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

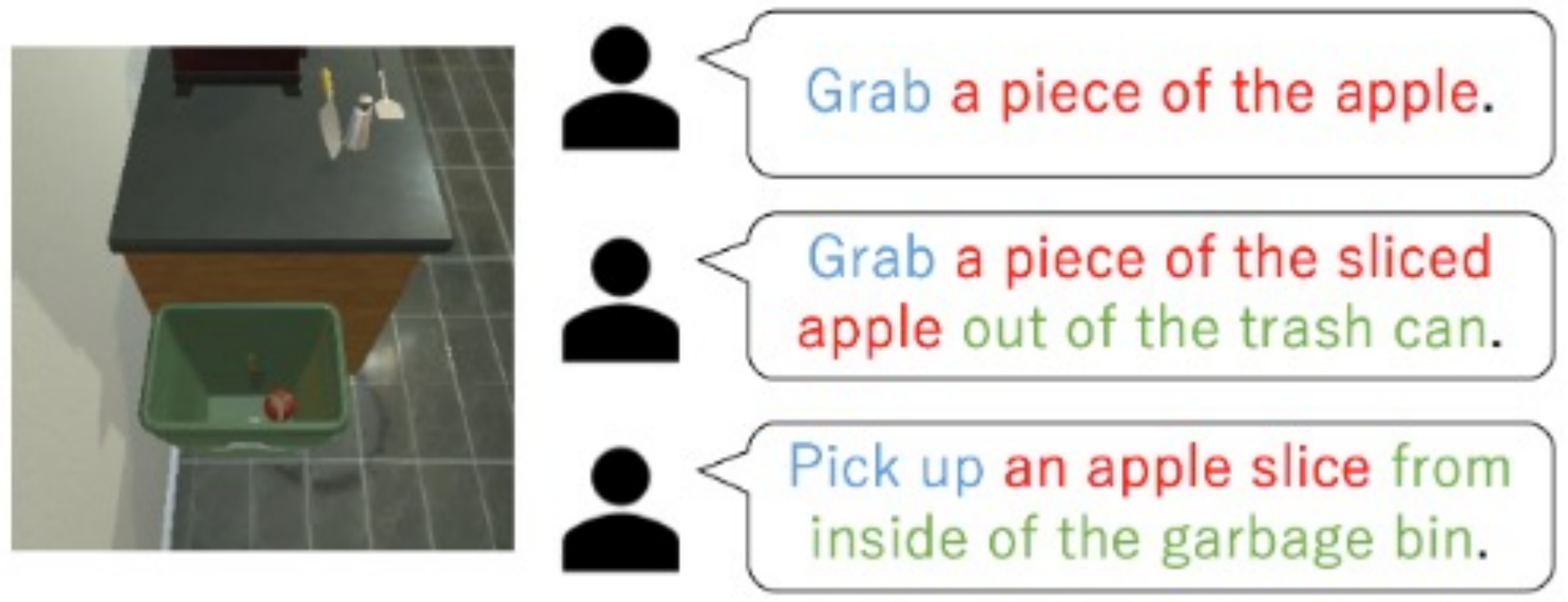
Kazutoshi Shinoda shinoda@is.s.u-tokyo.jp
Yuki Takezawa yuki-takezawa@ml.ist.kyoto-u.ac.jp
Masahiro Suzuki, Yusuke Iwasawa, Yutaka Matsuo {masa,iwasawa,matsuo}@weblab.t.u-tokyo.ac.jp

Summary

- We propose Neuro-Symbolic Instruction Follower (NS-IF), which introduces object detection and semantic parsing 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.



Grab a piece of the apple.

Grab a piece of the sliced apple out of the trash can.

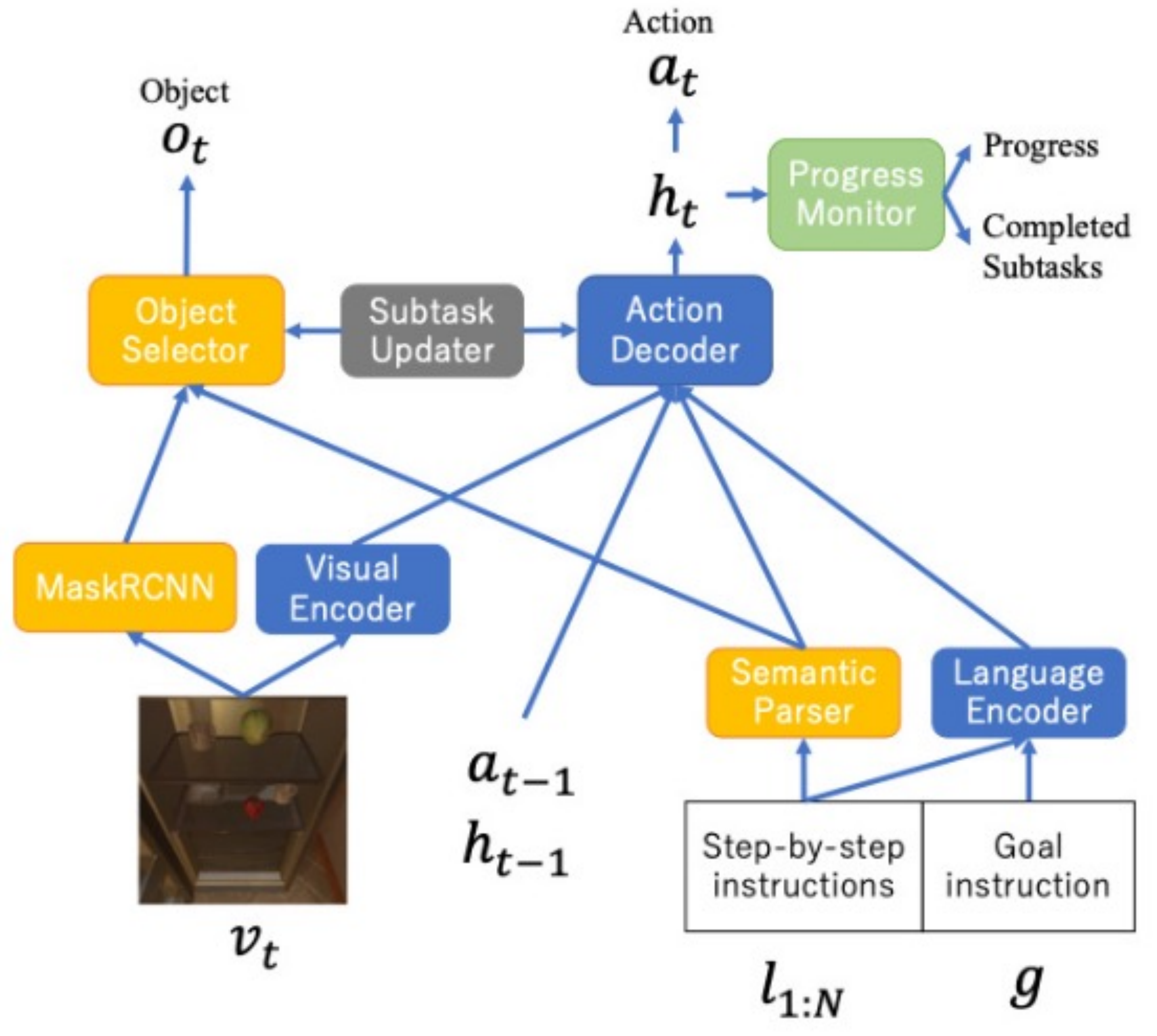
Pick up an apple slice from inside of the garbage bin.

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.

Introducing High-level Symbolic Representations

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



In this study, **we use the ground-truth high-level symbolic representations for the output of Subtask Updater and Semantic Parser**.

Example of Symbolic Representation

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 Parser

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

MaskRCNN

Objects:
DishSponge,
ButterKnife,
Fork, Pot, …

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 (Paper)	- (51)	- (32)	- (25)	- (100)
	S2S+PM (Ours)	**55** (46)	37 (32)	20 (15)	**100** (100)
	NS-IF	42 (35)	**70** (64)	**73** (59)	**100** (99)
Unseen	S2S+PM (Paper)	- (22)	- (21)	- (12)	- (32)
	S2S+PM (Ours)	26 (15)	14 (11)	3 (3)	34 (28)
	NS-IF	**28** (17)	**66** (54)	**76** (52)	**52** (52)

The proposed NS-IF model outperforms the existing model by 18, 52, and 73 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. This study is still in progress.