2D Relational Choice Task with Image and Graph Feature Learning with Graph Attention

Minjune Hwang [1]

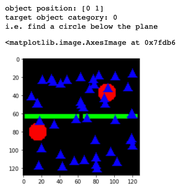
[1]Department of Computer Science, Stanford University

Overview

Visual search task is one of the fundamental tasks in Embodied AI. In the problem, an agent is asked to navigate to a target instance, using visual inputs of the current scene. As such, previous methods often focus in leverage different modalities with vision approaches, such as raw image input, object detection, or depth estimation. However, an agent can learn relations between objects in the scene and actively update and use this relational information when navigating to the target object, using a structured representation like graphs with different types of nodes and edges. This is especially true when there exists similar objects with different relational state (e.g. above or below a table). In this work, we created a 2D relational object choice environment in which the agent has to choose a relevant object given by a relational goal specification. In the environment, this paper shows that if an agent can update and learn from this additional modality of relational graphs, when combined with other visual modalities, can help itself to navigate to target objects more effectively. Also, experimental results demonstrated that leveraging graph attention is beneficial when multiple occlusion and distractor objects are present in the observation.

Problem Statement & Dataset: Relational Choice Task

For the experiment, I have designed a 2D environment for relational choice tasks. The agent is presented with a symmetric 2D environment with two circles and two rectangles: one circle above and the other below their respective rectangle. The agent is given a goal specification of "circle above rectangle" or "circle below rectangle" and must choose one of two possible actions (`left`/`right`) to select the side of the environment that satisfies the goal definition. The agent is given a reward of 1 if it selects the action that matches the goal descriptor, and a reward of 0 otherwise. The episode terminates after one step (bandit problem). This 2D environment is an abstracted version of relational choice tasks in the real-world simulation (iGibson).



(a) An example of a randomly generated RGB of the the environment with dummies

(b) A sample task in iGibson simulation environment.

To study the effect of high scene complexity, we simulate a heavily populated scene by adding a random number (up to 75) of triangles as distractors. The distractor objects occlude other relevant objects in the scene, making harder for the CNN model to learn features and the GNN model to learn from dynamically sized graph size.

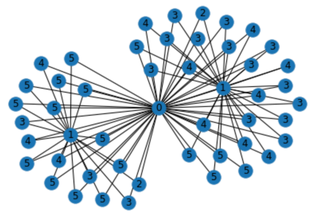


Figure 2. An example of a scene graph of the environment with distractors (dummies).

Model Design

In our tasks, the agent mainly receives RGB observation of the enviroment (In 3D simulation environment, depth data is additionally given). These observations represent the local, high resolution features. I aim to use convolutional neural networks to extract image features from these observations. On top of this, I also have a graph that represents the relations between objects. I aim to use graph convolutional networks to extract global features from this data. The graph is dynamically updated as the robot navigates in the scene, and thus it contains episodic memory that is required for high level planning. Figure 1 shows the three main components of the model: a convolution encoder for RGB input, a convolution encoder for depth input, and a graph neural network for scene graphs.

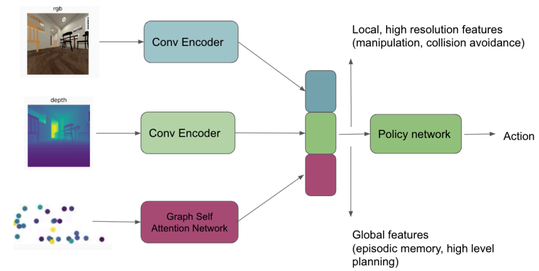


Figure 3. A model diagram

Scene Graph and Graph Neural Network

Unlike raw sensory inputs, scene graphs are object-centric representations with pairwise connections that easily encode the information necessary. For the model, I used HGT with different attenions.

1. **Heterogeneous graph transformer (HGT)**: To compute a per node embedding, the graph is passed through three heterogeneous graph transformer layers with ReLU activations. The HGT convolutions use distinct edge-based matrices for each edge type when computing attention, allowing the model to learn representations conditioned on different edge types, rather than connectivity alone.
2. **Graph attention pooling**: We additionally introduce an attention mechanism at the final pooling layer in order to effectively aggregate task-relevant nodes. The task-driven attention mask is constructed by assigning a value of 1 to all nodes for which the semantic category matches any semantic category found in the current episodic goal description and 0 otherwise. The mask is used as the weights for a weighted mean pooling of the node embeddings to extract a graph embedding. For this attention, we have two possible designs. First is the task-driven (TD) attention, which fixes the weight of 1 to objects of interest and 0 to other distractor objects. Second is the learned (LN) attention, which has the separate GNN model to learn the weights for graph attention.

Experimental Design

We design our experiments to answer the following questions.

- **1**: Does the scene graph help the agent to learn faster to choose correct objects in relational choice tasks?
- **2**: Does the additional attention mechanism help GNN models to learn faster? If so, should we leverage learned attention model or fixed / heuristic-based attention model?

Model Lists

We test the following combinations to attest the effect of scene graph and graph attention.

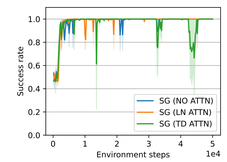
- CNNs (RGB)
- CNNs (RGB) + GNNs (SG)
- CNNs (RGB) + GNNs (SG) + Task-Driven (TD) Attention
- CNNs (RGB) + GNNs (SG) + Learned (LN) Attention

Experiment Result

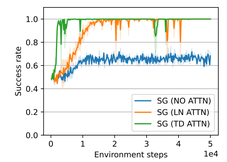
We observed that without attention, the performance of our model suffers significantly when there are many distractors. We observed that attention model performs much better than RGB only, suggesting that scene graphs are effective representation for tasks that require relational reasoning.

Model	SR↑ with No Distractors	SR↑ with Distractors
RGB	0.992±0.001	0.458±0.004
RGB + SG	0.997± 0.003	0.659±0.002
RGB + SG + TD ATTN	0.991±0.009	**0.993±0.007**

Table 1. Success Rate (SR) for Task without with Distractors

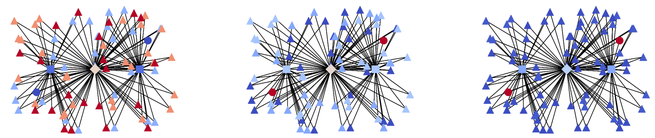


(a) Success Rate (SR) versus environment steps over training for the 2D relational choice tasks without dummies



(b) Success Rate (SR) versus environment steps over training for the 2D relational choice tasks without dummies

Additionally, we visualized evolution of learned attention over training in the experiment with distractors. The attention shifts gradually towards task-relevant objects (circles and rectangles) and away from distractors (triangles). In comparison, our task-driven attention would focus on the two planes (squares) and the circles.



Conclusion

- Designed a novel environment for 2D relational choice task with RGB and Scene Graph.
- Results showed the benefits of the scene graph as a representation for relational choice tasks that require reasoning about object relations.
- In large, populated scenes, having a task-driven attention mechanism is essential in aggregating task-relevant information and achieving a high success rate.