11-777 Spring 2021 Class Project
Analysis of Baselines

Christian Deverall* **Jingyuan Li** * **Artidoro Pagnoni** *
{cdeveral, jingyua4, apagnoni}@andrew.cmu.edu

1 Introduction

As detailed in the Baselines and Metrics chapter, the baselines we evaluate are the "Modular Object-Centric Approach" (MOCA) (Singh et al., 2020) model and the "Sequence-to-Sequence with Progress Monitor" model (baseline of the ALFRED proposal paper) (Shridhar et al., 2020). We often compare in two settings: "seen" and "unseen". "Seen" refers to when all the test tasks take place in rooms that have been previously seen in the training set. This chapter is split into a analysis of the action-sequences comparing baselines to the ground-truth sequences, an analysis of the predicted segmentation masks for interactive actions, an analysis of the errors leading to invalid actions, and an analysis of the performance of unimodal baselines.

We use the published results from the two baselines to discuss the unimodal ablations (we did not retrain the unimodal models). However, we provide our results for the MOCA full model baseline confirming they are in line with the published results (Table 1). We use this model to obtain the statistics discussed below. Setting up the environment (in particular AI2Thor) was not trivial, we used this assignment to successfully complete the setup.

Through our analysis we identify areas that could lead to performance improvements for the MOCA baseline.

2 Action-Sequence Analysis

Action Length In Table 2, we compare the average length of the predicted action sequences for the seen and unseen test cases. In this case, the "true" baseline refers to the human labelled action sequence. Furthermore, we analyze the MOCA baseline in greater depth by comparing how sequence length changes for successful and failed action sequences. The results are displayed in Table 3. The two main insights from both graphs are that the MOCA sequences contain a relatively large number of actions and that sequence length for failed tasks are much larger than successful tasks.

Distribution of Individual Actions Firstly, we provide an analysis into the distribution of individual actions depending on the baseline model. Table 4 displays this for the seen and unseen tasks as well as for the MOCA and ALFRED baselines. Interestingly, across both baselines there are more "MoveAhead" instructions in the unseen dataset than in the seen dataset. Additionally, for the MOCA baseline specifically we analyze the distribution of actions for successful and failed tasks as shown in table 5. For both seen and unseen tasks, there were far more "MoveAhead" actions predicted for failed tasks than for successful tasks. For the MOCA baseline, we plot the most frequent actions in figure 1.

Action Dependencies For the MOCA model, we analyze the effect that the previous action prediction has on the distribution of the next prediction in figure 3. To allow for comparison we also display the action-action dependencies for the human-labelled sequences in figure 4 and for the ALFRED baseline model in 2 . In every heatmap, the previous action is shown on the left axis and the predicted action is shown on the vertical axis. Interestingly, the heatmaps show the fact that both the MOCA and ALFRED baseline models struggle to correctly predict the "CloseObject" action after the "PickupObject" action.

3 Segmentation

Interaction tasks (non-navigational) use an output segmentation mask to specify the object to inter-

*Everyone Contributed Equally – Alphabetical order

		Seen		Unseen	
		Task (PW)	Goal-Cond (PW)	Task (PW)	Goal-Cond (PW)
	Seq2Seq + PM	3.7 (2.1)	10.0 (10.0)	0.0 (0.0)	6.9 (5.1)
	MOCA	19.2 (13.6)	28.5 (22.3)	3.8 (2.0)	13.4 (8.3)
	MOCA (**our trial**)	20.8 (14.4)	29.7 (22.6)	2.8 (1.4)	12.8 (7.6)
ALFRED	No Language	0.0 (0.0)	5.9 (3.4)	0.0 (0.0)	6.5 (4.7)
	No Vision	0.0 (0.0)	5.7 (4.7)	0.0 (0.0)	6.8 (6.0)
	Goal-Only	0.1 (0.0)	6.5 (4.3)	0.0 (0.0)	6.8 (5.0)
	Instruction-Only	2.3 (1.1)	9.4 (6.1)	0.0 (0.0)	7.0 (4.9)
MOCA	No Language	2.7 (1.3)	9.6 (6.3)	0.0 (0.0)	6.7 (3.0)
	No Vision	0.2 (0.1)	5.9 (4.7)	0.0 (0.0)	7.2 (6.1)
	Goal-Only	5.4 (2.7)	14.3 (10.0)	0.2 (0.1)	8.5 (4.8)
	Instruction-Only	2.3 (1.1)	12.8 (9.6)	0.5 (0.3)	7.5 (5.3)

Table 1: Results Table (validation dataset). The unimodal results are taken from the original papers. We tested the MOCA paper to verify the results were in line with the published ones and to perform additional evaluations.

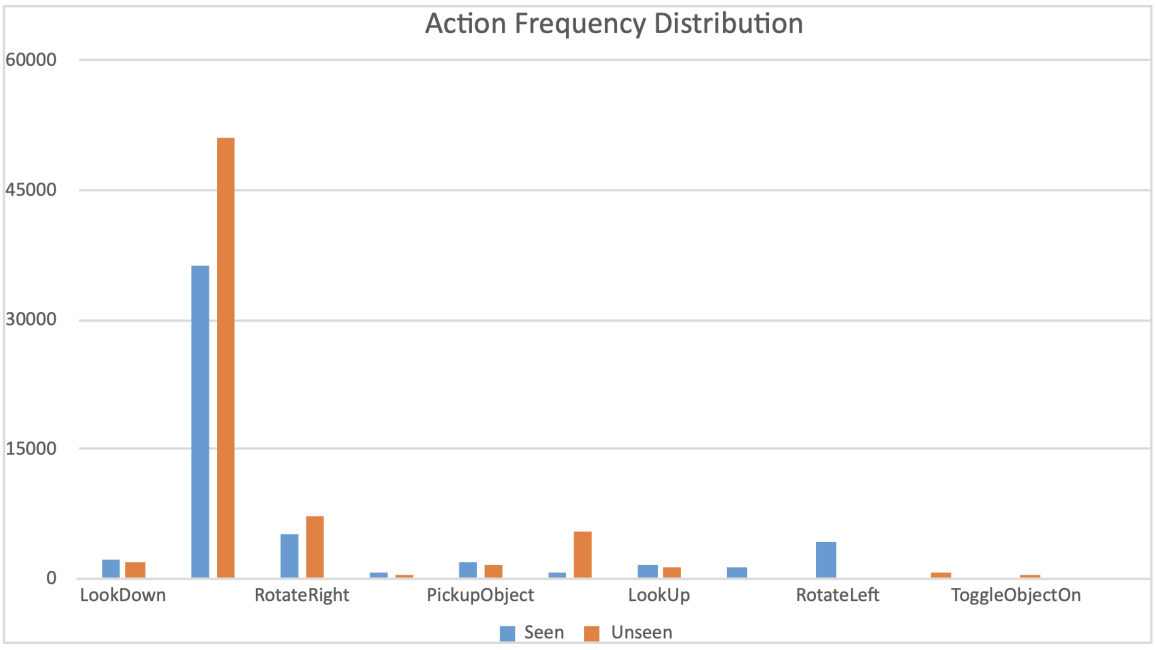

Figure 1: Action frequency distribution. Some bars on the x-axis are missing their label. The following are the labels from left to right: LookDown, MoveAhead, RotateRight, OpenObject, PickupObject, CloseObject, LookUp, PutObject, RotateLeft, SliceObject, ToggleOn, ToggleOff

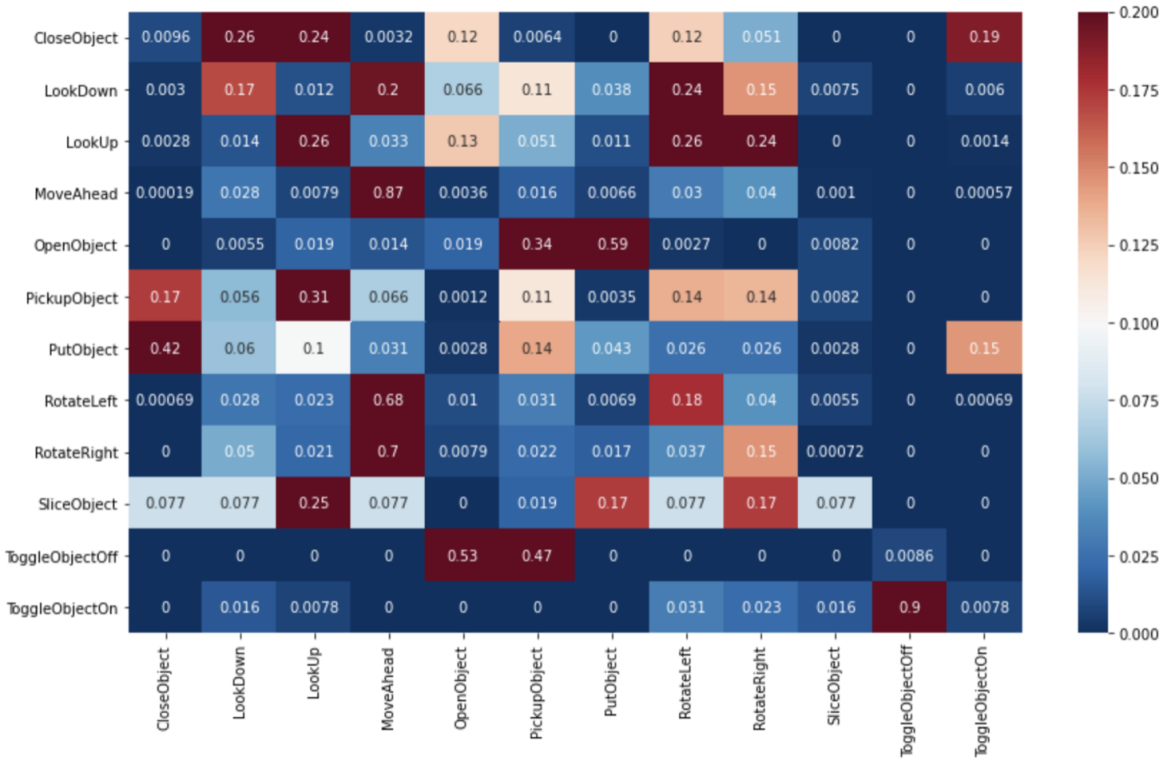

Figure 2: Action dependency heatmap for the ALFRED baseline

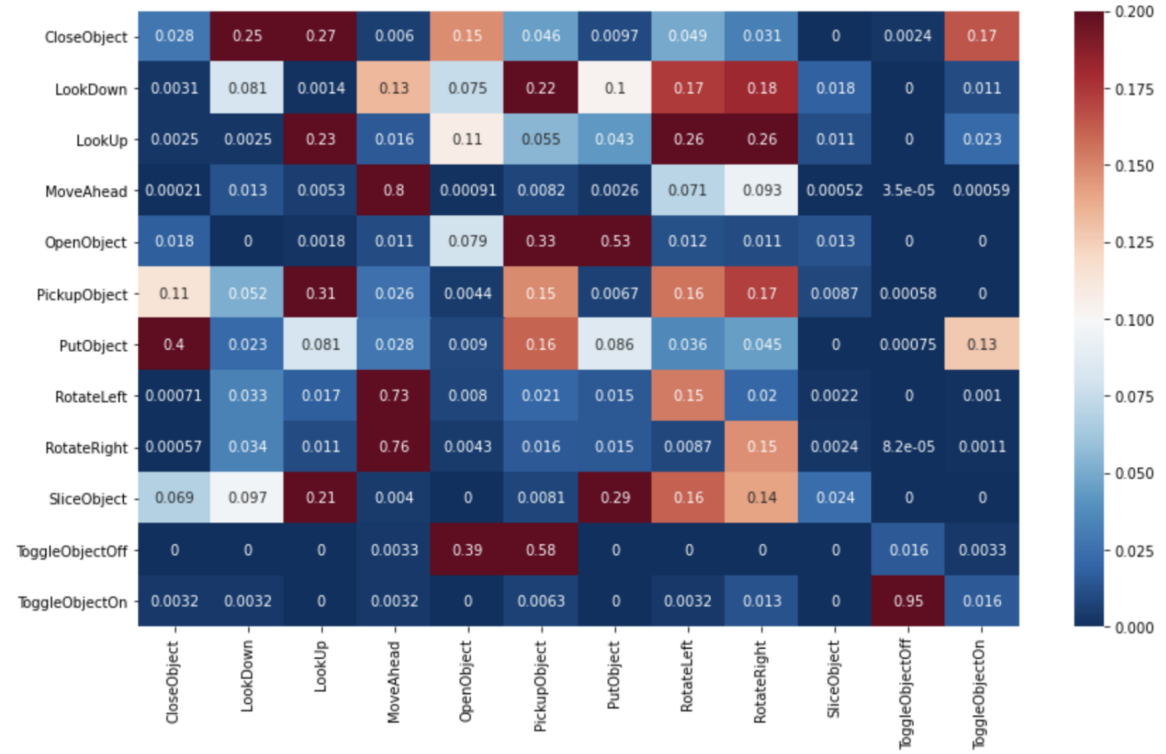

Figure 3: Action dependency heatmap for the MOCA model

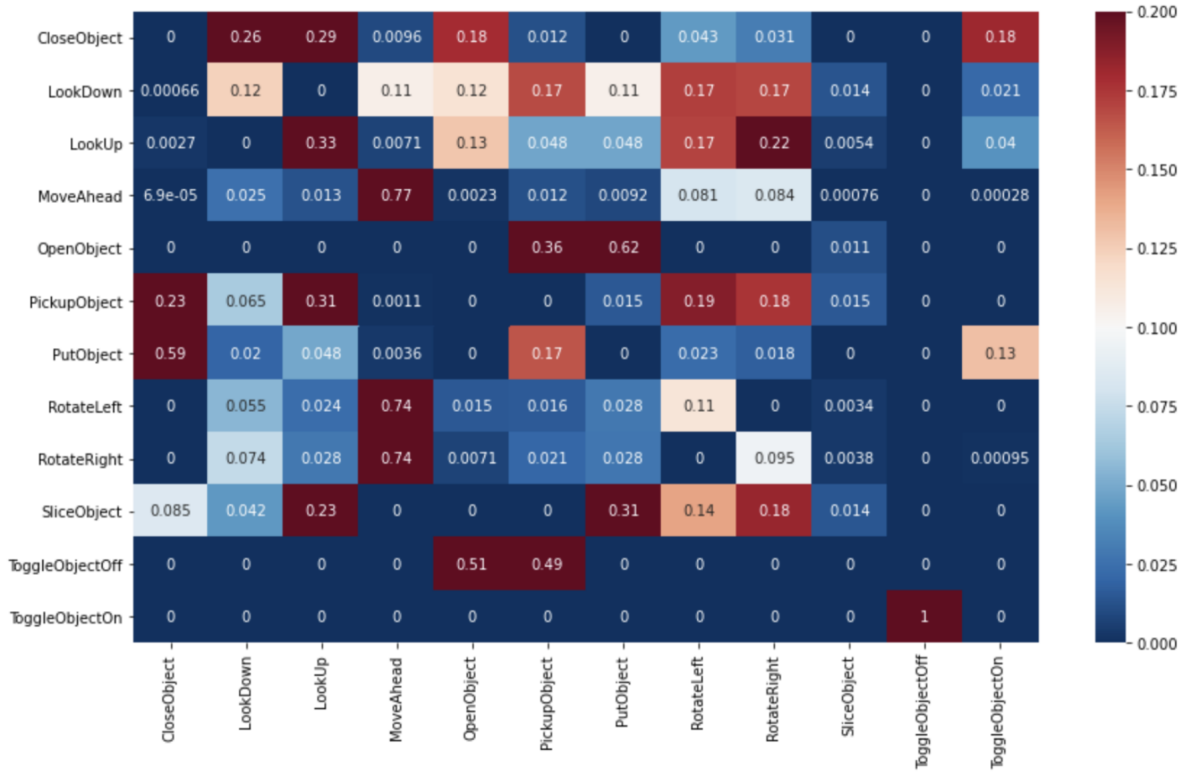

Figure 4: Action dependency heatmap for human-labelled sequences

	Seen	Unseen
ALFRED	48.55	46.22
MOCA	66.16	86.42
True	50.12	46.98

Table 2: Average action length

	Seen	Unseen
Success	49.45	31.87
Fail	70.56	87.99

Table 3: Action length in successful vs failed tasks (MOCA)

	MOCA		ALFRED		Ground	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
LookDown	0.04	0.03	0.06	0.06	0.06	0.57
MoveAhead	0.67	0.72	0.62	0.73	0.60	0.06
RotateRight	0.09	0.10	0.08	0.04	0.09	0.08
OpenObject	0.01	0.01	0.02	0.01	0.02	0.02
PickupObject	0.03	0.02	0.04	0.03	0.04	0.04
CloseObject	0.01	0.08	0.02	0.01	0.02	0.02
LookUp	0.03	0.02	0.04	0.03	0.04	0.05
PutObject	0.02	0.00	0.03	0.02	0.03	0.03
RotateLeft	0.08	0.00	0.07	0.05	0.08	0.08
SliceObject	0.00	0.01	0.00	0.00	0.00	0.00
ToggleOn	0.00	0.01	0.01	0.01	0.01	0.01
ToggleOff	0.00	0.00	0.01	0.00	0.01	0.01

Table 4: Distribution of individual actions

	Seen		Unseen	
	Fail	Success	Fail	Success
LookDown	0.04	0.05	0.03	0.08
MoveAhead	0.68	0.62	0.72	0.40
RotateRight	0.09	0.09	0.10	0.12
OpenObject	0.01	0.02	0.01	0.02
PickupObject	0.03	0.04	0.02	0.07
CloseObject	0.01	0.02	0.01	0.02
LookUp	0.03	0.03	0.02	0.04
PutObject	0.02	0.03	0.01	0.07
RotateLeft	0.08	0.08	0.08	0.14
SliceObject	0.00	0.00	0.00	0.00
ToggleOn	0.00	0.01	0.00	0.02
ToggleOff	0.00	0.01	0.00	0.02

Table 5: Distribution of actions for successful and failed tasks for the MOCA baseline

	Unseen	Seen
Avg IoU	0.474	0.602
Acc@0.0	0.467	0.684
Acc@0.1	0.396	0.651
Acc@0.2	0.355	0.616
Acc@0.3	0.318	0.582
Acc@0.4	0.288	0.541
Acc@0.5	0.247	0.486

Table 6: Average IoU and accuracy at 6 different IoU thresholds for the MOCA model.

act with. In this section, we analyze the performance of these mask predictions. Intersection over union (IoU) is a metric for the overlap between two regions. We calculate the IoU between the output masks predicted by the MOCA model and the ground truth masks. We also explore how the proportion of output segmentation masks with IoUs changes at various threshold of IoU. The results are displayed in table 6. In general, the IoU is significantly larger for seen task environments than for unseen task environments. This highlights the problem of generalization of the vision component of the model.

In addition to the IoU we observe in our error analysis table for the MOCA model (Table 7) that a large portion of invalid actions (about 40%) are due to the object being visible but incorrectly located in the image by the interaction mask. This indicates that improving the component of the model that generates the interaction masks could lead to significant improvements in the overall performance of the model.

4 Error Analysis

In the Alfred world, actions are not always valid. This might be due to a variety of reasons, for example, objects could be blocking the movement of the agent, or an object the agent wants to interact with might not be present in the image. In Table 7, we report statistics on the reason why certain actions are considered invalid for the MOCA model. Our first observation is that in the seen scenes the vast majority (80%) of invalid actions are from two categories: the agent is unable to move (blocked) and the object was visible but not located correctly. In the unseen scenes this trend in the errors is even more accentuated with 87% of the errors being from these two categories (47% for the agent being blocked, and 40% for the object not being located correctly). These results indicate possible areas for improvement.

Furthermore, in the second part of the Table 7, we observe that 28% and 40% of trajectories are interrupted because they reach the limit of 10 errors (for seen and unseen scenes respectively). This indicates that the presence of many invalid actions causes a large portion of failures. The two major areas for improvement in this direction are: 1. spatial detection of the validity of an action and 2. visual detection of objects.

We also manually observed such examples, and it appeared that in a significant portion of these examples the agent appeared to be stuck in between objects with repeated movement obstructions. This indicates that the agent is frequently unable to get unstuck.

5 Unimodal Analysis

Both the MOCA and ALFRED baselines are tested in 4 identical ablation studies which restrict the input to the model. We report the results of the ablation studies performed by authors in Table 1. "No Visual" refers to not having access to any visual input. The "No Language" setting removes all textual input to the model so that the model has to guess the task based on the visual input. "Goal only" refers to only having access to the visual input and the overall task objective without the step-by-step instructions. "Instruction Only" refers to having access to the visual input and the low-level step-by-step instruction without the overall task objective. We observe that for the ALFRED Seq2Seq baseline in the seen scenes there is benefit in having access to different modalities. However, in the unseen scenes, the performance on the single modalities is equal (in one case even greater) than the performance using all modalities. This indicates that

In the MOCA baseline we observe that especially in the unseen scenes all modalities are important to obtain high performance. This indicates that there is a certain degree of cross-modality reasoning that the model is able to perform. One of the key insights from the ablation studies is that the MOCA paper makes far superior predictions when textual input is limited (No language and goal-only). An explanation is that the MOCA model is able to memorize the relationship between objects and the actions that are frequently performed upon them.

	Seen		Unseen	
	Failure	Success	Failure	Success
Agent blocked	1514	139	2639	7
Agent state not allowed	127	2	79	0
Object visible but not located correctly	1569	28	2191	3
Object not found in scene	96	0	202	4
Object property not allowed	170	9	102	2
Others	163	0	124	0
Target state not allowed	105	0	46	0
No valid target	40	0	25	0
Object not visible	109	3	153	0
Object state not allowed	21	0	18	2
Frequency of error limit (10 errors) reached	0.28		0.40	
Number of Errors (when less then 10 errors)	4.44		4.91	
Average number of errors	6.03	1.05	6.99	0.78
Average length of sequences of errors	2.954	1.19	3.57	1.17

Table 7: Analysis of types of error messages returned by Thor for invalid actions. We compare failed and successful tasks as well as unseen and seen scenes. The first part of the table contains counts of errors of different types across the validation dataset. The second part computes statistics about the number of errors per trajectory.

6 Discussion and Conclusion

In this report we provide the results of the analysis of two baseline models for the ALFRED task, with a particular emphasis on MOCA. We believe that the following conclusion from our analysis could inform future work to improve the MOCA model:

1. Unseen scene generalization is an issue. The performance of the models drops significantly between seen and unseen scenes highlighting the problem of generalizing to new environments.

2. Trajectories reaching the invalid action limit of 10 correspond to a large portion of errors (30% for seen, 40% for unseen tasks).

3. The two major reasons for invalid actions are the agent being blocked, and the object being incorrectly located in the image even though it is visible.

4. The agent fails to have a good model of the validity of movements, especially in unseen scenes. This might be due to a mismatch in the training and testing environments. We hypothesize that during training the agent might have seen few cases of being stuck or having an obstruction.

5. The agent could improve its ability to recognize objects in an image.

References

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kunal Pratap Singh, Suvaansh Bhambri, Byeonghwi Kim, Roozbeh Mottaghi, and Jonghyun Choi. 2020. Moca: A modular object-centric approach for interactive instruction following. *arXiv preprint arXiv:2012.03208*.