An interpretable representation of dialog history in referential visual dialog

Mauricio Mazuecos[1,2] and **Franco Luque**[1,2] and **Jorge Sánchez**[2]
Hernán Maina[1,2] and **Thomas Vadora**[1] and **Luciana Benotti**[1,2]
[1]Universidad Nacional de Córdoba
[2]CONICET, Argentina
{mmazuecos,hernan.maina,thvadora}@mi.unc.edu.ar
{francolq,jorge.sanchez,luciana.benotti}@unc.edu.ar

Abstract

Visual Dialog is assumed to require the dialog history to generate correct responses during a dialog. However, it is not clear from previous work how dialog history is needed for visual dialog. In this paper we define what it means for visual questions to require dialog history and we propose a methodology for identifying them. We release a subset of the Guesswhat?! questions for which their dialog history completely changes their responses. We propose a novel *interpretable* representation that visually grounds dialog history: the *Region under Discussion*. It constrains the image's spatial features according to a semantic representation of the history inspired by the information structure notion of *Question under Discussion*. We evaluate the architecture on task-specific multimodal models and the visual transformer model LXMERT and show that there is still room for improvement. Here we present published work (Mazuecos et al., 2021).

1 Introduction

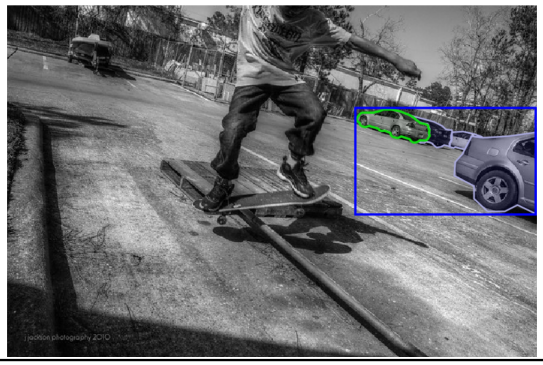
Visual Dialog (VD) is a task that combines natural language understanding grounded in vision with dialog. Being *visual*, VD is closely related to the area of Visual Question Answering (VQA). On VQA, important progress has been obtained recently with models that connect vision and language and are pre-trained on a variety of tasks (Tan and Bansal, 2019). Arguably, less progress has been made on the *dialog* part of VD, which is the topic of this paper. Currently, the two most popular datasets for visual dialog are VisDial (Das et al., 2017) (chit-chat) and GuessWhat?! (de Vries et al., 2017) (task-oriented).

Visual Dialog is assumed to require the dialog history to generate correct responses. However, it is not clear from previous work how dialog history is used for VD (Agarwal et al., 2020). In this paper we define history dependence in terms of a representation that is interpretable as a region of the visual common ground shared between dialog participants (Traum, 1994; Clark, 1996). This representation, which we call *Region under Discussion* (RuD). In this paper we define RuD and use it to connect a question to its visual dialog history; we make the following contributions:

- We define what it means for a visual question to require dialog history considering intrinsic and relative visual properties.
- We design a methodology for annotating a subset of the Guesswhat?! questions for which their dialog history is required because it completely changes their responses.
- We propose an interpretable representation of history based on the *Question under Discussion (QuD)* theory; we call our representation *Region under Discussion (RuD)*.
- We extend the Oracle model by (de Vries et al., 2017) and the LXMERT-based model of (Testoni et al., 2020) with our RuD.
- We find that RuD summarizes dialog history in an interpretable visual way which is linguistically well founded and improves responses for history dependent questions.

2 Region Under Discussion

Question under Discussion (QuD) (Ginzburg, 2012; De Kuthy et al., 2020) is an analytic tool that has become popular among linguists and language philosophers as a way to characterize how a sentence fits in its context (Velleman and Beaver, 2016). The idea is that each sentence in discourse is interpreted with respect to a QuD. The QuD is defined by the dialog or discourse history. The linguistic form and the interpretation of an utterance, in turn, may depend on the QuD that provides the constraints that define the utterance's context. Similarly, we define a *Region under Discussion* (RuD) for visual dialog as a representation of the



Question	Human response
1. It is a person?	no
2. It is a car?	yes
3. Is it in the back?	yes
4. Are there two together?	yes
5. *Is it on the left?*	yes

Figure 1: Human-human dialog from the Guesswhat?! dataset (de Vries et al., 2017). The example illustrates our definition of history dependent question. Question 5 can be correctly answered with *no* if asked at the beginning of the dialog, when the dialog history is empty because the target (marked in green) is not to the left of the picture. However, when the RuD (depicted in blue) is constrained by the initial turns then the correct answer to the same question is *yes*.

constraints that the dialog history establishes. The interpretation of a question depends on its RuD.

Figure 1 shows a dialog from the GuessWhat?! visual dialog dataset (de Vries et al., 2017). In the figure, the target is highlighted in green and the RuD marked in blue. Previous oracle models (de Vries et al., 2017) fail to answer question 5 because *question 5 is the only question for which the dialog history modifies the response*. Most questions in this dialog *can* be correctly answered independently of the dialog: they do not need the history. In effect, except for one turn, Figure 1 is just visual question answering. In this paper we model dialog history as constraints that represent the part of the image on which the dialog partners have agreed is the RuD. The rest of the questions are to be interpreted over this RuD. For our example, with respect to the blue box, the correct answer of *Is it on the left?* is yes since the car is on the left of the agreed RuD.

We model in the RuD the constraints that are related to intrinsic properties of the target that have been previously agreed upon between the dialog participants. An *intrinsic* property is one that is inherent and inseparable to the target and is not dependent on the visual context that the target is put in whereas *relative* properties are dependent on the context. In Figure 1 an intrinsic property is the fact that the target is a car. A relative property would be that "it is on the letf". We decided to represent in the RuD only intrinsic history motivated by literature from robot dialog, where intrinsic properties are plentiful and stable constraints (Tan et al., 2020).

3 Methodology

GuessWhat?! (GW) (de Vries et al., 2017) is a two player visual dialog game in which a *Questioner* tries to guess a target object in an image by asking yes/no questions and an *Oracle* answers them.

We annotated a subset of the GW test set to spot history dependent questions. We first sample a set of relational questions (questions that use another object in the image for reference) that follow a positively answered object question (questions that asks about the type of object the target is). We ask annotators to answer those questions with "yes", "maybe yes", "maybe no", "no" or "I don't know" using only the image as grounding. Then we compare the answers with the ones from the corpus for that particular question. If the answers do not coincide, we mark the question as history dependent. From the 1658 questions analyzed, two annotators agreed that 204 questions are history dependent. We call these 204 questions our **GWHist testset**.

3.1 Semantic history

To build the RuDs, we perform a parsing of the questions to find relations of the types *"is a"* and *"is the"* (and their negations) between a noun phrase (NP) and the target object using regular expressions for the most common syntactic patterns. Then we lemmatize and match the NPs with the 80 categories from the COCO dataset. We use WordNet's (Fellbaum, 1998) hypernym relations for *supercategories*, that is, nouns that cover several COCO categories (e.g. "food", "vehicle", etc).

The parsing and matching processes result in an ordered list of positive and negative relations to (super)categories found in the previous turns (e.g. [(*pos*, " vehicle"), (*neg*, "car")] which means that the target is a vehicle but it is not a car). We call this list a *semantic history*. We use the semantic history to filter the set of candidate objects.

We keep only the last positive history, assuming

Question	HR	CMO	+RuD
1. is it human?	no	no	no
2. is it food?	no	no	no
3. is it on the gas stove?	no	no	no
4. is it on the nearby counter top?	yes	yes	yes
5. is it red?	no	no	no
6. is the yellow spoon in the plate?	no	no	no
7. is a bottle?	yes	no	no
8. *the big one near the white plate?*	yes	no	yes
1. it is a sign?	no	no	no
2. it is a car?	yes	yes	yes
3. it is grey?	no	no	no
4. *it is brown?*	yes	no	yes
5. it is front the other car?	yes	no	no

Figure 2: The questions in italics are history-dependent. They illustrate how different kinds of questions may need to be interpreted respect to the RuD. CMO does not answer these questions correctly, but CMO+RuD does. The RuDs are in blue. The targets are in green. HR is the human response.

that it is the most specific one. For the negative history, our policy is to remove all the objects in the negated (super)categories from the candidates. If the history removed the target from the candidates, we force the inclusion of the target as an *ad-hoc* policy. Despite the simplicity of our approach, there is an important coverage of the questions, with more than 60% having semantic history.

We extend previous proposed oracle models: a Question+Category+Spatial (QCS) baseline (de Vries et al., 2017) and Cross-Modal Oracle (CMO) (Testoni et al., 2020) with a representation of the RuD as a constraint or shift of the spatial information respectively. We define the RuD as the smallest bounding box that encloses all the objects in the set of candidates. If no history is available we set the RuD to match the whole image.

4 Results

We implement both of our models as three-way classifiers using MLPs and a cross-entropy loss, accordingly with the relevant literature. For QCS we use the same model as de Vries et al. (2017). For CMO (Testoni et al., 2020) we use a simpler setup with just one layer on top of the cross-modality output of a pre-trained LXMERT model from the Transformers library (Wolf et al., 2019).

We report empirical results for the Oracle task of the GuessWhat?! benchmark (de Vries et al., 2017) and for the history dependent subset GWHist described in Section 3. We evaluate the RuD-augmented models and compare them with their respective RuD-less baselines. We found that RuD improves the performance in 1.5% and 1.1% in the GW testset and 41% and 45.9% in our GWHist testset respectively, as seen in Table 1.

Type	QCS	QCS+RuD	CMO	CMO+Rud
GW	0.733	0.744	0.809	0.813
GWHist	0.285	0.402	0.285	0.416

Table 1: Test response accuracy for the Oracle models discussed in Section 3 with and without Region under Discussion (RuD).

Not only questions about location are improved (like question 5 in Figure 1), but CMO+RuD shows improvements on questions about size and color, among others, as shown in Figure 2.

5 Conclusions

We proposed a novel *interpretable* representation for visual dialog history: Region Under Discussion (RuD). We release a challenging test set for history dependency. Similarly to the results from Agarwal et al. (2020), we found a low percentage of history dependency in GuessWhat?!. This may result in current dialog models not learning history dependence. Our experiments suggest that our implementation of RuD leads to improvements in performance of history dependent questions.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. ACL.

Herbert Clark. 1996. *Using Language*. Cambridge University Press, New York.

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

Kordula De Kuthy, Madeeswaran Kannan, Haemanth Santhi Ponnusamy, and Detmar Meurers. 2020. Towards automatically generating questions under discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5786–5798.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4475. IEEE Computer Society.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford Press.

Mauricio Mazuecos, Franco M. Luque, Jorge Sánchez, Hernán Maina, Thomas Vadora, and Luciana Benotti. 2021. Region under Discussion for visual dialog. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4745–4759, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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)*, Hong Kong, China. Association for Computational Linguistics.

Xiang Zhi Tan, Sean Andrist, Dan Bohus, and Eric Horvitz. 2020. Now, over here: Leveraging extended attentional capabilities in human-robot interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 468–470, New York, NY, USA. Association for Computing Machinery.

Alberto Testoni, Claudio Greco, Tobias Bianchi, Mauricio Mazuecos, Agata Marcante, Luciana Benotti, and Raffaella Bernardi. 2020. They are not all alike: Answering different spatial questions requires different grounding strategies. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 29–38, Online. Association for Computational Linguistics.

David Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Computer Science Dept., U. Rochester, USA. Supervised by James Allen.

Leah Velleman and David Beaver. 2016. Question-based models of information structure. In Caroline Féry and Shinichiro Ishihara, editors, *The Oxford Handbook of Information Structure*. Oxford University Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.