Commonsense and Semantic-Guided Navigation through Language in Embodied Environment

Dian Yu[1]*, Chandra Khatri[2], Alexandros Papangelis[2], Andrea Madotto[3], Mahdi Namazifar[2], Joost Huizinga[2], Adrien Ecoffet[2], Huaixiu Zheng[2], Piero Molino[2], Jeff Clune[2], Zhou Yu[1], Kenji Sagae[1], Gokhan Tur[2]

[1]University of California, Davis
[2]Uber AI
[3]Hong Kong University of Science and Technology
[1]{dianyu, joyu, sagae}@ucdavis.edu
[2]{chandrak, apapangelis, mahdin, jhuizinga, adrienle, huaixiu.zheng, piero, jeffclune, gokhan}@uber.com
[3]{amadotto}@connect.ust.hk

Abstract

Grounded language navigation tasks such as room navigation (e.g. “go to the kitchen“) and embodied question answering (e.g. “what color is the car?") in realistic 3D environments require the agent to be generalizable to unseen environments. However, previous research suggests that vision inputs do not contribute to multi-modality performances. Humans, on the other hand, utilize commonsense and semantic understanding of both language instruction and vision to navigate in new environments. In this work, we address the challenges with the room navigation task by: (1) Building and incorporating commonsense about layouts of environments through observations made during training for path planning and (2) Enforcing semantic understanding of scenes in each step through semantic grounding as auxiliary tasks. In addition, we apply self-supervision to fine-tune navigation actions in unseen environments through semantic understanding. Our results show that commonsense and semantic understanding facilitate grounded navigation across a variety of metrics. Moreover, through self-supervision we showcase that once an agent is taught to perform semantic understanding, it can further update the semantically grounded navigator on the unseen environments to perform even better navigation leading to generalization.

1 Introduction and Related Work

Visually grounded language understanding (Harnad (1999); Hermann et al. (2017)) requires an agent to follow language inputs and interact with a simulated 3D environment to complete tasks such as object and room navigation (Wijmans et al. (2019); Wu et al. (2018a)), embodied question answering (Das et al. (2017, 2018); Gordon et al. (2017); Manolis Savva et al. (2019); Mirowski et al. (2018)), and instruction following (Anderson et al. (2017); Fried et al. (2018); Wang et al. (2018)). Although researchers have significantly advanced the state of the art in these tasks, most efforts have been on bridging language and visual inputs rather than understanding of the environments and acting based on learned semantics. Indeed, some research show that using language instruction alone outperforms adding visual features (Anand et al. (2018); Hu et al. (2019)). The fundamental question of how language and semantics facilitate navigation and how

[*] work done during internship at Uber AI

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

navigation helps in semantic understanding is yet to be studied in more depth. When humans are left in an unseen environment, it is through commonsense and semantic understanding that they perform navigation; in fact, they further update their common sense and semantic knowledge as they explore the environment through continual learning. In this work, we demonstrate how an agent can similarly leverage commonsense and semantic understanding to learn to navigate in different environments.

Similar to our hypothesis, Hermann et al. (2017) analyzed the problem of how agents learn to interpret instructions and how they generalize in one-shot and multi-task settings through reinforcement and unsupervised learning. Their study showcased the importance of prior semantic knowledge and curriculum learning for better generalization and faster task completion. These experiments were performed in simple one or two room settings (Beattie et al. (2016)) with few objects within the environment. In comparison, House3D (Wu et al. (2018b)) or MatterPort3D (Chang et al. (2017)) environments consist of realistic indoor houses with multiple rooms. These environments contain visually, structurally, and semantically a wide variety of objects, some of which can be obstacles during navigation. Semantic understanding can therefore benefit the agent when navigating this visually and structurally complex environment.

In our work, we focus on the Room Navigation (*RoomNav*) task by asking an agent to navigate to a specific target room in unseen complex environments. Instead of focusing on task completion, we are addressing several research questions related to grounded language understanding such as (1) how can semantic information from the environment help with navigation in unseen complex settings, (2) how can an agent learn generic layouts for path planning from commonsense, and (3) how to adjust policies learned from experience to unseen environments for generalization. We also address the challenge of multiple possible targets (for example, a house can have multiple bedrooms or bathrooms). This problem which has been widely ignored in previous works as they filter out such scenarios to avoid ambiguity (Das et al. (2017)).

To answer these research questions, we design semantic grounding and commonsense planning modules to guide the agent for navigation. We learn the sequential and structural information of rooms and objects in the house by incorporating semantics observed during training and use it to guide the action decoder in a hierarchical fashion (Kulkarni et al. (2016); Wernsdorfer and Schmid (2014)). For knowledge grounding, we introduce two auxiliary tasks: grounding during navigation (by asking the agent to predict current and nearby rooms), and grounding after the agent stops (by asking questions like "did you see a bathroom on your way?"). Inspired by Wang et al. (2018), we also leverage the idea of self-supervision on unseen environments for better exploration by fine-tuning the model parameters through semantic understanding loss in unseen environments. Results on realistic environments show that the proposed model significantly outperforms the baseline, indicating the importance of path planning through semantics and commonsense.

2 Agent Architecture

There are three kinds of inputs for the model: (i) task specific instruction (e.g. "go to the kitchen") (ii) RGB values of visual observations for each state, and (iii) semantic information such as room annotations (used only during training). Following previous works on embodied navigation (Fried et al. (2018), Wang et al. (2018)), we extract panoramic image features using a fixed pretrained ResNet-152 (He et al. (2015)). Specifically, we turn the agent 90 degrees to the left and right, respectively, to get a 270-degree view. We extract and concatenate features for the left, front, and right images and pass through a single layer Feed Forward Neural Network to obtain the visual representation.

Our proposed model consists of a Semantically Grounded Navigator, a Commonsense Planning Module, and a Semantic Grounding Module. Figure 1 provides an overview of the architecture.

Semantically Grounded Navigator (*SGN*): An $LSTM$ model to predict one of four possible actions at each step: go forward (0.25m), turn left (10 degrees), turn right (10 degrees) and stop. The navigator takes the $RoomNav$ instruction (which remains static throughout the task) and visual representation (which changes at each turn or step) as input.

Commonsense Planning Module (CP): For destinations in a distance, the agent may not be able to utilize a static instruction for route planning. Instead, we design a Scene-based Next

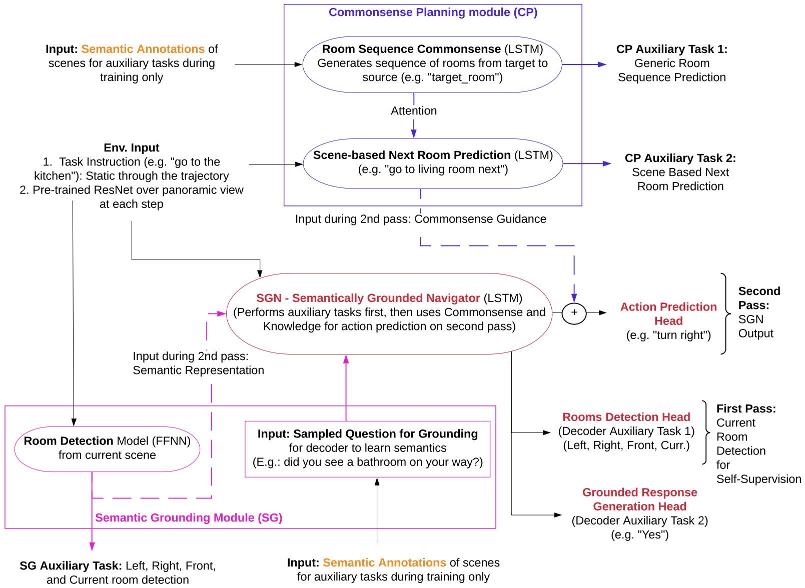
Commonsense Planning module (CP)
Input: Semantic Annotations of scenes for auxiliary tasks during training only
Room Sequence Commonsense (LSTM) Generates sequence of rooms from target to source (e.g. "target_room")
CP Auxiliary Task 1: Generic Room Sequence Prediction
Attention
Env. Input
1. Task Instruction (e.g. "go to the kitchen"): Static through the trajectory
2. Pre-trained ResNet over panoramic view at each step
Scene-based Next Room Prediction (LSTM) (e.g. "go to living room next")
CP Auxiliary Task 2: Scene Based Next Room Prediction
Input during 2nd pass: Commonsense Guidance
SGN - Semantically Grounded Navigator (LSTM) (Performs auxiliary tasks first, then uses Commonsense and Knowledge for action prediction on second pass)
+
Action Prediction Head (e.g. "turn right")
Second Pass: SGN Output
Input during 2nd pass: Semantic Representation
Rooms Detection Head (Decoder Auxiliary Task 1) (Left, Right, Front, Curr.)
First Pass: Current Room Detection for Self-Supervision
Room Detection Model (FFNN) from current scene
Input: Sampled Question for Grounding for decoder to learn semantics (E.g.: did you see a bathroom on your way?)
Grounded Response Generation Head (Decoder Auxiliary Task 2) (e.g. "Yes")
Semantic Grounding Module (SG)
SG Auxiliary Task: Left, Right, Front, and Current room detection
Input: Semantic Annotations of scenes for auxiliary tasks during training only

Figure 1: Components for Commonsense and Knowledge Guided Navigation model. Red components correspond to baseline navigator model. Purple components are introduced to incorporate Commonsense while pink components are for Semantic Understanding. Semantically Grounded Navigator (SGN) is designed to perform semantic understanding for action prediction, while commonsense is fed as guidance for better planning. Dotted lines indicate input in the second path in the model. Detailed architecture illustration with loss functions can be found in Figure 2 in the Appendix.

Room Prediction module (CP_Nxt) to help the agent to navigate to an intermediate target on the way. To do this, the Navigator LSTM model is trained simultaneously to capture generic room sequence patterns across the houses through sequences in training trajectories. However, such an intermediate target prediction model does not have explicit information of what rooms are near the target room so that the prediction may suggest a surrounding room of the current room instead. Therefore, we design a Room Sequence Commonsense module (CP_RS) using LSTM to generate backward room sequences starting from the target room. We get contextual representation by attending to the output of the CP_RS module and predict the next room target.

Semantic Grounding Module (*SG*): Human navigators can glimpse semantic information, such as a dining room on the path, which may help with finding the target such as a kitchen. Similarly, this module consists of a Room Detection model (SG_RD) (for detecting the current room and what rooms are to the left, right, front, from the panoramic image) and a Post Navigation grounding model (SG_PN) which generates the question (e.g. "did you see a bathroom on your way?") to incorporate semantic understanding within the agent. We use the ground truth semantic annotations provided during training for room detection and question generation (using templates for positive and negative sampling). We also pass the hidden representation of the room detection model to the SGN for action prediction.

Self-supervision for Unseen Environment Fine-tuning (*SS*) Inspired by self-supervision (Wang et al. (2018)), we take the room detection model in the SG_RD as ground truth and fine-tune the action prediction by updating the SGN model parameters. During training, we perform two passes over the SGN at each time step. In the first pass, the Room Detection task on top of SGN is performed without the knowledge of nearby rooms (from the room detection module). In the second pass, commonsense knowledge together with the room prediction represenation from SG_RD (depicted via dotted lines in Figure 1) are also fed to the decoder for action prediction. During inference, we take the prediction from SG_RD as ground truth and fine-tune the Grounded Response Generation Head together with the SGN for action prediction. The agent moves towards the target according to pretrained policy for t steps, calculates losses, and updates the model parameters. Then the model is spawned to the original source coordinate and start navigating towards the target room.

3 Experiments and Results

Data and Environment: We use the Habitat environment (Manolis Savva et al. (2019)) with the MatterPort3D dataset for all of our tasks. Habitat's task is Point navigation ($PointNav$), where an agent needs to navigate from a source coordinate to a uniformly sampled target coordinate. We adapt this task to form RoomNav by replacing the target coordinates with the corresponding 27 room types such as bedroom, kitchen, etc. Out of the training houses and tasks provided in Habitat for the $PointNav$ task, we filter out tasks where the source and target are in the same room and end up using 48 for training (5423 games) and validation (582 games), and five for unseen environment testing (402 games). We use the same measure as the $PointNav$ task to define the complexity of the game, easy, medium, and hard, by the geodesic distance between the source and the target. We use the shortest path (SP) for imitation learning and SP serves as an oracle path when evaluating the action policy.

Success Metric and Loss: We use Success (the agent enters the target room), Success Per Length (SPL) (a Success metric normalized with respect to the shortest path distance (Anderson et al. (2017)) while the last step action is stop), and non-stop SPL (the requirement for the last step action is relaxed in SPL for evaluation). Results are shown in Table 1.

Model	SPL	non-stop SPL	Overall Succ.	Easy Succ.	Medium Succ.	Harrd Succ.
$SP(Baseline)$	0.0495	0.2778	0.3109	0.4426	0.3203	0.1732
CP_Nxt	0.0784	0.2745	0.3333	0.4672	0.3464	0.1604
CP_Nxt+CP_RS	0.0583	**0.2888**	**0.3532**	**0.5082**	**0.3529**	**0.2047**
SG_RD	0.0539	0.2681	0.3308	0.4836	0.3333	0.1811
SG_RD+SG_PN	0.0424	0.2718	0.3333	0.4757	**0.3529**	0.1732
CP_Nxt+SG_RD	**0.0972**	0.2761	0.3333	0.5	0.3137	0.1969
$CP\_Nxt+SG\_RD+CP\_RS$	0.0897	0.2804	0.3308	0.4508	**0.3529**	0.189
SG_CR ($SS-Baseline$)	0.0573	0.2743	0.3458	**0.5738**	0.281	0.2047
SS on $SG-CR$ (20 $Steps$)	0.0638	**0.2913**	**0.3557**	0.5574	0.3137	**0.2126**
SS on $SG-CR$ (40 $Steps$)	0.0447	0.2643	0.3184	0.5082	0.281	0.1811

Table 1: Results on Supervised Learning and Self-Supervision on Test Environments (easy, medium, and hard games).

Discussion From Table 1, it can be observed that CP helps significantly in obtaining higher success rates across all the difficulties. Both next-room guidance (Nxt) and generic room sequence (RS) help, however CP_Nxt alone leads to early stopping and hence higher SPL but worse performance on hard games. RS helps avoid early stopping through attention layer and makes the agent move towards the rooms closer to target, thereby leading to best results across most metrics. SG with room detection ($+RD$) and grounding post navigation ($+PN$) help in general but leads to longer trajectories as it prefers exploration over stopping, hence doesn't help in harder games. CP when combined with SG leads to best SPL scores, that is able to find the target faster, however due to opposite effects success rates are not on par with CP_Nxt+CP_RS. To perform self-supervision (SS), we introduced another head in SGN model for current-room prediction (SG_CR), which also ensures better grounding for SGN. SG_CR significantly outperforms most other models, i.e. non semantically grounded navigators, in fact in case of easy games the performance is close to 57%, which is 10-20% better than all the models. After performing SS for 20 steps (using the loss obtained from SG_RD model) on unseen environments, SG_CR head helps the overall accuracy by 15% with respect to the SG_CR baseline. With more SS steps (40) performance degrades due to introduction of noise, as the current room prediction from SG_RD model is not perfect/ground truth. Through self-supervision, we showcase that agent can be made to update the semantic understanding through navigation.

4 Conclusion

The only thing which humans have when they navigate in unseen environments is commonsense and semantic understanding obtained through past experiences. Inspired by this idea, this work introduces commonsense planning and makes the agent semantically aware when performing the room navigation task in a realistic and complex 3D environments. We showcase that if agent is taught to perform commonsense and semantic understanding with self-supervision, then it can be more generalizable on unseen environments to perform better grounded navigation.

References

Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron C. Courville. 2018. Blindfold baselines for embodied QA. *CoRR*, abs/1811.05013.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2017. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *CoRR*, abs/1711.07280.

Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. 2016. Deepmind lab. *CoRR*, abs/1612.03801.

Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2017. Embodied question answering. *CoRR*, abs/1711.11543.

Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Neural modular control for embodied question answering. *CoRR*, abs/1810.11181.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. *CoRR*, abs/1806.02724.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2017. IQA: visual question answering in interactive environments. *CoRR*, abs/1712.03316.

Stevan Harnad. 1999. The symbol grounding problem. *CoRR*, cs.AI/9906002.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. 2017. Grounded language learning in a simulated 3d world. *CoRR*, abs/1706.06551.

Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. 2019. Are you looking? grounding to multiple modalities in vision-and-language navigation. *CoRR*, abs/1906.00347.

Tejas D. Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. 2016. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3675–3683.

Manolis Savva, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2018. Learning to navigate in cities without a map. *CoRR*, abs/1804.00168.

Xin Wang, Qiuyuan Huang, Asli Çelikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2018. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. *CoRR*, abs/1811.10092.

Mark Wernsdorfer and Ute Schmid. 2014. Grounding hierarchical reinforcement learning models for knowledge transfer. *CoRR*, abs/1412.6451.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. 2019. Embodied question answering in photorealistic environments with point cloud perception. *CoRR*, abs/1904.03461.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018a. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018b. House3d: A rich and realistic 3d environment. *arXiv preprint arXiv:1801.02209*.

5 Appendix

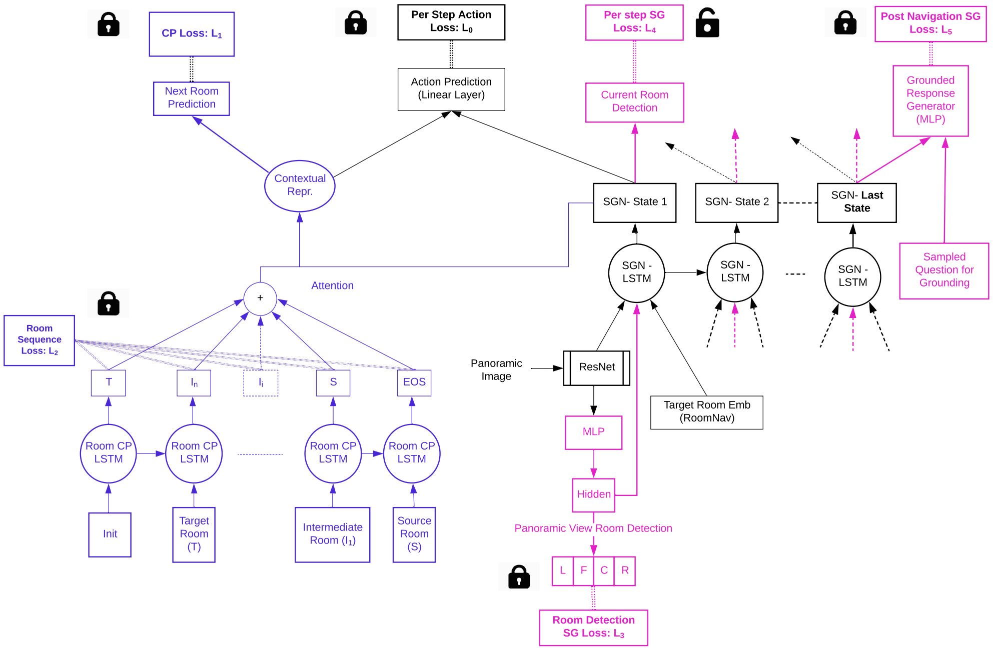

Figure 2: Detailed architecture for Commonsense and Knowledge Guided Navigation model. Black components correspond to the baseline navigator model. Purple components are introduced to incorporate Commonsense while pink components are for Semantic Understanding. Semantically Grounded Navigator (SGN) is designed to perform semantic understanding for action prediction, while commonsense is fed as guidance for better planning. There are six losses, five of them are locked during inference, except L_4 is unlocked for self-supervision.