Improving the Robustness to Variations of Objects and Instructions with a Neuro-Symbolic Approach for Interactive Instruction Following

Kazutoshi Shinoda, Yuki Takezawa, Masahiro Suzuki, Yusuke Iwasawa, Yutaka Matsuo

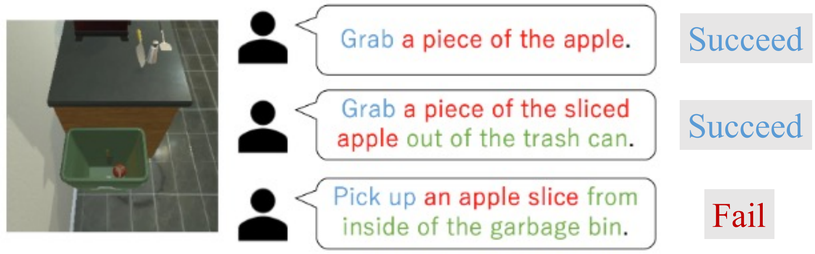
@shino__c shinoda@is.s.u-tokyo.ac.jp

Summary

- We propose Neuro-Symbolic Instruction Follower (NS-IF), which introduces high-level symbolic feature extraction and reasoning modules to improve the robustness to variations of objects and language instructions for the interactive instruction following task.
- In subtasks requiring interaction with objects, our NS-IF significantly outperforms an existing end-to-end neural model in the success rate while improving the robustness to the variations of vision and language inputs

Lack of Robustness to Variations of Vision and Language Inputs

We find that an existing end-to-end neural model for interactive instruction following lacks robustness to variations of language instructions and attributes of objects as shown below.

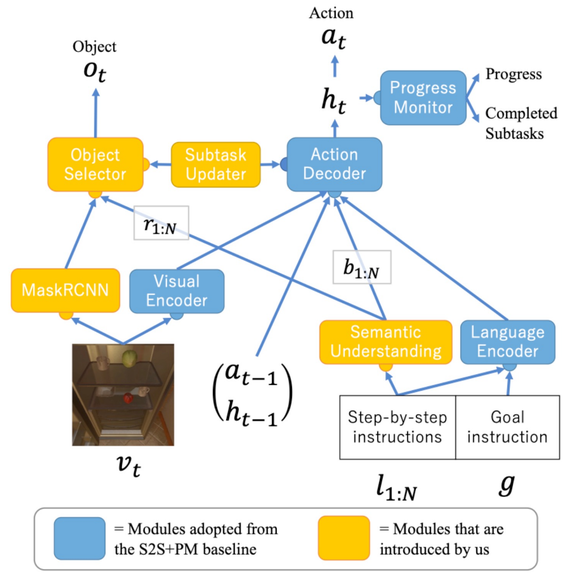


Different language instructions are given by different annotators to the same action, taken from ALFRED.

Four apples with different attributes such as color, texture, and shape, taken from ALFRED.

High-level Symbolic Representations

The proposed Neuro-Symbolic Instruction Follower (NS-IF) utilizes symbolic representations obtained from MaskRCNN and Semantic Understanding to improve the robustness to variations of inputs.



In this study, we use the ground-truth for the output of Subtask Updater.

Example of Symbolic Representation

Law Inputs

n	Step-by-step instructions l_n
0	Turn right then head to the counter beside the microwave
1	Pick up the knife on the counter
2	Turn left then head to the sink
3	Slice the apple in the sink

Semantic Understanding

MaskRCNN

Symbolic Repr.

High-level action b_n	Argument r_n
GotoLocation	countertop
PickupObject	knife
GotoLocation	apple
SliceObject	apple

Objects:
DishSponge,
ButterKnife, …

Subtask Evaluation

We evaluate the performance on each subtask here.
Dataset: ALFRED (Shridhar et al., 2020)
Metrics: Success rate (path length weighted score)

	Model	Goto	Pickup	Slice	Toggle
Seen	S2S+PM [19]	- (51)	- (32)	- (25)	- (100)
	S2S+PM (Reproduced)	55 (46)	37 (32)	20 (15)	**100** (100)
	MOCA [17]	**67** (54)	**64** (54)	67 (50)	95 (93)
	NS-IF	43 (37)	**64** (58)	**71** (57)	83 (83)
	NS-IF (Oracle)	43 (37)	69 (63)	73 (59)	100 (100)
Unseen	S2S+PM [19]	- (22)	- (21)	- (12)	- (32)
	S2S+PM (Reproduced)	26 (15)	14 (11)	3 (3)	34 (28)
	MOCA [17]	**50** (32)	**60** (44)	68 (44)	11 (10)
	NS-IF	32 (19)	**60** (49)	**77** (66)	**43** (43)
	NS-IF (Oracle)	32 (19)	65 (53)	78 (53)	49 (49)

Our NS-IF outperforms S2S+PM (w/o symbolic repr.) by 9, 46, and 74 points in the success rate on the Toggle, Pickup, and Slice subtasks in unseen environments respectively.

Conclusion

High-level symbolic representations are effective to improve the robustness to small changes in vision and language inputs.