Embodied Question Answering in Photorealistic Environments with Point Cloud Perception

Erik Wijmans[1*], Samyak Datta[1*], Oleksandr Maksymets[2], Abhishek Das[1], Georgia Gkioxari[2], Stefan Lee[1], Irfan Essa[1], Devi Parikh[2, 1], Dhruv Batra[2, 1]
[1]Georgia Institute of Technology, [2]Facebook AI Research
[1]{etw, samyak, abhshkdz, steflee, irfan}@gatech.edu
[2] {maksymets, gkioxari, dbatra, dparikh}@fb.com

Abstract

To help bridge the gap between internet vision-style problems and the goal of vision for embodied perception we instantiate a large-scale navigation task – Embodied Question Answering [2] in photo-realistic environments (Matterport 3D). We thoroughly study navigation policies that utilize 3D point clouds, RGB images, or their combination. Our analysis of these models reveals several key findings. We find that two seemingly naive navigation baselines, forward-only and random, are strong navigators and challenging to outperform, due to the specific choice of the evaluation setting presented by [2]. We find a novel loss-weighting scheme we call Inflection Weighting to be important when training recurrent models for navigation with behavior cloning and are able to out perform the baselines with this technique. We find that point clouds provide a richer signal than RGB images for learning obstacle avoidance, motivating the use (and continued study) of 3D deep learning models for embodied navigation.

1 Introduction

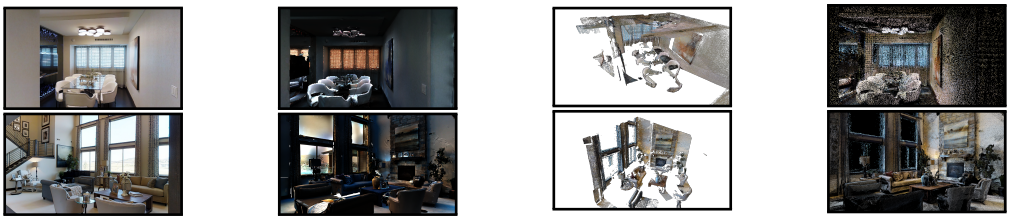
Imagine asking a home robot *'Hey - can you go check if my laptop is on my desk? And if so, bring it to me.'* In order to be successful, such a agent would need a range of artificial intelligence (AI) skills – visual perception, language understanding and navigation. Much of the recent success in these areas is due to large neural networks trained on massive human-annotated datasets collected from the web. However, this static paradigm of *'internet vision'* is poorly suited for training embodied agents. What are needed then are richly annotated, photo-realistic environments where agents may learn about the consequence of their actions on future perceptions while performing high-level goals.

To this end, a number of recent works have proposed goal-driven, perception-based tasks situated in simulated environments to develop such agents. While these tasks are set in semantically realistic environments most are based in synthetic environments and these problems are typically approached with 2D perception (RGB frames) despite the widespread use of depth-sensing cameras (RGB-D) on actual robotic platforms

Contributions. We address these points of disconnect by instantiating a large-scale, language-based navigation task in photorealistic environments and by developing end-to-end trainable models with point cloud perception – from raw 3D point clouds to goal-driven navigation policies.

Specifically, we generalize the recently proposed Embodied Question Answering (EmbodiedQA) [2] task (originally proposed in synthetic SUNCG scenes [5]) to the photorealistic 3D reconstructions from Matterport 3D (MP3D) [1].

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.



(a) RGB Panorama (b) Mesh Reconstruction (c) Point Cloud (d) RGB-D Render

Figure 1: Illustration of mesh construction errors and what point clouds are able to correct. Notice the wrapping of flat surfaces, the extreme differences in color, and texture artifacts from reflections.

We introduce the MP3D-EQA dataset, consisting of 1136 questions and answers grounded in 83 environments. We present a large-scale exhaustive evaluation of design decisions, training a total of 16 navigation models (2 architectures, 2 language variations, and 4 perception variations), 3 visual question answering models, and 2 perception models – ablating the effects of perception, memory, and goal-specification. We find that point clouds provide a richer signal than RGB images for learning obstacle avoidance, motivating continued study of utilizing depth information in embodied navigation tasks.

We find a novel weighting scheme we call *Inflection Weighting* – balancing the contributions to the cross-entropy loss between *inflections*, where the ground truth action differs from the previous one, and non-inflections – to be an effective technique when performing behavior cloning with a shortest path expert.

To the best of our knowledge, this is the first work to explore end-to-end-trainable 3D perception for goal-driven navigation in photo-realistic environments. With the use of point clouds and realistic indoor scenes, our work lays the groundwork for tighter connection between embodied vision and goal-driven navigation, provides a testbed for benchmarking 3D perception models, and hopefully brings embodied agents trained on simulation one step closer to real robots equipped with 2.5D RGB-D cameras.

2 Questions in Environments

In this work, we instantiate the Embodied Question Answering (EQA) [2] task in realistic environments from the Matterport3D dataset [1]. Our question generation pipeline follows [2]. We automatically generate templated questions. In total, we generate $\sim$1136 questions across 83 home environments (7 environments resulted in no questions after filtering). We use the same train/val/test split of environments as in MINOS [4]. We restrict agent start locations to lie on the same floor as question targets and limit episodes to single floors.

Learning Point Cloud Representations

Consider a point cloud $P \in \mathcal{P}$ which is an unordered set of points in 3D space with associated colors, *i.e.* $P = \{(x_m, y_m, z_m, R_m, G_m, B_m)\}_{m=1}^{M}$. To enable a neural agent to perceive the world using point clouds, we must learn a function $f : \mathcal{P} \to \mathbb{R}^d$ that maps a point cloud to an observation representation. To do this, we leverage the recently proposed PointNet++ [3] architecture.

PointNet++. At a high-level, PointNet++ alternates between spatial clustering and feature summarization – resulting in a hierarchy of increasingly coarse point clusters with associated feature representations summarizing their members. This approach draws a direct analogy to convolution and pooling layers in standard convolutional neural network architectures for RGB images.

Auxiliary Task Pretraining. To train the encoder architecture to extract semantically and spatially semantic representations of agent views, we introduce three pretraining tasks based on the annotations provided in Matterport3D.

RGB Image representations. We utilize ResNet50 trained using an analogous set of tasks (semantic segementaion, autoencoding, and depth from single images) to learn a representation for RGB images.

Navigator	Navigation																		QA		
	$\mathbf{d_0}$ (For reference)			$\mathbf{d_T}$ (Lower is better)			$\mathbf{d_{min}}$ (Lower is better)			$\mathbf{d_\Delta}$ (Higher is better)			$\%_{\mathbf{collision}}$ (Lower is better)			$\mathbf{IoU_T}$ (Higher is better)			$\mathbf{Top-1}$ (Higher is better)		
	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}
R/Fwd	0.354	1.898	3.547	0.933	1.330	2.154	**0.011**	0.346	1.397	−0.579	0.568	1.393	82.707	69.577	64.970	0.062	0.050	0.030	0.384	0.369	0.372
R+Q/Fwd	0.354	1.898	3.547	0.933	1.330	2.154	**0.011**	0.346	1.397	−0.579	0.568	1.393	82.707	69.577	64.970	0.062	0.050	0.030	0.384	0.369	0.372
R+RGB	0.354	1.898	3.547	1.194	1.617	2.340	0.040	0.375	1.349	−0.840	0.281	1.207	59.959	51.460	48.425	0.077	0.058	0.031	**0.395**	**0.396**	0.372
R+RGB+Q	0.354	1.898	3.547	1.407	1.740	2.521	0.034	0.340	1.332	−1.053	0.157	1.026	54.879	48.151	45.624	0.111	0.070	0.054	0.390	0.385	0.379
R+PC	0.354	1.898	3.547	1.428	1.754	2.352	0.021	**0.320**	1.164	−1.074	0.144	1.195	53.794	45.467	45.079	0.070	0.067	0.047	0.390	0.379	0.373
R+PC+Q	0.354	1.898	3.547	1.514	1.812	2.394	0.033	0.325	**1.160**	−1.160	0.085	1.153	50.479	40.307	42.048	0.059	0.052	0.043	0.380	0.376	0.378
R+PC+RGB	0.354	1.898	3.547	1.547	1.791	2.336	0.020	0.322	1.211	−1.193	0.107	1.211	48.747	38.792	39.968	0.084	0.077	0.044	0.385	0.386	0.373
R+PC+RGB+Q	0.354	1.898	3.547	1.539	1.843	2.420	0.032	0.323	1.170	−1.185	0.055	1.127	46.125	38.420	40.244	0.067	0.072	0.055	0.383	0.387	**0.387**
M	0.354	1.898	3.547	**0.366**	**0.830**	1.833	0.090	0.505	1.460	**−0.012**	**1.068**	1.714	11.665	15.410	26.463	0.128	0.091	0.081	0.384	0.376	0.378
M+Q	0.354	1.898	3.547	0.508	0.933	1.920	0.052	0.426	1.421	−0.154	0.965	1.627	24.516	25.856	36.614	0.147	0.109	0.068	0.368	0.368	0.369
M+RGB	0.354	1.898	3.547	0.637	1.157	2.177	0.099	0.538	1.479	−0.283	0.741	1.370	18.641	20.569	29.645	0.188	0.136	0.075	0.391	0.380	0.377
M+RGB+Q	0.354	1.898	3.547	0.707	1.171	2.194	0.071	0.423	1.386	−0.353	0.727	1.353	21.028	21.404	29.077	0.189	0.141	0.083	0.386	0.381	0.376
M+PC	0.354	1.898	3.547	0.494	1.020	1.817	0.098	0.484	1.236	−0.140	0.878	1.730	13.763	14.671	22.069	0.163	0.114	0.083	0.384	0.387	0.382
M+PC+Q	0.354	1.898	3.547	0.502	1.030	1.910	0.081	0.497	1.272	−0.148	0.868	1.637	12.135	**13.843**	**19.362**	0.184	0.158	**0.118**	0.374	0.395	0.368
M+PC+RGB	0.354	1.898	3.547	0.461	0.940	**1.791**	0.103	0.513	1.269	−0.107	0.958	**1.756**	**9.973**	14.348	22.249	**0.209**	**0.179**	0.111	0.379	0.387	0.374
M+PC+RGB+Q	0.354	1.898	3.547	0.574	1.044	1.898	0.083	0.431	1.203	−0.220	0.854	1.649	14.915	16.401	23.492	**0.209**	0.148	0.112	0.387	0.387	0.371
Random	0.354	1.898	3.547	0.912	1.275	2.652	0.048	0.797	2.262	−0.558	0.623	0.895	16.954	12.715	12.024	0.097	0.072	0.040	0.376	0.368	0.382
ShortestPath	0.354	1.898	3.547	0.005	0.005	0.005	0.005	0.005	0.005	0.349	1.893	3.542	0.000	0.000	0.000	0.581	0.581	0.581	0.394	0.394	0.394

Table 1: Evaluation of EmbodiedQA agents trained with inflection weighting on navigation and answering metrics for the MP3D-EQA v1 test set. RGB models perceive the world via RGB images and use ResNet50. PC models perceive the world via point clouds and use PointNet++. PC+RGB models use both perception modalities and their respective networks.

Imitation Learning from Expert Trajectories

Navigation models. All navigation models are trained with Behavior Cloning where they are made to mimic the ground truth , shortest path agent trajectories. For example, at a given time step, the Reactive models (Reactive and Reactive+Q) are shown the most recent *ground truth* point cloud and trained to predict the next ground truth action taken by the agent. Similarly, the sequence-based navigation models (LSTM, LSTM+Q) are trained to predict the ground-truth action at every time step by looking at the *ground-truth* point cloud at the current time step, (possibly) question embedding and *ground-truth* action embedding for the previous time step.

Inflection weighting. We add a novel weighting term to our loss called inflection weighting. When following the shortest path, there are a very few number of times the model needs to predict an action that differs from the previous action. We therefore find it critical to increase the importance of these actions and weight them by twice as much. More concretely, we uses a weighted average over the batch where w_t is 1.0 when $a_t = a_{t-1}$ and 2.0 when they differ. We define the first action as an inflection.

3 Experiments and Results

We closely follow the experimental protocol proposed by Das *et al*. [2] for their EmbodiedQA experiments on House3D. All results here are reported on novel test environments. Agents are evaluated by spawning 10, 30, or 50 primitive actions away from target (in the question), which corresponds to distances of approximately 0.35, 1.89, and 3.54 meters from target respectively, denoted by $\mathbf{d_0}$ in Tab. 1.

Our experiments found that **attention** answering module worked best (compared to **question-only** and **last-frame**) and thus, all results here are reported with **attention** answering module.

Question Answering. For measuring question answering performance, we report the top-1 accuracy, *i.e*. did the agent's predicted answer match the ground truth or not. We propose a new metric, $\mathbf{IoU_T}$ (higher is better), to evaluate the quality of the view of the target the agent obtains at the end of navigation. We compute the intersection-over-union (IoU) score between the ground-truth target segmentation and the same centered bounding box used to select views during dataset generation.

Navigation. For navigation, we report the distance to the target object from where the agent is spawned ($\mathbf{d_0}$) for reference, and measure distance to the target object upon navigation completion $\mathbf{d_T}$ (lower is better), and the change in distance from to the target object from initial to final position

$d_\Delta = \mathbf{d_T} - \mathbf{d_0}$ (higher is better). All the distances are geodesic, *i.e.* measured along the shortest path.

Table 1 shows all preliminary navigational and question-answering metrics for all approaches. Based on current state of experiments we noticed next observations:

Forward-only is a strong baseline. One of the side-effects of the evaluation procedure proposed in [2] is that the agent is commonly facing the correct direction when it is handed control. As a result, a forward-only navigator does quite well. Our vision-less reactive models with and without the question (`R/Fwd` and `R+Q/Fwd` respectively) learn to only predict forward.

Inflection weighted training is effective. We find inflection weighting to be crucial for training navigation models with behavior cloning of a shortest-path expert. Without it, few of our learned models beat forward-only and random baselines.

Memory helps. Models with memory are better navigators than their reactive counter parts. Surprisingly, a vision-less navigator with memory performs well at distance based navigation metrics.

Vision helps gaze direction metrics. The addition of vision leads to improvements on $\mathbf{IoU_T}$ and $\mathbf{QA}$, however, the improvements in $\mathbf{IoU_T}$ do not translate directly improvement on $\mathbf{QA}$. Models with vision also tend to collide with the environment less often.

Vision hurts distance metrics. Surprisingly, adding vision hurts distance based navigation metrics ($\mathbf{d_T}$). For reactive models, adding vision causes the models to collide significantly less frequently, resulting in a loss of the 'functional stop' that forward-only uses. For memory models, the story isn't as clear; however, memory models with vision stop less often.

Question somewhat helps. Interestingly, we do not see much of an improvement when providing models with the question. We suspect that because of limitations of behavior cloning.

`PC+RGB` **provides the best of both worlds.** The general tend is that point clouds provided a richer signal for obstacle avoidance (corresponding to lower $\%_{\mathbf{collision}}$ values), while RGB provides richer semantic information (corresponding to a higher $\mathbf{IoU_T}$ and $\mathbf{QA}$). Combining both point clouds and RGB provides improvements to both obstacle avoidance and leveraging semantic information.

4 Conclusion

We present an extension of the task of EmbodiedQA to photorealistic environments utilizing the Matterport 3D dataset and propose the MP3D-EQA v1 dataset. We then present a thorough study of 2 navigation baselines and 2 different navigation architectures with 8 different input variations. We provide analysis and insight into the factors that affect navigation performance and propose a novel weighting scheme – *Inflection Weighting* – that increases the effectiveness of behavior cloning. We demonstrate that two the navigation baselines, random and forward-only, are quite strong under the evaluation settings presented by [2]. Our work serves as a step towards bridging the gap between ***internet vision***-style problems and the goal of ***vision for embodied perception***.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[2] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *CVPR*, 2018.

[3] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017.

[4] Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017.

[5] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017.