Learning to Navigate in Interactive Environments with the Transformer-based Memory

Weiyuan Li
East China Normal University
10162100162@stu.ecnu.edu.cn

Ruoxin Hong
East China Normal University
51205904033@stu.ecnu.edu.cn

Jiwei Shen
East China Normal University
sjwcee@gmail.com

Yue Lu
East China Normal University
ylu@cs.ecnu.edu.cn

Abstract

Substantial progress has been achieved in embodied visual navigation based on reinforcement learning (RL). These studies presume that the environment is stationary where all the obstacles are static. However, in real cluttered scenes, interactable objects (e.g. shoes and boxes) blocking the way of robots makes the environment non-stationary. We formulate this interactive visual navigation as a Partial Observed Markov Decision Problem. To handle it, we propose a transformer encoder to learn a belief state which captures the long spatial-temporal dependencies of the aggregated observations in the memory. However, leveraging the transformer architecture in the RL settings is highly unstable. We propose a surrogate objective to predict the next waypoint, which facilitates the representation learning and bootstrap the RL. We demonstrate our method in the iGibson environment and experimental results show a significant improvement over the interactive Gibson benchmark and the related recurrent RL policy both in the validation seen scenes and the test unseen scenes.

1. Introduction

Traditional navigation tasks require the agent to reach the destination under the premise of avoiding obstacles in a static environment which is applicable in some empty scenarios such as outdoors or in factories. However, the dynamic scenarios inherent to real human environments, such as offices and coffee shops, contain a large number of interactive objects, such as furniture, toys, shoes, etc. The robot has to interact with the environment by pushing or moving the obstacles away to clear the path. Learning to navigate in the interactive environment considering the physical interaction remains several challenge.

In the interactive navigation task, the main challenge is the partial observation problem that arises from the non-stationary environment where the interactable objects are randomly placed blocking the way to the destination. Therefore, the robot has to figure out whether it can push the obstacles out of the way by leveraging the history information. To overcome it, it is popular to use recurrent neural networks (RNNs) such as LSTM or GRU [2] to aggregate the past observation and actions [5, 7]. However, RNNs' inability to capture the long-term dependencies of the memory is not suitable for this task. We apply a transformer encoder, which is improved from the scene memory transformer [3] with residual connection and local attention mechanism, to handle the long sequence and ensure the stability of training simultaneously.

In general, to tackle interactive visual navigation, We propose a deep reinforcement learning model with the transformer-based memory, which can learn to navigate in a cluttered room and interact with the obstacles based on the accumulated experience. Besides, a surrogate objective is proposed to optimize the transformer encoder for a stable and sufficient state representation. We train and evaluate the model in the Interactive Gibson Environment [8] rendering based on real-world homes and additional interactable objects from the Google Scanned Objects dataset [4]. Our approach outperforms the Interactive Gibson Benchmark with over 17% improvement in success rate and 9.2% in success weighted by shortest path (SPL) [1].

2. Approach

As described above, to alleviate the partial observed problem, we define a state representation ϕ_Z as a stochastic mapping from the historical observation sequences to a representation space Z: $p(Z = z_t|O_0, O_1, \dots, O_t)$. We formulate the interactive navigation task as a partially observed Markov decision problem (POMDP). Formally, the POMDP is defined as a tuple (Z, A, τ, r, γ), where Z, A, r are the belief state, action, and reward. τ is the belief state transition function and $\gamma \in [0, 1]$ is the discount factor.

2.1. Belief state encoder

Memory The memory is initialized as an empty set at the beginning of each episode. During the robot exploration period, we maintain the M_t in fixed-length l by storing the embedding of the current observation O_t.
Transformer encoder Motivated by GTrXL [6], based on the Attention function [9], we modify the position of the layer normalization (LN) to reconstruct the $\mathrm{AttBlock}$. The $\mathrm{AttBlock}$ takes the $e_t \in \mathbb{R}^{1\times d_e}$, $M_t \in \mathbb{R}^{l\times d_e}$ as the input, where l is the size of memory, d_e is the dimension of the embedding. Therefore, the embedding of the current observation can directly flow through the $\mathrm{AttBlock}$ without any transformation, which stabilizes the transformer-based RL algorithm substantially.

$$\begin{gathered}\mathrm{AttBlock}(e_t, M_t) = \mathrm{FC}\left(\mathrm{LN}\left(H\right)\right) + H,\\ \text{where } H = \mathrm{Att}(e_t W^Q, M_t W^K, M_t W^V, \mathrm{Mask}(M_t)) + e_t\end{gathered} \tag{1}$$

where the Mask is the function calculated according to the dimension of M_t. $W^Q \in \mathbb{R}^{d_e\times d_k}$, $W^K \in \mathbb{R}^{d_e\times d_k}$ and

The transformer encoder comprises a stack of P $\mathrm{AttBlock}$, where the output of each $\mathrm{AttBlock}$ can be viewed as the query over the next $\mathrm{AttBlock}$.

2.2. Learning objective of the state representation

The optimal shortest path between the agent and the goal is discretized into several waypoints. Intuitively, if the state representation ϕ_Z is sufficient for the optimal navigation policy $\pi^*(a_t|z_t)$, it should preserve enough information for the prediction of the next waypoint. Accordingly, the prediction of the next waypoint can be regarded as the surrogate objective to optimize the state representation ϕ_Z.

During the training stage, the waypoints sampled from the shortest path can be calculated by A^* algorithm on the global map information S. Therefore, we can minimize the KL divergence between the first optimal waypoint w given the state s_t and the probability distribution of the output $\hat{w}$ of the prediction network given z_t.

$$\begin{aligned}\mathrm{KL}(p(w|s_t)||p(\hat{w}|z_t)) &= \mathbb{E}_{p(w|s_t)}[\log p(w|s_t) - \log p(\hat{w}|z_t)]\\ &= \mathcal{H}(w) - \mathbb{E}_{p(w|s_t)}[\log p(\hat{w}|z_t)],\end{aligned} \tag{2}$$

where $\mathcal{H}(w)$ is the entropy of the optimal waypoint and it dose not influence on the optimization process. Then, the optimzation objective of the state representation is to maximize $\mathbb{E}_{p(w|s_t)}[\log p(\hat{w}|z_t)]$.

Accordingly, the loss function of the surrogate objective to predict the next optimal waypoint contains two components specifying: 1) the angle prediction of the waypoint $\mathcal{L}_\theta$, and 2) the distance prediction of the waypoint $\mathcal{L}_d$.

$$\mathcal{L} = \sum_{i=s1}^{L}(1 - \cos(\theta_i - \hat{\theta}_i)) + \sum_{i=1}^{L}(d_i - \hat{d}_i)^2, \tag{3}$$

where L is the length of the sequence, θ and d are the angle and distance of the optimal waypoint relative to the robot respectively. $\hat{\theta}$ and $\hat{d}$ are the corresponding estimation.

3. Experiment

	Validation Seen				*Test Unseen*			
Methods	SR	SPL	EE	INS$_{0.5}$	SR	SPL	EE	INS$_{0.5}$
Baselines:								
Random	1.0	0.4	-	-	0.6	0.2	-	-
Xia et al. [10]	68.8	41.5	92.5	67.0	59.7	37.0	92.3	64.6
Savva et al. [7]	72.3	43.1	**95.7**	69.4	66.7	40.3	**95.6**	67.9
ours	**82.8**	**48.7**	94.5	**71.6**	**76.7**	**46.2**	93.9	**70.0**
Ablations:								
ours w/o $\mathcal{L}_\theta$	70.8	43.0	94.4	68.7	61.0	38.8	92.4	65.6
ours w/o BSE	78.2	45.6	93.4	69.5	68.3	41.7	92.7	67.2

Table 1. **Performance comparison.** We show the result of the comparison of the performance of our method with the previous state-of-the-art method along with the ablation study.

Representation study. According to the performance of RL agents trained with different state representations, which are optimized with the corresponding surrogate objective. Optimizing the prediction of the angle between the agent and the next waypoint produces sufficient representation while estimating the distance may have difficulty predicting the waypoint. The end-to-end RL performs poorly.
Comparative experiment. We involve the performance of several baselines including Random, Xia et al. [10] and Savva et al. [7] The results are shown in Table 1. Randomly sampling the policy from the action space is impossible to work. Notably, our method outperforms the baselines in SR, SPL, and INS while the EE is slightly lower than Savva et al.'s work because of the increasing interaction. Particularly, the success rate increases by over 10% compared to the best of the baseline. The comparison with baselines justifies the effectiveness of our method.
Ablation study. The performance of 'ours w/o $\mathcal{L}_\theta$' drops noticeably, which indicates that the angle prediction is important when the robots planning the route. The performance of 'ours w/o Belief State Encoder (BSE)' ablation decreases because the temporal information in the memory is essential for the robot to fully understand the environment and overcome the problem of partial observation.

4. Conclusion

We propose a transformer-based reinforcement learning to tackle the interactive visual navigation task. With a transformer encoder, the robot can capture the long spatio-temporal information from the memory to decide whether to push the obstacles or not. The length of the memory is crucial as the longer memory contains more details while shorter memory can ensure the faster convergence of the transformer encoder. Furthermore, the proposed methods can also be generally applicable to point navigation, object navigation, or even vision language navigation.

References

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 1

[2] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. 1

[3] Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–547, 2019. 1

[4] GoogleResearch. Google scanned objects, August. 1

[5] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andy Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. In *International Conference on Learning Representations*, 2016. 1

[6] Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. Stabilizing transformers for reinforcement learning. In *International Conference on Machine Learning*, pages 7487–7498. PMLR, 2020. 2

[7] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 1, 2

[8] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martın-Martın, Linxi Fan, Guanzhi Wang, Shyamal Buch, Claudia D'Arpino, Sanjana Srivastava, Lyne P Tchapmi, Kent Vainio, Li Fei-Fei, and Silvio Savarese. igibson, a simulation environment for interactive tasks in large realistic scenes. *arXiv preprint*, 2020. 1

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[10] Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters*, 5(2):713–720, 2020. 2