets for dRER: Matterport3D and StreetView. The results show that HOLM outperforms the state-of-the-art methods on both datasets, demonstrating its effectiveness in handling partial observability and generalizing to different domains. The paper also conducts ablation studies and qualitative analysis to highlight the contributions of each component and the challenges of the task.

Results:

The authors measure their success by comparing their FoV accuracy vs current models

“Our main evaluation metric for methods is FoV accuracy: the percentage of the time the target location is visible in the final FoV. The FoV accuracy sets an upper bound on the localization accuracy for predicting the pixel location of the target point, i.e., if the target is not visible, it is impossible to predict the exact location. Thus, we focus on this metric to compare systems.

