

Learning hierarchical relationships for object-goal navigation

Yiding Qiu *
UC San Diego
yiqiu@eng.ucsd.edu

Anwesan Pal *
UC San Diego
a2pal@eng.ucsd.edu

Henrik I. Christensen
UC San Diego
hichristensen@eng.ucsd.edu

Abstract: Direct search for objects as part of navigation poses a challenge for small items. Utilizing context in the form of object-object relationships enable hierarchical search for targets efficiently. Most of the current approaches tend to directly incorporate sensory input into a reward-based learning approach, without learning about object relationships in the natural environment, and thus generalize poorly across domains. We present Memory-utilized Joint hierarchical Object Learning for Navigation in Indoor Rooms (MJOLNIR), a target-driven navigation algorithm, which considers the inherent relationship between target objects, and the more salient contextual objects occurring in its surrounding. Extensive experiments conducted across multiple environment settings show an 82.9% and 93.5% gain over existing state-of-the-art navigation methods in terms of the success rate (SR), and success weighted by path length (SPL), respectively. We also show that our model learns to converge much faster than other algorithms, without suffering from the well-known overfitting problem. Additional details regarding the supplementary material and code are available at <https://sites.google.com/eng.ucsd.edu/mjolnir>.

Keywords: Visual navigation, Object recognition, Learning

1 Introduction

Human beings are able to perform complex tasks such as object-goal navigation efficiently, with implicit memorization of the relationships between the different objects. For example, to navigate to a target such as a toaster in the kitchen, a natural thing to do is to start by looking around a set of larger candidate objects which are likely to be nearby, such as a stove or microwave. Unfortunately, this type of *hierarchical relationship* is rarely used in robot navigation. As a result, the agent usually fails to develop any intuition about the goal location when they are further away. In this work, we refer to the object goal as *target*, and those larger candidates which have either a spatial and semantic relationship with the target as *parent*. This is depicted in Figure 1. Two main challenges are addressed here - (i) Correctly associating the robot’s current observation with some prior intuition about object relationships into the model. (ii) Efficiently utilizing this hierarchical relationship for the visual navigation problem.

Existing research in this area tends to aggregate sensory input into a meaningful state, before sending it to a Reinforcement Learning (RL) framework, with the expectation that the robot can implicitly learn the navigation problem through recursive trials. Zhu *et al.*[1] proposed to solve this problem by finding the similarity between current observation and the target observation through a trained Siamese Network [2]. The work of Wortsman *et al.*[3] incorporated a meta-learning approach where the agent learns a self-supervised interaction loss during inference to avoid collision. However, none of these methods use any prior information or semantic context. Sax *et al.*[4] highlight that a set of mid-level visual priors such as depth and edge information, surface normals, keypoints, etc. can be useful for the navigation task. However, a different encoder is required for learning each of these representations, and thus, the model does not scale very well. Moreover, the learnt features are expected to vary with the test setting, thereby limiting the scope of the method in unseen dissimilar

*Equal contribution

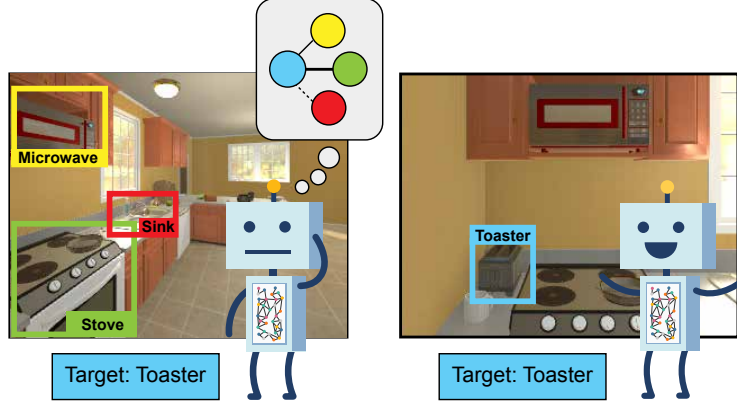


Figure 1: Illustration of the parent-target relationship. Upon seeing a set of parent objects (left image), the agent learns to associate the correct parent (Stove or Microwave, here) to the target object (Toaster, here) from the knowledge graph, and tries to search in its neighborhood to successfully locate the target object (right image). [Best when viewed in color]

environments. Instead, using the knowledge about object relationships is more robust since they are domain-independent. Two approaches which are similar to ours are Yang *et al.*[5], and Druon *et al.*[6]. In [5], semantic priors in the form of a knowledge graph are used to capture object-object connectivity, but this connectivity is defined only in terms of their spatial proximity, not the inter-object dependency. For instance, an object such as pillow in the bedroom might be visible next to both a bed and an alarmclock. Yet, we know from experience that to find a pillow, one should always start by looking around the bed. This type of common-sense knowledge is missing from their method. In contrast, our proposed approach assigns a sub-goal reward to force the agent to learn this key hierarchical relationship. Also, important information pertaining to an object’s spatial location in the scene is missing as they learn a single context for the whole scene. Instead, we utilize an object detector to capture this information explicitly. Recently, [6] introduced the concept of a *context grid* where they modeled the spatial (through an object detector) and language (through a word-embedding) similarity between the target and other objects as a 16×16 grid. However, this relationship does not update during the training stage, thereby making the model less adaptable. Additionally, their action space is much bigger, thus simplifying the navigation task.

A major issue with RL algorithms in a real-world setting is efficient modeling of the continuous high-dimensional state space in the agent’s surrounding [7]. While implementing other algorithms [1, 5], it was observed that even though the convergence during training stage was fast, they gave a relatively poor performance during testing. A possible reason for this is that due to their non-representative state space, the agent is perhaps simply learning to memorize the training setting after a certain number of episodes rather than understanding the underlying object relationships, and thus it fails to generalize to a novel scene. The adaptive gradient of Wortsman *et al.*[3] alleviates this problem to some extent. Nevertheless, it doesn’t remove it completely since it does not consider any prior memory into account. In contrast, our representative state leverages the correct balance of prior context and current observation to provide sufficient information about the surrounding during training, while simultaneously being abstract enough to generalize to a different room layout during testing.

Contributions We propose a hierarchical object relationship learning approach to the object-goal navigation problem by understanding the role of semantic context. This is done through the proposed novel *Context Vector* as a node embedding in the graph convolutional neural network. We also emphasize the role object detection plays as opposed to a scene representation from traditional image classification networks. Finally, we incorporate the mentioned parent-target object relationship through a new reward shaping mechanism.

The remainder of the paper is organized as follows. In Section 2, we discuss related work, followed by the task definition in Section 3. The core ideas behind our approach are described in Section 4. In Section 5 we discuss the dataset used and the overall experimental design and results, before we summarize the work and outline future challenges in Section 6.

2 Related Work

Map vs map-less navigation approaches - Traditional approaches in visual navigation involved formalizing this as an obstacle avoidance problem, where the agent learns to navigate in its environment through a collision-free trajectory. This is done either in the form of offline maps [8, 9, 10], or online maps [11, 12, 13, 14] generated through Simultaneous Localization and Mapping (SLAM) techniques. Given this map as input, the typical approach was to employ a path planning algorithm such as A* [15] or RRT* [16] to generate a collision-free trajectory to the goal. The limitation of these algorithms is that it might not be possible to have a pre-computed map of the environment. Generating a rich semantic map online is also a non-trivial task. With the advent of deep learning, the focus has instead shifted towards methods which are map-less [17, 18, 19, 20, 21], meaning that the representations can be learnt over time through interactions. However, most of these algorithms are not suited for finding specific target objects in a previously unseen environment. In contrast, our approach solves the target-driven navigation problem entirely using only visual inputs without the need of a pre-computed map.

Point-goal navigation vs object-goal navigation - Point-goal navigation [22] refers to the problem where an agent starts from a randomly chosen pose and learns to navigate to a specific target point, usually specified in terms of 2D/3D coordinates relative to the agent. Lately, there has been some research [23, 24] in this area. However, it is still an ill-defined problem in a realistic setting, thereby making comparison difficult [25]. Object-goal navigation refers to the problem where the agent instead learns to navigate to a specified target object while successfully avoiding obstacles. These tasks usually require some prior knowledge about the environment which can be useful for navigation [1, 3, 5, 6]. Our approach falls in this category, but involves learning a robust contextual object-object relationship.

Role of learning semantic context - Learning semantic context of the surrounding world is an important research topic in the computer vision and robotics community. However, most of the existing work [26, 27, 28, 29, 30, 31] surrounds static settings such as object detection, semantic segmentation, activity recognition etc. Recently, object-object relationship modeling has been studied for tasks such as image retrieval [32] using scene graphs, visual relation detection [33], visual question-answering [34, 35], place categorization [36], and driver saliency prediction [37]. In this paper, we propose an algorithm which successfully learns to exploit hierarchical object relationships for object-goal visual navigation.

Reward shaping for policy networks - Reward shaping is a method in reinforcement learning for engineering a reward function in order to provide more frequent feedback on appropriate behaviors [38]. In general, defining intermediate goals or sub-rewards for an interaction-based learning algorithm is non-trivial, since the environment model is not always known. However, a divide-and-conquer approach is often imperative to exploit the latent structure of a task to enable efficient policy learning [39, 40, 41]. Specifically, for the object-goal visual navigation problem, it is important to learn the inherent “parent-target” object relationship for providing a meaningful feedback to the end-to-end training. For the policy network, we use the Asynchronous Advantage Actor-Critic (A3C) [42] algorithm to sample the action and the value at each step as per the approach of other models [1, 3, 5].

3 Task Definition

The goal of the object-goal navigation problem is to find a target object, defined through a set $T = \{t_1, \dots, t_N\}$, in a given environment. The problem is defined purely from a vision perspective, and therefore any information about the environment in the form of a semantic or a topological map is not provided. The agent is spawned at a random location in the environment at the beginning of an episode. The input to the model is current observation in the form of RGB images, and the target object’s word-embedding, rolled over each time-step. Using this, the agent has to sample an action a from its trained policy given a set of actions A , where $a \in A = \{\text{MoveAhead}, \text{RotateLeft}, \text{RotateRight}, \text{LookUp}, \text{LookDown}$ and $\text{Done}\}$. The `MoveAhead` action takes the agent forward by 0.25 meters, while the `RotateLeft` and `RotateRight` actions rotate it by 45 degrees. Finally, the `Look` action tilts the camera up/down by 30 degrees. An episode is considered a “success”, if the target object is visible, meaning the agent can detect it in the current frame, and is within a

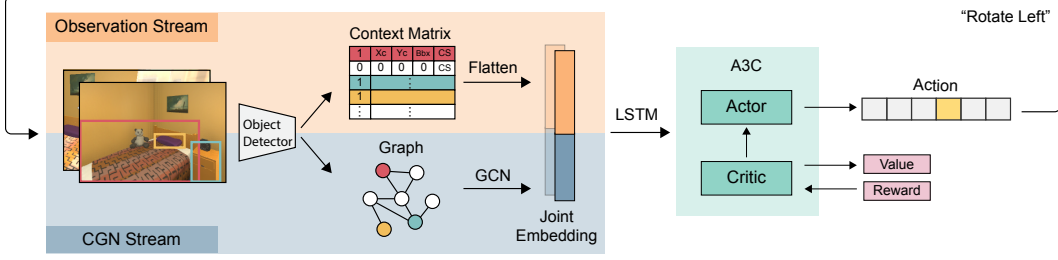


Figure 2: The entire MJOLNIR architecture. In the Observation stream, we use the ground-truth object detector to construct the context vector for all objects in our environment. This observation vector (orange) is concatenated with the graph embedding (blue) from the CGN stream to form a joint embedding. This is then sent to an LSTM cell and finally fed to the A3C model. [Best when viewed in color]

distance of 1.5 meters from it. When the “Done” action is sampled by the agent, the episode ends and the model checks if this criterion is met.

4 Memory-utilized Joint hierarchical Object Learning for Navigation in Indoor Rooms (MJOLNIR)

Parent-Target object relationship - In addition to the target objects, we introduce a new set of object classes, defined by the set $P = \{p_1, \dots, p_M\}$. These “parent objects” consist of the larger objects present in a room, which also happen to be spatially/semantically related to the target object. For example, CounterTop is a parent object in the Kitchen and Bathroom scenes, while Shelf is a parent object in Living room and Bedroom. The set of parent objects, P , are manually picked for each room based on the strong correspondences they have with the target object list, T , in the knowledge graph (explained below). The aim of the navigation agent is to start by exploring the area around $p \in P$, eventually leading to the target object $t_i \in T$.

Construction of Knowledge Graph and the Context vector - Similar to [5], our knowledge graph is also constructed using the objects and relationships extracted from the image-captions of the Visual Genome (VG) dataset [43]. However, by pruning a lot of the object (for instance, “armchair” vs “arm chairs”) and relationship (for instance, “near” vs “next to”) aliases, we were able to build a cleaner adjacency matrix for the graph convolution network, containing strong relationship correspondences between those VG objects, which also appear in our experimental setting.

In addition to the newly constructed graph, we also introduce a novel *context vector* \mathbf{c}_j for each object $o_j \in O$, where O is the list of all the 101 objects in our environment. This 5-D vector gives information regarding the state of o_j in the current frame, and can be represented as $\mathbf{c}_j = [b, x_c, y_c, bbox, CS]^T$. The first element, b , is a binary vector specifying whether o_j can be detected in the current frame. The next two elements, (x_c, y_c) , and $bbox$ correspond to the center (x, y) coordinates of the bounding box of o_j , and its covered area, both normalized with respect to the image size. Finally, CS is a number giving the cosine similarity between the respective word embeddings of o_j , and the target object $t \in T$. This is expressed as:

$$CS(\mathbf{g}_{o_j}, \mathbf{g}_t) = \frac{\mathbf{g}_{o_j} \cdot \mathbf{g}_t}{\|\mathbf{g}_{o_j}\| \cdot \|\mathbf{g}_t\|}$$

where \mathbf{g} denotes the word embeddings in the form of GloVe vectors [44].

End-to-end model - In this section, we talk about our entire network, shown in Figure 2. We adopt a two-stream network approach, consisting of (i) the Observation stream, which encodes the agent’s current observation in the environment, and (ii) the Contextualized Graph Network (CGN) stream, which embeds the prior memory obtained through a knowledge graph $G = (V, E)$.

For the Observation stream, we try two variants - (i) We use the ResNet-18 [45] conv features obtained from the current frame to give a holistic representation of the scene. This feature map is then combined with the target object word embedding using point-wise convolution, and flattened to obtain the observation vector. (ii) We replaced the conv features with the 5-D *context vector*

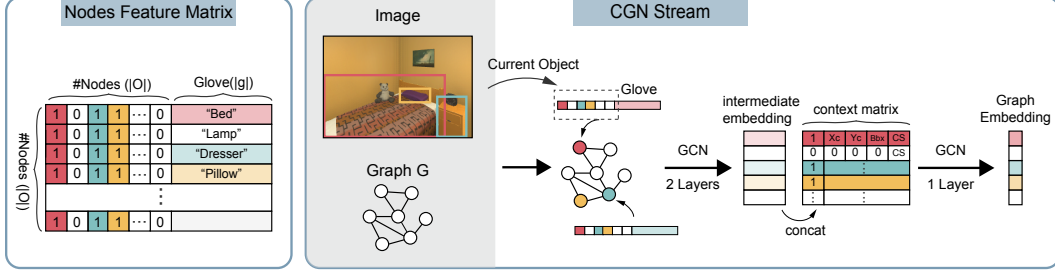


Figure 3: The novel CGN architecture. The input node feature constitutes the current detected object list and the node object’s glove embedding. This is passed through two layers of GCN. This intermediate embedding is concatenated with the context vector and another layer of GCN is trained on it. [Best when viewed in color]

(described above) for every object in our environment. The resulting *context matrix* $\in R^{|O| \times 5}$ is then flattened and forms the observation vector.

Even though the input knowledge graph, G , provides a strong initial guidance to the agent, this information, by itself, can be insufficient due to the domain difference between VG [43] and our navigation environment. The CGN stream is used to diminish this gap. We use a graph convolution network (gcN) [46] to learn the node embeddings. Given G , we design the node feature vector $X \in R^{|O|+|g|}$ which is the concatenation of the output of an object detector on the current frame, specified by the $|O| = 101$ -dimensional vector having 1 for the current frame objects, and 0 for others, along with each node object’s word-embedding. “101” is the length of the exhaustive list of object types in our simulator, which is analogous to the number of object classes in a trained object detector. This is different from the 1000-dimensional class probability used by Yang *et al.*[5]. The reason for this change is two-fold - (i) the probability vector obtained from ResNet-18 [45] is pre-trained on the 1000 classes of ImageNet [47], which are quite different from the object list present in our environment, and (ii) since the pre-trained network primarily learns to classify a single object, in the multi-object setting, it is more likely to generate noisy labels². The combined input node features are passed through two layers of gcN, to generate the intermediate embedding. This new feature is then concatenated with the *context vector*, and then fed to another layer of gcN to generate the final graph node embedding. The concatenation of the observation vector and the graph embedding results in the joint embedding (shown in Figure 2), which is the input to the A3C model. The CGN stream is detailed in Figure 3.

To separately highlight the contribution of the changes made to each of the streams, we present two variants of our algorithm. MJOLNIR-r uses ResNet and word embedding in the current observation stream, while using CGN as the graph stream. MJOLNIR-o uses the flattened *context matrix* as the observation vector, along with our CGN stream.

Reward - We tune our reward function to correctly learn to utilize the parent-target relationship for the navigation task. The agent in our model receives a “partial reward”, R_p , when a parent object $p \in P$ is visible. This is given by $R_p = R_t * Pr(t|p) * k$, where R_t is the target reward and $k \in (0, 1)$ is a scaling factor. We choose $R_t = 5$ and $k = 0.1$ for our experiments. $Pr(t|p)$ is taken from the partial reward matrix M ³, where each row has the probability distribution of the relative “closeness” of all the parent objects, to a given target object. The closeness is defined based on the relative spatial distance (measured in terms of the L2 distance) between a pair of objects in our floorplans. If multiple parent objects are visible, we only choose the one with the maximum R_p . Moreover, the agent does not get this reward the next time it sees the same parent object. This encourages it to explore different parent objects in the room until the target is located. If the “Done” action is sampled, *i.e.* the termination criterion occurs, and t is visible, the agent gets the goal reward, which is the sum of R_t and R_p . In this way, it learns to associate parent objects with a target, as well as the current state. Since the entire network is trained end-to-end, this shaped reward propagates back to the gcN layers, and tunes them to correctly learn the “parent-target” hierarchical relationship from the input knowledge graph. Finally, if neither the parent nor the target object is

²An illustration of this is shown in the supplementary videos in Appendix A

³Please refer to Appendix B of the supplementary for details about the construction of M .

visible, the agent gets a negative step penalty of 0.01. The reward for the state s , and action a , is therefore given by:

$$R(s, a) = \begin{cases} R_p, & \text{if } p \text{ is visible} \\ R_t, & \text{if } t \text{ is visible at termination} \\ R_t + R_p, & \text{if both are visible at termination} \\ -0.01, & \text{otherwise} \end{cases}$$

Algorithm 1 summarizes the above explained process.

5 Experiments and Results

5.1 Experimental setting

We primarily use the AI2-THOR (The House Of inteRactions) [48] environment as our platform for the navigation tasks. It is a challenging simulation platform, consisting of 120 photo-realistic floorplans categorized into 4 different room layouts - Kitchen, Living room, Bedroom, and Bathroom. Each scene is populated with real-world objects which the agent can observe and interact with, thereby enabling algorithms trained here to be easily transferable to real-robot settings. Out of the 30 floorplans for each scene layout, we consider the first 20 rooms from each scene type for the training set, and the remaining 10 rooms as the test set in our experiments. The list of target and parent objects can be found in Appendix C of the supplementary.

5.2 Comparison Models

Random - In this model, as each step, the agent randomly samples its actions from the action space with a uniform distribution. **Baseline** - This model closely resembles that of Zhu *et al.*[1], as it comprises of the current observation (in the form of the ResNet features of the current RGB frame) and the target information (in the form of glove embedding of the target object) as its state. **Scene Prior (SP)** - The publicly available implementation of Yang *et al.*[5] has been used here. This uses the prior knowledge in the form of a knowledge graph, but does not utilize the hierarchical relationships between objects. **SAVN** - In this model [3], the agent keeps learning about its environment through an interactive loss function even during inference time.

Metrics - For fair comparison with other state-of-the-art algorithms, we use the evaluation metrics proposed by [22], and adopted by other target-driven visual navigation algorithms [1, 3, 5, 6]. The Success Rate (SR) is defined as $\frac{1}{N} \sum_{i=1}^N S_i$, while the Success weighted by Path Length (SPL) is given by $\frac{1}{N} \sum_{i=1}^N S_i \frac{l_i}{\max(l_i, e_i)}$. Here, N is the number of episodes, S_i is a binary vector indicating the success of the i -th episode. e_i denotes the path length of an agent episode, and l_i is the optimal trajectory length to any instance of the target object in a given scene from the initial state. We evaluate the performance of all the models on the trajectories where the optimal path length is at least 1 ($L \geq 1$), and at least 5 ($L \geq 5$).

5.3 Implementation details

We built our models on the publicly available code of [3], using the PyTorch framework. The agent was trained for 3 million episodes on the offline data from AI2-THOR v1.0.1 [48]. During evaluation, we used 250 episodes for each of the 4 room types, resulting in 1000 episodes in total. In each episode, the floorplan, target, and the initial agent position was randomly chosen from the test set defined in Section 5.1. Additional implementation details can be found in Appendix D of the supplementary.

Algorithm 1: Reward Shaping for MJOLNIR

Input: state s , action a , target $t \in T$, SeenList

Data: target reward R_t , partial reward matrix M

Function Judge (s, a, t):

if $a \neq \text{"DONE"}$ **then**

 reward = Partial (s, t)

else if $a == \text{"DONE"}$ and t is visible **then**

 SeenList = [];

 reward = $R_t + \text{Partial}(s, t)$;

return reward;

Function Partial (s, t):

foreach parent $p_i \in P$ **do**

if p_i is visible and $p_i \notin \text{SeenList}$ **then**

$p \leftarrow \text{argmax } M[t]$;

 SeenList $\leftarrow p$;

$R_p = M[t][p] * R_t * k$

end

return R_p ;

Model	$L \geq 1$		$L \geq 5$	
	SR(%)	SPL(%)	SR(%)	SPL(%)
Random	11.2	5.1	1.1	0.50
Baseline [1]	35.0	10.3	25.0	10.5
Scene-prior [5]	35.4	10.9	23.8	10.7
SAVN [3]	35.7	9.3	23.9	9.4
MJOLNIR-r (our)	54.8	19.2	41.7	18.9
MJOLNIR-o (our)	65.3	21.1	50.0	20.9

Table 1: Comparison with state-of-the-art visual navigation algorithms on the unseen test set

5.4 Results

Table 1 and Figure 4a shows the performance of each of the models on the test environments. It is to be noted that the test environments consist of rooms that the agent has not previously seen during training, and therefore the location of the different objects in completely unknown. It can be seen that both of our models significantly outperform the current state-of-the-art. MJOLNIR-o has 82.9% increase in SR. This supports our hypothesis that incorporating *context vector* into the observation state is indeed a better idea than directly using ResNet and GloVe features. This is because the semantic information extracted from the scene in this manner is more object-centric, thereby making the target-driven navigation problem easier. It is also important to note here, that the graph convolutional network of [5], which does not include the *context vector* as its node feature, performs poorer than even the baseline [1] and SAVN [3].

It is interesting to note that even though MJOLNIR-r cannot beat the performance of MJOLNIR-o, it is still a substantial improvement over the other state-of-the-art methods (an observed gain of 53.5%). This highlights the importance of our proposed CGN stream which can better capture the contextual information extracted via ResNet conv features from the current input image. Moreover, using reward shaping to tune the model parameters ensures that not only is the prior memory preserved, but the current information containing the parent-target object relationship hierarchy is also utilized.

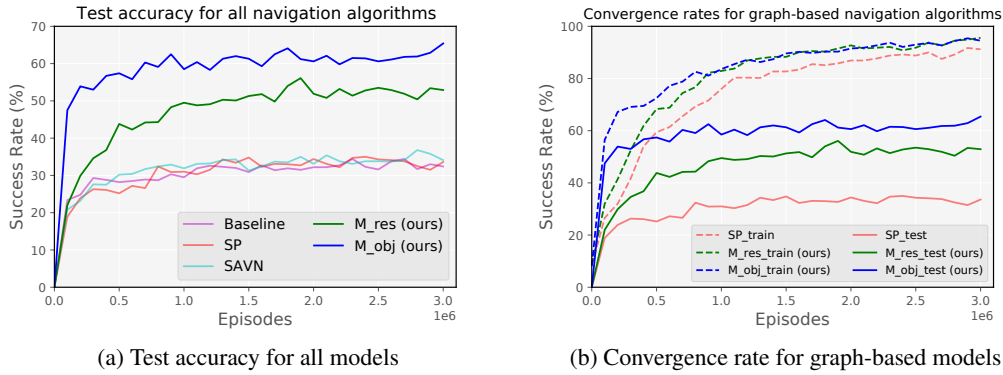


Figure 4: Test accuracy and convergence rates for all the algorithms

In Figure 4b, we show the convergence rate of our MJOLNIR with Yang *et al.*[5]. It can be seen that the training and testing SR of our algorithms rapidly increase within the first 5 million episodes itself, before saturating. This shows that our models learn to correctly locate targets much faster than others. In contrast, for [5], even though the training SR is quite high ($\approx 90\%$), there is huge drop during the testing performance ($\approx 35\%$), signifying severe overfitting. Finally, Table 2 gives the comparison of room-wise results.

Models	Bathroom		Bedroom		Kitchen		Living Room		Average	
	SR(%)	SPL(%)	SR(%)	SPL(%)	SR(%)	SPL(%)	SR(%)	SPL(%)	SR(%)	SPL(%)
Baseline [1]	53.2	13.4	28.8	9.0	32.4	10.9	35.2	10.0	37.4	10.8
Scene Prior [5]	41.6	13.3	33.6	10.4	26.4	9.1	36.0	9.9	34.4	10.7
SAVN [3]	47.6	14.6	21.6	6.7	34.8	8.3	40.0	9.0	36.9	9.7
MJOLNIR-r (our)	72.8	24.3	41.2	16.9	56.4	21.2	50.8	15.9	55.3	19.6
MJOLNIR-o (our)	82.4	25.1	43.2	14.4	74.8	22.9	50.0	17.9	62.6	20.1

Table 2: Evaluation results on a per-room basis

5.5 Ablation study

We show a number of ablations on MJOLNIR in Table 3. In MJOLNIR-o (no_g), we remove the third graph convolution layer of MJOLNIR-o, where we concatenated the intermediate embedding with the *context vector*. Instead, here, we directly feed the intermediate embedding to the joint embedding. The slightly lower performance shows that the context vector can indeed help in learning meaningful node embeddings for the graph. In MJOLNIR-o (w), we used a weighted adjacency matrix for the graph convolutional network. This does not affect the performance significantly as we believe that the object-object relationship is inherently learnt by MJOLNIR-o.

DONE action	Model	$L \geq 1$		$L \geq 5$	
		SR(%)	SPL(%)	SR(%)	SPL(%)
only sampled	MJOLNIR-r	54.8	19.2	41.7	18.9
	MJOLNIR-o	65.3	21.1	50.0	20.9
	MJOLNIR-o (no_g)	59.0	16.6	41.0	16.9
	MJOLNIR-o (w)	64.7	21.6	46.4	20.6
sampled + env	SAVN [3]	54.4	35.55	37.87	23.47
	MJOLNIR-o	83.1	53.9	71.6	36.9

Table 3: Ablation study for MJOLNIR

We also evaluated our model with a different stopping criteria. In this case, the agent does not rely only on its sampled “DONE” action to learn the termination action. Instead, it stops even when the environment gives the signal that the target object has been found. We compare our model with the SAVN [3] model using this stopping criteria. The results show that we perform significantly better.

Some failure case analysis provides insights into our models, and also opens up research in further directions. We identified mainly two such cases. Firstly, the agent gets stuck at a particular state, when the target is visible, but the straight path to it blocked by some obstacle. We hypothesize that a planner module which checks for collision might help to overcome this. Secondly, since “DONE” is one of the candidate actions, (i) it may be wrongly sampled even when the goal hasn’t been reached, or (ii) it might not be sampled even after reaching the goal. A workaround is to have the termination criterion provided by the environment. As seen from our ablation study, this boosts the success rate from 65.3% to 83.1%. However, it is at the cost of increased episode length as the agent is encouraged to explore more of the environment. Moreover, in a real-robot setting, it might not be possible to have the environment signal the termination.

6 Conclusion

We introduced MJOLNIR, a novel object-goal visual navigation algorithm that utilizes prior knowledge, and also learns to associate object “closeness” in the form of parent-target hierarchy during training. This is done through the proposed *context vector* which can be easily derived from the output of an object detector. We show that besides the modified state space, knowledge graph and reward shaping too play a significant role in guiding the agent to search for the target. Extensive experiments show that the agent can successfully find small target objects using the larger parent object as an anchor. Our model has the ability to generalize across different unseen scenes, and current state-of-the-art models.

Acknowledgments

The authors would like to thank Army Research Laboratory (ARL) W911NF-10-2-0016 Distributed and Collaborative Intelligent Systems and Technology (DCIST) Collaborative Technology Alliance for supporting this research.

References

- [1] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, pages 3357–3364. IEEE, 2017.
- [2] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.
- [3] M. Wortsman, K. Ehsani, M. Rastegari, A. Farhadi, and R. Mottaghi. Learning to Learn How to Learn: Self-Adaptive Visual Navigation Using Meta-Learning. In *IEEE CVPR*, June 2019.
- [4] A. Sax, J. O. Zhang, B. Emi, A. Zamir, S. Savarese, L. Guibas, and J. Malik. Learning to navigate using mid-level visual priors. In *Proceedings of the Conference on Robot Learning*, pages 791–812. PMLR, 30 Oct–01 Nov 2020.
- [5] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*, 2018.
- [6] R. Drun, Y. Yoshiyasu, A. Kanezaki, and A. Watt. Visual object search by learning spatial context. *IEEE RAL*, 5(2):1279–1286, April 2020.
- [7] G. Dulac-Arnold, D. Mankowitz, and T. Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.
- [8] J. Borenstein, Y. Koren, et al. The vector field histogram-fast obstacle avoidance for mobile robots. *IEEE T-RO*, 7(3):278–288, 1991.
- [9] D. Kim and R. Nevatia. Symbolic navigation with a generic map. *Autonomous Robots*, 6(1): 69–88, 1999.
- [10] Y. Kantaros and G. J. Pappas. Optimal temporal logic planning for multi-robot systems in uncertain semantic maps. In *IROS*, pages 4127–4132, 2019.
- [11] A. DAVISON. Real-time simultaneous localization and mapping with a single camera. In *IEEE ICCV*, pages 1403–1410, 2003.
- [12] R. Sim and J. J. Little. Autonomous vision-based exploration and mapping using hybrid maps and rao-blackwellised particle filters. In *IROS*, pages 2082–2089. IEEE, 2006.
- [13] M. Tomono. 3-D Object Map Building Using Dense Object Models with SIFT-based Recognition Features. In *IROS*, pages 1885–1890, Oct 2006.
- [14] D. Wooden. A guide to vision-based map building. *IEEE RAM*, 13(2):94–98, 2006.
- [15] P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE T-SSC*, 4(2):100–107, 1968.
- [16] S. Karaman and E. Frazzoli. Sampling-based algorithms for optimal motion planning. *IJRR*, 30(7):846–894, 2011.
- [17] C. Chen, A. Seff, A. Kornhauser, and J. Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *IEEE ICCV*, pages 2722–2730, 2015.
- [18] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. In *IEEE CVPR*, pages 2616–2625, 2017.
- [19] A. Giusti, J. Guzzi, D. C. Cireşan, F.-L. He, J. P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro, et al. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE RAL*, 1(2):661–667, 2015.
- [20] C. Linegar, W. Churchill, and P. Newman. Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera. In *ICRA*, pages 787–794. IEEE, 2016.
- [21] P. Saeedi, P. D. Lawrence, and D. G. Lowe. Vision-based 3-D trajectory tracking for unknown environments. *IEEE T-RO*, 22(1):119–136, 2006.
- [22] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.
- [23] D. Mishkin, A. Dosovitskiy, and V. Koltun. Benchmarking classic and learned navigation in complex 3D environments. *arXiv preprint arXiv:1901.10915*, 2019.
- [24] N. Kojima and J. Deng. To learn or not to learn: Analyzing the role of learning for navigation in virtual environments. *arXiv preprint arXiv:1907.11770*, 2019.

- [25] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, et al. Habitat: A platform for embodied AI research. In *IEEE ICCV*, pages 9339–9347, 2019.
- [26] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *IEEE ICCV*, 2003.
- [27] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *IEEE ICCV*, volume 1, pages 654–661. IEEE, 2005.
- [28] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *IEEE ICCV*, pages 1–8. IEEE, 2007.
- [29] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE CVPR*, pages 891–898, 2014.
- [30] A. Shrivastava and A. Gupta. Contextual priming and feedback for Faster R-CNN. In *ECCV*, pages 330–348. Springer, 2016.
- [31] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, pages 2929–2936. IEEE, 2009.
- [32] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *IEEE CVPR*, pages 3668–3678, 2015.
- [33] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *IEEE CVPR*, pages 5532–5540, 2017.
- [34] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE CVPR*, pages 2901–2910, 2017.
- [35] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE CVPR*, pages 3195–3204, 2019.
- [36] A. Pal, C. Nieto-Granda, and H. I. Christensen. DEDUCE: Diverse scene Detection methods in Unseen Challenging Environments. In *IEEE IROS*, pages 4198–4204, Nov 2019.
- [37] A. Pal, S. Mondal, and H. I. Christensen. “Looking at the Right Stuff” - Guided Semantic-Gaze for Autonomous Driving. In *IEEE CVPR*, June 2020.
- [38] E. Wiewiora. *Reward Shaping*, pages 863–865. Springer US, Boston, MA, 2010.
- [39] D. Tang, X. Li, J. Gao, C. Wang, L. Li, and T. Jebara. Subgoal discovery for hierarchical dialogue policy learning. In *EMNLP*, pages 2298–2309, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D18-1253>.
- [40] B. Bakker, J. Schmidhuber, et al. Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. In *Proceedings of IAS*, pages 438–445, 2004.
- [41] S. Goel and M. Huber. Subgoal discovery for hierarchical reinforcement learning using learned policies. In *AAAI Conference on Artificial Intelligence*, 2003.
- [42] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937, 2016.
- [43] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.
- [44] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.
- [45] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, June 2016.
- [46] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255. Ieee, 2009.
- [48] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017.
- [49] I. Kostrikov. Pytorch implementations of asynchronous advantage actor critic. <https://github.com/ikostrikov/pytorch-a3c>, 2018.

Appendices

Appendix A: Object detector vs ResNet features

One of the significant changes made in our algorithm from existing works [5] is to use object detection features instead of a classification probability from a CNN such as ResNet as the input node embedding in the graph convolution network. This is because detection of multiple objects is integral to learning the “parent-target” hierarchical object relationships effectively. To showcase this, we run a YOLOv3 detection, `yolo.mp4`, along with a Grad-cam visualization of ResNet-18, `resnet.mp4`, on a cell phone recording of an indoor kitchen scene. As seen from the videos, the multi-object detector can generate much less false-detections as compared to the 1000 class probability from ResNet, which generates noisy labels such as “prison” and “home_theatre”.

Parent \ Target	Fridge	StoveBurner	Microwave	TableTop	Sink	CounterTop	Shelf
Toaster	0.15	0.29	0.15	0.04	0.15	0.23	-
Spatula	0.03	0.31	0.22	0.02	0.19	0.22	0.02
Bread	-	0.16	0.13	0.16	0.20	0.36	-
Mug	-	0.19	0.17	0.11	0.30	0.23	-
CoffeeMachine	0.08	0.10	0.10	0.06	0.38	0.28	-
Apple	0.12	0.12	0.12	0.12	0.25	0.23	0.02

(a) Kitchen

Parent \ Target	Drawer	Shelf	TableTop	Sofa	FloorLamp
Painting	0.86	0.14	-	-	-
Laptop	0.27	0.14	0.36	0.23	-
Television	0.40	0.2	0.35	-	0.05
RemoteControl	0.25	0.05	0.25	0.4	0.05
Vase	0.21	0.47	0.26	-	0.05
ArmChair	0.09	-	0.27	0.18	0.45

(b) Living room

Parent \ Target	Shelf	Dresser	NightStand	Drawer	Desk	Bed
Blinds	1	-	-	-	-	-
DeskLamp	0.16	0.2	0.12	0.24	0.28	-
Pillow	-	0.04	0.21	0.17	-	0.58
AlarmClock	0.19	0.11	0.26	0.21	0.04	0.19
CD	0.21	0.1	0.08	0.33	0.23	0.06

(c) Bedroom

Parent \ Target	CounterTop	Cabinet	Drawer	ShowerDoor	Toilet	Bathtub
Mirror	0.51	0.26	0.21	0.03	-	-
ToiletPaper	0.19	0.19	0.10	0.06	0.46	-
SoapBar	0.29	0.18	0.14	0.04	0.2	0.16
Towel	0.15	0.04	0.15	0.44	0.07	0.15
SprayBottle	0.35	0.20	0.16	0.02	0.22	0.06

(d) Bathroom

Table 4: Partial reward matrices for each of the 4 room-type in AI2-THOR.

Appendix B: Construction of Partial reward matrix M

As mentioned in Section 4 of the main paper, the conditional probability of a target being found, given that a parents was observed, $Pr(t|p)$, is obtained from the Partial reward matrix M . We utilized the training split of our environment, AI2-THOR, for creating M . For every floorplan, the 3D position of each object present there was plotted. Then, we counted the occurrence of different target objects which were located within a euclidean distance of 1 meter from a parent object (from Table 5). It is to be noted that even though the same parent object might be present in more than

one room type, its relationship with the target objects might be different. Thus, we computed M for every room type, as shown in Table 4. Normalizing each row provides the probability distribution of a target object $t \in T$, given a parent object $Pr(t|P)$. For the final parent reward R_p , we also used a scaling factor, $k = 0.1$ to ensure that the agent receives lesser reward for the parent object as compared to the target reward $R_t = 5$.

Appendix C: Parent and target object list

Room type	Target Objects T	Parent Objects P
Kitchen	Toaster, Spatula, Bread, Mug, CoffeeMachine, Apple	Fridge, StoveBurner, Microwave, TableTop, Sink, CounterTop, Shelf
Living room	Painting, Laptop, Television, RemoteControl, Vase, ArmChair	Drawer, Shelf, TableTop, Sofa, FloorLamp
Bedroom	Blinds, DeskLamp, Pillow, AlarmClock, CD	Shelf, Dresser, NightStand, Drawer, Desk, Bed
Bathroom	Mirror, ToiletPaper, SoapBar, Towel, SprayBottle	CounterTop, Cabinet, Drawer, ShowerDoor, Toilet, Bathtub

Table 5: Parent and target object list

Table 5 provides the list of target objects, T , and parent objects, P , used in our experiments.

Appendix D: Implementation Details

For the word embeddings, we used the 300-D GloVe vectors pretrained on 840 billion tokens of Common Crawl [44]. The A3C model is based on [49]. The model hyperparameters used for our experiments are tabulated below.

Parameters	Baseline [1]	Scene Prior [5]	SAVN [3]	MJOLNIR-r (our)	MJOLNIR-o (our)
Learning rate	0.0001	0.0001	0.0001	0.0001	0.0001
Optimizer	SharedAdam	SharedAdam	SharedAdam	SharedAdam	SharedAdam
Discount Factor	0.99	0.99	0.99	0.99	0.99
max #workers	8	8	6	8	8
Observation stream encoder	ResNet	ResNet	ResNet	ResNet	Object detector
Max training episodes	3×10^6	3×10^6	3×10^6	3×10^6	3×10^6

Table 6: Summary of Hyperparameters