Towards Navigation by Reasoning Over Spatial Configurations

Yue Zhang (zhan124@msu.edu), Quan Guo (guoquan@msu.edu), Parisa Kordjamshidi (kordjams@msu.edu)
Michigan State University

MOTIVATION

- ❖ Investigate the influence of the spatial semantic structure of the instructions on the navigation agent's reasoning ability.
- ❖ Using semantic representation of instruction to improve both interpretability and generalizability of VLN deep learning model.

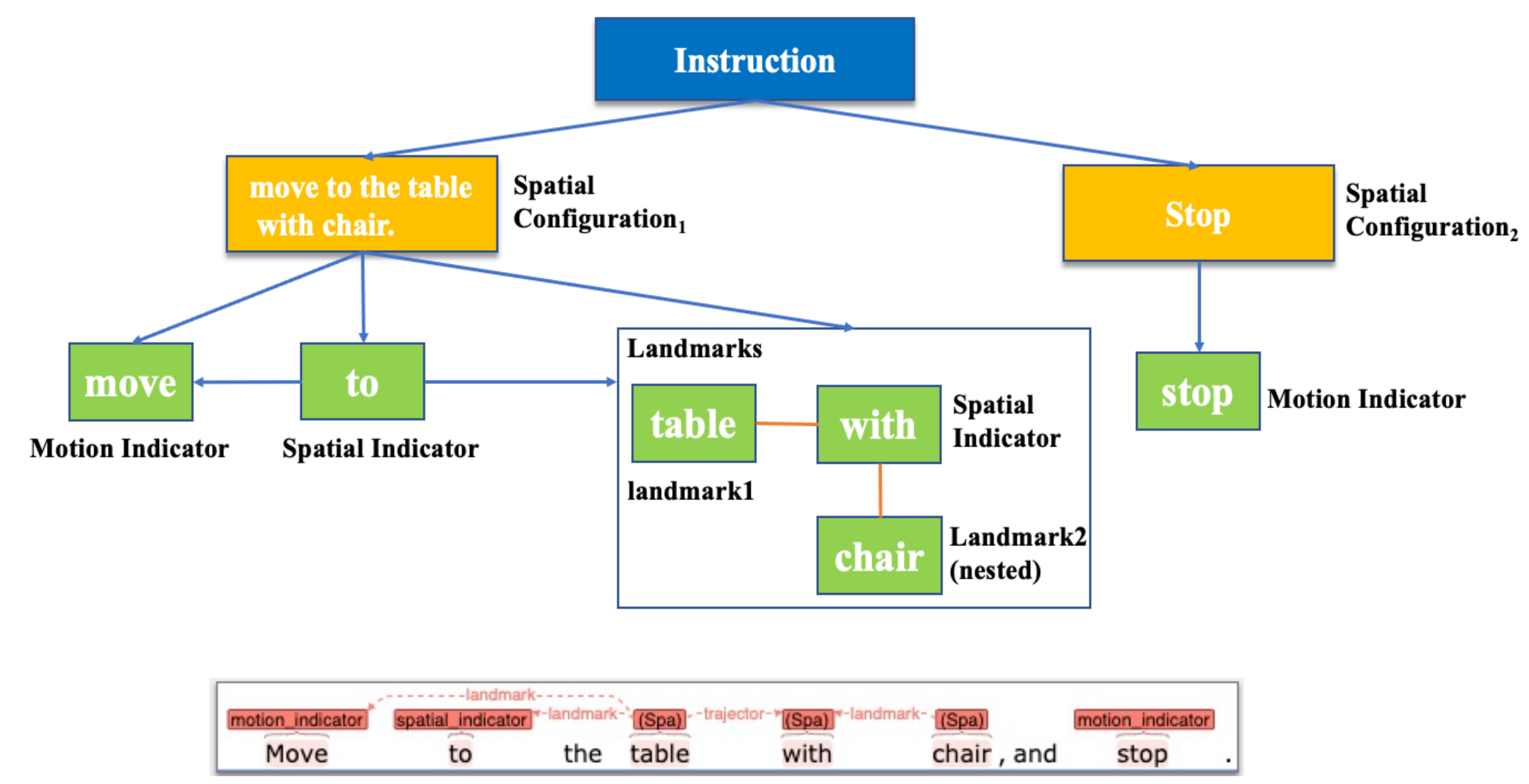
VLN TASK

A task to deal with the navigation problem where the agent follows natural language instructions while observing the photo-realistic simulated environment. The task is to select the next viewpoints or current viewpoint (indicating stop) to generate the trajectory that takes the agent close to an intended goal location.

CONTRIBUTION

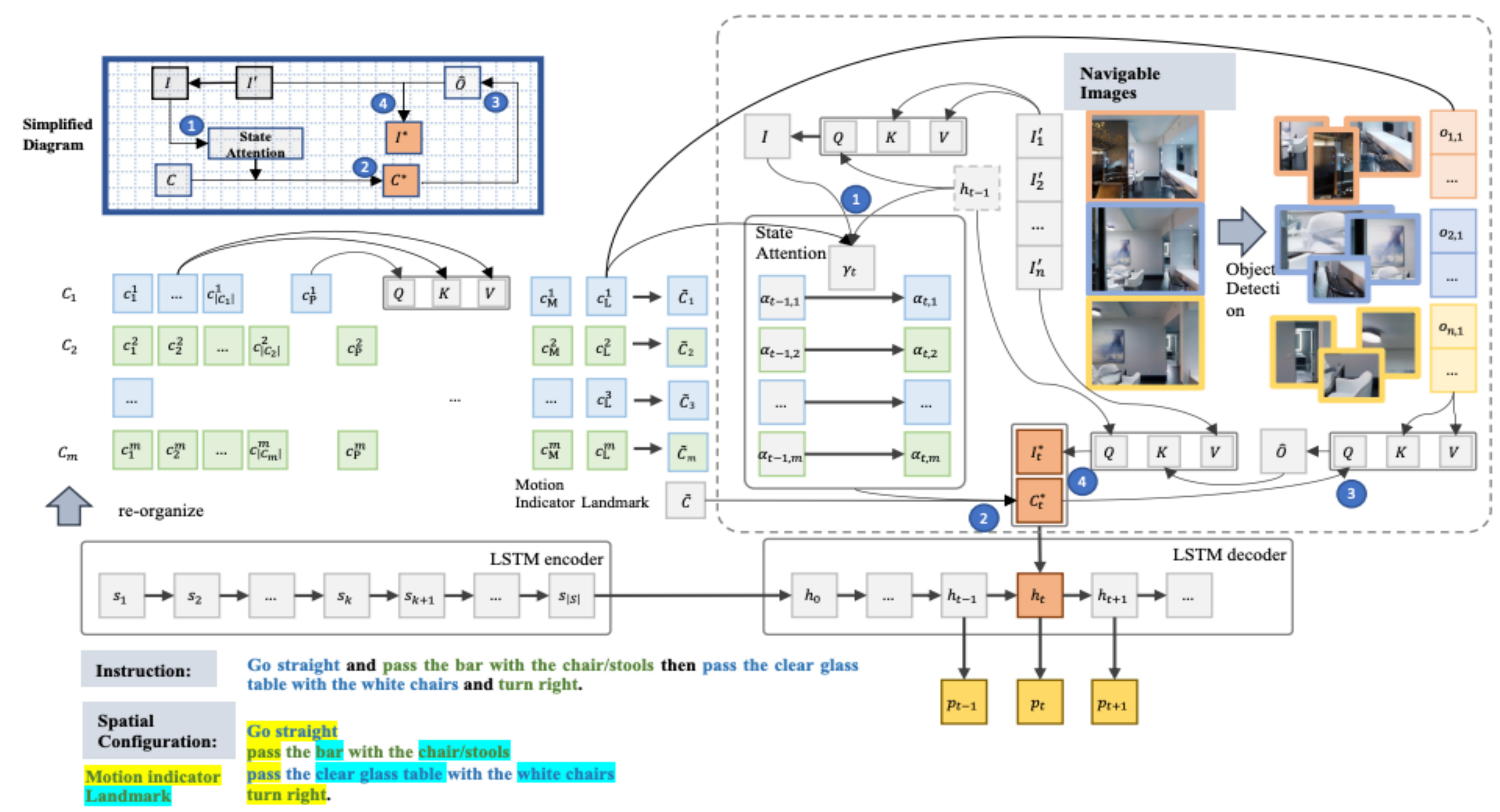
- ❖ We consider the spatial semantic structure of the instructions explicitly in terms of spatial configurations and their spatial semantic elements, i.e., spatial/motion indicators, and landmarks.
- ❖ We introduce a state attention to guarantee that configurations are executed sequentially.
- ❖ We utilize the grounding between the extracted spatial elements and the object representation to help control the transitions between configurations.
- ❖ Our model improves the strong baselines significantly in the seen environments and yields competitive results in the unseen environments.

SPATIAL CONFIGURATION



Instruction: Move to the table with chair, and stop.

MODEL



Encoder

$$\bar{s}_j = LSTM_{encode}\ s_j$$

Decoder

$$h_t = LSTM_{decode}([C_t^*, I_t^*])$$

C_t^*: Grounded Config Representation
I_t^* : Aligned Visual Representation

State Attention

$$\alpha_0 = [1, 0, 0, \cdots] \quad \gamma_0 = [1, 0]$$

$$\alpha_{t,i} = \sum_{j=i-1}^{i} \alpha_{t-1,j}\gamma_{t,i-j}$$

$$\gamma_t = FC_\gamma([h_{t-1}; \bar{I}; sim(C_L, O)])$$

C_L: Landmark Representation
O : Object Representation

RESULTS

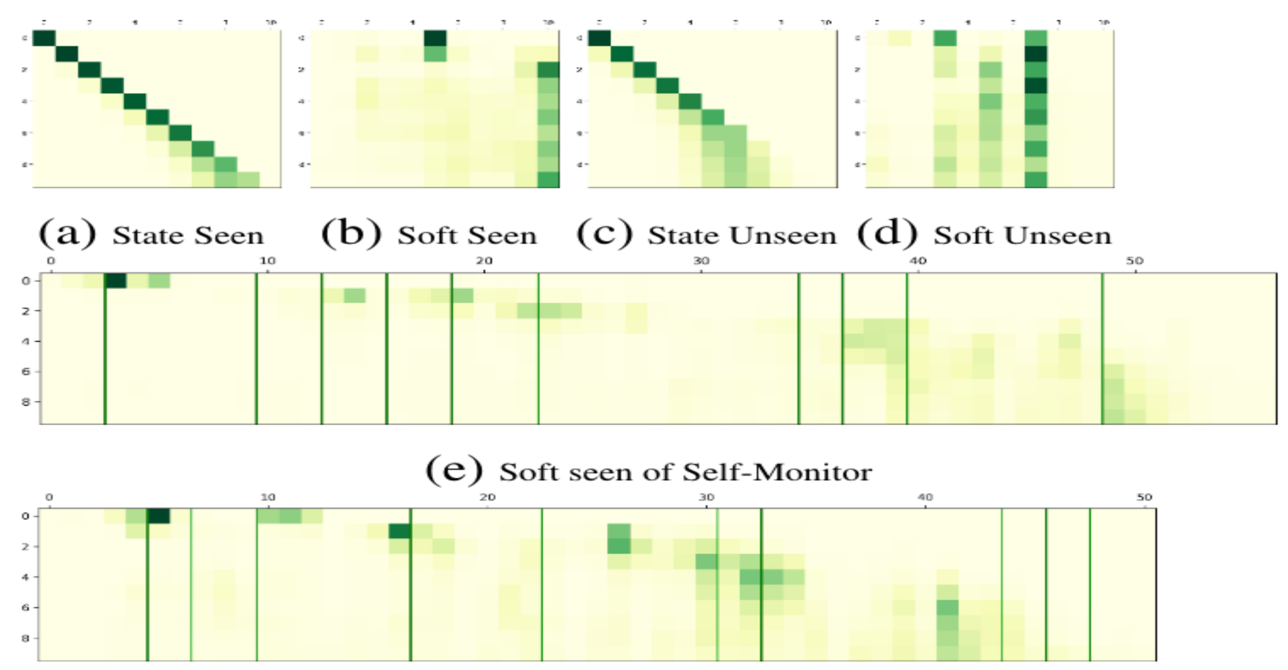
		Validation-Seen			Validation-Unseen			Test(Unseen)		
	Method	NE ↓	SR ↑	SPL ↑	NE ↓	SR ↑	SPL ↑	NE ↓	SR ↑	SPL ↑
1	Random (Anderson et al., 2018)	9.45	0.16	-	9.23	0.16	-	9.77	0.13	0.12
2	Student-forcing (Anderson et al., 2018)	6.01	0.39	-	7.81	0.22	-	7.85	0.20	0.18
3	Speaker-Follower (Fried et al., 2018)	4.36	0.54	-	7.22	0.27	-	-	-	-
4	Speaker-Follower*	3.66	0.66	0.58	6.62	0.36	-	6.62	0.35	0.28
5	Self-Monitor* (Ma et al., 2019)	**3.22**	**0.67**	0.58	5.52	0.45	0.32	**5.67**	0.48	0.35
6	Environment Dropout* (Tan et al., 2019)	4.19	0.58	0.55	**5.43**	**0.48**	**0.44**	-	**0.52**	**0.47**
7	Environment Dropout + BERT*	4.40	0.61	0.57	5.54	0.46	0.43	-	-	-
8	SpC-NAV*	4.09	0.65	**0.61**	5.92	0.45	0.42	6.22	0.46	0.44

Table1 Experiment Result comparing with baseline models.
* means data augmentation

	Validation-Seen			Validation-Unseen		
Model	NE↓	SR↑	SPL↑	NE↓	SR↑	SPL↑
1 SpC-NAV	4.11	0.62	0.53	6.49	0.39	0.29
2 SpC-NAV$_{M}$	**3.88**	0.62	0.53	**6.21**	**0.40**	0.28
3 SpC-NAV$_{M+L}$	4.01	0.62	0.54	6.27	0.39	0.29
4 SpC-NAV$_{M+L+S}$	3.95	**0.65**	**0.59**	6.51	0.39	**0.32**

Table2 Ablation Study with different spatial semantics. M: *motion indicator*; **L**: *landmark*; **S**: *similarity score*.

STATE ATTENTION VISULIZATION



(a) State Seen (b) Soft Seen (c) State Unseen (d) Soft Unseen

(e) Soft seen of Self-Monitor

(f) Soft unseen of Self-Monitor

EXAMPLE

Instruction: Turn right, and walk past the couch.

with similarity between landmarks and objects

without similarity between landmarks and objects

REFERENCE

- Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3674-3683.
- Ma C Y, Lu J, Wu Z, et al. Self-monitoring navigation agent via auxiliary progress estimation[J]. arXiv preprint arXiv:1901.03035, 2019.
- Tan H, Yu L, Bansal M. Learning to navigate unseen environments: Back translation with environmental dropout[J]. arXiv preprint arXiv:1904.04195, 2019.

MICHIGAN STATE UNIVERSITY