RTransNav:Relation-wise Transformer Network for More Successful Object Goal Navigation

Extended Abstract

Kang Zhou
School of Computer Science, Wuhan University
Wuhan, China
2019102110013@whu.edu.cn

Chi Guo
Institute of Artificial Intelligence,Wuhan University,Hubei Luojia Laboratory
Wuhan , China
guochi@whu.edu.cn

Huyin Zhang
School of Computer Science,Institute of Artificial Intelligence, Wuhan University
Wuhan, China
zhy2536@whu.edu.cn

Wenfei Guo
Global Navigation Satellite System Research Center, Wuhan University
Wuhan, China
wf.g@whu.edu.cn

ABSTRACT

The task of object goal navigation is to drive an embodied agent to finding the location of given target only using visual observation. The mapping from visual perception of observation determines the navigation actions. We consider the problem of generalization for the agent across scenes to be lacking good visual perception and spatial reasoning ability. The mutual relationships between edges and objects in the observation is the essential part of scene graph, which reflect the deep understanding of visual perception. Despite recent advances, such as visual transformer and contextual information embedding, the visual perception of graph representation remains a challenging task. In this work, we propose a novel Heterogeneous Zone Graph Visual Transformer formulation for graph representation and visual perception. It consists of two key ideas:1)Heterogeneous Zone Graph (HZG) that explore the heterogeneous target related zones graph and spatial information. It allows the agent to navigate efficiently. 2) Relation-wise Transformer Network (RTN) that transforms the relationship between previously observed objects and navigation actions. RTN extracts rich nodes and edges features as pay more attention on the target-related zone. We model self-attention on the node-to-node encoder and cross-attention on the edge-to-node decoder. The HZG-based model and RTN are shown to improve the agent's policy and to achieve SOTA results on the commonly-used datasets.

KEYWORDS

Visual navigation, Knowledge graph, Reinforcement learning, Relation-wise transformer network, Indoor mobile robot

ACM Reference Format:
Kang Zhou, Chi Guo, Huyin Zhang, and Wenfei Guo. 2023. RTransNav:Relation-wise Transformer Network for More Successful Object Goal Navigation: 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.

1 INTRODUCTION

Several recent works have resorted to explore scene graph from different perspectives[7, 10]. Scene context information is exchanged either globally [2] or across neighborhoods [5]. Yang et al. [4] first proposed the use of graph convolutional networks (GCNs) to encode the object information. However, their fixed object graph is too generic to adapt to specific environments. Du et al. [3] used a graph attention layer to improve the adaptability of object relation graphs. Zhang et al. [9] put forward a hierarchical object-to-zone (HOZ) graph to guide agents search the target. The above methods used object features to better understand complete scenes and predict possible target locations.

However, some irrelevant latent information, such as "countertop" zone, "table" zone information is missing for locating possible position of the target. Besides, the edge information between objects is also important for reasoning directions for the agent. This spatial reasoning is vital for guiding the agent selecting actions, especially in deadlock.

To address above problem, learning and interacting among objects and their corresponding edges through contextual attention for scene graph is necessary. This paper proposed a novel transformer-based formulation for the object goal navigation task. In the following, we rationalize the self-attention and cross-attention of the proposed RTN in context of ObjNav task. HZG provides the agent with scene priors that can generalize across scenes. RTN extract HZG spatial features and concentrate more on target related area, thus making the navigation more efficient. Experiments show our navigation framework can get SOTA results in AI2THOR environment.

2 PROPOSED VISUAL NAVIGATION METHOD

We build the HZG during the training begins using object detectors DETR [1]. If the current view contains many objects belonging to the same category, it is recorded only once in the object triple. N represents the number of object categories and

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.

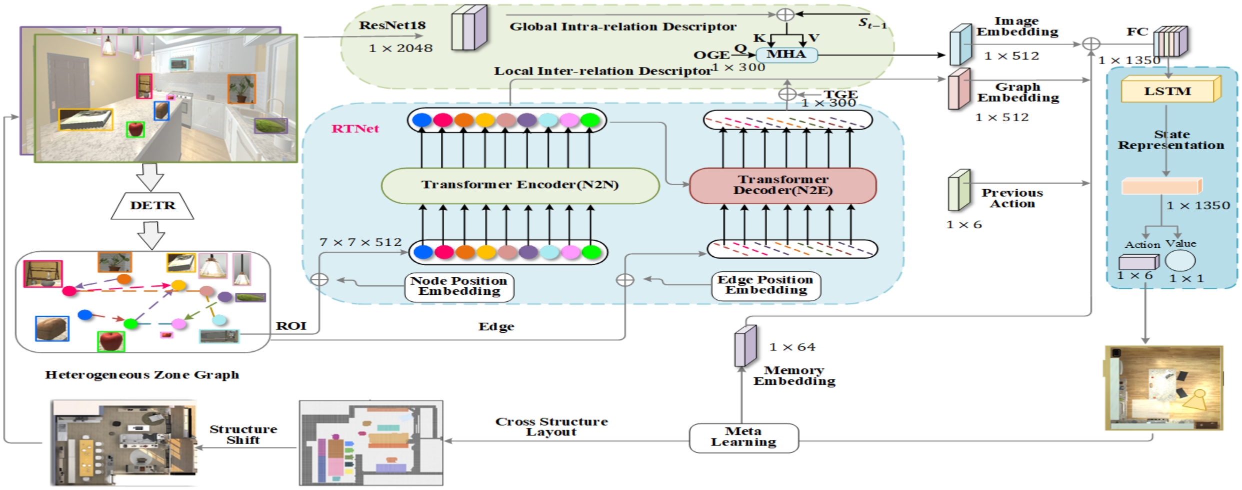

Figure 1: Overview of our proposed navigation framework. Visual features are obtained from Resnet18, target word embedding features extracted from Glove and HZG represents spatial features. Graph attention network is used to reason the primary references (objects) to make knowledge reasoning in guiding RL action sample. Meanwhile, meta learning supervise the RL's training process and providing adaptation in HZG.

$l = \{x, z, \theta_{yaw}, \theta_{pitch}, ROI\}$ represents the observation position, where x, z represent the horizontal coordinates and θ_{yaw}, θ_{pitch} represent the yaw and pitch angles of the agent. ROI is the fusion region of interest features for all objects' (n) ROI in the zone, where $ROI = Concat(roi_n)$. Then k-means clustering feature f is used to obtain K regions and form the zone-level HZG $Z_m(V_m, E_m)$.

The structure of RTN is shown in Figure. 2. RTN consists of encoder for N2N(node to node) and decoder for E2N(edge to node). N2N is represented using $f_i^{in} = W_{node}([v_i; s_i; b_i])$ where (W_{node}) is linear projection to get i^{th}node of the initial node eigenvector (f_i^{in}) In addition, for i^{th}node, we added a positional feature vector $(pos(n_i))$ and the characteristics of its initial f_i^{in}. It takes the categorical position of the i^{th} node in linear ordering of all nodes, and converts it to a continuous sinusoidal vector $f_i^{final} = encoder(f_i^{in} + pos(n_i))$ as described in [8].Since our edges are directed, we assume that the proposed edge position embedding is to distinguish the source node from all the different nodes. The goal is to accumulate the necessary global context (all the different object instances) without losing focus on the local context. We define $PE_{e_{ij}}(k, k+1) = [\sin(\frac{p_i}{m^{\frac{2k}{d}}}), \cos(\frac{p_i}{m^{\frac{2k}{d}}})]$, where $p_i$and $p_j$is the location of the node n_i and n_j, m is a maximum number of nodes in the sequence, $d = 2048$, k denotes positional coding feature vector of k^{th}. Finally, the output of RTN fused the node and edge information in the scene graph with high efficient node message. The overal of our proposed navigation framework is shown in Figure. 1. The framework uses Glove [6] to generate a 300-dimensional semantic embedding of target and graph objects, in total of 92 objects. The input of our actor-critic network represents 512 hidden states LSTM network and two fc layers representing actor and critic. It is concatenated with the target object as a 300-dimensional vector, observation features, as a 1024-dimensional feature vectors and HZG with 92 nodes is fed into GAT producing 92-dimensional vectors. Meanwhile, GAT is also used to make knowledge inference for producing a single value which is appended in our critic. The actor outputs a 6-dimensional actions distribution $\pi(a_t|x_t)$. The critic estimates a single value using softmax.

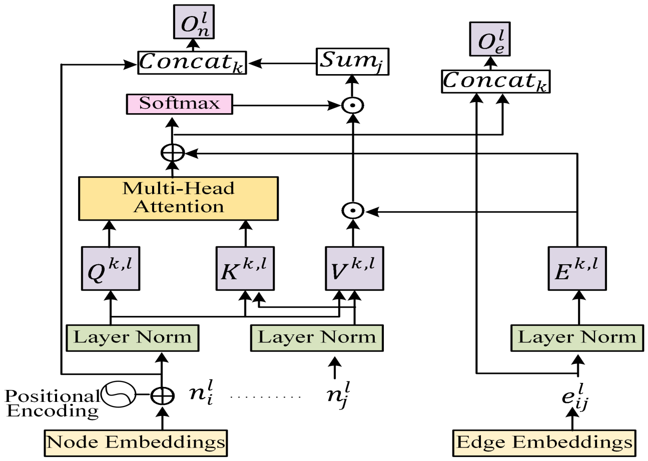

Figure 2: The structure of Relation-wise Transformer Network

Extensive experiment and ablation study demonstrate the efficiency of our approach in benefiting from our HZG when necessary, which allows for adapting to the new environment. We show for the first time that knowledge distillation from RTN and HZG can improve the performance of navigation agents.

ACKNOWLEDGMENTS

We acknowledge the support of the National Key Research and Development Program of China under Grant 2022YFB3903801, Key Science and Technology Program of Hubei Province, and the Open Fund of Hubei Luojia Laboratory.

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

[2] R. Druon, Y. Yoshiyasu, A. Kanezaki, and A. M. Watt. 2020. Visual Object Search by Learning Spatial Context. *IEEE Robotics and Automation Letters* PP, 99 (2020), 1–1.

[3] Heming Du, Xin Yu, and Liang Zheng. 2020. Learning Object Relation Graph and Tentative Policy for Visual Navigation. In *European Conference on Computer Vision*. Springer, 19–34.

[4] R. Krishna, Y. Zhu, O. Groth, J. Johnson, and F. F. Li. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017).

[5] Yunlian Lv, Ning Xie, Yimin Shi, Zijiao Wang, and Heng Tao Shen. 2020. Improving target-driven visual navigation with attention on 3D spatial relationships. *arXiv preprint arXiv:2005.02153* (2020).

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[7] Kaili Sun, Chi Guo, Huyin Zhang, and Yuan Li. 2022. HVLM: Exploring human-like visual cognition and language-memory network for visual dialog. *Information Processing & Management* 59, 5 (2022), 103008.

[8] 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).

[9] Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakui Chu, and Shuqiang Jiang. 2021. Hierarchical Object-to-Zone Graph for Object Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 15130–15140.

[10] Kang Zhou, Chi Guo, Huyin Zhang, and Bohan Yang. 2023. Optimal Graph Transformer Viterbi knowledge inference network for more successful visual navigation. *Advanced Engineering Informatics* 55 (2023), 101889.