Blindfold Baselines for Embodied QA

Ankesh Anand[1] **Eugene Belilovsky**[1] **Kyle Kastner**[1] **Hugo Larochelle**[2,1] **Aaron Courville**[1,3]
[1]Mila [2]Google Brain [3]CIFAR Fellow

Abstract

We explore blindfold (question-only) baselines for Embodied Question Answering. The EmbodiedQA task requires an agent to answer a question by intelligently navigating in a simulated environment, gathering necessary visual information only through first-person vision before finally answering. Consequently, a blindfold baseline which ignores the environment and visual information is a degenerate solution, yet we show through our experiments on the EQAv1 dataset that a simple question-only baseline achieves state-of-the-art results on the EmbodiedQA task in all cases except when the agent is spawned extremely close to the object.

1 Introduction

Recent breakthroughs in static, unimodal tasks such as image classification [16] and language processing [18] has prompted research towards multimodal tasks [1, 8] and virtual environments [4, 15, 25]. This is substantiated by embodiment theories in cognitive science that have argued for agent learning to be interactive and multimodal, mimicking key aspects of human learning [9, 17]. To foster and measure progress in such virtual environments, new tasks have been introduced, one of them being Embodied Question Answering (EmbodiedQA) [5].

The EmbodiedQA task requires an agent to intelligently navigate in a simulated household environment [25] and answer questions through egocentric vision. Concretely, an agent is spawned at a random location in an environment (a house or building) and asked a question (e.g. 'What color is the car?'). The agent perceives its environment through first-person egocentric vision and can perform a few atomic actions (move-forward, turn, strafe, etc.). The goal of the agent is to intelligently navigate the environment and gather visual information necessary for answering the question. Subsequent to the introduction of the task, several methods have been introduced to solve the EmbodiedQA task [5, 6], using some combination of reinforcement learning, behavior cloning and hierarchical control. Apart from using the question and images from the environment, these methods also rely on varying degrees of expert supervision such as shortest path demonstrations and subgoal policy sketches.

In this work, we evaluate simple question-only baselines that never see the environment and receive no form of expert supervision. We examine whether existing methods outperform baselines designed to solely capture dataset bias, in order to better understand the performance of these existing methods. To our surprise, blindfold baselines achieve state-of-the-art performance on the EmbodiedQA task, except in the case when the agent is spawned extremely close to the object. Even in the latter case, blindfold baselines perform surprisingly close to existing state-of-the-art methods. We note that this finding is reminiscent of several recent works in both Computer Vision and Natural Language Processing, where researchers have found that statistical irregularities in the dataset can enable degenerate methods to perform surprisingly well [11, 12, 14, 21].

Our findings suggest that current EmbodiedQA models are ineffective at leveraging the context from the environment, in fact this context or embodiment in the environment can negatively hamper them. We hope comparison with our baseline results can more effectively demonstrate how well a method

Preprint. Work in progress.

is able to leverage embodiment in the environment. Upon further error analysis of our models and qualitative inspection of the dataset, we find that there exist biases in the EQAv1 dataset that allow blindfold models to perform so well. We acknowledge the active effort of Das et al. [5] in removing some biases via entropy-pruning but note that further efforts might be necessary to fully correct these biases.

2 Related Work

EmbodiedQA Methods: Das et al. [5] introduced the PACMAN-RL+Q model which is bootstrapped with expert shortest-path demonstrations and later fine-tuned with REINFORCE [24]. This model consists of a hierarchical navigation module: a planner and a controller, and a question answering module that acts when the navigation module has given up control. In a later work, Das et al. [6] introduce Neural Modular Control (NMC) which is a hierarchical policy network that operates over expert sub-policy sketches. The master and sub-policies are initialized with Behavior Cloning (BC), and later fine-tuned with Asynchronous Advantage Actor-Critic (A3C) [19].

Dataset Biases and Trivial Baselines: Many recent studies in language and vision show how biases in a dataset allow models to perform well on a task without leveraging the meaning of the text or image in the underlying dataset. A simple CNN-BoW model was shown to achieve state-of-the-art results [12] on the Visual7W [26] task while also performing surprisingly well compared to the most complex systems proposed for the VQA dataset [1] and other joint vision and language tasks [2, 10]. Simple nearest neighbor approaches have been shown to perform well on image captioning datasets [7]. This phenomenon has also been observed in language processing tasks. On the Story-cloze task which was presented to evaluate common-sense reasoning, Schwartz et al. [23] achieved state-of-the-art performance by ignoring the narrative and training a linear classifier with features related to the writing style of the two potential endings, rather than their content. Similar observations were found on the Natural Language Inference (NLI) datasets, where methods ignoring the context and relying only on the hypothesis perform remarkably well [11, 21]. Most recently, question-only and passage-only baselines on several QA datasets highlighted similar issues [14].

3 Methods

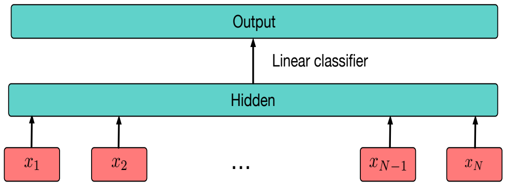

Figure 1: Model architecture for a sentence with N word vectors $x_1, \ldots, x_N$. The embeddings are averaged to form the hidden variable. Figure adapted from Joulin et al. [13].

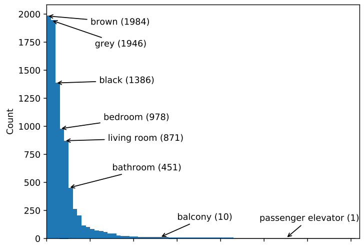

Figure 2: Frequency of each answer in the entire EQAv1 dataset. We observe that answers do not appear equally in the dataset, and are biased toward a select few.

Average BOW Embedding We use a simple linear classifier as described in [3, 13, 22], which takes word level embeddings and averages them to construct the question representation. We first perform a look-up over an embedding matrix for each word to get individual word representations. These word representations are then averaged into a text representation, which is in turn fed to a linear classifier. This architecture is similar to the `fastText` model of [13]. It is also a common and strong baseline in language and vision and language tasks [13, 22]. We use the softmax function f to compute the probability distribution over the predefined classes. The training criterion minimizes the negative log-likelihood over the classes.

Nearest Neighbor Answer Distribution (NN-AnswerDist) This method attempts to answer purely based on the per-question answer distribution of the training set. For an input question we find either the identical question in the training set or if one doesn't exist the nearest matching question (based on number of shared words). We then select the most likely answer for the training set. Performance on this baseline is directly indicative of the bias in answer distributions in the dataset. We note that for EQAv1 almost all questions in the validation and test sets are present in the training set.

4 Experiments

EQAv1 Dataset The EQAv1 dataset consists of 8985 questions split across 7190, 862 and 933 questions among training, validation and test sets respectively. The questions are generated via functional templates and are of following forms:

`location`: *'What room is the <OBJ> located in?'*

`color`: *'What color is the <OBJ>?'*

`color_room`: *'What color is the <OBJ> in the <ROOM>?'*

`preposition`: *'What is <on/above/below/next-to> the <OBJ> in the <ROOM>?'*

The answers span across 72 different categories of color, location and objects. We note that there are only 2 questions in the validation set, and 6 questions in the test set that are not in the training set. This limits the ability to test how well an agent generalizes across unseen combinations of rooms/objects/colors. To get rid of peaky answers, an entropy pruning method was applied by [5] where questions with normalized entropy below 0.5 were excluded. However this still leaves an uneven answer distribution that can be exploited.

Training Details[1] We evaluate the efficacy of our proposed baselines on the EQAv1 dataset. For the BoW model, we initialize the embeddings with Glove vectors [20] of size 100, which are allowed to be fine-tuned during the training procedure. We use the Adam optimizer (batch-size of 64) with a learning rate of $5e^{-3}$ which is annealed via a scheduling mechanism based on plateaus in the validation loss. The training procedure is run for 200 epochs and we use the checkpoint with minimum validation loss to compute accuracy on the test set. The NN-AnswerDist and the Majority baselines are self-descriptive and there are no specific training details that we apply. We also train the [5] text embedding model (an LSTM) with the optimization settings described in [5] for 200 epochs.

Results Detailed results are reported in Table 4. Following Das et al. [6], we report the agent's top-1 accuracy on the test set when spawned 10, 20 and 50 steps away from the goal, denoted as T_{10} , T_{20} and T_{50} respectively. Since the performance of blindfold baselines are not affected based on where the agent is spawned, their accuracy is same across T_{10} , T_{20} and T_{50}. We observe that the BoW model outperforms all existing methods except NMC(BC+A3C) in the case where agent is spawned very close to the target. The Nearest Neighbour method also does pretty well, and only falls behind to PACMAN (BC+REINFORCE) and NMC(BC+A3C) in the T_{10} case. The difference in performance b/w the Nearest Neighbour method and BoW is primarily due to the fact that the BoW method leverages validation metrics more effectively, uses distributed word representations and differs in optimization. We also observe that the majority baseline achieves an accuracy of only 17.15%, suggesting that the other question-only baselines leverage dataset biases separate from class modes. For completeness, we also include a question only baseline derived directly from the EmbodiedQA codebase, which uses only the Question LSTM in the PACMAN model, termed as PACMAN Q-only (LSTM). Note that we only compare the top-1 accuracy of different methods here, and not the navigation performance since it's not directly applicable to these blindfold baselines.

To better understand the exact bias exploited by the text only models we observe that (a) The questions from training set are largely repeated in the validation and test set, with only 2 and 6 questions being unique to them respectively. As noted earlier, this means that models don't need to generalize across unseen combinations of rooms/objects/colors to perform well on this task (b) Despite entropy-pruning, there is a noticeable bias in the answer distribution of EQAv1 questions (see [5, Appendix A]). Our results on the Nearest Neighbour baseline confirm this source of bias and explain largely the text model performance.

Viewing these results holistically, we conclude that current methods for the EmbodiedQA task are not effective at using context from the environment, and in fact this negatively hampers them. This

[1] Code for reproducing the experiments is available at https://github.com/ankeshanand/blindfold-baselines-eqa

shows that there is room for building new models that leverage the context and embodiment in the environment.

Oracles: We now examine whether the EQAv1 dataset and the proposed oracle navigation can improve over pure text baselines, to leverage visual information in the most ideal case. We reproduce the settings for training the VQA model[2]. Specifically we train the VQA model described in [6] on the last 5 frames of oracle navigation for 50 epochs with ADAM and a learning rate of $3e - 4$ using batch size 20. We observe the accuracy is improved over text baselines in this unrealistic setting, but the use of this model with navigation in PACMAN reduces performance to below the text baselines. For completeness we benchmark an oracle with our BoW embedding model in place of the LSTM with all other settings kept constant. As noted in [5], we re-iterate that these oracles are far from perfect, as they may not contain the best vantage or context to answer the question.

	T_{10}	T_{20}	T_{50}	T_{any}
Navigation + VQA				
PACMAN (BC) [5]	48.48	40.59	39.87	N/A
PACMAN (BC+REINFORCE)[5]	50.21	42.26	40.76	N/A
NMC (BC) [6]	43.14	41.96	38.74	N/A
NMC (BC+A3C) [6]	**53.58**	46.21	44.32	N/A
Question only				
Majority	17.15	17.15	17.15	17.15
Nearest Neighbor Answer	48.45	48.45	48.45	48.45
BOW	50.34	**50.34**	**50.34**	**50.34**
PACMAN Q-only (LSTM) (*)	46.07	46.07	46.07	46.07

Oracle VQA system	
PACMAN VQA-Only [5] (*)	55.9
BOW-CNN VQA-Only	56.5

Table 1: We compare to the published results from [6] for agent spawned at various steps away from the target: 10, 30, 50, and anywhere in the environment. Question-only baselines outperform Navigation+VQA methods except when spawned 10 steps from the target object. A VQA-only system with oracle navigation can improve on a pure text baseline but isn't effective when combined with navigation. (*) indicates our reproduction of the model described in [5]

Error Analysis: To better understand the shortcomings and limitations, we perform an error analysis of the one of the runs of the BoW model on different question types: Here, the color category subsumes `color` and `color_room` both. The particularly low accuracy on preposition questions is due to the fact that there exist very few questions of this type in the training set (2.44%), and the entropy of answer distribution in this class is much higher compared to color and location question types.

Preposition	Location	Color
9.09	51.72	53.31

Table 2: Accuracy of the BoW model on different question types

5 Conclusion

We show that simple question only baselines largely outperform or closely compete with existing methods on the EmbodiedQA task. Our results indicate existing models are not able to convincingly use sensory inputs from the environment to perform question answering, although they have been demonstrated some ability navigate toward the object of interest. Besides providing a benchmark score for future researchers working on this task, our results suggest considerations for future dataset and task construction in EQA and related tasks.

[2] We use the software provided by the authors https://github.com/facebookresearch/EmbodiedQA

Acknowledgements

We are grateful for the collaborative research environment provided by Mila. We also thank CIFAR for research funding and computing support. We further thank NVIDIA for donating a DGX-1 and Tesla K80 used in this work. Lastly, we thank Sandeep, Rithesh, Abhishek and Catalina for reviewing earlier drafts of this work.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] E. Belilovsky, M. Blaschko, J. R. Kiros, R. Urtasun, and R. Zemel. Joint embeddings of scene graphs and images. *ICLR Workshop*, 2017.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[4] S. Brodeur, E. Perez, A. Anand, F. Golemo, L. Celotti, F. Strub, J. Rouat, H. Larochelle, and A. Courville. Home: A household multimodal environment. *ICLR Workshop*, 2018.

[5] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 5, page 6.

[6] A. Das, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Neural modular control for embodied question answering. In *Conference on Robot Learning (CoRL)*, 2018.

[7] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[8] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426. ACM, 2015.

[9] K. Fisher, K. Hirsh-Pasek, and R. M. Golinkoff. Fostering mathematical thinking through playful learning. *Contemporary debates on child development and education*, pages 81–92, 2012.

[10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[11] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

[12] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

[13] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[14] D. Kaushik and Z. C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*, 2018.

[15] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[17] B. Landau, L. Smith, and S. Jones. Object perception and object naming in early development. *Trends in cognitive sciences*, 2(1):19–24, 1998.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[19] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[20] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[21] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*, 2018.

[22] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[23] R. Schwartz, M. Sap, I. Konstas, L. Zilles, Y. Choi, and N. A. Smith. Story cloze task: Uw nlp system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 52–55, 2017.

[24] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[25] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018.

[26] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.