Cognitive Mapping for Object Searching in Indoor Scenes

—— Master Thesis Defense

Presenter:
Shibin Zheng
Committees:
Yezhou Yang, Chair
Wenlong Zhang, member
Ren Yi, member
Date: 11/01/2019

Introduction

- Robot helps to search objects in a house
  - Household robot is a type of service robots that is designed to help human with household chores and provide assistances especially for elder, and it is promising industry product.
  - Object searching is a fundamental skill for a household robot.

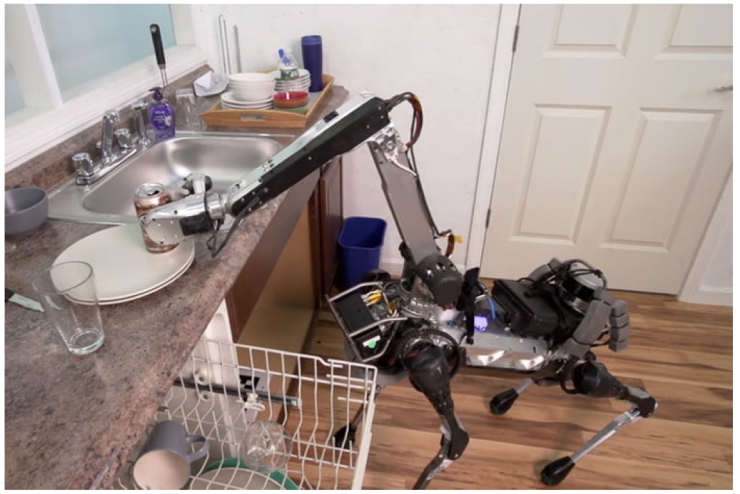


Image source from DIGITAL TREND

Introduction

- Object Searching
  - The robot needs to distinguish the target object and figure out how to approach.
- Methods
  - Supervised Learning

    A possible solution is to let robot learn from expert. We teach robot what optimal actions to take in different situations.
  - Reinforcement Learning

    Train an autonomous robot based on “rewarding desired behaviors” to help it obtain an action policy that maximize rewards while the robot interacting with the environment .

Introduction

- Deep Reinforcement Learning for Object Searching.
  - Deep Reinforcement Learning: Model policy with deep neural networks, from pixels to action.
  - Object Searching Setting: Under the “robot with vision that finds objects” setting, the goal state is the location of the target object with a high reward assigned.

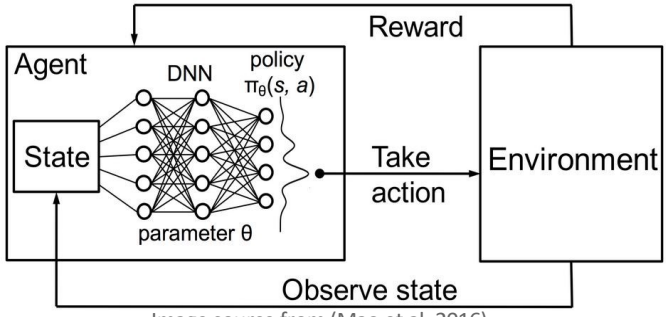
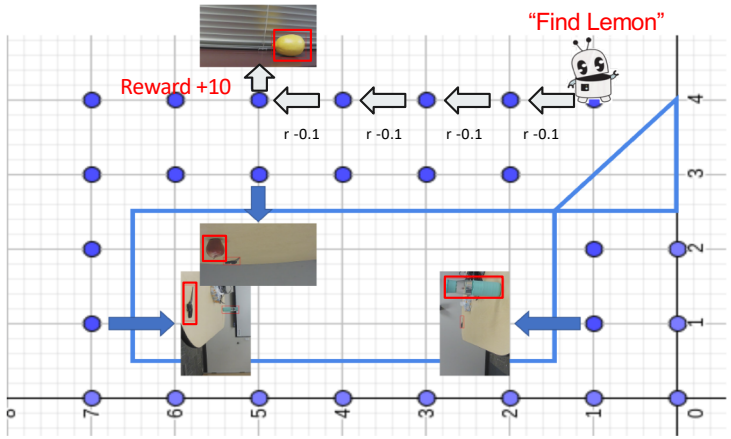


Image source from (Mao et al. 2016)

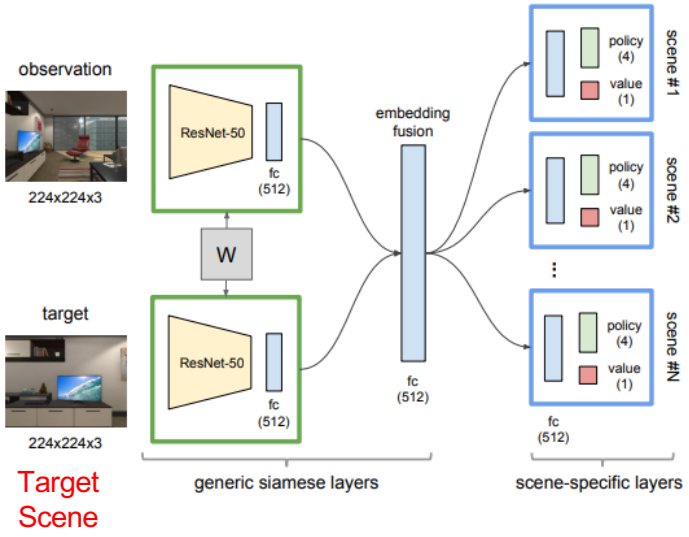
$$S_t = O(I_t)$$
$$a_t = \pi_\theta(s_t, a_t)$$



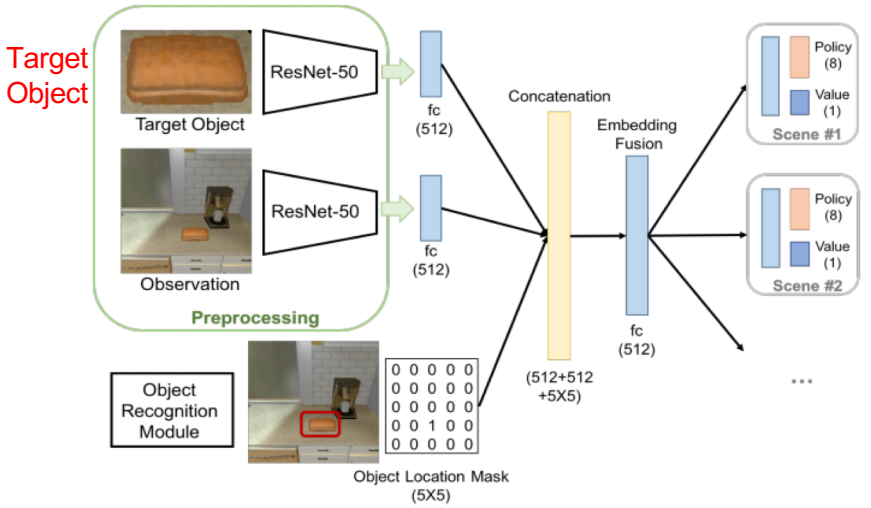
Previous Works

- Deep Reinforcement Learning for Visual Navigation.

Zhu, et al. 2017



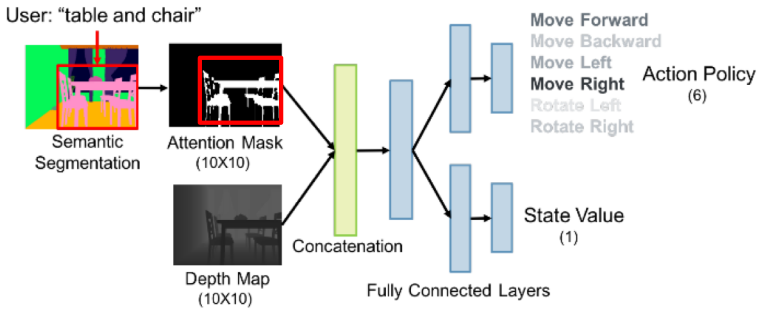
Ye, et al. 2018



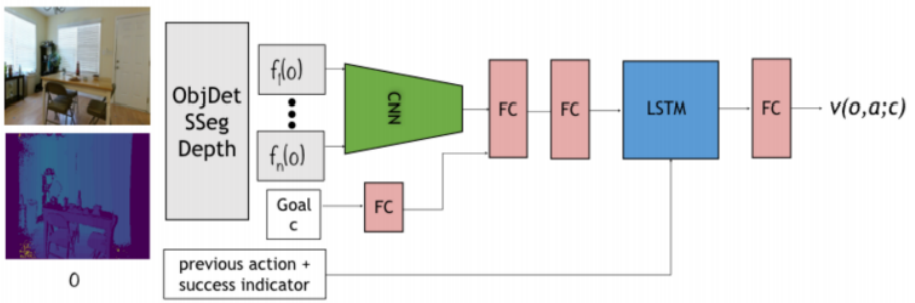
Motivations

- Generalizability of Deep Reinforcement Learning
  - Unsatisfiable generalization capability when transferring to new environment or target.
  - Follow the end-to-end training manner, when training in the environments, some high-level scene dependent features are more easily learnt by the agent rather than general scene representations.
  - Time-consuming re-training process is needed every time the target or the environment alters.

Ye, et al. 2019

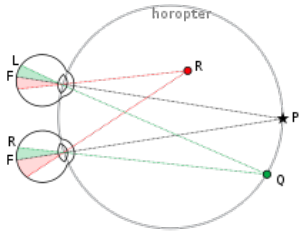


Mousavian, et al. 2019

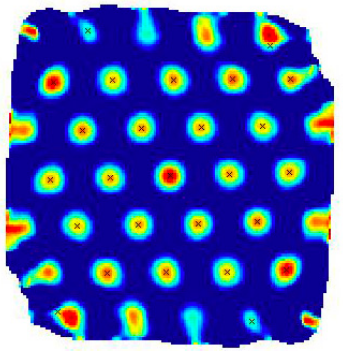


Motivations

- General Representation of environment
  - As human navigating in environments, we reason about the geometry and **topology** of the environment, locate the target, move around blocking objects and approach the destination.
  - As a matter of fact and the knowledge from stereo vision, with binocular vision, the furthest accurate depth estimation can only reach the end of an arm.
  - The further topology reasoning is based on **grid cell** in brains, it is a type of neuron in the brains of human that functions as a GPS and allows people to understand their position in space.

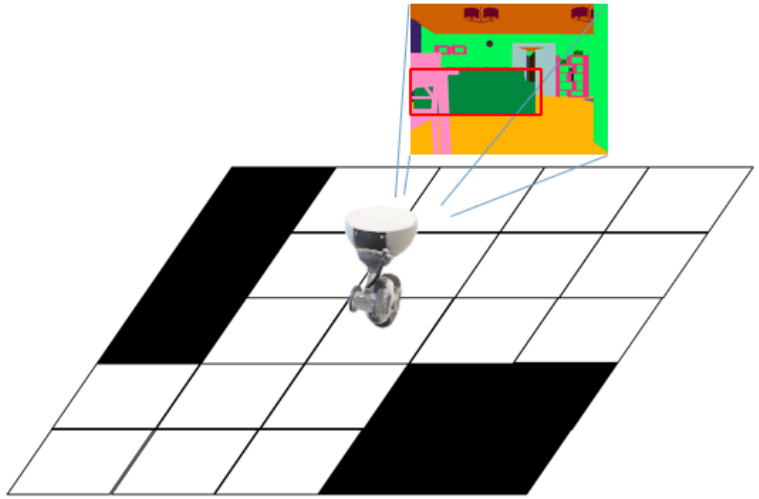


Binocular Vision

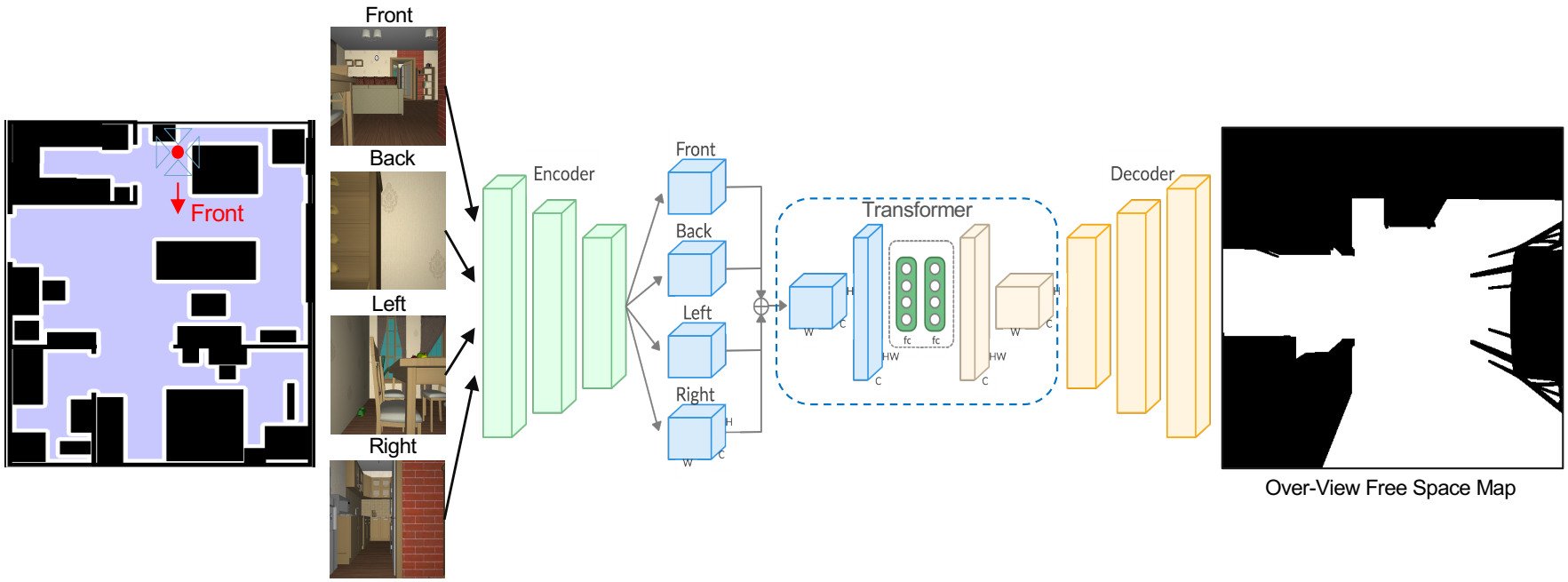

Activity of grid cells in a brain
(source from Wikipedia)

Motivations

- General Representation of environment
  - Reasonably, household robot will appear in houses first as wheel-based robot. So the interacting space for the robot is actually the **top-down free space**.
  - Meanwhile providing the robot with the semantic mask so as to giving a sense of **direction**.
  - With the geometric information of surrounding and the sense of direction. The robot is ought to learn how to approach the target and maneuver around the obstacles at the same time.

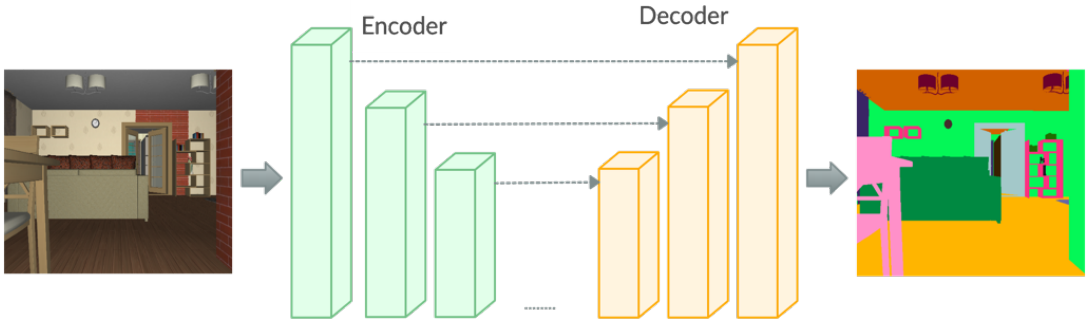


Cognitive Mapper

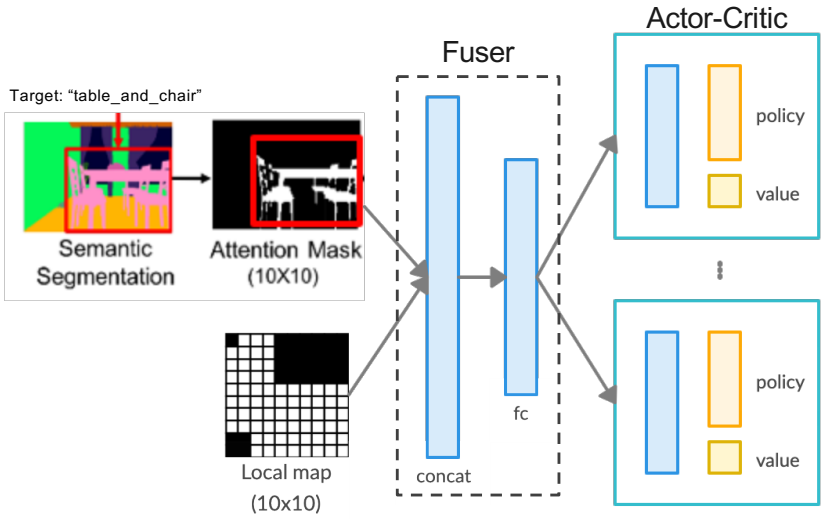


Semantic Segmentation

- Deeplab v3+ *(chen, et al. 2018)* is employed for semantic segmentation.



Action Policy Network



- Policy $\pi_\theta(s)$: a set of action probability outputs.
- Value function $V_\theta(S)$: how good a certain state is to be in.

- Asynchronous Advantage Actor-Critic Network (A3C) (Mnih, et al. 2016)

Experimental Settings

- Dataset
  - House3D (Wu, et al. 2018) has rich, extensible and efficient environment that contains 45,622 human-designed 3D scenes of visually realistic houses, ranging from single-room studios to multi-storied houses, equipped with a diverse set of fully labeled 3D objects

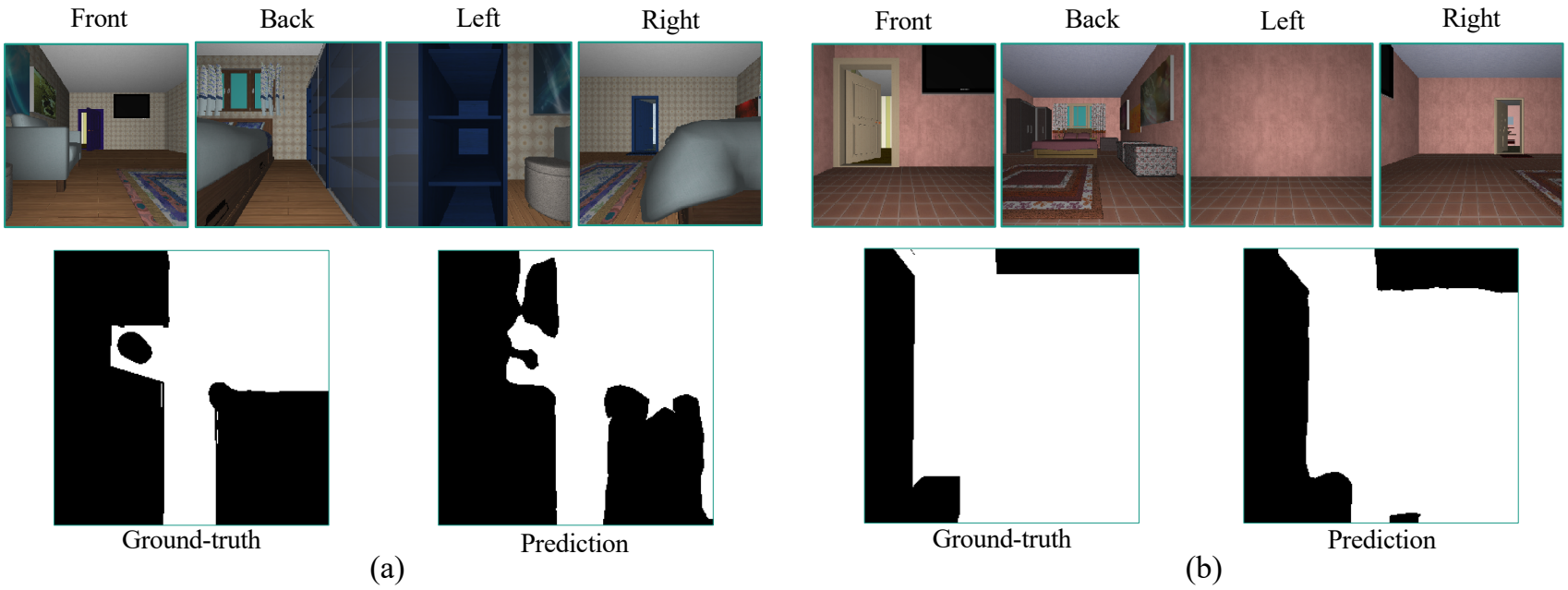
Experimental Settings

- Cognitive Mapper
  - 410 scenes selected from the House3D dataset, and further split the dataset into training set that contains 340 scenes and validation set that contains 70 scenes. The training set has nearly 256k pairs of data and the validation set has nearly 52k pairs of data.
- Semantic Segmentation
  - 56k data pairs for training, and shrink the semantic channel from original 94 labels to 77 labels with the rest being classified as “background”.
- Action Policy Network
  - 248 simulated environments that are suitable for testing. to avoid ambiguity, we select the objects that only have one instance in an environment as the target objects for the robot to approach.

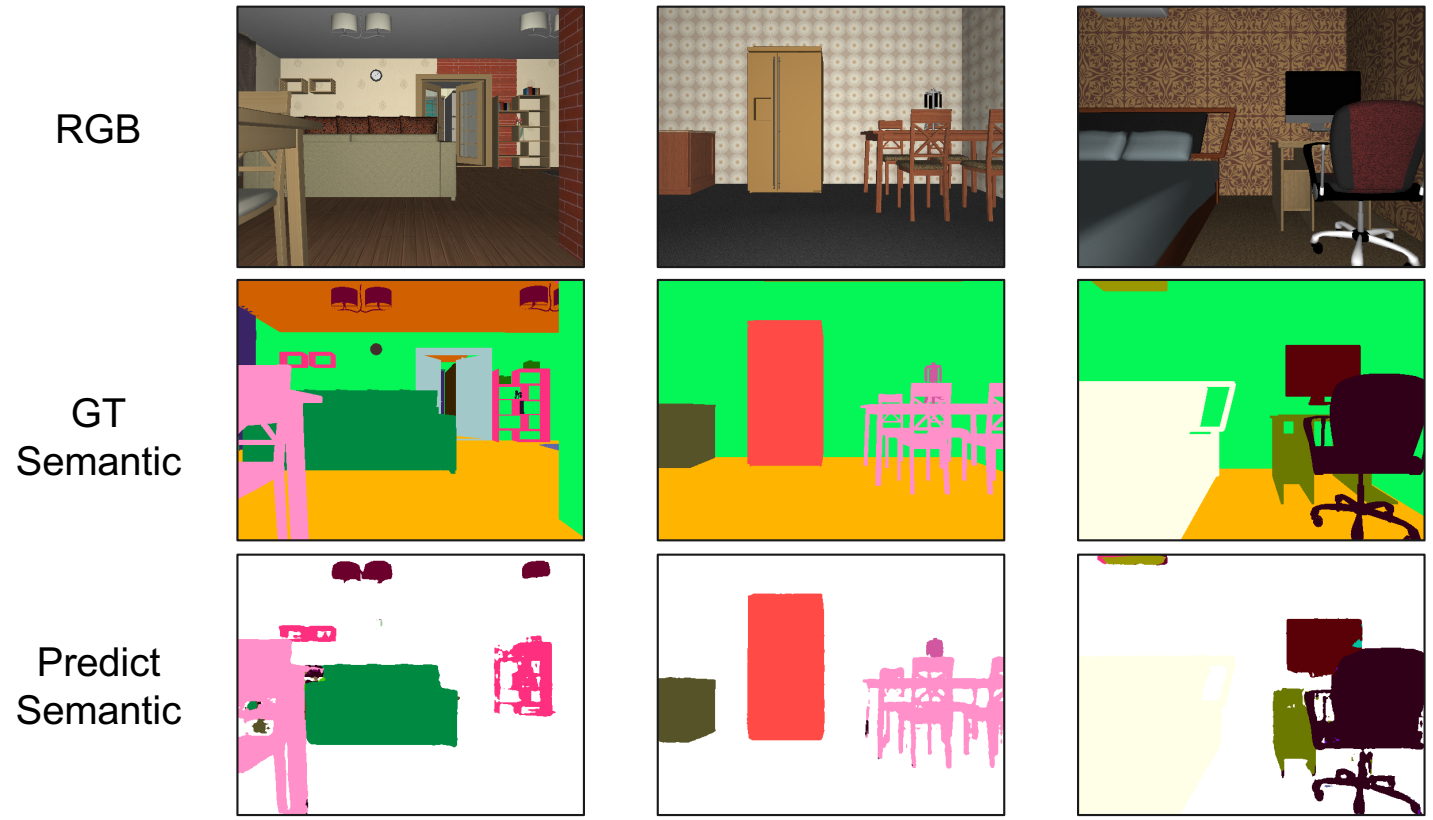
Experiment Results

- Cognitive Mapper
  - Cross-Entropy Loss Function:
    - $\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{N} y^{(i)} \log \hat{y}^{(i)}$
  - The model achieves 0.**246** in Mean Intersection Over Union (mean IOU) and **75.7%** in Pixel Accuracy.
- Semantic Segmentation
  - Cross-Entropy Loss Function:
    - $\mathcal{L}(p, p^*) = -\sum_{i=1}^{N} p \log p_i^*$
  - The model achieves 0.**436** in Mean IOU.

Experiment Qualitative Results



Experiment Qualitative Results



Experiment Results

- Generalizability across targets in 1 scene.

Succ. Rate	New Targets [%age]					Trained Targets [%age]				
Minimal Steps*	1 x ms	2 x ms	3 x ms	4 x ms	5 x ms	1 x ms	2 x ms	3 x ms	4 x ms	5 x ms
a)	3.0	5.4	7.2	10.6	12.4	3.5	5.2	8.2	10.5	12.5
b)	24.8	30.0	37.2	44.4	48.2	50.0	80.0	89.0	91.5	93.0
c)	10.7	**25.1**	**44.4**	**47.4**	**51.2**	30.1	**48.3**	**49.9**	51.8	51.8
d)	**14.2**	24.3	39.6	44.5	47.0	**32.5**	40.8	45.4	**52.1**	**55.0**
e)	25.6	**46.2**	**55.2**	**63.6**	**70.2**	49.3	**79.8**	**89.0**	**91.7**	**93.3**
f)	**28.0**	40.1	55.0	59.2	63.5	**51.2**	62.3	70.0	70.0	75.5

a) Random.

b) AOP (Ye, et al. 2018).

c) GAPLE: with predicted semantic segmentation and depth.

d) Ours: with predicted semantic. segmentation and free-space map.

e) GAPLE: with ground-truth semantic segmentation and depth.

f) Ours: with ground-truth semantic. segmentation and free-space map.

* n x ms: within n times of minimal steps from the starting location to the target location, the agent successfully reaches the target.

Experiment Results

- Generalizability across scenes.

Succ. Rate	New Scenes [%age]					Trained Scenes [%age]				
Minimal Steps*	1 x ms	2 x ms	3 x ms	4 x ms	5 x ms	1 x ms	2 x ms	3 x ms	4 x ms	5 x ms
a)	9.2	14.8	18.8	21.9	25.6	7.5	12.0	15.5	18.1	20.6
b)	29.6	34.0	37.0	39.5	42.2	30.6	39.6	46.2	49.2	52.0
c)	24.7	32.3	34.3	36.2	36.4	24.2	34.3	36.3	38.1	38.8
d)	**33.0**	**39.9**	**43.4**	**51.6**	**55.3**	**32.2**	**40.0**	**45.4**	**50.6**	**56.5**
e)	29.3	34.0	37.0	39.5	42.2	30.6	39.6	46.2	49.2	52.0
f)	**51.2**	**63.3**	**74.3**	**76.0**	**77.0**	**50.1**	**69.0**	**72.9**	**74.4**	**76.2**

a) Random.

b) AOP (Ye, et al. 2018).

c) GAPLE: with predicted semantic segmentation and depth.

d) Ours: with predicted semantic. segmentation and free-space map.

e) GAPLE: with ground-truth semantic segmentation and depth.

f) Ours: with ground-truth semantic. segmentation and free-space map.

* n x ms: within n times of minimal steps from the starting location to the target location, the agent successfully reaches the target.

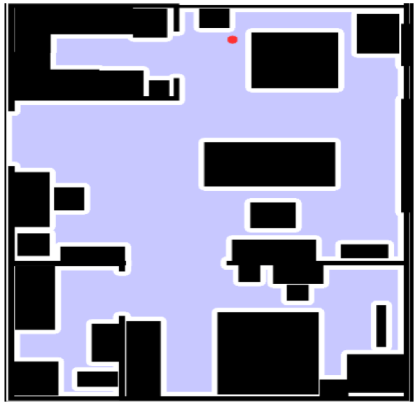
Conclusions

- Conclusion
  - Shown by the significant improvement of performances, we argue that by inherently predicting the local free-space map, the model generalizes better. And for a wheel-based robot whose actual movable space is on the 2D map, the top-down map is a more general representations.
- Thoughts
  - Now the robot learns the free-space map and the semantic segmentation from the first-person view images. Is there any more general information as a data representation?
  - What if we combine the free-space map and the semantic information?

Motivations

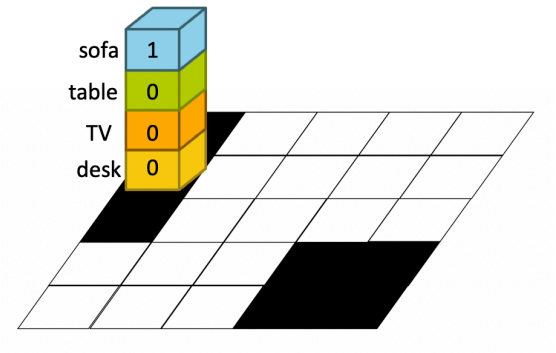
- What if we not only encode the occupancy information in the map, but also encode the semantic information of the objects that occupy the space.

- Object relation is also encoded. In human houses, objects are placed with common-sense rules: Objects with concurrent usages are more likely to put together.



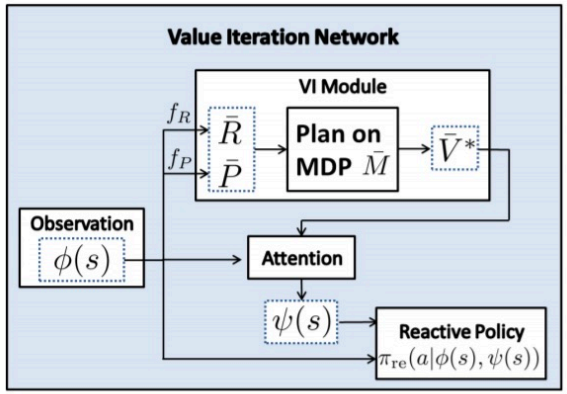
Free-space map of a house

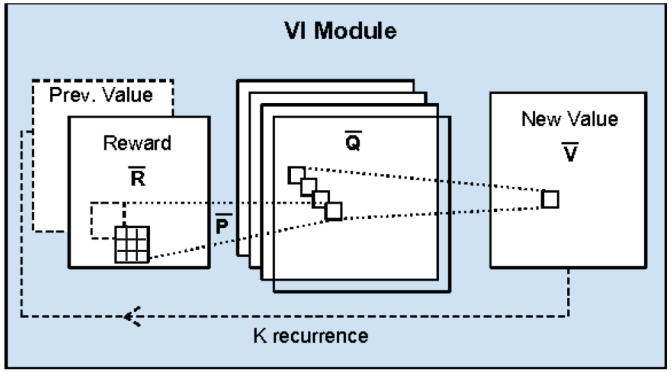
Semantic map of a house visualization

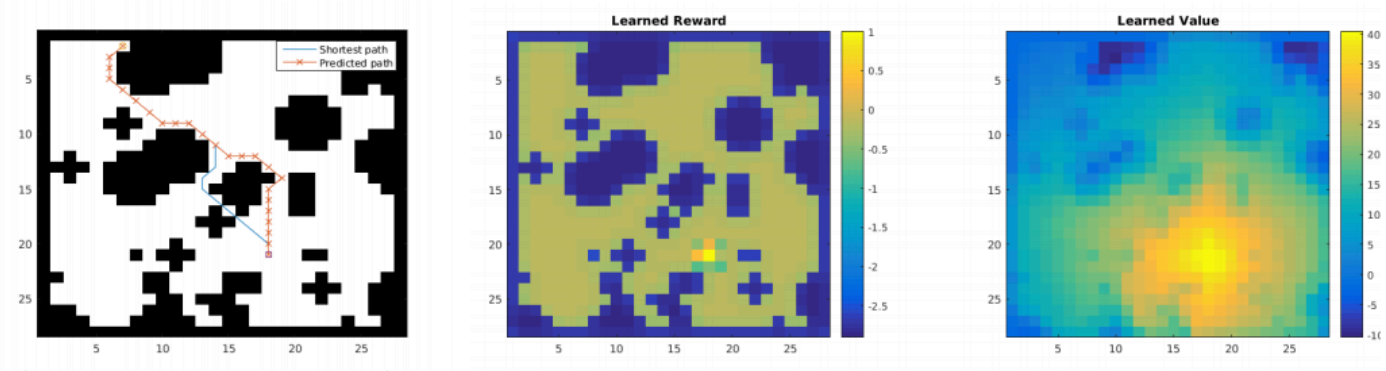


Semantic map

Value Iteration Network

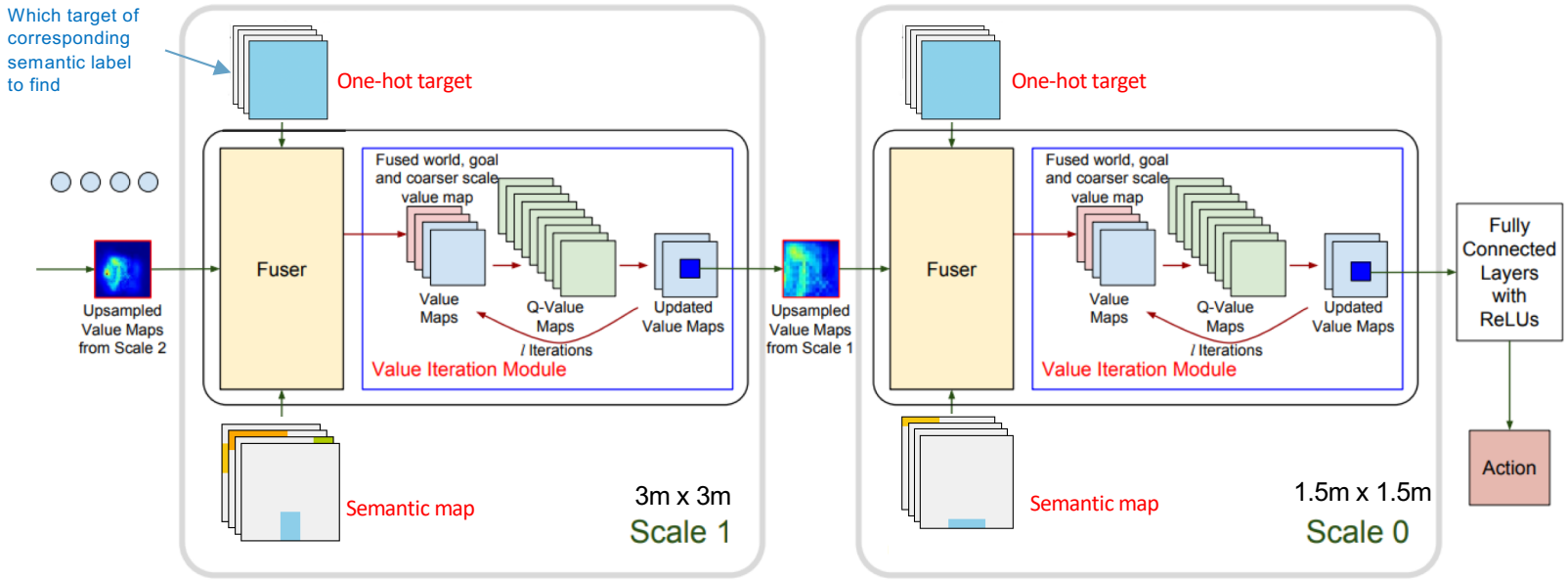






(Tammar, et al. 2016)

Network Structure



With this network structure, it has

i) a relatively large map at scale 2 for long-range planning and

ii) a detailed map at scale 0 for better actions policy learning.

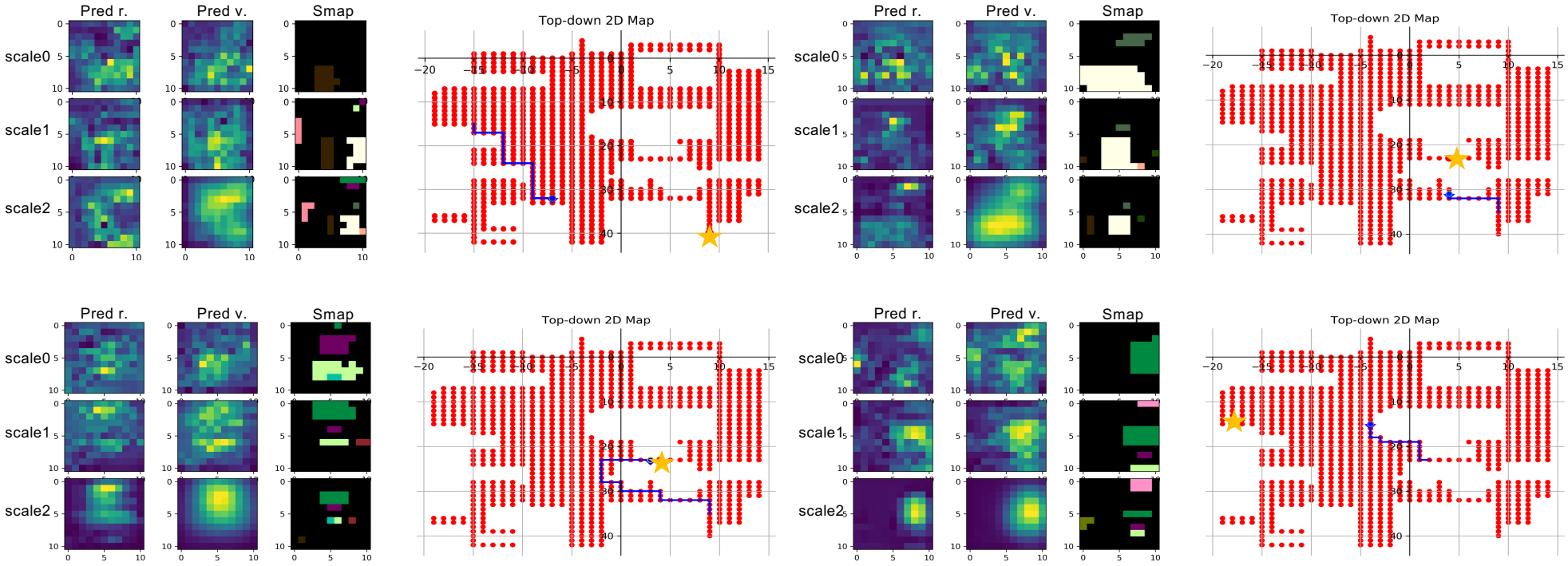
Experiment Results

- One house multi-targets
  - Supervised Learning with cross-entropy loss.
  - Trained in 1 house finding 4 targets and test the success rate of new targets in the same house.
- Multi-house multi-targets
  - Trained in 5 houses finding 4 targets and test the success rate of same targets in 1 house.

Max_step = 500	New Targets		Trained Targets	
	Succ. Rate	Ave. length	Succ. Rate	Ave. length
Random	21.35%	470.13	19.71	480.32
Ours	**29.51%**	**338.24**	**100.0**	**34.27**

Max_step = 500	New Scenes		Trained Scenes	
	Succ. Rate	Ave. length	Succ. Rate	Ave. length
Random	**25.62%**	463.18	21.46	433.45
Ours	18.75%	**370.18**	**99.0**	**25.58**

Experiment Results



* Pred v. means predicted value map; Pred r. means predicted reward map. Smap means semantic map

Experiment Results

- Analysis
  - From the visualization of learnt value and reward, the sign is quite clear that both the value map and reward map tend to highlight all the objects or the center of the objects rather than the target location.
  - The visualization and previous quantitative results shows that the model actually learns and overfits to the holistic geometric features of the entire training houses rather than the target object or the objects have close relation with the target object.
  - When testing the model on finding new objects in the same houses, the success rate drops sharply, we figure that it is because the channel-wise correlation of the semantic map and one-hot target inputs are not learnt.

Conclusion and Future work

- Conclusion
  - We presented a deep reinforcement learning network which intermediately predict the free-space map and the semantic segmentation. It is proved that using the map as the representation of environment improves the generalizability.
  - Presented a novel approach of using semantic map for object searching. However, so far the result isn't promising and further developments need to be taken.
- Future Work
  - **Increase the training set**.
  - Extend the latter model to real-world scenes.
  - Transfer learning to real-world scenes for both models.

Reference

- Mao, H., Alizadeh, M., Menache, I., & Kandula, S. (2016, November). Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks* (pp. 50-56). ACM.
- Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Fei-Fei, L., & Farhadi, A. (2017, May). Target-driven visual navigation in indoor scenes using deep reinforcement learning. In 2017 IEEE international conference on robotics and automation (ICRA) (pp. 3357-3364). IEEE.
- Ye, X., Lin, Z., Li, H., Zheng, S., & Yang, Y. (2018, October). Active object perceiver: Recognition-guided policy learning for object searching on mobile robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 6857-6863). IEEE.
- Ye, X., Lin, Z., Lee, J. Y., Zhang, J., Zheng, S., & Yang, Y. (2019). GAPLE: Generalizable Approaching Policy LEarning for Robotic Object Searching in Indoor Environment. IEEE Robotics and Automation Letters, 4(4), 4003-4010.
- Mousavian, A., Toshev, A., Fišer, M., Košecká, J., Wahid, A., & Davidson, J. (2019, May). Visual representations for semantic target driven navigation. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 8846-8852). IEEE.
- Wu, Y., Wu, Y., Gkioxari, G., & Tian, Y. (2018). Building generalizable agents with a realistic and rich 3d environment. arXiv preprint arXiv:1801.02209.
- Tamar, A., Wu, Y., Thomas, G., Levine, S., & Abbeel, P. (2016). Value iteration networks. In Advances in Neural Information Processing Systems (pp. 2154-2162).
- Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In International conference on machine learning (pp. 1928-1937).