

Visually Grounded Follow-up Questions: a Dataset of Spatial Questions Which Require Dialogue History

Tianai Dong¹, Alberto Testoni², Luciana Benotti³, Raffaella Bernardi^{1,2}

¹ CIMEC, University of Trento, Italy ² DISI, University of Trento, Italy

³ Universidad Nacional de Córdoba, CONICET, Argentina

{tianai.dong|alberto.testoni|raffaella.bernardi}@unitn.it
luciana.benotti@unc.edu.ar

Abstract

In this paper, we define and evaluate a methodology for extracting history-dependent spatial questions from visual dialogues. We say that a question is history-dependent if it requires (parts of) its dialogue history to be interpreted. We argue that some kinds of visual questions define a context upon which a *follow-up spatial question* relies. We call the question that restricts the context: *trigger*, and we call the spatial question that requires the trigger question to be answered: *zoomer*. We automatically extract different trigger and zoomer pairs based on the visual property that the questions rely on (e.g. color, number). We manually annotate the automatically extracted trigger and zoomer pairs to verify which zoomers require their trigger. We implement a simple baseline architecture based on a SOTA multimodal encoder. Our results reveal that there is much room for improvement for answering history-dependent questions.

1 Introduction

The development of multimodal conversation agents is a long standing challenge (e.g. (Winograd, 1972)). In recent years, much has been achieved on the challenge of Visual Question Answering (VQA) (e.g. (Antol et al., 2015; Goyal et al., 2017).) The rapid advancements have brought researchers to further increase the difficulty of the task by proposing Visual Dialogue datasets (e.g. (Das et al., 2017; de Vries et al., 2017)) suitable to train multimodal dialogue systems. With this switch from VQA to Visual Dialogue, the challenge has increased in difficulty. First of all, while VQA involves only understanding the multimodal input (image and question), Visual Dialogues also require visual question generation and the acquisition of a dialogue strategy. Moreover, while VQA involves visual grounding of the question to be answered, Visual Dialogues require grounding the question against

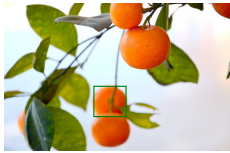
	Questioner	Oracle
	Q1. Is it a fruit?	Yes
	Q2. Is it in the foreground? No	
	Q3. Are there two of them on the branch? Yes	Yes
	Q4. Is it the top one? Yes	Yes

Figure 1: Through the dialogue the focus shifts from all the mandarins to just one. To answer Q4 (the *zoomer*), “top” needs to be interpreted relatively to the group of two mandarins identified by Q3 (the *trigger*).

both the visual and language contexts. As such this multi-folded challenge is rather ambitious. Our work focuses on identifying Follow-up Questions (FuQs) in Visual Dialogue. Namely, our goal is to construct a dataset of questions that we know require grounding both on the visual input and the dialogue history.

Work carried out on modeling the role of dialogue history in visual dialogue (Agarwal et al., 2020) has used the chit-chat dialogues of VisDial (Das et al., 2017) as a case study. However, it has been shown that in this dataset the role of grounding the question on the dialogue history is limited: models that take history into account do better, but the dataset contains a small percentage of questions that require dialogue history to be interpreted correctly. Based on these findings, Agarwal et al point out the need for data which captures dialogue history dependence. Our work is a contribution to this data collection challenge. We aim to identify FuQs which require (part of) the dialogue history to be interpreted.

Schlangen (2019) claims that goal-oriented settings will contain more dialogue phenomena. Following this claim, we run our analysis on Guess-What?! (de Vries et al., 2017), a multimodal dataset in which the goal of the dialogues is to identify a referent in an image.

In referential dialogues, the questions aim to collect information so as to narrow the set of potential candidates and univocally identify the referent among them. Interestingly, in referential multimodal game, this progressive refinement happens both through the language and visual contexts by incrementally zooming the joint attention to the conjectured referent. For instance in Figure 1, Q1 focuses on the full image and all objects are potential candidates. As the dialogue proceeds the attention is moved on the mandarins (Q1), then on those two mandarins in background (Q2) and finally on the mandarin on the top of the group of mandarins in the lower branch (Q3 and Q4). In this view, the data collection challenge launched in (Agarwal et al., 2020) can be rephrased by looking for a method to extract FuQs which require to zoom on a specific region of the image by narrowing the set of entities on which the dialogue focuses. We claim that FuQs requiring multimodal grounding can be extracted by identifying patterns of `trigger-zoomer` questions.

By manual inspection of the human dialogues, we have observed that often, after a positively answered question, the questioner tries to narrow down the choice by asking further details that discriminate the candidate. This happens in particular, when first a question (the `trigger`) identifies a group of objects that share some property and then the FuQ (the `zoomer`) focuses on one or more of the members of the identified referential set. In most cases, the `zoomer` requires the dialogue history to be answered. For instance, in Figure 1, the positively answered Q3 acts as a trigger which identifies the group of oranges under discussion and Q4 zooms on one of those. Notice that the question Q4 would be answered incorrectly if answered without considering Q3 and Q2 because the referent is not at the top of the picture. In this paper, we investigate the role of spatial questions in the identification of such patterns and focus on the evaluation of the Oracle player of the Guess-What?! game. We show that the method we propose facilitates data collection of follow-up questions that need to be grounded on the visual and dialogue context to be answered correctly or at least with higher confidence. The dataset is publicly available at <https://github.com/tianaidong/2021SpLU-RoboNLP-VISPA> for future model developments and evaluations.

2 Related Work

Clark (1996) defines dialog *common ground* to be the commitments that the dialog partners have agreed upon during the dialog. An important part of the common ground is the *Question under Discussion (QuD)* (Ginzburg, 2012; De Kuthy et al., 2020). QuD 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. We reinterpret this theory to analyse referential visual dialogue: we take the QuD to be the objects conjectured to be the target. The interpretation of a question depends on its QuD.

Most of the work on the GuessWhat?! game has focused attention on the Questioner player; as a consequence, the issue of dialogue history needed by the Oracle has never been considered. Since the first baseline model (de Vries et al., 2017), the Oracle receives just the question without the previous turns. Furthermore, this baseline model is blind: it takes the question, the target’s category and its location as inputs. This simple model has been widely used as the *Oracle* agent by all work on the *Questioner* (eg. (Strub et al., 2017; Shekhar et al., 2019; Pang and Wang, 2020).) Testoni et al. (2020) compared the LSTM baseline with a visually grounded LSTM (V-LSTM) and with an adaptation of LXMERT (Tan and Bansal, 2019). They show LXMERT based Oracle improves over the baseline achieving a new SOTA for the Guess-What?! Oracle. Yet the model does not use dialog history as an input. We evaluate LSTM, V-LSTM and LXMERT against our dataset of context-dependent questions.

Agarwal et al. (2020) argues that although complex models that encode history for visual dialogs have been proposed (Yang et al., 2019), such work has not demonstrated that history matters for visual dialogs. Agarwal et al. propose and apply a new methodology for evaluating history dependence of questions in visual dialog. They show crowdsourcers a question with its image without the dialog history and ask the crowdsourcer “would you be able to answer this question by looking at

the image only or you need more information from the previous conversation?”. However, it could happen that workers could be confident in answering the question just by looking at the image, but that they would give a different answer if the dialogue history is provided. This difference is crucial for studying context-dependent questions. In this paper, we proposed a new methodology for detecting history-dependent visual questions.

3 Dataset

We aim to identify Follow-up Questions (FuQs) that need the previous turn to be answered correctly or at least with higher confidence. We claim that FuQs which zoom in a specific region of the image to identify an object (or a set of objects) in it satisfy this request. This might hold in particular when the region contains more objects of the same category (e.g., more instances of mandarins, as in Figure 1) and the question refers to one (or more) member(s) of such a group. Moreover, we conjecture that most of such questions might also need to be visually grounded since the answer to it could change if the specific visual region they refer to is not properly identified and the question is mistakenly grounded over the full image. These challenging questions that zoom into a group are usually triggered by a question that refers to the whole group, the latter is identified by its location, the number or the color of its members. For instance, in Figure 1 the question that zooms on the target object of the game, “Is the top one?”, is triggered by the previous question that identifies the group itself by referring to the number of its members “Are there two of them on the branch?”. Interestingly, the zoomer question would be answered incorrectly without the previous turn since the target is at the top of the zoomed region and not on the top of the full image.

We focus on games in the test set in which there are more candidates of the same category of the target; we obtain 13,024 unique games containing 57,241 questions. We refer to these questions as the full test set. Shekhar et al. (2019) has classified GuessWhat?! questions into entity and attribute questions, the latter are subdivided into spatial, color, action, size, texture and shape. Testoni et al. (2020) further divided the spatial questions into group, absolute and relational questions. We build on these classifications to extract trigger and zoomer pairs. We see group and color questions as potential triggers for collecting history-dependent

questions: for instance, group questions that contain explicit numbers indicate groups (e.g., “One of the three oranges?” refers to a group containing 3 members) and color questions might identify a group of objects which differ with respect to the color (e.g., “Is it blue?” may refer to a group of objects one of which is blue). For the zoomer questions, we consider group and absolute questions. Absolute questions are those spatial questions that contain an absolute location adjective (e.g., “Is it in the middle?” contains “middle”). Other types of questions, such as size (“Is it one of the big bottles?” which contains “big”) and shape (“Is it kind of round?” which contains “round”) could be used as triggers and zoomers as well. In this paper we do not use them because they are not frequent in the Guesswhat?! dataset and a preliminary analysis showed we would not extract sufficient trigger zoomer pairs through them.

Using the automatic annotation of Testoni et al. (2020), which is based on keyword matching, we extract group and absolute questions, 4342 and 11,743, respectively. Moreover, we extract the dialogues containing context-dependent group questions using the following patterns: a positively answered group question followed by another group question (Group-Group) and a positively answered color question followed by a group question (Color-Group); we obtained 364 and 145 pairs, respectively; and similarly for absolute questions obtaining 919 context-dependent absolute questions (530 from the Group-Absolute and 389 from the Color-Absolute patterns).

We randomly retrieve 200 samples for each subset¹ and manually checked them. We filtered out those pairs in which the zoomer question could be correctly interpreted without the dialogue history. We also removed samples that were noisy (the image was blurry or the target was too small, the question was not clear, etc). Each datapoint was annotated by two annotators (the four authors), and we maintained only those on which there was an agreement between the two annotators. After this filtering, we obtained in total 271 context-dependent questions manually checked: 164 group questions (103 group-group and 61 color-group) and 107 absolute questions, the latter are all from the group-absolute pattern.²

¹For the Color-Group we took all the 145 datapoints.

²We are not considering questions extracted by color-absolute pattern in our evaluation, because the manual inspection of 200 samples randomly chosen from the automatically

We will refer to the set of visually grounded spatial questions that are context-dependent as VISPA. To gain a better understanding of the linguistic features of our dataset, we collect the statistics of question length, nouns and function words (*prepositions, pronouns, determiners, conjunctions, auxiliaries*) for questions in each subset.³ As we can see in Table 1 and Table 2, the context-dependent group and absolute questions do not show distinguishing surface features from the questions of the same type. Therefore they would have not be captured by using surface heuristics, such as searching for pronouns.

3.1 Examples

Figure 2 and Figure 3 report examples of context-dependent questions we have identified through our automatic process and further manual filtering. As we can see, when the previous turn is given, we can be much more confident in providing a correct answer. The previous turn is the question we have used to trigger the context-dependent FuQ, in one case its a group question (“is it between the two players in black?” “Yes”) and in the second case it is a color question (“one of the two gray ones?” “Yes”). The example on the upper part (group-group) is particularly interesting since the FuQ further specifies the previous turn, hence it should be properly integrated with it and interpreted as saying “Is it between the two players in black closest to the bat?”. Only models that truly ground questions within the previous linguistic context can properly answer it. The latter example requires the Oracle to understand the group of objects the question refers to, the previous turn identifies this group through the color of its members.

Figure 3 provides an example of absolute questions in our manually filtered subset; the zoomer question would be answered negatively if the previous turn is not given, since “middle” would refer to the middle of the image. When the previous turn is given, “middle” should be instead interpreted as the middle of the 3 planes in front. This FuQ should be grounded on the linguistic and visual context to be properly answered.

4 Models

LSTM The first model we consider is the language-only baseline model proposed in

extracted color-absolute questions provided too few cases of history-dependent questions.

³We utilize NLTK Python Package for the analysis

(de Vries et al., 2017). This Oracle model receives as input the embeddings of the target object’s category, its spatial coordinates, and the question to be answered encoded via an LSTM network. These three embeddings are concatenated and fed to a Multi-Layer Perceptron that gives the answer (Yes, No, or N/A).

V-LSTM We also consider a multimodal Oracle model. V-LSTM (Testoni et al., 2020) receives as input the embeddings of the target object’s crop features, its spatial coordinates, the features of the image, and the question to be answered encoded via an LSTM network. All these embeddings are concatenated as in LSTM. The visual features are extracted with the frozen ResNet-152 network pre-trained on ImageNet (Russakovsky et al., 2015). Differently from LSTM, this model does not have access to the target object category.

LXMERT We additionally considered the Oracle model proposed in Testoni et al. (2020). This model is based on LXMERT (Learning Cross-Modality Encoder Representations from Transformers)(Tan and Bansal, 2019), a powerful multimodal transformer-based model. LXMERT represents an image by the set of position-aware object embeddings for the 36 most salient regions detected by a Faster R-CNN, and it processes the text input by position-aware randomly-initialized word embeddings. Both the visual and linguistic representations are processed by a specialized transformer encoder based on self-attention layers; their outputs are then processed by a cross-modality encoder that through a cross-attention mechanism generates representations of the single modality (language and visual output) enhanced with the other modality as well as their joint representation (cross-modality output). LXMERT uses the special tokens CLS and SEP. Testoni et al. (2020) fine-tuned the pre-trained version of LXMERT on the GuessWhat?! Oracle task by feeding the visual features and the spatial coordinates of the target object as the last region in the visual input. They took the representation corresponding to the special token CLS and fed it to a Multi-Layer Perceptron to obtain the answer to the input question. The authors show that this model outperforms the baseline model to a large extent.

	Nr	Length	Nouns	Function W	Pronoun
All questions	57,241	4.89	1.23	3.08	0.75
Group Q	4342	7.27	1.51	4.31	0.61
Absolute Q	11,743	5.79	1.52	3.57	0.64
CD group Q	509	7.28	1.45	4.28	0.62
Group-Group	364	7.03	1.36	4.16	0.58
Color-Group	145	7.92	1.66	4.59	0.71
CD absolute Q	919	5.95	1.46	3.75	0.65
Group-Absolute	530	5.84	1.43	3.73	0.63
Color-Absolute	389	6.10	1.50	3.77	0.68

Table 1: Automatically extracted datapoints: Length: average question length; Nouns: average number of nouns per question; Function W: average number of function words per question; Pronouns: average number of pronouns per question

	Nr	Length	Nouns	Function W	Pronoun
CD Group Q					
Group-Group	103	6.89	1.26	4.10	0.66
Color-Group	61	7.50	1.49	4.55	0.72
CD Absolute Q					
Group-Absolute	107	5.59	1.31	3.76	0.57

Table 2: Manually filtered questions: Length: average question length; Nouns: average number of nouns per question; Function words: average number of function words per question; Pronouns: average number of pronouns per question

5 Experiments

We evaluated the models described above when receiving just the question or the question and the previous QA turn, we refer to the latter setting by marking the model names by -DH. We run each model three times (seed: 1, 50 and 100) and report their average together with the significance test results about the difference across runs. Table 3 and Table 4 report the model task accuracy on automatically extracted sets and manually filtered sets, respectively.

We claim the FuQs identified through the trigger-zoomer patterns need (at least) the previous turn to be answered properly or at least with higher confidence, this need should be even stronger for the manually filtered subsets.

As expected, LXMERT is the model that reaches the highest accuracy of the full test set (Table 3). Our results confirm what had been noticed by Testoni et al. (2020), namely that spatial questions are harder than average, and group questions are harder than absolute questions. This is reflected both by the baseline and the SOTA model: LSTM drops from 77.31 (All) to 70.45 (Absolute) to 67.11 (Group) and similarly does LXMERT – from 82.40 to 79.42 to 74.48. Even the accuracy of the best

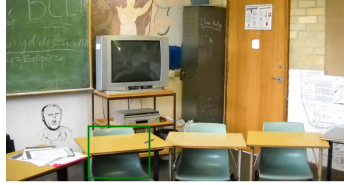
performing model, LXMERT-DH, further drops on the context-dependent questions reaching 74.43 and 71.12 for absolute and group questions, respectively.

When looking at the context-dependent questions, the standard-deviation among the accuracies reached by the three runs is rather high, hence in order to understand its effect on the comparison between models when receiving just the question and the question together with the previous turn, we have run a statistical significance paired t-test (following the suggestions in Dror et al. (2018).) The result shows that the difference between the two settings is never significant, except for LSTM/LSTM-DH on the absolute questions (p-value < 0.05). This shows that model performance is rather unstable and hence the selection of the binary answer is not properly grounded. This instability is not due to the size of the set: we have computed accuracy of the three runs of LXMERT on subsets of 500 and 100 randomly chosen questions and obtained a very low standard deviation.

Since the questions we have accurately selected are context-dependent, ideally a model should increase confidence in its answers when receiving the context (we simplify by giving just the the previ-



Questioner **Oracle**
 Q10. Is it between the two players in black? Yes
 Q11. The two players closest to the bat? Yes



Questioner **Oracle**
 Q4. one of the three gray ones? Yes
 Q5. first one counting from left to right? Yes

Figure 2: Context-dependent group questions: group-group (up) color-group (down)



Questioner **Oracle**
 Q2. One of the 3 planes in front? Yes
 Q3. is the one in the middle? Yes

Figure 3: Context-dependent absolute questions: group-absolute

ous turn). To verify this hypothesis, we computed the confidence of LXMERT/LXMERT-DH (see Table 5) by using the average probability assigned to the answers in the manually filtered set. In our results, we find rather the opposite of our assumption: while both LXMERT and LXMERT-DH show relatively high confidence (>0.80) in providing correct answers, LXMERT-DH’s confidences do not increase on LXMERT with the addition of the previous turn. On the positive side, we observe that for those cases where the model failed to provide the correct answer, LXMERT-DH is usually more uncertain than LXMERT about its own predictions. We consider this as a positive behavior of the model, since it suggests it is “aware” of what it does not know.

5.1 LXMERT attention

To understand the possible reasons that prevent the model from learning to exploit the dialogue history, we have analysed how LXMERT-DH puts attention to different parts of the input sequence through the computation of the cross-attention layers from language to vision (Figure 4). Ideally, context-dependent questions would require the model to put more attention to the trigger questions compared to questions that could be answered without the context. Model’s attention on the previous turn in the manually selected subset should therefore

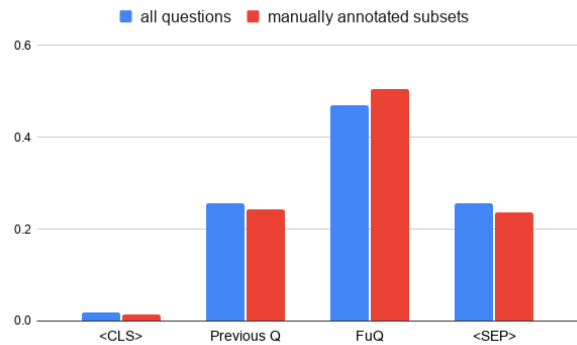


Figure 4: LXMERT-DH attention: all questions vs. context-dependent FuQs and the previous turn.

be higher than in the full test set, if the model takes advantage of it as it should. However, this does not happen with LXMERT-DH: its attention on the previous turn does not change, and it actually slightly increases its attention on the zoomer question instead. This result confirms our claim that the current Oracle architecture fails in exploiting the trigger question while answering context-dependent questions and suggests that the model should be designed and trained to better attend the dynamically changing multimodal context.

	Controlled sets			Context Dependent	
	All	Absolute	Group	Absolute	Group
LSTM	77.31	70.45	67.11	60.57*	59.39
LSTM-DH	77.88	71.03	67.75	64.09*	59.06
V-LSTM	74.65	70.87	67.42	58.43*	62.15
V-LSTM-DH	73.82	70.13	65.12	63.33*	62.27
LXMERT	82.40	79.42	74.48	74.21	70.01
LXMERT-DH	82.79	80.19	74.49	74.43	71.12

Table 3: Models task accuracy when they receive vs. do not receive the previous turn. Context-Dependent Absolute questions are the only one for which statistically significant difference is found when the DH is taken into account (t-test among the runs of LSTM/LSTM-DH and V-LSTM/V-LSTM-DH, p-value < 0.05).

	Group-Group	Color-Group	Group-Absolute
LXMERT	65.84	71.58	76.95
LXMERT-DH	66.34	74.32	76.63

Table 4: Task accuracy on manually filtered sets of (271) Context-Dependent questions.

5.2 Qualitative Analysis

We have looked into the errors LXMERT does in three runs and compared them with those made by LXMERT-DH runs. Figure 5 illustrates the trigger-zoomer pairs in images that contain a color trigger question followed by a zoomer group question. We report three examples, in the first one all three runs of LXMERT-DH answers correctly while in two runs LXMERT does not; while in the others two examples both models fail in all runs.

The first example includes the spatial question *1 that is in the left?* that is answered incorrectly by LXMERT without history. Without history, we suspect it is answered with “Yes” since the target is indeed on the left of the image. However, if the previous trigger turn, *are there 2 black cars? Yes*, is considered, the objects that are relevant to answering the spatial question are the 2 black cars; within this group, the target is not on the left but on the right. LXMERT with our simple history encoding is able to answer this spatial question correctly.

The second and third examples include spatial questions that our simple history encoding cannot capture. The second one (*is it the 1st one from right?*) is a spatial question that orders the objects in the inverse order. Usually objects are ordered from left to right but this question counts from right to left. The third example includes a group with four objects in the question *is it in the center row of 4 birds?*. We hypothesize that larger numbers are harder to interpret and answer correctly for LXMERT.

Some of the questions in our history-dependent dataset VISPA could be answered correctly by a human without reading the trigger question since they (being an Oracle) have access to the identity of the target and its attributes (such as category, color, etc). For instance, in the second game in Figure 5 the target is the red light on the right of the image (in the green box). A human Oracle can correctly answer the question *is it the 1st one from the right* assuming it considers only the red lights in the image, but without being sure the questioner is also making this assumption.. We also consider these questions to be history-dependent because they can be answered with more certainty considering the trigger question and its answer. We think that investigating whether history-dependent models become more certain of their correct answers (for the wrong reasons) is an interesting line for future research.

6 Discussion and Conclusions

Visual Dialogues are an interesting challenge because of the interplay between the language and visual modality. When focusing on answering visually grounded questions in dialogues, the main challenge they pose in addition to visual question answering is the need of grounding the question against the dialogue history. In our work we define and evaluate a methodology for extracting visually grounded history-dependent spatial questions from visual dialogues.

Our methodology does not capture all history dependencies in the dataset but it assures that

	Color-Group		Group-Group		Group-Absolute	
	Succeeded	Failed	Succeeded	Failed	Succeeded	Failed
LXMERT	80.97	72.72	84.68	83.45	85.95	73.23
LXMERT-DH	81.37	63.95	80.54	71.37	82.30	65.51

Table 5: Confidence of models in answering FuQs in the manually filtered set. Succeeded: computed over the subsets in which the model provides the correct answer; Failed: computed over the subsets in which the model gives the wrong answer.

those pairs that are identified are indeed history-dependent.

The "trigger-zoomer" methodology we propose is evaluated here on the Guesswhat?! dataset. One possible question is how generic and applicable this model is in longer and open-world dialogues. We think that this method can be extended to longer dialogues by making the "trigger-zoomer" recursive. Moreover, it could be extended to datasets that not only contain questions, but also other forms of language. As far as a "trigger" affects a zoomer question by requiring the dynamic change of the multimodal attention to properly interpret it, the trigger can take any form. For instance, the trigger could be provided in the form of a caption referring to a specific region of an image. We believe that the "trigger-zoomer" methodology would be applicable to all open-word subdialogues that focus on reference resolution. Reference resolution is a frequent task in dialogue which takes up a large part of the turns in domains that are complex or need search. See for instance da Silva Rocha and Paraboni (2020).

We release both the automatically extracted question pairs as well as the subset of such questions which have been manually verified for context dependence. Some of these questions cannot be answered correctly without the previous trigger turn or at least confidence in answering them should be higher when the previous turns are provided. We evaluate the simple oracle models proposed so far in the literature and show that the architecture does not profit from the previous turn as it should. We pose the problem of interpreting follow-up questions as an open problem for the community.

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. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

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 (CVPR)*.

Kordula De Kuthy, Madeeswaran Kannan, Haemant 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.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11831–11838. AAAI Press.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition




	<p>Questioner</p> <ol style="list-style-type: none"> 1. is it a car? 2. is it white color? 3. is it red? 4. black? 5. <i>are there 2 black cars?</i> 6. <i>1 that is in the left?</i> 	<p>Oracle</p> <p>no no no yes yes no</p>
	<ol style="list-style-type: none"> 1. is it a car? 2. is it something on a building? 3. is it some light? 4. is it the street light? 5. can you see 5 or 6 of them on the right side? 6. <i>ok..so it is a light but may be the red lamps?</i> 7. <i>is it the 1st one from right?</i> 	<p>no no no no no yes yes</p>
	<ol style="list-style-type: none"> 1. is it a bird? 2. <i>is it white?</i> 3. <i>is it in the center row of 4 birds?</i> 4. Is it second from the front? 5. is it third from the front? 6. from the top white birds it is in second? 7. is it the top first? 	<p>yes yes yes no yes yes yes</p>

Figure 5: The two questions in italics in each dialogue correspond to pairs that start with a color question and continue with a group question. The first is an example in which LXMERT-DH answers correctly while LXMERT does not. The second and third ones illustrate kinds of spatial questions that are too challenging for our simple history encoding.

challenge. *International journal of computer vision*, 115(3):211–252.

David Schlangen. 2019. [Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings](#). *CoRR*, abs/1908.11279.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. [Beyond task success: A closer look at jointly learning to see, ask, and Guess-What](#). 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 2578–2587.

Danillo da Silva Rocha and Ivandr e Paraboni. 2020. [Building referring expression corpora with and without feedback](#). *Lang. Resour. Evaluation*, 54(4):875–891.

Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of international joint conference on artificial intelligence (IJCAI)*.

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 5103–5114.

Alberto Testoni, Claudio Greco, Tobias Bianchi, Maurizio 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.

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

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 Con-*

ference on Computer Vision and Pattern Recognition, pages 5503–5512.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3:1–191.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2561–2569. IEEE.