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
Table 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.

<table>
<thead>
<tr>
<th>Question</th>
<th>Human response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. It is a person?</td>
<td>no</td>
</tr>
<tr>
<td>2. It is a car?</td>
<td>yes</td>
</tr>
<tr>
<td>3. Is it in the back?</td>
<td>yes</td>
</tr>
<tr>
<td>4. Are there two together?</td>
<td>yes</td>
</tr>
<tr>
<td>5. Is it on the left?</td>
<td>yes</td>
</tr>
</tbody>
</table>

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.

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 left”. 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) hyponym 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
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.

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


