Look and Answer the Question: On the Role of Vision in Embodied Question Answering

Nikolai Ilinykh and Yasmeen Emampoor and Simon Dobnik

Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Sweden
nikolai.ilinykh@gu.se, gusemampya@student.gu.se,
simon.dobnik@gu.se

INLG 2022

More research on embodied multi-modal agents is needed

GÖTEBORGS UNIVERSITET

- In recent years, the call for building systems that can act in real world became much more prominent in the NLP field.
  - This is partially reflected in the increasing number of papers on multi-modality (Vinyals et al., 2015), grounding (Clark and Brennan, 1991), and interaction (Hill et al., 2021).

More research on embodied multi-modal agents is needed

GÖTEBORGS UNIVERSITET

- In recent years, the call for building systems that can act in real world became much more prominent in the NLP field.
  - This is partially reflected in the increasing number of papers on multi-modality (Vinyals et al., 2015), grounding (Clark and Brennan, 1991), and interaction (Hill et al., 2021).
- Many of such models employed in real-world scenarios need to perceive the environment, understand physics of the objects and reason about the events in order to correctly describe them (Lake et al., 2017).
- However, many existing multi-modal tasks do not have "embodied" component: Visual Question Answering (Antol et al., 2015), Visual Dialogue (Das et al., 2017).

More research on embodied multi-modal agents is needed

GÖTEBORGS UNIVERSITET

- In recent years, the call for building systems that can act in real world became much more prominent in the NLP field.
  - This is partially reflected in the increasing number of papers on multi-modality (Vinyals et al., 2015), grounding (Clark and Brennan, 1991), and interaction (Hill et al., 2021).
- Many of such models employed in real-world scenarios need to perceive the environment, understand physics of the objects and reason about the events in order to correctly describe them (Lake et al., 2017).
- However, many existing multi-modal tasks do not have "embodied" component: Visual Question Answering (Antol et al., 2015), Visual Dialogue (Das et al., 2017).
  - At the same time, generation is a crucial component in all of these tasks, because these models are required to produce correct and plausible answer or reply.

More research on embodied multi-modal agents is needed

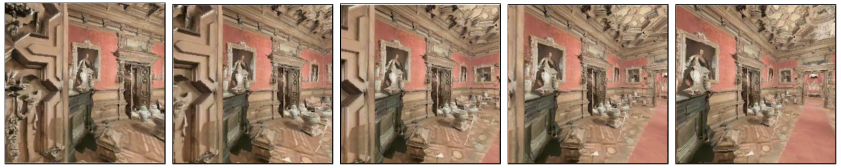
GÖTEBORGS UNIVERSITET

- In recent years, the call for building systems that can act in real world became much more prominent in the NLP field.
  - This is partially reflected in the increasing number of papers on multi-modality (Vinyals et al., 2015), grounding (Clark and Brennan, 1991), and interaction (Hill et al., 2021).
- Many of such models employed in real-world scenarios need to perceive the environment, understand physics of the objects and reason about the events in order to correctly describe them (Lake et al., 2017).
- However, many existing multi-modal tasks do not have "embodied" component: Visual Question Answering (Antol et al., 2015), Visual Dialogue (Das et al., 2017).
  - At the same time, generation is a crucial component in all of these tasks, because these models are required to produce correct and plausible answer or reply.

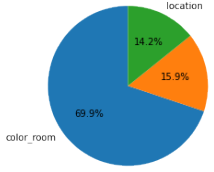
This paper focuses on the task of Embodied Question Answering (Das et al., 2018), which brings the aspects of **generation**, **real-world simulation** and **multi-modal interaction** together. We will specifically look at the **question answering** component of the task.

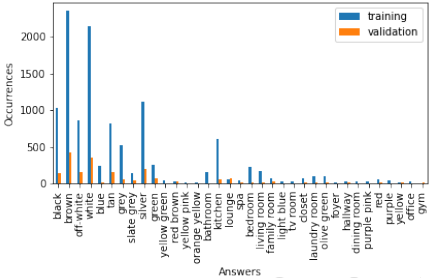
More about Question Answering in the EQA

The task: given the last five (5) images in the end of the navigation, answer the question: *What colour is the fireplace?*



- Dataset: Matterport3D-EQA (Wijmans et al., 2019)
  - Images: Matterport 3D scenes
  - 3 types of questions: Colour Room, Colour, Location



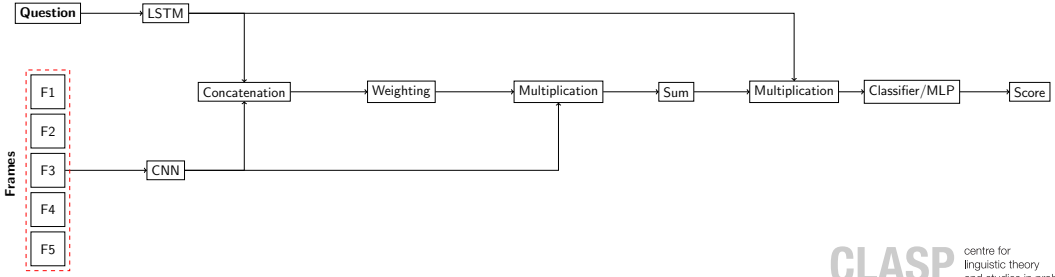


CLASP centre for linguistic theory and studies in probability

More about Question Answering in the EQA

The task: given the last five (5) images in the end of the navigation, answer the question: *What colour is the fireplace?*

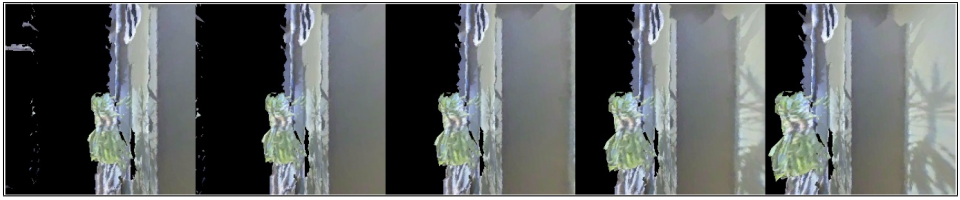
The model:



EQA is hard because

- The quality of rendered scenes is often poor.

Q: What colour is the plant in the kitchen?
Ground truth answer: Green.
Predicted answer: Olive green.



EQA is hard because

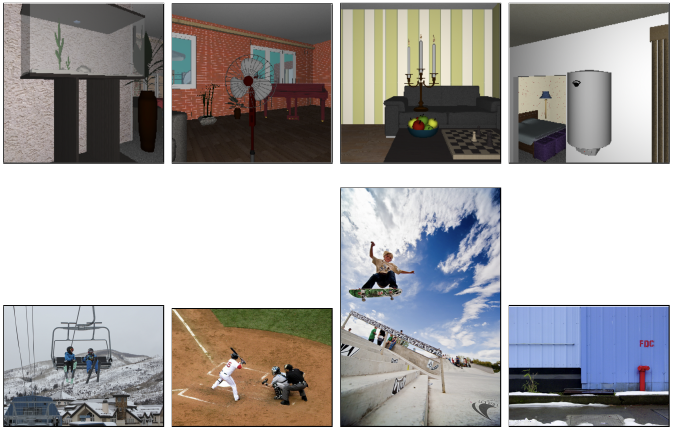
- The QA dataset is generated automatically.
  - Many answers are limited and fixed, e.g. there are only 24 colour answers.

KELLY'S 22 COLOURS OF MAXIMUM CONTRAST

Colour serial or selection number	Colour sample matched visual to ISCC-NBS centroid colour	General colour name	ISCC-NBS centroid number	ISCC-NBS colour name (abbreviation)	Munsell notation of ISCC-NBS centroid colour
1		white	263	white	2.5PB 9.5/0.2
2		black	267	black	N 0.8
3		yellow	82	v.Y	3.3Y 8.0/14.3
4		purple	218	s.P	6.5P 4.3/9.2
5		orange	48	v.O	4.1YR 6.5/15.0
6		light blue	180	v.l.B	2.7PB 7.9/6.0
7		red	11	v.R	5.0R 3.9/15.4
8		buff	90	gy.Y	4.4Y 7.2/3.8
9		grey	265	med.Gy	3.3GY 5.4/0.1
10		green	139	v.G	3.2G 4.9/11.1
11		purplish pink	247	s.pPk	5.6RP 6.8/9.0
12		blue	178	s.B	2.9PB 4.5/10.4
13		yellowish pink	26	s.yPk	8.4R 7.0/9.5
14		violet	207	s.V	0.2P 3.7/10.1
15		orange yellow	66	v.OY	8.6YR 7.3/15.2
16		purplish red	255	s.pR	7.3RP 4.4/11.4
17		greenish yellow	97	v.gY	9.1Y 8.2/12.0
18		reddish brown	40	s.rBr	0.3YR 3.1/9.9
19		yellow green	115	v.YG	5.4GY 6.8/11.2
20		yellowish brown	75	deep yBr	8.8YR 3.1/5.0
21		reddish orange	34	v.rO	9.8R 5.4/14.5
22		olive green	126	d.OlG	8.0GY 2.2/3.6

EQA is hard because

- All questions are about house environments which are more similar to each other in terms of structure and present objects than photographs used for VQA.



EQA requires fine-grained visual understanding

GÖTEBORGS UNIVERSITET

- House environments are visually similar and consist of many instances of the same object classes (e.g., sofas, plants) that often share the same attributes (e.g., sofas are brown, plants are green).

EQA requires fine-grained visual understanding

GÖTEBORGS UNIVERSITET

- House environments are visually similar and consist of many instances of the same object classes (e.g., sofas, plants) that often share the same attributes (e.g., sofas are brown, plants are green).

EQA requires fine-grained visual understanding

GÖTEBORGS UNIVERSITET

- House environments are visually similar and consist of many instances of the same object classes (e.g., sofas, plants) that often share the same attributes (e.g., sofas are brown, plants are green).

- However, previous research has shown that EQA models often struggle to learn from vision, learning mostly from language alone (Anand et al., 2018; Thomason et al., 2019).
- Given that the models do not exploit vision as much as they should, **we are going to examine what is it exactly that makes them learn so little from vision.**

Experiment I: is vision really needed?

GÖTEBORGS UNIVERSITET

We train three kinds of models to predict the most probable answer $a^\star$:

Given both question and images, predict the answer; Vis-L:

$$a^\star = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathrm{P}(a | \mathbf{Q}, \mathbf{I}_{T-N}, \ldots, \mathbf{I}_T). \quad (1)$$

Experiment I: is vision really needed?

GÖTEBORGS UNIVERSITET

We train three kinds of models to predict the most probable answer $a^\star$:

Given both question and images, predict the answer; Vis-L:

$$a^\star = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathrm{P}(a | \mathbf{Q}, \mathbf{I}_{T-N}, \ldots, \mathbf{I}_T). \tag{1}$$

Given question and black images, predict the answer; Blind-L:

$$\mathbf{I}_t = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \quad \mathbf{I}_t \in \mathbb{R}^{3 \times 256 \times 256}. \tag{2}$$

Experiment I: is vision really needed?

GÖTEBORGS UNIVERSITET

We train three kinds of models to predict the most probable answer $a^\star$:

Given both question and images, predict the answer; Vis-L:

$$a^\star = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathrm{P}(a|\mathbf{Q}, \mathbf{I}_{T-N}, \ldots, \mathbf{I}_T). \tag{1}$$

Given question and black images, predict the answer; Blind-L:

$$\mathbf{I}_t = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \quad \mathbf{I}_t \in \mathbb{R}^{3 \times 256 \times 256}. \tag{2}$$

Given question, predict the answer; Ø-L:

$$a^\star = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \, \mathrm{P}(a|\mathbf{Q}). \tag{3}$$

Experiment I: vision improves accuracy for a wrong reason

Metric	Vis-L	Blind-L	Ø-L
↓ Overall Mean Rank (MR)	4.352	4.454	3.685
MR, Color Room Questions	3.611	3.157	3.247
MR, Color Questions	2.693	2.261	2.304
MR, Location Questions	10.137	13.667	7.611
↑ Overall Accuracy (A)	**0.38**	0.323	0.362
A, Color Room Questions	**0.374**	0.348	0.337
A, Color Questions	**0.528**	0.478	0.522
A, Location Questions	0.222	0	0.278
Kappa Score	**-0.005**	0.014	0.024

Experiment I: vision is not used for every question type

GÖTEBORGS UNIVERSITET

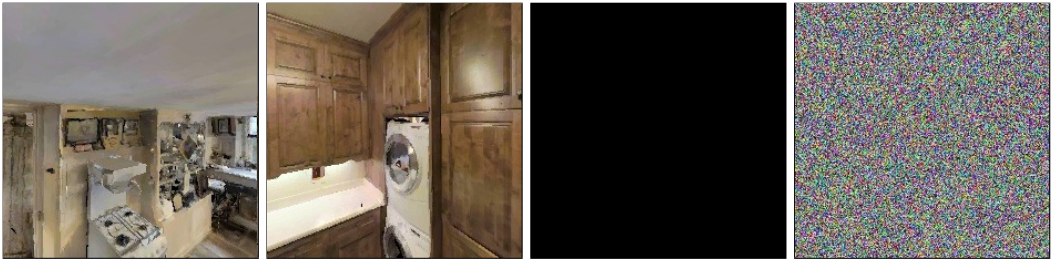
Metric	Vis-L	Blind-L	Ø-L
↓ Overall Mean Rank (MR)	4.352	4.454	3.685
MR, Color Room Questions	3.611	3.157	3.247
MR, Color Questions	2.693	**2.261**	2.304
MR, Location Questions	10.137	13.667	**7.611**
↑ Overall Accuracy (A)	0.38	0.323	0.362
A, Color Room Questions	0.374	0.348	0.337
A, Color Questions	**0.528**	0.478	0.522
A, Location Questions	0.222	0	**0.278**
Kappa Score	-0.005	0.014	0.024

Experiment I: model's approximations are better without vision

GÖTEBORGS UNIVERSITET

Metric	**Vis-L**	**Blind-L**	**Ø-L**
↓ Overall Mean Rank (MR)	4.352	4.454	**3.685**
MR, Color Room Questions	3.611	**3.157**	3.247
MR, Color Questions	2.693	**2.261**	2.304
MR, Location Questions	10.137	13.667	**7.611**
↑ Overall Accuracy (A)	0.38	0.323	0.362
A, Color Room Questions	0.374	0.348	0.337
A, Color Questions	0.528	0.478	0.522
A, Location Questions	0.222	0	0.278
Kappa Score	-0.005	0.014	0.024

Experiment II: "how much" vision is required?



- We take the original model (**Vis-L**) that sees both question and original images and evaluate it on different types of vision.
- From left to right (given the question "What colour is the stove in the kitchen?"):
  - **Original**, structure, content and context are present
  - **Eval-Shuffled**, structure and content are present, but context is incorrect
  - **Eval-Blind**, no content and context, but structure
  - **Eval-Random**, most disturbed representation

Experiment II: Vis-L confused by novel input

GÖTEBORGS UNIVERSITET

Metric	Vis-L	Eval-Shuffled	Eval-Blind	Eval-Random
↓ Overall Mean Rank (MR)	4.352	5.145	5.508	6.899
MR, Color Room Questions	3.611	4.157	4.562	5.512
MR, Color Questions	2.693	3.035	3.087	3.319
MR, Location Questions	10.137	12.722	13.278	18.33
↑ Overall Accuracy (A)	0.38	0.266	0.246	0.211
A, Color Room Questions	0.374	0.264	0.258	0.258
A, Color Questions	0.528	0.307	0.217	0.194
A, Location Questions	0.222	0.222	0.222	0
Kappa Score	-0.005	0.013	0.004	-0.005

What did we learn about question answering in EQA?

GÖTEBORGS UNIVERSITET

- The model can extract general patterns from images (but these patterns have to be present there). However, it cannot utilise images fully and acquire deeper understanding of what is in the image.
- We have identified multiple dataset and model biases through our experiments which suggest directions for future research:
  - Improve model's vision my implementing cognitive attention (Kruijff-Korbayová et al., 2015; Dobnik and Kelleher, 2016)
  - Split question answering into several subtasks
  - Using pre-trained multi-modal transformer such as LXMERT (Tan and Bansal, 2019) could tell us whether these models are able to overcome problems related to dataset construction and image selection for the task.

GÖTEBORGS UNIVERSITET

Thank you!

References I

Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron C. Courville. 2018. Blindfold baselines for embodied QA. *CoRR*, abs/1811.05013.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, New Brunswick, NJ. SEMDIAL.

References II

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2021. Grounded language learning fast and slow. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ivana Kruijff-Korbayová, Francis Colas, Koen Hindriks, Mark Neerincx, Petter Ögren, Mario Gianni, Tomáš Svoboda, and Rainer Worst. 2015. TRADR Project: Long-Term Human-Robot Teaming for Robot Assisted Disaster Response. *KI - Künstliche Intelligenz*, 29(2):193–201.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

References III

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the baseline: Single modality performance on visual navigation & QA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1977–1983, Minneapolis, Minnesota. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. 2019. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kappa

$$\kappa = (p_0 - p_e)/(1 - p_e) \tag{4}$$

where p_0 is the observed agreement, in this case accuracy, and p_e is the expected agreement, calculated by:

$$p_e = \sum_{k \in K} P(k|classifier) \cdot P(k|ground\_truth) \tag{5}$$