Modeling Dynamic Environments with Scene Graph Memory

Extended Abstract

Andrey Kurenkov
Stanford University
Stanford, United States
andreyk@stanford.edu

Michael Lingelbach
Stanford University
Stanford, United States
mjlbach@stanford.edu

Tanmay Agarwal
Stanford University
Stanford, United States
tanmayx@stanford.edu

Chengshu Li
Stanford University
Stanford, United States
chengshu@stanford.edu

Emily Jin
Stanford University
Stanford, United States
emilyjin@stanford.edu

Ruohan Zhang
Stanford University
Stanford, United States
zharu@stanford.edu

Li Fei-Fei
Stanford University
Stanford, United States
feifeili@cs.stanford.edu

Jiajun Wu
Stanford University
Stanford, United States
jiajunwu@cs.stanford.edu

Silvio Savarese
Salesforce AI Research
San Francisco, United States
silvio.savarese@gmail.com

Roberto Martín-Martín
University of Texas at Austin
Austin, United States
robertomm@cs.utexas.edu

ABSTRACT

Embodied AI agents operating in dynamic environments often need to predict object locations to make informed decisions. We propose a method for doing this via link prediction on partially observable dynamic graphs. We represent the agent's accumulated set of observations in a data structure called a Scene Graph Memory (SGM), combine this data structure with a neural net architecture we call Node Edge Predictor (NEP), and show that it can be trained to predict the locations of objects in a variety of environments with diverse object movement dynamics. To evaluate our method, we implement the Dynamic Household Simulator, a novel benchmark which enables sampling of diverse dynamic scene graphs that follow the semantic patterns typically seen at peoples' homes. We demonstrate that our method outperforms baselines both in terms of quickly adapting to the dynamics of a new scene and in terms of its overall accuracy.

KEYWORDS

Robotics, Graph Neural Network, Embodied AI, Scene Graph, Link Prediction, State Representation

ACM Reference Format:
Andrey Kurenkov, Michael Lingelbach, Tanmay Agarwal, Chengshu Li, Emily Jin, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, and Roberto Martín-Martín. 2023. Modeling Dynamic Environments with Scene Graph Memory: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

1 INTRODUCTION

The use of relational object-centric representations, such as scene graphs [3, 6, 8, 14], has gained momentum in addressing the active research question of how to represent the state of a large environment to enable agents to make better decisions in AI and robotics. Scene graphs are efficient at representing object and relational information and scale well to large natural scenes, making them useful for downstream decision-making tasks such as navigation, manipulation, and search [1, 2, 10, 13, 15, 20].

However, in dynamic environments that are only partially observable to the agent, scene graphs are often incomplete and unreliable. To address this challenge, we propose new state representation named **Scene Graph Memory** (SGM), which encodes all nodes and edges observed by the agent, including those that may no longer be true, in a single graph. To predict the relationships between pairs of objects (including locational relationships such as "inside of", or "on top"), all that is needed is to predict the likelihood of a given edge, which is an instance of the link prediction problem [9]. To perform link prediction on scene graphs, the proposed method combines SGMs with a new neural net architecture names **Node Edge Predictor**, which allows for better generalization and enables agents to operate in dynamic, partially observable environments that are unseen during training. Lastly, we address the lack of existing benchmarks for link prediction for partially observable graphs by introducing the **Dynamic House Simulator**, a new simulator that produces plausible scene graphs of household environments that change dynamically as simulated humans would act on them. The approach proposed in this study will enable embodied AI agents to make efficient decisions in daily tasks such as navigation, object search, rearrangement, and household chores.

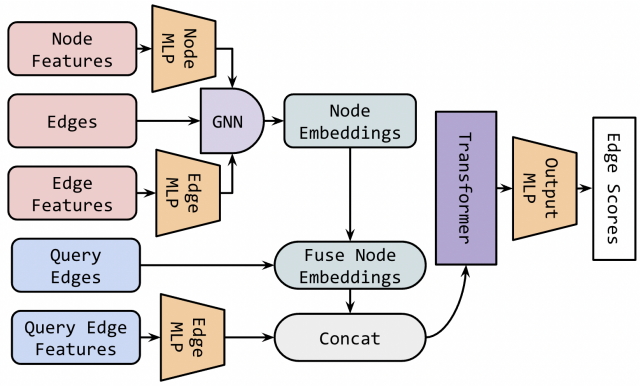

Figure 1: Node Edge Selector (NEP) model architecture.

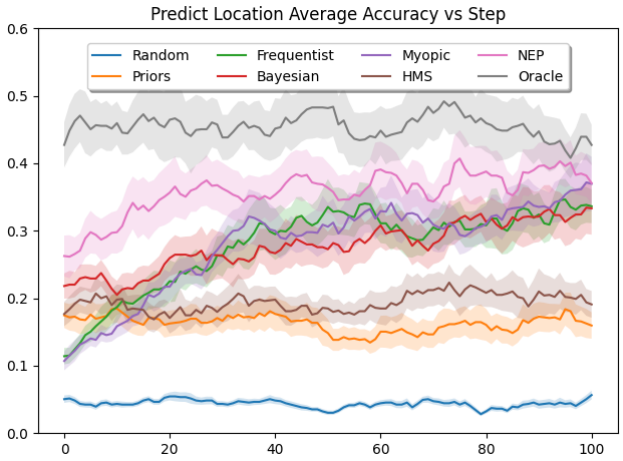

Figure 2: The average accuracy and variance for the Predict Object Location task, average across 100 different runs/environments. The task is Predict Object Location (with dynamic nodes and coarse environment priors).

2 METHOD

While there are many simulation frameworks built for evaluating embodied agents in household environments [4, 5, 12, 16, 17], none of them support simulating realistic object movement over time. So, we implemented a Dynamic Household Simulator that supports sampling a wide variety of household environments, where each environment has a distinct set of initial objects, and a distinct pattern of object placement and movement. These environments are represented with scene graphs [3], with nodes being either locations or objects and edges being furniture locations or object placements.

We propose the Scene Graph Memory (SGM) data structure to enable the agents to estimate environment-specific dynamics. An instance of a Scene Graph Memory $SGM_t = (V_t^{SGM}, E_t^{SGM})$ is composed of a set of nodes and edges of the same type as in the environment scene graphs. The SGM nodes $V_t^{SGM} = \bigcup_{n=0}^{t}(V_n^O \bigcup Q_n)$ are made up of all the observed nodes and the queried nodes (Q_n) up until timestep t. The SGM edges $E_t^{SGM} = (\bigcup_{n=0}^{t} E_n^O) \bigcup E_n^H$ are made up of all the observed edges, and hypothetical edges $E_t^H = f(Q_t)$ up until timestep t. Each node and edge in the SGM is associated with features reflecting the semantic properties of the object or relationship it represents as well its observed dynamics.

Using SGMs, we can predict the location of an object by estimating the likelihood of its edges. We propose the Node Edge Predictor model to do so, shown in Figure 1. It is similar to other GNN neural networks used for link prediction, with two main novel elements. The first key novel aspect of the model is that a set of query edges are part of the input, and the model is optimized to predict the likelihood of just these edges, as opposed to predicting the likelihood of all edges in the graph. Secondly, we add a transformer [18] prior to outputting edge probabilities, so as to enable the model to evaluated them relative to each other rather than independently. This is most similar to GraphFormer from Yang et al. [19] Heterformer from Jin et al. [7], but differs in that NEP uses the transfomer only on the embedded query edge, as opposed to the costlier use of transformers in alteration with GNN layers.

3 RESULTS

We design our experiment to test whether the proposed NEP model can outperform alternative approaches on the following task: at every step, the agent must predict the location (furniture node) of a queried object. The query object is sampled at random either from the the set of objects that has moved since the last step or from all objects nodes. The agent is then able to observe the node it has predicted and its associated object nodes, regardless of whether its output is correct. We simulate observation error by randomly skipping 25% of the objects during observation. Once the agent has received the observation, objects are moved and the next step begins.

We compare to these baselines: **Random:** Randomly chooses an edge. **Frequentist:** Chooses the option with the highest ratio of true observations to total observations. **Priors:** Chooses the most likely option according to priors. **Myopic:** Always chooses the last location each object was observed at, or at random if the object has not been observed. **Bayesian:** Treats each edge in the SGM as having a distinct beta-binomial probability distribution. **HMS:** The model from [11], which addresses a very similar problem to ours. **Oracle:** Uses ground truth knowledge about the dynamics of the scene as well as memory to make the best choice possible.

The main results can be seen in Figure 2. NEP significantly outperforms all the baselines from the very beginning to the end. There are two main reasons. First, compared to models that completely rely on the prior information and do not adapt during test time (Priors), NEP has the capability to adapt to the noisy dynamics of the particular environment it is deployed in. Second, NEP is able to leverage its starting priors and learned scene statistics from the training phase before testing. This gives NEP a head start compared to the models that purely rely on observations made during test time (Myopic and Frequentist).

The results support our initial hypothesis that combining the proposed representation (SGM) with a learning-based approach is suitable for the temporal link prediction tasks in dynamic partially observable graphs, since this combination allows generalization to unseen environments as well as on-line adaptation.

REFERENCES

[1] Christopher Agia, Krishna Murthy Jatavallabhula, Mohamed Khodeir, Ondrej Miksik, Vibhav Vineet, Mustafa Mukadam, Liam Paull, and Florian Shkurti. 2022. Taskography: Evaluating robot task planning over large 3D scene graphs. In *Conference on Robot Learning*. PMLR, 46–58.

[2] Saeid Amiri, Kishan Chandan, and Shiqi Zhang. 2022. Reasoning with scene graphs for robot planning under partial observability. *IEEE Robotics and Automation Letters* 7, 2 (2022), 5560–5567.

[3] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 2019. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5664–5673.

[4] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. 2022. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. *arXiv preprint arXiv:2206.06994* (2022).

[5] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, et al. 2020. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954* (2020).

[6] Nathan Hughes, Yun Chang, and Luca Carlone. 2022. Hydra: A Real-time Spatial Perception Engine for 3D Scene Graph Construction and Optimization. *arXiv preprint arXiv:2201.13360* (2022).

[7] Bowen Jin, Yu Zhang, Qi Zhu, and Jiawei Han. 2022. Heterformer: A Transformer Architecture for Node Representation Learning on Heterogeneous Text-Rich Networks. *arXiv preprint arXiv:2205.10282* (2022).

[8] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3668–3678.

[9] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. 2020. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications* 553 (2020), 124289.

[10] Gulshan Kumar, N Sai Shankar, Himansu Didwania, Ruddra Dev Roychoudhury, Brojeshwar Bhowmick, and K Madhava Krishna. 2021. GCExp: Goal-Conditioned Exploration for Object Goal Navigation. In *IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 123–130.

[11] Andrey Kurenkov, Roberto Martín-Martín, Jeff Ichnowski, Ken Goldberg, and Silvio Savarese. 2021. Semantic and geometric modeling with neural message passing in 3d scene graphs for hierarchical mechanical search. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11227–11233.

[12] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. 2021. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272* (2021).

[13] Son Nguyen, Ozgur S Oguz, Valentin N Hartmann, and Marc Toussaint. 2020. Self-Supervised Learning of Scene-Graph Representations for Robotic Sequential Manipulation Planning.. In *CoRL*. 2104–2119.

[14] Zachary Ravichandran, Lisa Peng, Nathan Hughes, J Daniel Griffith, and Luca Carlone. 2022. Hierarchical representations and explicit memory: Learning effective navigation policies on 3D scene graphs using graph neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[15] Iury B de A Santos and Roseli AF Romero. 2022. A Deep Reinforcement Learning Approach with Visual Semantic Navigation with Memory for Mobile Robots in Indoor Home Context. *Journal of Intelligent & Robotic Systems* 104, 3 (2022), 1–21.

[16] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9339–9347.

[17] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. 2021. iGibson 1.0: a simulation environment for interactive tasks in large realistic scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7520–7527.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. 2021. GraphFormers: GNN-nested transformers for representation learning on textual graph. *Advances in Neural Information Processing Systems* 34 (2021), 28798–28810.

[20] Yifeng Zhu, Jonathan Tremblay, Stan Birchfield, and Yuke Zhu. 2021. Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6541–6548.