They Are Not All Alike: Answering Different Spatial Questions Requires Different Grounding Strategies

Alberto Testoni[1], **Claudio Greco**[1], **Tobias Bianchi**[3], **Mauricio Mazuecos**[2], **Agata Marcante**[4], **Luciana Benotti**[2], **Raffaella Bernardi**[1]
[1] University of Trento, Italy [2] Universidad de Córdoba, Conicet Argentina
[3] ISAE-Supaero, France [4] Université de Lorraine, France
{alberto.testoni|claudio.greco|raffaella.bernardi}@unitn.it
{mmazuecos|luciana.benotti}@unc.edu.ar
tobias.bianchi@student.isae-supaero.fr
agata.marcante7@etu.univ-lorraine.fr

Abstract

In this paper, we study the grounding skills required to answer spatial questions asked by humans while playing the GuessWhat?! game. We propose a classification for spatial questions dividing them into absolute, relational, and group questions. We build a new answerer model based on the LXMERT multimodal transformer and we compare a baseline with and without visual features of the scene. We are interested in studying how the attention mechanisms of LXMERT are used to answer spatial questions since they require putting attention on more than one region simultaneously and spotting the relation holding among them. We show that our proposed model outperforms the baseline by a large extent (9.70% on spatial questions and 6.27% overall). By analyzing LXMERT errors and its attention mechanisms, we find that our classification helps to gain a better understanding of the skills required to answer different spatial questions.

1 Introduction

Visual Dialogues are a useful testbed to study how models ground natural language and in particular how they ground spatial language, which is the focus of our analysis. Visual Dialogues have been the aim of early work on natural language understanding (NLU) (Winograd, 1972) and are now studied by a very active community at the interplay between computer vision and computational linguistics (e.g. Baldridge et al. (2018); Ilinykh et al. (2019); Haber et al. (2019)). Recently, important progress has been made on visual dialogue systems thanks to the release of datasets like VisDial (Das et al., 2017) and GuessWhat?! (de Vries et al., 2017). The former contains chit-chat conversations about an image whereas the latter is a visual game, hence its dialogues are goal-oriented. In both cases, one agent asks questions and the

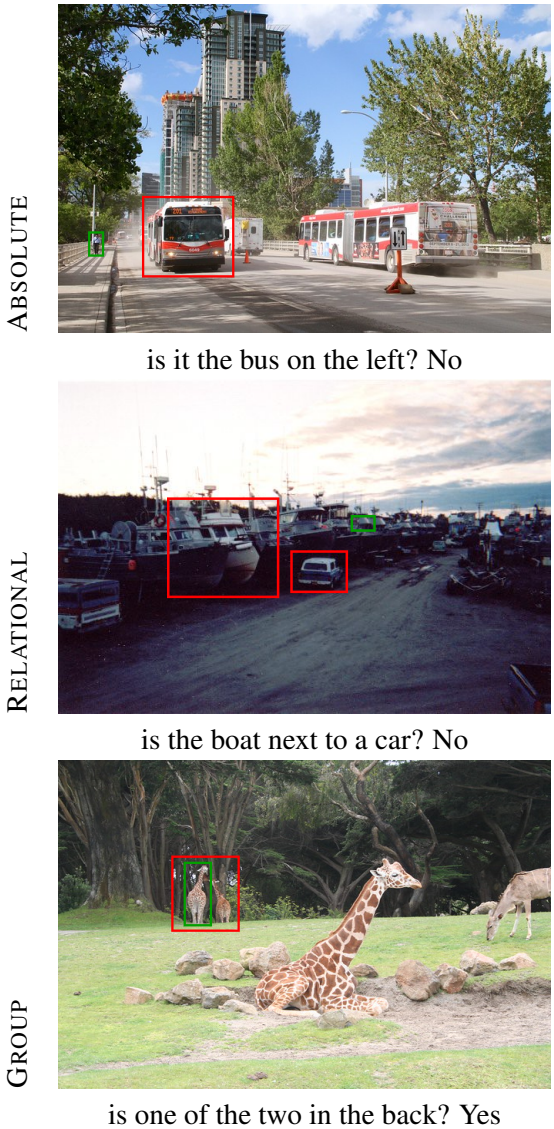


Figure 1: A vast amount of questions asked by humans in the GuessWhat?! game (de Vries et al., 2017) are spatial. We classify them as *absolute*, *relational*, and *group* based on how they many objects are involved and how they are related. The red box marks the object(s) involved in the question, while the green box marks the target of the game. *Relational* and *group* questions need more than one object, whereas absolute do not.

Proceedings of the Third International Workshop on Spatial Language Understanding (SpLU 2020), pages 29–38

https://doi.org/10.18653/v1/P17

other, which we call the Oracle, answers. For VisDial most of the work focused on the answerer, but in-depth evaluation has been carried out on the questioner too (eg., Murahari et al. (2019); Testoni et al. (2019)). For GuessWhat?!, instead, work has been done mostly, if not only, on the questioner. Current models trained with reinforcement learning achieve high task success; they adapt to the oracle limitations and end-up asking questions that are linguistically simpler than those asked by humans (Shekhar et al., 2019; Pang and Wang, 2020).

It is interesting to understand where current multimodal NLU models stand with respect to this task: answering questions asked by humans in a goal oriented visual dialogue. Our paper addresses this question by evaluating how the Oracle model of the GuessWhat?! game answers questions asked by humans while playing the game.

GuessWhat?! is a cooperative game where two players talk in order to identify an object in an image. The player known as the *Questioner* has to guess the target by asking yes/no questions. The other player, the *Oracle*, knows the target object and answers the questions. Shekhar et al. (2019) show that most of the questions in the dataset are about the entity of the target ("Is it a female?") or its location ("is it the first one?"). Mazuecos et al. (2020) show that the baseline model, commonly used for the Guesswhat?! task since its introduction in de Vries et al. (2017), has almost human-like accuracy on the entity questions and a much lower accuracy on questions about attributes. In this paper, we focus on spatial questions and classify them into three groups: *absolute*, *relational*, and *group* questions as illustrated in Figure 1.

An unpleasant aspect of the baseline model is that it receives the gold standard entity of the target (that is, the category label, e.g. "giraffe" or "boat") as input. Furthermore, it answers questions without seeing either the image or the visual features of the target, but instead it simply relies on the category label of the target and its coordinates. Important progress on multimodal encoders has been obtained since the GuessWhat?! release; hence, we study the effect of using models that ground the question into the image and do not have access to the gold standard category label of the target. We adapt a multimodal universal encoder, LXMERT (Tan and Bansal, 2019), to play the role of the Oracle and compare it with the baseline model.

It is known that grounding spatial expressions is challenging for neural networks since quite often they require models to put attention on more regions simultaneously and spot the relation holding among them (e.g., the car and the boat in Figure 1, middle). LXMERT is a transformer-based neural network and as such it heavily exploits attention-based mechanisms. In this paper, we run a qualitative analysis of the attention LXMERT exhibits for the different types of location questions and run an in-depth error analysis of its results. To sum up, we make the following contributions:

- We adapt LXMERT to play the role of the Oracle of the GuessWhat?! game obtaining an overall accuracy of 82.21%, an increase of 6.27% with respect to the usual baseline.
- We find that LXMERT improves over the baseline also on spatial questions (+9.70%), but they remain a large source of errors also for this model – with 77.00% accuracy.
- We classify spatial questions into three subtypes and use this classification to annotate the subset of spatial questions in the GuessWhat?! test set. The fine-grained evaluation shows that the hardest spatial questions are the *relational* and *group* ones.
- We run an in-depth qualitative analysis of LXMERT cross-modal attention and an analysis of its errors on each question sub-type. The analysis shows that LXMERT attention differs between *absolute* and *relational* questions as expected, and that some spatial questions need the dialogue history to be interpreted correctly.

The paper proceeds as follows. Section 2 reviews previous work on visual question answering and on spatial referring expressions. Section 3 presents the models providing information on how we adapt LXMERT for the Oracle task. Section 4 describes the dataset and our classification of spatial questions. In Section 5 we compare the accuracy of the models reporting a fine-grained evaluation by question type and zoom into the subset of spatial questions. We further analysed this subset through a manual inspection of LXMERT attention and errors in Section 6, before drawing our conclusions in Section 7. The code of our work is available at: https://github.com/albertotestoni/unitn_unc_splu2020.

2 Related Work

Answering visual questions is a task that has received increasing attention during the last years. Interesting exploratory analysis has been carried out to understand Visual Question Answering (VQA) systems which highlight their weaknesses and strengths, e.g. (Johnson et al., 2017; Shekhar et al., 2017; Suhr et al., 2017; Kafle and Kanan, 2017). VQA datasets contain both wh- and Y/N-questions. But the kind of Y/N visual questions the Oracle needs to answer are different than those of the VQA datasets: it has to check whether the target has or does not have the questioned property. Hence, it has to compare the target's properties with those of the entity the question refers to and answer accordingly. Moreover, differently from VQA, the GuessWhat?! dataset has been collected in a more naturalistic environment, by letting humans play the games. We adapt LXMERT (Tan and Bansal, 2019), a multimodal universal encoder State-of-the-Art in VQA, to accomplish the Oracle's challenge.

After the introduction of the supervised baseline models (de Vries et al., 2017), several models have been proposed for the Questioner, which are mostly based on reinforcement learning (Sang-Woo et al., 2019; Zhang et al., 2018b; Zhao and Tresp, 2018; Zhang et al., 2018a; Gan et al., 2019; Yang et al., 2019; Pang and Wang, 2020). For these models, the role of the Oracle is even more salient than for models based on supervised or cooperative learning (Shekhar et al., 2019) since they are reinforced to ask those questions the Oracle is good at answering. Despite this important role of the Oracle, no work has been carried out to evaluate and improve it. We aim to fill this gap.

Shekhar et al. (2019) show that GuessWhat?! human players ask quite a lot spatial questions. It has been observed that capturing the spatial relation about objects is challenging for neural network models. Kelleher and Dobnik (2017) argue that Convolutional Neural Network (CNN) do not ground spatial information properly: since they discard location information through the pooling mechanism, their embeddings can only capture rough relative positions of objects within a scene. In line with this claim, Collell and Moens (2018) show that linguistic features are more spatially informative than CNN visual features. New multimodal models, like LXMERT, start from positional aware embeddings. We therefore study how well they handle the spatial questions asked by GuessWhat?! players.

Spatial expressions have been deeply studied within the referring expression generation community. In this area, earlier work (Paraboni et al., 2007) has suggested that, in ordered domains (e.g., a document divided into sections and subsections), referring expressions that include spatial information, even when redundant, lead to a significant reduction in the amount of search that is needed to identify the referent. It has been argued that spatial information reduces the cognitive load (measured by eye tracking) necessary for resolving a referring expression (Paraboni et al., 2017). This research area (Krahmer and van Deemter, 2012; Ghanimifard and Dobnik, 2017) distinguishes between spatial referring expressions that involve another object in the description (e.g. "the rabbit in the hat") from those that do not (e.g. "the rabbit on the left"). The first group of expressions is known as *relational*, while we shall refer to the second one as *absolute*. A further distinction is made between referring expressions that are singular (e.g. "the rabbit in the hat") and those that are plural (e.g. "the three rabbits on the table") and refer to a *group* (see e.g., Lønning (1997); Gatt and van Deemter (2007); Krahmer and van Deemter (2012)).

In this paper, we classify GuessWhat?! spatial questions using *absolute*, *relational* and *group* distinctions and examine how LXMERT performs for each type of spatial question. We also conduct an error analysis and an attention analysis taking these categories into consideration.

Recent work by Agarwal et al. (2020) shows that in current visual dialogue datasets the dialogue history rarely matters. The authors ask crowdsourcers whether they can confidently answer a question by looking at the image and the question, without seeing the dialogue history. In our qualitative analysis we check whether history plays a role for the spatial questions of the GuessWhat?! game that LXMERT fails to answer.

3 Models

In this section we present the models that we compare. We also explain how we adapted LXMERT to the Oracle task. The models are trained on successful games.

LSTM is the baseline model proposed in de Vries et al. (2017). It does not have access to the raw image features. It receives as input embeddings of the target object's category,

its spatial coordinates, and one question encoded by a dedicated LSTM. These three embeddings are concatenated and fed to a Multi-Layer Perceptron (MLP) that gives an answer (Yes or No).

V-LSTM We enhance the LSTM model described above with the visual modality and we remove the information about the target object category. We extract the visual vectors corresponding to the input image and the crop of the target object using a frozen ResNet-152 network pre-trained on ImageNet (He et al., 2016) and we pass them through a linear layer and a *tanh* activation function. We concatenate these scaled representations to the embeddings of the target object's spatial coordinates and the question: the resulting vector is fed to an MLP to obtain the answer, as it happens in the LSTM model.

LXMERT To evaluate the performance of a universal multimodal encoder, we employ LXMERT (Learning Cross-Modality Encoder Representations from Transformers) (Tan and Bansal, 2019). It represents an image by the set of position aware object embeddings for the 36 most salient regions detected by Faster R-CNN and it processes the text input by position aware randomly initialized word embeddings. We fill the 36th position with the visual features of the target object. 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; the latter is used to separate sequences and to denote the end of the textual input. LXMERT has been pre-trained on five tasks.[1] It has 19 attention layers: 9 and 5 self-attention layers in the language and visual encoders, respectively and 5 cross-attention layers. We process the output corresponding to the CLS token. We consider both the pre-trained version (**LXMERT**) and the one trained from scratch (**LXMERT-S**).[2]

[1]Masked cross-modality language modeling, masked object prediction via RoI-feature regression, masked object prediction via detected-label classification, cross-modality matching, and image question answering.

[2]We have also evaluated a simplified version of LXMERT-S in which we use 6 self (4 language and 2 visual) and 2 cross-modal attention layers. The model behaves similarly to the more complex version trained from scratch.

		Single label	Multi labels
	Entity	39269	39269
	Not classified	7925	7925
ATTRIBUTE	Spatial	29845	39250
	Color	7145	15403
	Action	3063	7645
	Size	532	1364
	Texture	538	901
	Shape	166	301

Table 1: Question type distribution in successful games following the classification proposed in (Shekhar et al., 2019) where a question can be assigned to more than one attribute type (multiple labels); the Single label column reports the number of questions which have been assigned to only one type.

4 The Dataset

The GuessWhat?! dataset is composed of more than 150k human-human dialogues containing an average of 5.3 questions in natural language created by turkers playing the game on MS COCO images (Lin et al., 2014). Humans have succeeded on 85% of the games. Not successful games may contain errors made by the human oracle which lead to task failure, we discard questions that belong to human dialogues that were not successful. The remaining set contains around 672K questions which are grounded on about 63K unique images and belong to about 135K dialogues.

Shekhar et al. (2019) propose a classification of the questions based on their focus distinguishing questions which ask about the entity of the target ("Is it an animal?" or "Is it a dog?") or an attribute of it. A question can focus on just one attribute (e.g., "Is it the black dog"? or "Is it black?") in which case it is assigned just to one attribute question type (color in the examples) or about more attributes (e.g., "does it have orange pillows on it?") in which case it is assigned to more attribute question types (to both color and spatial information in the example.) Table 1 reports their distribution in the human-human dialogues giving the numbers of questions assigned to one or more types (multi label) or to just one type (single label).

We conjecture that the spatial question type includes questions posing different challenges to multimodal models. Krahmer and van Deemter (2012) divide spatial expressions into *relational* (e.g. "the

rabbit in the hat"), that specifies the location of the referent of a noun phrase (the target, "rabbit") relative to another object (the landmark, "hat"), and *absolute* that focus only on the target by providing locative information about it (e.g. "the rabbit on the left"). A third spatial expression that has received attention within the REG community are group referring expressions whose target is a group of entities (e.g. "the three rabbits on the table") or some specific entity of a the group to which the expression refers by ordering them (e.g. "the second rabbit from the left").

We adapt such classification to the GuessWhat?! spatial questions and classify them into four types: relational, absolute, group and other. To distinguish these types we have leveraged syntactic and lexical characteristics specific to each. Relational questions usually include a prepositional phrase followed by a noun phrase that includes either a pronoun (e.g. "Is there a sink directly above it?") or an object word (e.g. "is it the pen behind the laptop?"). Absolute spatial questions (e.g. "the one on the left?") instead contain a location word either in the x axis (e.g. right, middle, left), or the y (top, bottom), or the z (e.g. front, back) axis. We also consider absolute those questions that include a spatial adjective in its superlative form (e.g. "the leftmost one?"). Finally, we consider group questions those containing a number which may indicate order (e.g. "right to left, is it the first one?") or groups (e.g. "in the back among four women?"). We have automatically annotated spatial questions by identifying nouns, prepositions and number using the Part of Speech tagger Stanza (Qi et al., 2020). When a question is not assigned to any of the three groups, we include it in the "Other" category.[3] We tried identifying objects using the entity recognizers included in Stanford core NLP (Manning et al., 2014) and Stanza (Qi et al., 2020) but the coverage was not good.

In the next section, we will first compare models using the multi-label classification reported in Table 1, then we will zoom into the spatial questions which together with the entity questions constitute the large majority of questions asked by humans. In order to understand strength and limits of multimodal models in answering spatial questions, we focus on those which are assigned only to the spatial question type to avoid confounding effects. Table 2 reports number of such sub-set.

[3]Examples of questions following into the "Other" category are: "Is it the tree outside?" – i.e. an elliptical question which could be completed as "Is it the tree outside the fenced garden?" – or "Can you sleep on it?" which is not about a spatial property that occurs in the image but an afforded one.

	%	Example
Relational	31.9	Is it the pen behind the PC?
Absolute	31.8	Is it the one on the left?
Group	17.3	Is it among the 4 women?
Other	19.0	Can you sleep on it?

Table 2: Sub-type spatial questions distribution in successful games of questions annotated with only the spatial label in the test set (total: 29845).

5 Experiments

5.1 Evaluation by Question Type

de Vries et al. (2017) shows that the "blind" version of the LSTM model performs better that the version receiving the visual features. This result is heavily dependent on the question type distribution in human-human dialogues. As we have seen, entity questions are a great proportion of the questions humans ask. The "blind" baseline model is facilitated in answering them, since it is given the category of the target object. Following Mazuecos et al. (2020), we evaluate models accuracy by question types. As we can see from Table 3, the higher overall accuracy reached by the "blind" LSTM model is indeed mostly due to the "entity" questions for which it reaches 94% (questions like: "is it a vehicle?"). As expected, when removing the category (V-LSTM) the accuracy on answering questions about entities decreases to a large degree, but the use of visual features helps the model to answer color questions better. The replacement of the LXMERT architecture, together with the use of positional aware embedding representations of the image, bring an important boost in the accuracy: LXMERT trained from scratch outperforms the LSTM based model on all types of questions. The pre-training phase further increases the performance in important ways.

5.2 Evaluation on Spatial Questions

Above we have seen that LXMERT outperforms the other models on the spatial questions. Our fine-grained classification sheds light on an interesting point: its main advantage comes from the *relational* questions (Table 4). *Absolute* questions require cross-modal attention only to align a word with its referent, whereas *relational* questions are

	LSTM	V-LSTM	LXMERT-S	LXMERT
Entity	93.37	83.24 (-10.13)	88.64 (-4.73)	91.09 (-2.28)
Spatial	67.30	66.40 (-0.90)	71.31 (+4.01)	77.00 (+9.70)
Color	61.64	68.06 (+6.42)	70.51 (+8.87)	76.42 (+14.78)
Action	64.32	65.44 (+1.12)	70.23 (+5.91)	77.16 (+12.84)
Size	60.41	62.76 (+2.35)	67.23 (+6.82)	75.44 (+15.03)
Texture	69.92	66.15 (-3.77)	71.92 (+2.00)	77.47 (+7.55)
Shape	68.44	64.12 (-4.32)	70.76 (+2.32)	74.42 (+5.98)
Not classified	75.02	70.45 (-4.57)	74.94 (-0.08)	82.18 (+7.16)
Total	75.94	72.70 (-3.24)	77.41 (+1.47)	82.21 (+6.27)

Table 3: Accuracy of the models on the successful games by question type based on the multi label assignment. Values in parenthesis report the comparison with LSTM.

	Absolute	Relational	Group
LSTM	76.4	67.1	63.3
V-LSTM	75.2	63.5	62.8
LXMERT-S	80.5	69.6	68.4
LXMERT	83.4	77.2	71.6

Table 4: Accuracy of the sub-type of spatial questions (successful games, questions assigned only one type)

more challenging: the model has to locate the regions corresponding to the two related words and understand the relation holding among them. The *group* questions may require "counting" skills that go beyond the scope of this paper.

6 Qualitative Analysis

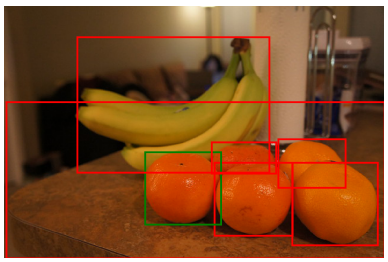
As a first step towards a deeper understanding of LXMERT performance, we use a linear logistic regression model for the task of predicting whether a question was answered correctly. In Shekhar et al. (2018) it has been shown that unsuccessful games contain more objects in the image than successful ones, and that the target size area is smaller. We use these two features as predictor variables together with the length of the question and the turn in which it was asked in the full dialogue. We observe that the number of objects in the image and the question turn play a significant role in predicting the model behaviour. This might be due to the fact that models do not receive the dialogue history as input. Below we run an error analysis based on the three spatial sub-type questions described above to check whether indeed this could be a source of error. After the error analysis, we study whether LXMERT uses its cross-modal attention differently across these three groups of questions.

6.1 Error Analysis

We did a manual error analysis of 20% of LXMERT errors on spatial questions. We tagged emergent error categories by following a qualitative annotation methodology. Below we describe our findings by classifying them in the three types of spatial questions that we consider throughout the paper.

We found that absolute and group questions have more errors related to the missing dialogue history than relational questions even though we explicitly allow for relational questions that include anaphoric pronouns. For these two categories, around 50% of errors are related to missing dialogue history. Dialogue history dependency in the dataset is generally not lexicalized with explicit pronouns but left implicit through ellipsis (e.g. "in the middle?"). Figure 2 shows an example of this. Question 5 could be answered with "yes" if asked at the beginning of the dialogue ("middle" would refer to the middle of the image) but its answer is "no" due to history ("middle" refers to the middle of the group of oranges). In most of these dialogues, the category of the target is left implicit because it is established in previous questions (e.g., "orange"). But also other information is implicit. For example, "the last single one?" does not say that the search is evolving from right to left. In these cases, the meaning of the question is only correctly interpretable in the dialogue context.

History dependence, as illustrated in Figure 2, is hard to detect even for human annotators. Using the presence of the pronoun to detect whether a question needs the history in order to be properly answered, as it has been done in Agarwal et al. (2020), might be misleading. Our examples show that ellipses might create more context dependen-



Human question	Human answer
1. It is a fruit?	yes
2. It is the orange?	yes
3. One of them I suppose?	yes
4. Is it to our right?	no
5. In the middle?	no
6. The last single one?	yes

Figure 2: Sample image and dialogue from the GuessWhat?! dataset. The red boxes mark the objects involved in the questions, while the green box marks the actual referent. LXMERT incorrectly answers "yes" to question 5. LXMERT, like all Oracles, does not have access to the dialogue history. It probably interprets the question as "is the target in the middle of the picture?". The image and dialogue illustrates the history dependence of questions.

cies and that there are questions which could be apparently answered even when given in isolation but they would be answered differently based on the context they are in.

For absolute only questions, we found the following errors. Questions related to the z-axis of the picture (e.g. “is it in the background?”) seem to be harder for the model than those questions related to the x-axis of the picture (e.g. “is it on the left?”). The errors that do occur on the x-axis are either related to the fact that the dialogue history is necessary in order to interpret the question as in Figure 2, or that the target is neither on the left nor on the right of the x-axis. In this dataset the adjective left and right behave as vague adjectives. Questions that include superlatives (e.g.“the rightmost book?) cause many errors. As well as questions that combine two or more of these characteristics (e.g. “is it the animal at the very front on the left ?”). Finally, the ambiguity of the word “middle”, which could be used for any axis, seems to confuse the model.

For group questions, the second most frequent errors corresponds to questions grouping in one of the three axes. The term "row" is often used to group the target with other objects, especially when images are overcrowded with objects belonging to the same category. However, the term is an ambiguous one, as it can refer to any of the three axes and its meaning is often dependent on which interpretation is more salient in the image. Furthermore, inverse x-axis properties (e.g., "third girl from right?") also seem to be problematic. Another frequent error type includes questions that require counting above three (e.g., "seventh bus from the left?"). People can immediately and precisely identify that an image contains 1, 2, 3 or 4 items by a simple glance, this ability is called subitizing (Kaufman et al., 1949; Piazza et al., 2002). Identifying

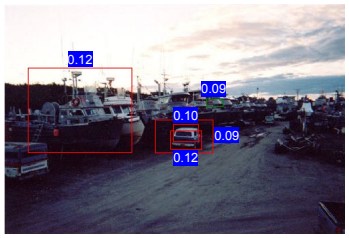
Layer	Absolute	Relational	Group
0	3.9	4.1	3.3
1	4.2	4.6	4.1
2	3.8	4.5	4.0
3	3.7	4.0	3.7
4	1.3	2.2	1.9

Table 5: Language to Vision attention in LXMERT: Number of regions of the image considered salient in the last layer from the CLS token – viz. regions with an attention value higher than the 0.05 threshold.

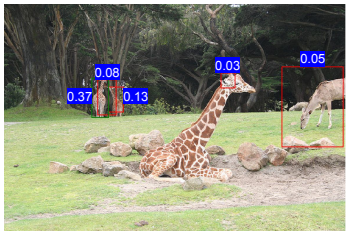
the quantity of a larger number of objects takes considerably longer and involves counting for humans. It seems models such as LXMERT are able to do subitizing, but not counting. Other problematic group questions are multi-type ones, for instance belonging also to the relational type (e.g., "are there two of them on the branch?"); and questions using entities outside the image as reference, such as the viewers (e.g., "is it in the first room closer to us?").

For relational questions we find that a source of errors is when the target and the landmark bounding boxes overlap or one is included in the other (“is it the clock behind the person?”). Also when the landmark is a part of another object instead of being an object with well delimited borders the model seems to get confused (“is it under his feet?”). Questions that include non projective prepositions seem harder (“is it the person near the bicycle?”) than those whose prepositions indicate the direction of the relation. Another source of errors are questions in which the landmark is large and no clear borders are visible (“is it on the water?”). Finally, those questions that require OCR (optical character recognition) are problematic (“does it have words on it?”).

is it the bus on the left? No



is it the boat next to a car? No



is it one of the two in the back? Yes

Figure 3: Attentions from the CLS: in absolute questions attention is mostly on the only object the question refers to (the left bus, 0.13) and the target object (0.64) (**left**); in the relational questions attentions spread between the two related objects (car and boat, 0.12 each) and the target object (the boat on the back, 0.9) (**middle**); in the group questions attentions goes to the entity of the referred group (0.08 and 0.13) and the target (0.37) (**right**).

6.2 LXMERT's Attention

Here we aim to understand how LXMERT uses attention mechanisms to answer spatial questions. We focus our analysis on the cross-attention layers from language to vision. Recall that, in our adaptation of LXMERT to the Oracle task, the crop of the target is given as the 36th visual embedding together with the most salient regions of the image detected by Faster R-CNN. We are interested in understanding how it exploits the target visual representation to guide attention.

The entropy of the attention maps shows that the model in the first attention layers distributes attention across all regions (its entropy is close to the maximum possible level), at layer 2 it learns to focus its attention on some regions of the image and on the crop of the target. Finally, at the last layer, the attention on the CLS (the embedding given to the classifier to select the answer) reveals an interesting difference among question types: the number of regions considered salient in the absolute questions is lower that the one of salient regions in the group and relational questions. Table 5 reports the numbers of regions with an attention value higher than 0.05.[4] We have used different thresholds to compute the number of top-valued regions and the same pattern emerges. From a manual inspection, we have seen that the higher number of salient regions in the relational questions often is due to the fact that they refer to more candidate objects, differently from the absolute ones which usually refer to fewer or even just one object.

Figure 3 illustrates how LXMERT uses its attention in three sub-type of spatial questions. As we can see, when it interprets relational questions involving two objects, it "looks" both at the target (the boat) and the landmark (the car); in the example it answers the question negatively since the target of the game is the boat marked by the green box and not the one to which the question refers to. Similarly, when interpreting a group question, it looks at the referred group (the two giraffes); in the example it answers the question positively since the target of the game is indeed within the referred group. By looking at the attention maps, we noticed that interesting patterns emerge when looking at the attentions from the CLS token (Figure 3 marks the regions considered more salient from the CLS token). Other tokens put attention mostly or only on the target object region.

7 Conclusion

In this paper we tackle the problem of grounding spatial questions in the GuessWhat?! visual dialogue game. We adapt LXMERT to play the role of the Oracle of the GuessWhat?! game reaching an overall accuracy of 82.21%. This result outperforms the widely used baseline model by 6.27%. The gain is even higher for spatial questions, where LXMERT outperforms the baseline by 9.70%. In order to perform an in-depth analysis, we classify spatial questions into three sub-types and use this classification to annotate the subset of spatial questions in the GuessWhat?! test set. The fine-grained evaluation shows that the hardest spatial questions are the relational and group ones. We perform an in-depth analysis of LXMERT cross-modal attention and an qualitative analysis of the errors on each question sub-type. First of all, we find out that LXMERT puts attention on more regions when processing relational questions compared to absolute and group questions. Secondly, the qualitative analysis highlights the importance of having access to the dialogue history in order to answer some spatial questions. We leave this for future work.

[4] If the attention is equally distributed among all the 36 regions, their attention value would be 0.02 (viz. 1/36).

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.

Jason Baldridge, Tania Bedrax-Weiss, Daphne Luong, Srini Narayanan, Bo Pang, Fernando Pereira, Radu Soricut, Michael Tseng, and Yuan Zhang. 2018. Points, paths, and playscapes: Large-scale spatial language understanding tasks set in the real world. In *Proceedings of the First International Workshop on Spatial Language Understanding*, New Orleans. Association for Computational Linguistics.

Guillem Collell and Marie-Francine Moens. 2018. Learning representations specialized in spatial knowledge: Leveraging language and vision. *Transactions of the Association for Computational Linguistics*, 6:133–144.

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.

Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6463–6474.

Albert Gatt and Kees van Deemter. 2007. Incremental generation of plural descriptions: Similarity and partitioning. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 102–111.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell Me More: A Dataset of Visual Scene Description Sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume abs/1612.06890.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973.

E. L. Kaufman, M. Lord, T. Reese, and J. Volkmann. 1949. The discrimination of visual number. *The American journal of psychology*, page 498–525.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *In CLASP Papers in Computational Linguistics. Proceedings of the Conference on Logic and Machine Learning in Natural Language*, pages 41–52.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer.

J. Lønning. 1997. Plurals and collectivity. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 1009–1054. Elsevier.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mauricio Mazuecos, Alberto Testoni, Raffaella Bernardi, and Luciana Benotti. 2020. On the role of effective and referring questions in GuessWhat?! In *Proceedings of the First Workshop on Advances in Language and Vision Research*, pages 19–25, Online. Association for Computational Linguistics.

Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Improving generative visual dialog by answering diverse questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1449–1454.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *Proceedings of 34th AAAI Conference on Artificial Intelligence*.

Ivandré Paraboni, Kees van Deemter, and Judith Masthoff. 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254.

Ivandré Paraboni, Alex Gwo Jen Lan, Matheus Mendes de Sant'Ana, and Flávio Luiz Coutinho. 2017. Effects of cognitive effort on the resolution of overspecified descriptions. *Computational Linguistics*, 43(2):451–459.

M. Piazza, A. Mechelli, B. Butterworth, and C.J. Price. 2002. Are subitizing and counting implemented as separate or functionally overlapping processes? *NeuroImage*, pages 435–446.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Lee Sang-Woo, Gao Tong, Yang Sohee, Yao Jaejun, and Ha Jung-Woo. 2019. Large-scale answerer in questioner's mind for visual dialog question generation. In *Proceedings of International Conference on Learning Representations, ICLR*.

Ravi Shekhar, Tim Baumgärtner, Aashish Venkatesh, Elia Bruni, Raffaella Bernardi, and Raquel Fernandez. 2018. Ask no more: Deciding when to guess in referential visual dialogue. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1218–1233, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! Find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265.

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 GuessWhat. 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.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 217–223, Vancouver, Canada. 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)*, pages 5103–5114.

Alberto Testoni, Ravi Shekhar, Raquel Fernández, and Raffaella Bernardi. 2019. The devil is in the detail: A magnifying glass for the GuessWhich visual dialogue game. In *Proceedings of the 23rd SemDial Workshop on the Semantics and Pragmatics of Dialogue (LondonLogue)*, pages 15–24.

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, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4466–4475. IEEE Computer Society.

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 *Proceedings of the International Conference on Computer Vision (ICCV)*.

Jiaping Zhang, Tiancheng Zhao, and Zhou Yu. 2018a. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 140–150.

Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2018b. Goal-oriented visual question generation via intermediate rewards. In *Proceedings of the European Conference of Computer Vision (ECCV)*, pages 186–201.

Rui Zhao and Volker Tresp. 2018. Improving goal-oriented visual dialog agents via advanced recurrent nets with tempered policy gradient. In *Proceedings of IJCAI*.