2021年度 人工知能学会全国大会 ロボットと実世界：要素技術
6月9日(水) 13:20 〜 15:00 J会場 (GS会場 5)
[2J3-GS-8b-03]

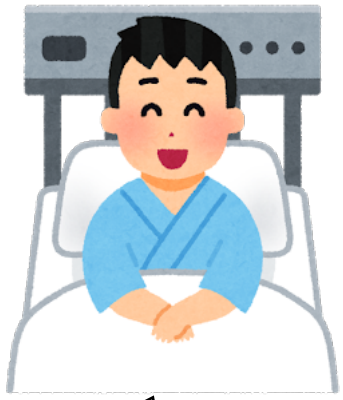
Interactive Instruction Followingのための Neuro-Symbolic手法による 多様な物体と言語指示への頑健性の向上

篠田 一聡[1]　竹澤 祐貴[2]　鈴木 雅大[1]　岩澤 有祐[1]　松尾 豊[1]

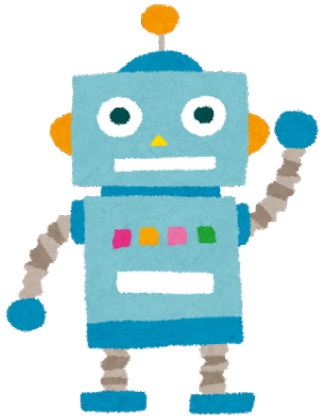
[1]東京大学　[2]京都大学

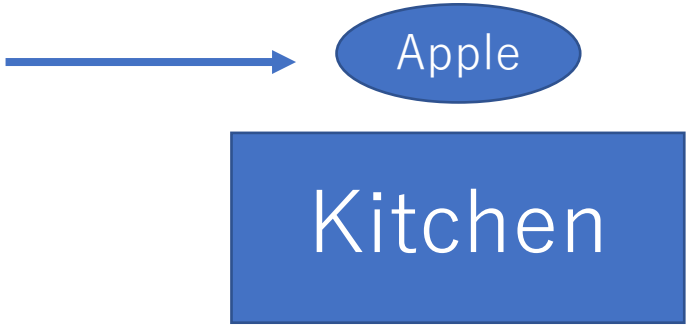
Instruction Followingとは?

●自然言語で記述された“指示”によってロボットを操作するための研究



Go to the kitchen and pick up the apple





✔非専門家でもロボットの操作が可能になり、身体が不自由な方への支援が可能
✔実世界に接地された言語理解を目指す
✖しかしこれまでは移動が主な目的

Interactive Instruction Following
——ALFRED dataset (Shridhar et al., 2020)

●移動だけでなく物体との相互作用が必要なInstruction Followingタスク

Goal Instruction	Put a chilled apple in the microwave.
Step-by-Step Instructions	Turn around and go to the counter by the sink. Pick up an apple on the counter. Turn around and go to the refrigerator. Chill the apple in the refrigerator. Turn and go to the microwave. Put the apple in the microwave.

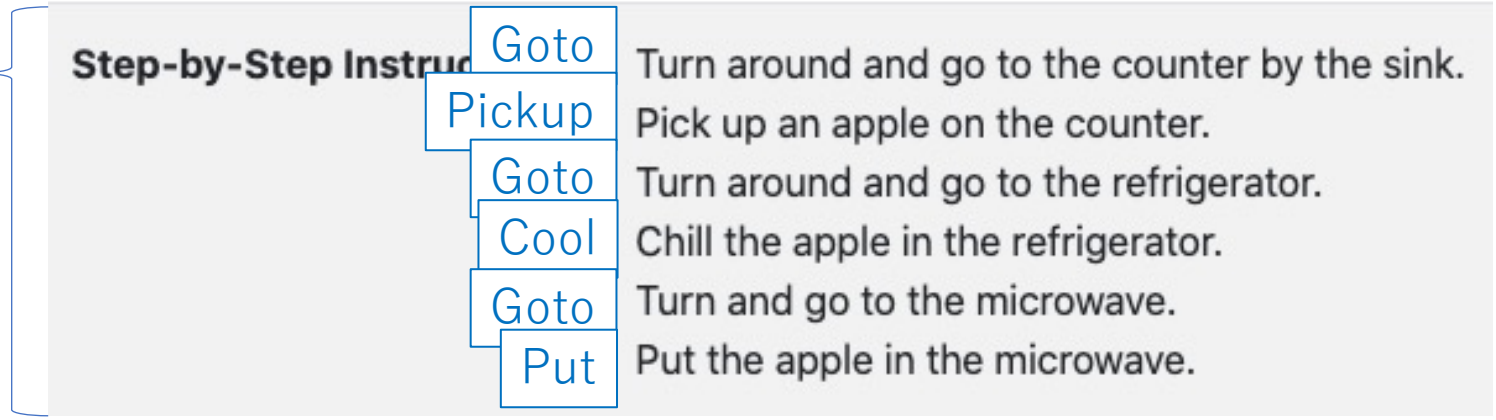
https://askforalfred.com/

Interactive Instruction Following
——ALFRED dataset (Shridhar et al., 2020)

●移動だけでなく物体との相互作用が必要なInstruction Followingタスク

タスク全体の目的を示す指示 →

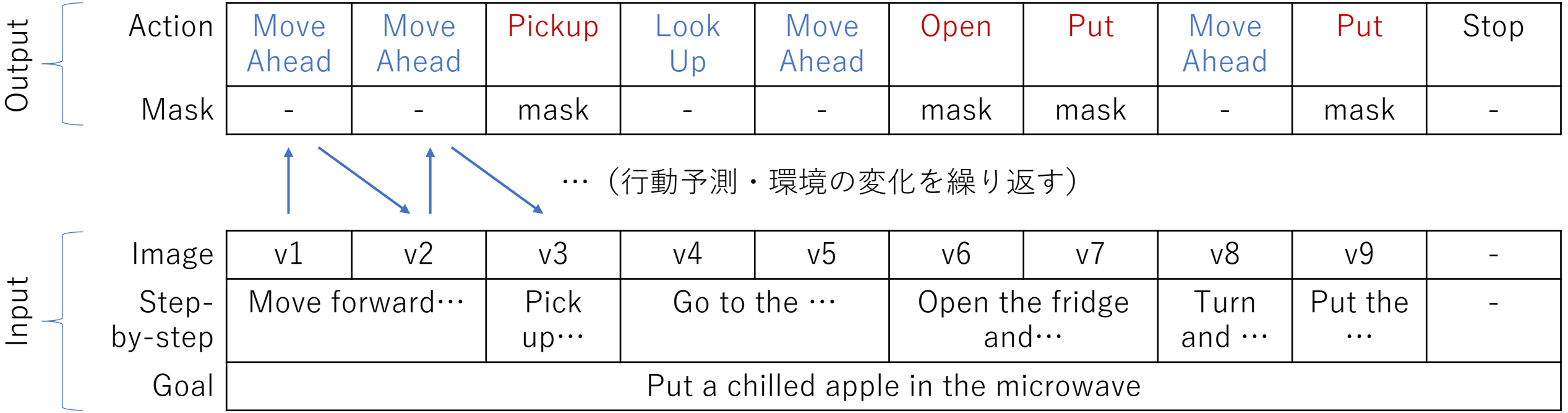
タスクを構成する
サブタスクごとの指示 →



✔複数のサブタスク(全8種類)を順番通りに連続して達成することでタスク全体の成功

✔本研究ではタスク全体ではなく、サブタスクごとの評価を行う

Interactive Instruction Following ——ALFRED dataset (Shridhar et al., 2020)



Navigation actions:
MoveAhead, RotateRight, RotateLeft, Lookup, LookDown
Interaction actions (ピクセルごとのマスクを指定する必要):
Pikuup, Put, Open, Close, ToggleOn, ToggleOff, Slice

課題

課題1. 多様な言語指示に対する頑健性の欠如

既存のEnd-to-endなニューラルモデルは多様な言語指示に頑健ではない

✔ALFREDでは各動作に対して3つの言語指示がアノテーションされている

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.

✔言語指示は述語・目的語（参照表現）・修飾語において多様になりうる
✔ニューラルモデルは言語指示の表現が変わるだけで失敗することがある

Related Work: Deep NLU models are not so robust to paraphrased textual inputs (Gan and Ng, 2019)

課題1. 多様な言語指示に対する頑健性の欠如

Seen			Clean	Cool	Goto	Heat	Pickup	Put	Slice	Toggle
	baseline	全ての言語指示で成功する	32	37	315	28	105	258	7	29
		少なくとも1つの言語指示で成功するが他の言語指示では失敗する	0	0	240	1	52	19	5	0
		いずれの言語指示でも失敗する	5	2	239	5	202	50	29	0
Unseen			Clean	Cool	Goto	Heat	Pickup	Put	Slice	Toggle
	baseline	全ての言語指示で成功する	0	28	147	36	42	102	1	13
		少なくとも1つの言語指示で成功するが他の言語指示では失敗する	0	5	99	5	21	45	0	10
		いずれの言語指示でも失敗する	36	3	513	1	281	143	31	30

✖特にGoto, Pickup, Putにおいて、言語指示によっては失敗することがある（赤枠）

課題2. 多様な物体に対する頑健性の欠如

既存のEnd-to-endなニューラルモデルは多様な物体に頑健ではない

✔形、色、質感が異なるりんごの例

Related Work: Deep CNNs tend to learn surface statistical cues in the dataset rather than higher-level abstract concepts (Jo and Bengio, 2017)

課題2. 多様な物体に対する頑健性の欠如

Sub-Goal Ablations - **Validation**

	Model	Goto	Pickup	Put	Cool	Heat	Clean	Slice	Toggle	Avg.
Seen	No Lang	28	22	71	**89**	**87**	64	19	90	59
	S2S	49	32	80	87	85	**82**	23	97	67
	S2S + PM	**51**	**32**	**81**	88	85	81	**25**	**100**	**68**
Unseen	No Lang	17	9	31	75	86	13	8	4	30
	S2S	21	20	**51**	**94**	88	21	**14**	**54**	45
	S2S + PM	**22**	**21**	46	92	**89**	**57**	12	32	**46**

Table 4: **Evaluations by path weighted sub-goal success.** All values are percentages. The highest values per fold and task are shown in **blue**. We note that the NO VISION model achieves less than 2% on all sub-goals. See supplemental material for more.

(Shridhar et al., 2020)

✖特に物体を拾う・切断するサブタスクで未知の物体に対して頑健ではない

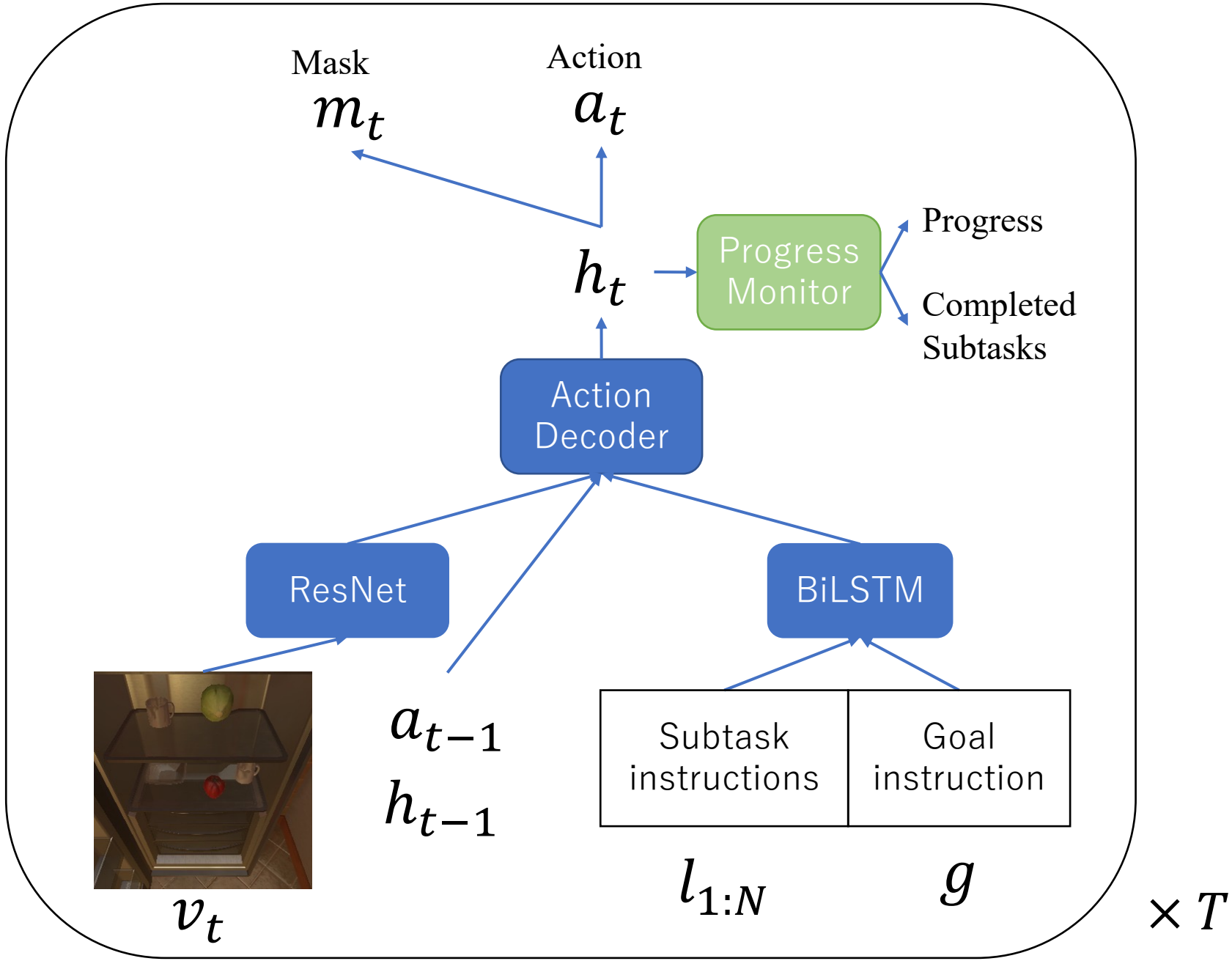
Research Question

画像・言語指示を抽象化された記号表現に変換して利用することで多様な物体・言語指示への頑健性を向上できるか？

- ニューラルモデルは入力の小さな変化に対して敏感
- 離散的な記号表現を用いれば、入力の多様性を吸収することで頑健性を向上できるのではないか？

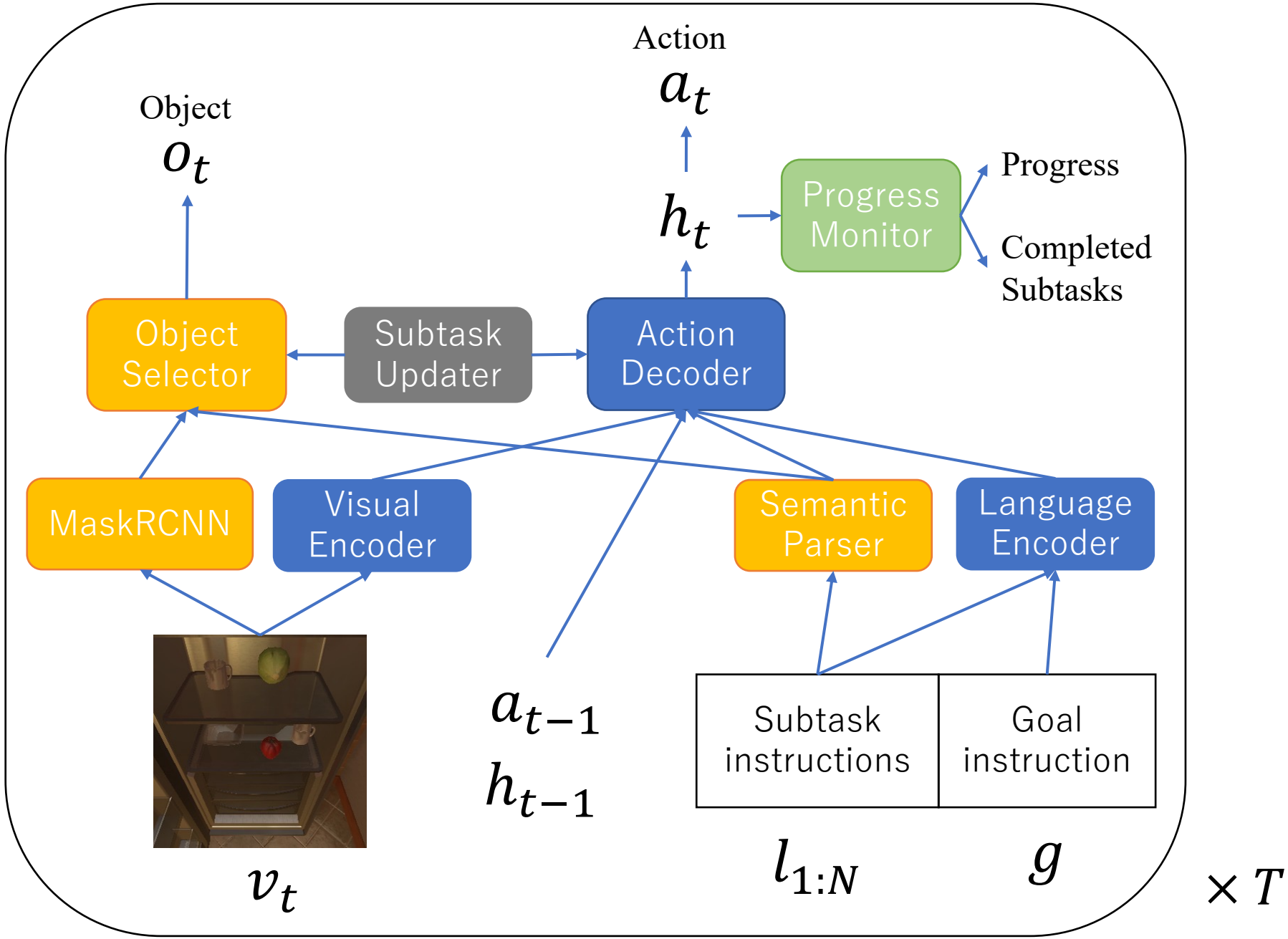
手法

Baseline model
SEQ2SEQ+PM
(Shridhar et al., 2020)

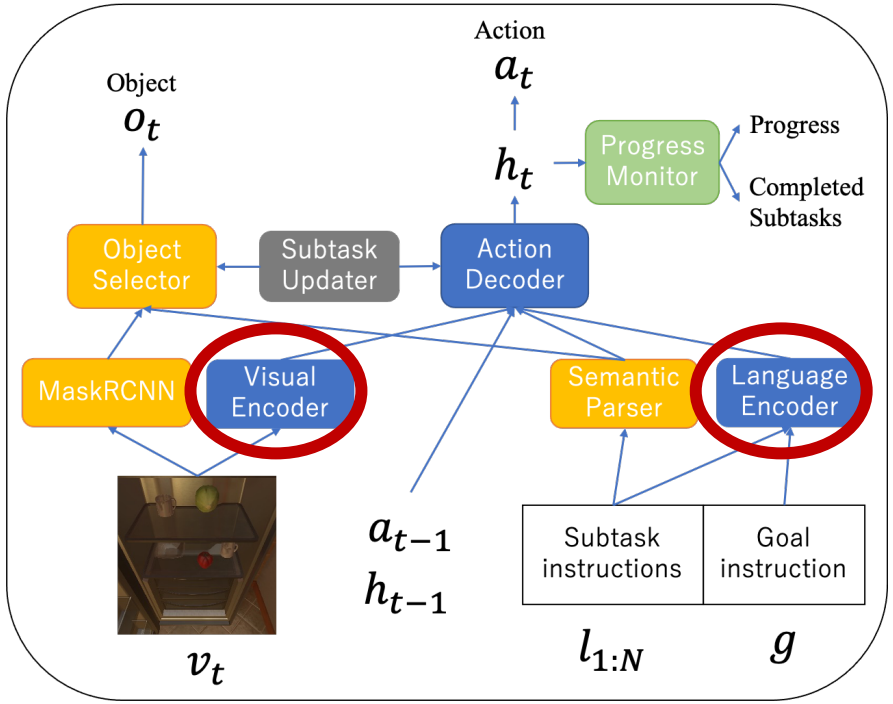


提案手法
NS-IF

ベースラインとの違い
1. Semantic Parser
2. MaskRCNN
3. Subtask Updater



連続的な特徴量の抽出



Visual Encoder (ResNet)

- 入力: 画像 v_t
- 出力: V_t

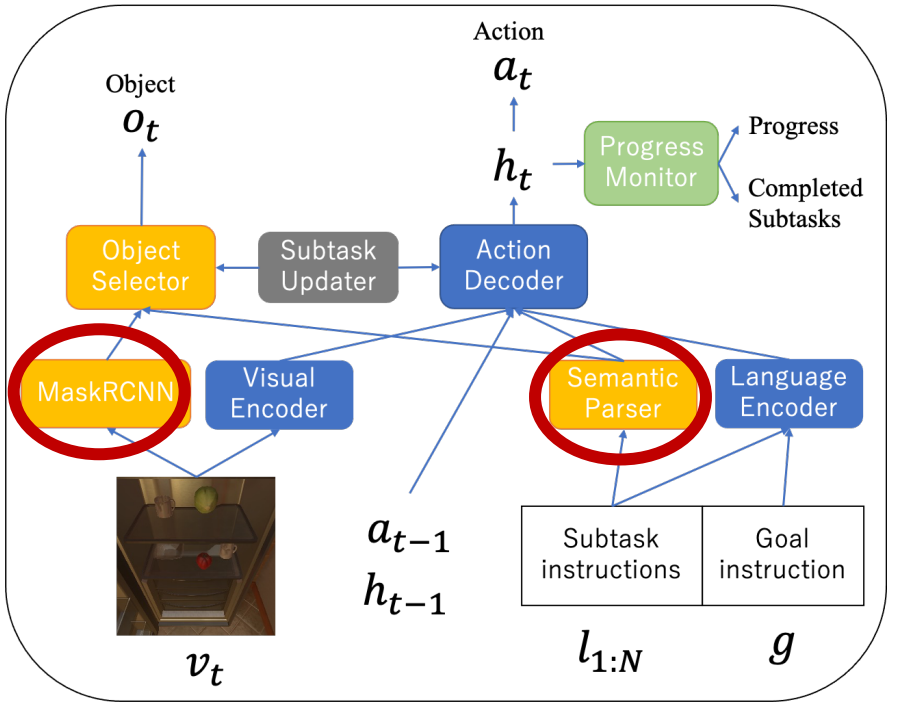
Language Encoder (BiLSTM)

- 入力: サブタスクの言語指示$l_{1:N}$、ゴールの言語指示g
- 出力: H（decoderのattentionの入力として使われる）

記号表現の獲得

Semantic Parser

- 入力: 言語指示 l_n
- 出力: (述語 b_n, 目的語 r_n)のペア
  - 例: (Clean, Cup), (Goto, Shelf)
- ここではOracleを使う



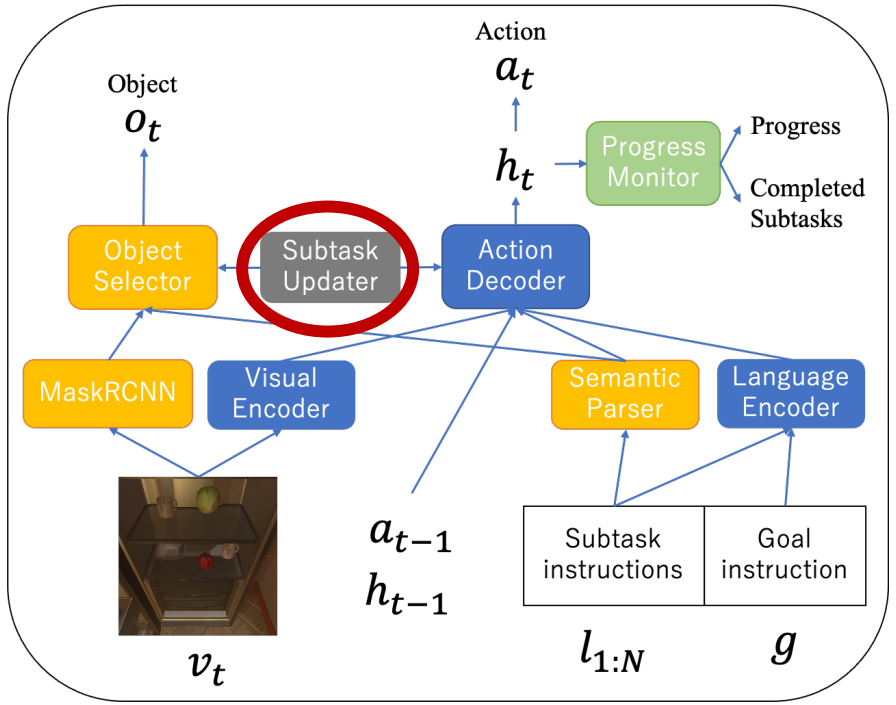
MaskRCNN

- 入力: 画像
- 出力: 各物体のマスク、クラス、（bbox、確信度）
- ALFREDデータセットで訓練済みのものを使う

Subtask Updater

獲得した記号表現をより有効に活用するために、今どのサブタスクを行っているかの予測を行う

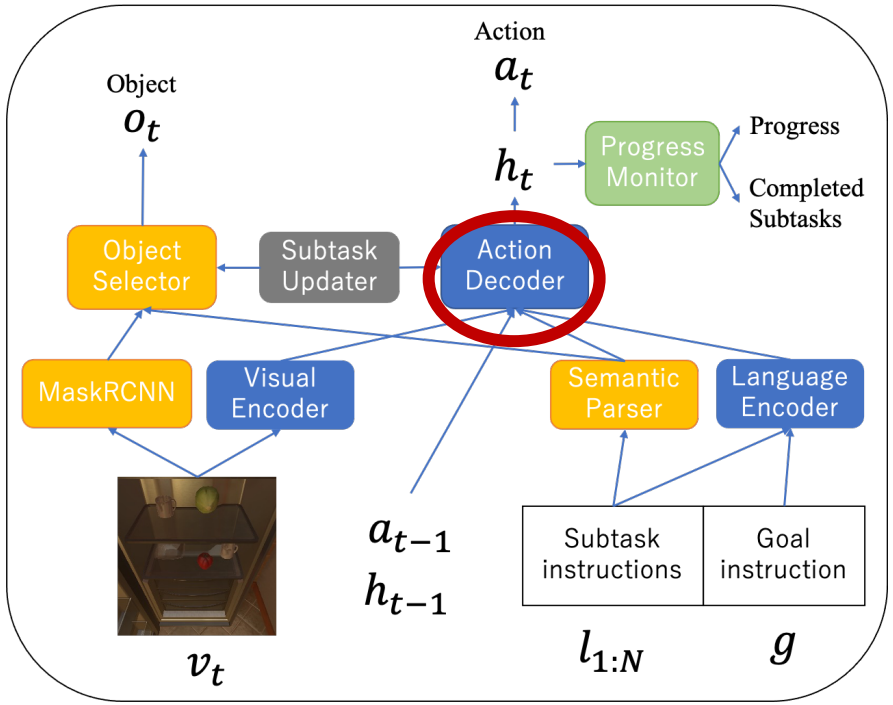
- 行っているサブタスクの確率分布$p(s_t)$を予測
- ここではOracleを使う



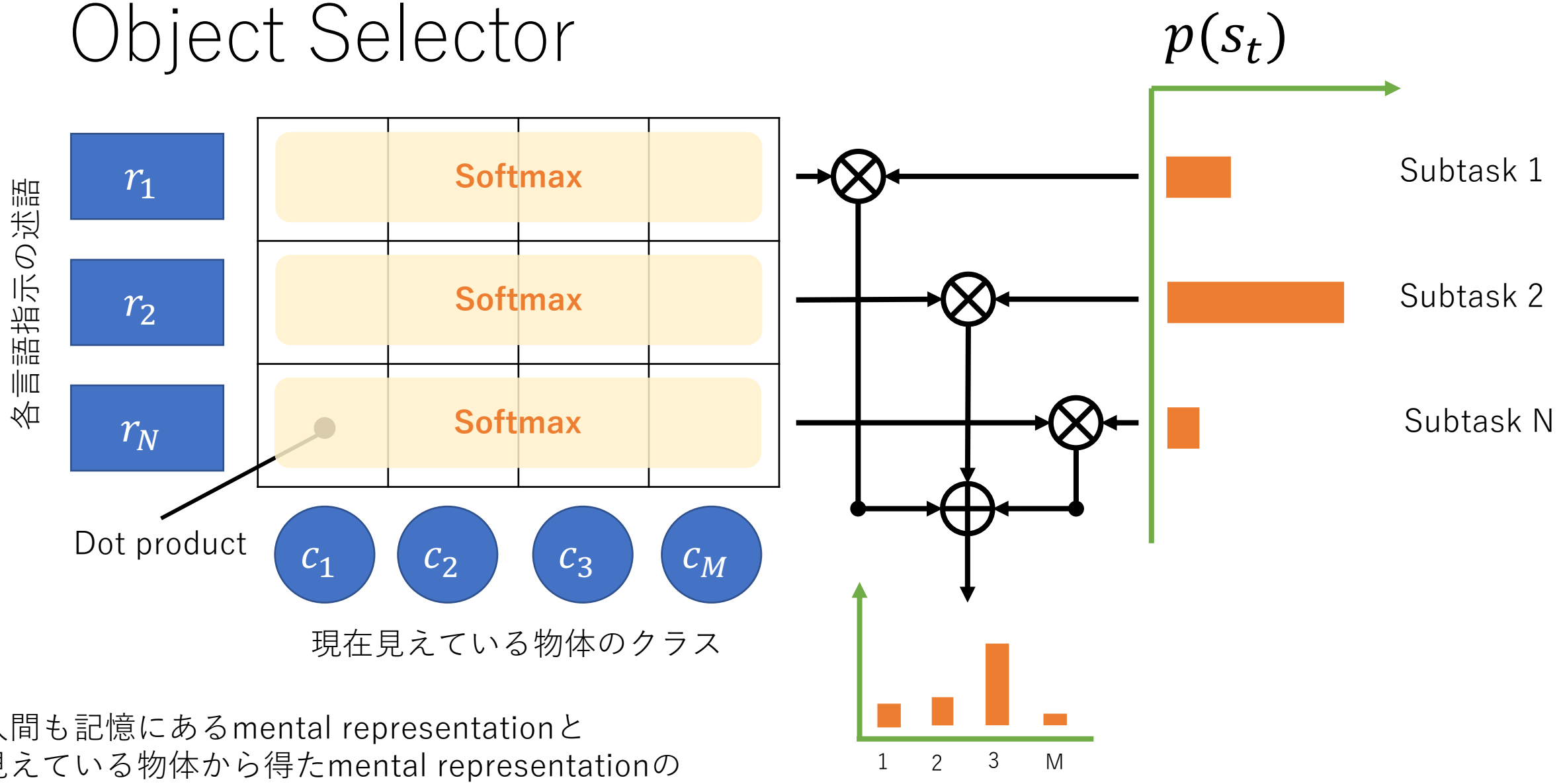
Action Decoder (LSTM)

ベースラインと違う部分
時刻tに行っているサブタスクの
高レベルな行動の埋め込みベクトル
の期待値

- 入力: $i_t = [V_t;\ Att(H, h_{t-1});\ E(a_{t-1});\ E(b_{1:N})^T p(s_t)]$
  - Att: アテンション, E: 埋め込み層
- $h_t = LSTM(i_t, h_{t-1})$
- 出力: $p(a_t) = softmax(w(i_t; h_t) + b)$



Object Selector



人間も記憶にあるmental representationと
見えている物体から得たmental representationの
対応をとっている（山鳥重『わかるとはどういうことか』）

実験

実験

- サブタスクごとの評価
  - あるサブタスクの直前までモデルを行動させたのちに予測を開始し、そのタブタスクの成否で評価

- 比較手法
  - SEQ2SEQ+PM (Shridhar et al., 2020)
  - Neuro-Symbolic Instruction Follower (NS-IF) (提案手法)

結果

表 1: サブタスク毎の成功率（%）.() 内には成功に要した行動の数を考慮したスコアを報告する.

	Model	Goto	Pickup	Slice	Toggle
Seen	S2S+PM (Paper)	- (51)	- (32)	- (25)	- (100)
	S2S+PM (Reproduce)	**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 (Reproduce)	26 (15)	14 (11)	3 (3)	34 (28)
	NS-IF	**28** (17)	**66** (54)	**76** (52)	**52** (52)

✔物を拾う・切断する・ボタンを押すサブタスクで、大幅に精度を向上
→多様な属性の物体に対する頑健性が向上

✖目的地に向かうサブタスクでは改善の余地がある

分析
——多様な言語指示に対して頑健か？

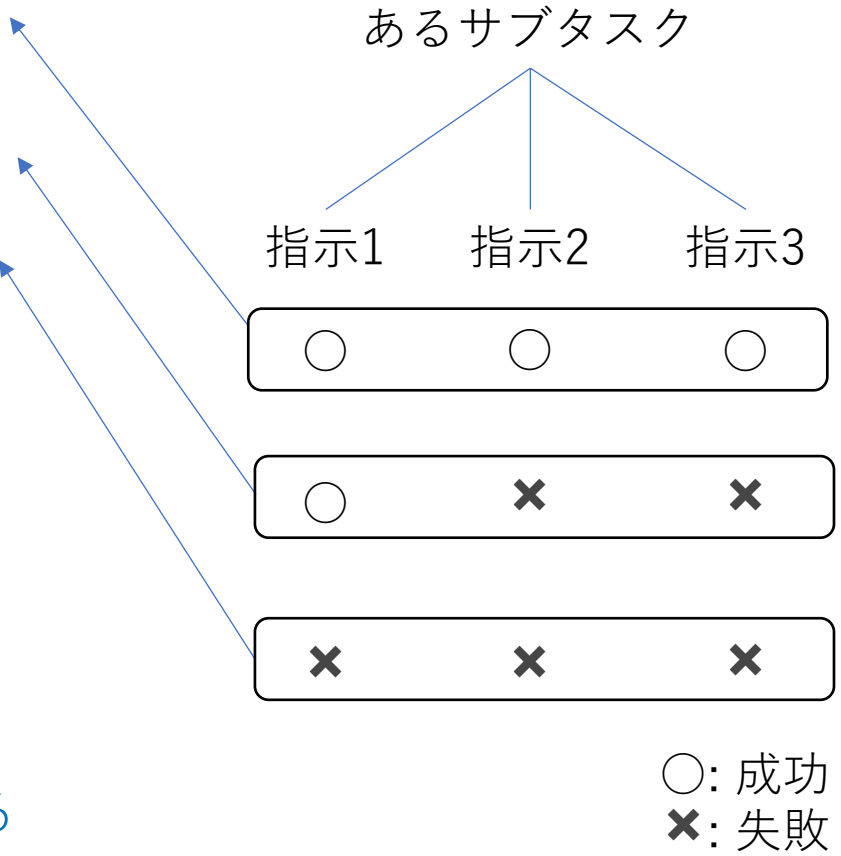
SEQ2SEQ+PM

	Goto	Pickup	Slice	Toggle
全ての言語指示で成功する	147	42	1	13
少なくとも1つの言語指示で成功するが他の言語指示では失敗する	99	21	0	10
いずれの言語指示でも失敗する	513	281	31	30

減少

NS-IF

	Goto	Pickup	Slice	Toggle
全ての言語指示で成功する	165	218	25	28
少なくとも1つの言語指示で成功するが他の言語指示では失敗する	89	12	0	0
いずれの言語指示でも失敗する	502	113	7	25



✔言語指示によっては失敗するケースが減っている
→提案手法(NS-IF)は多様な言語指示に対して頑健

Conclusion

貢献

- Interactive Instruction Followingのための、物体と言語指示の抽象化された記号表現を用いる手法を提案した
- 物体を選択するサブタスクにおいて多様な物体と言語指示に対する頑健性が大幅に向上することに成功した

今後の課題

- 意味解析、サブタスク予測モジュールの実装
- その他のサブタスク・タスク全体での精度向上

関連研究
Neuro-Symbolic手法

画像を見て質問に答えるタスク

- Mask RCNN + Semantic Parsing (Yi et al., 2018)
- Unsupervised representation learning of objects and semantic parsing (Mao et al., 2018)
- Scene Graph Prediction + Iterative Reasoning (Hudson et al., 2019)

動画を見て質問に答えるタスク

- 出力は答えのみ(Yi et al., 2020)

Interactive Instruction FollowingでNeuro-Symbolic手法を用いた研究はない

References

- Gan and Ng. 2019. Improving the Robustness of Question Answering Systems to Question Paraphrasing.
- Jo and Bengio. 2017. Measuring the tendency of CNNs to Learn Surface Statistical Regularities.
- Yi et al. 2018. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.
- Mao et al. 2018. The Neuro-Symbolic Concept Learner: In-terpreting Scenes, Words, and Sentences From Natural Supervision.
- Hudson et al. 2019. Learning by Abstraction: The Neural State Machine.
- Yi et al. 2020. CLEVRER: Collision Events for Video Representation and Reasoning.