"Yes" and "No": Visually Grounded Polar Answers

Claudio Greco[1*] **and Alberto Testoni**[2*] and **Raffaella Bernardi**[1,2]
CIMeC[1] and DISI[2], University of Trento
name.surname@unitn.it

Abstract

Modelling negation is a challenging goal and little is known about how neural models handle it. It has been shown that humans have harder time understanding negative sentences than positive ones, but that the processing cost of negation is mitigated by the presence of supportive context. Based on these findings, we argue that referential visual games are a good starting point for making progress towards the ambitious goal of modelling negation. In this paper, we study how a multimodal universal encoder, LXMERT, is able to encode negation when playing the GuessWhat?! referential visual game. We show that it profits from positively answered questions pretty well, but it struggles profiting from negatively answered questions even when they have been informative for humans to succeed in the game.

1 Introduction

Negation is often neglected by computational studies of natural language understanding, in particular when using the successful neural network models. Admittedly, modelling negation is an ambitious goal. Indeed, even humans have a harder time understanding negative sentences than positive ones (Clark and Chase, 1972; Carpenter and Just, 1975). However, it has been shown that the presence of supportive context mitigates the processing cost of negation, in particular within dialogues where (Dale and Duran, 2011) and (Nordmeyer and Frank, 2014) find that processing negation is easier for humans when a visual context is given. In (Kruszewski et al., 2016) it has been argued that conversational negation in distributional semantics models creates the alternative set of the negated expression, in line with what is claimed in (Oaksford, 2002) about how humans use negation. Based on these findings, we argue that Visual Dialogues and, in particular, referential grounded guessing ones, are a good starting point for making progress towards this ambitious goal: the guesser sees all the possible candidates, hence both the reference of the negated expression and the set of alternatives are at disposal. For instance, a multimodal encoder processing "Is it red? No" should focus its attention on all the candidates in the image that are not red (and that have not been excluded during the dialogue already) (Figure 1, 2nd turn in the example on the left).

Visual Dialogues have a long tradition (e.g. (Anderson et al., 1991)). Recently, several dialogue tasks have been proposed as referential guessing games in which an agent asks questions about an image to another agent and the referent they have been speaking about has to be guessed at the end of the game (de Vries et al., 2017; Das et al., 2017; He et al., 2017; Haber et al., 2019; Ilinykh et al., 2019; Udagawa and Aizawa, 2019). Among these games, GuessWhat?! and GuessWhich (de Vries et al., 2017; Das et al., 2017) are asymmetrical – the roles are fixed: one player asks questions (the Questioner) and the other (the Oracle) answers. The game is considered successful if the Guesser, which can be the Questioner itself or a third player, selects the correct target. (Greco et al., 2020) show that GuessWhat?! is suitable to be used as a diagnostic dataset to compare strengths and weaknesses of current State-Of-The-Art (SOTA) encoders in representing grounded dialogues. We argue that GuessWhat?! human dialogues are a suitable diagnostic dataset to analyse to which extent SOTA encoders properly represent visually grounded negation.

We focus on human dialogues of successful games, namely the games in which the target object has been guessed successfully at the end of the dialogue. Our work builds on the observation that within these dialogues a crucial role is played by

[*] Equal contribution. The first two authors are reported in alphabetic order.

Questioner	Oracle
1. Is it on a wooden surface?	Yes
2. Is it red?	No
3. Is it white?	No
4. Is it a scissor?	Yes
5. Is it the scissor on the left of the picture?	Yes

Questioner	Oracle
Q1. Is it an object?	No
Q2. Is it a person?	Yes
Q3. Does he have his right arm on the other's shoulder?	No

Figure 1: Two samples of GuessWhat?! human dialogues ending with a positive (left) and a negative (right) turn.

the question-answer pair in the last turn. As shown by the examples in Figure 1, this role is different when the question is answered positively or negatively. In the former case, the question tends to almost fully describe the target object, whereas in the latter case it conclusively identifies the target object by excluding those candidates which most likely are not the target – it creates a singleton alternative set. In the paper, we show that these features of the GuessWhat?! dialogues make them suitable data to shed some light on how well multimodal encoders interpret visually grounded negation. Our analysis shows that:

- LXMERT obtains a very high overall accuracy but most of its boost with respect to a simple multimodal LSTM model comes from dialogues ending with a positively answered question. Its pre-training phase let it encode the very informative last turn well;
- Its boost is moderate on grounding negatively answered questions.

2 Models

Following (Greco et al., 2020), we adapt LXMERT to the GuessWhat?! guessing game by pairing it with a Guesser as illustrated by the skeleton in Figure 2 which we use also for the baseline LSTM based models. For all models, the Guesser is the module proposed in (de Vries et al., 2017). Candidate objects are represented by the embeddings obtained via a Multi-Layer Perceptron (MLP) starting from the category and spatial coordinates of each candidate object. The representations so obtained are used to compute dot products with the hidden dialogue state produced by an encoder. The scores of each candidate object are given to a softmax classifier to choose the

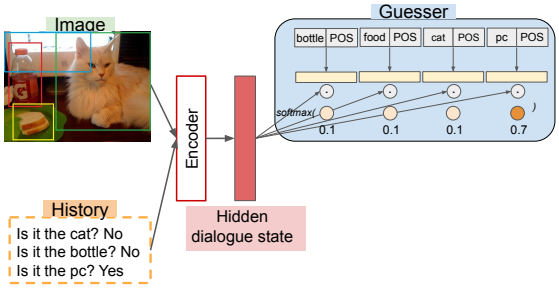

Figure 2: Shared Encoder-Guesser skeleton. Models differ in how they compute the hidden dialogue state.

object with the highest probability. The Guesser is trained in a supervised learning paradigm, receiving the complete human dialogue history at once. The models we compare differ in how the hidden dialogue state is computed.

LSTM As in (de Vries et al., 2017), the representations of the candidates are fused with the last hidden state obtained by an LSTM which processes only the dialogue history.

V-LSTM We enhance the LSTM model described above with the visual modality by concatenating the linguistic and visual representation and scaling its result with an MLP; the result is passed through a linear layer and a *tanh* activation function to obtain the hidden state which is used as input for the Guesser modules. We use a frozen ResNet-152 pre-trained on ImageNet (He et al., 2016) to extract the visual vectors.

LXMERT In order 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 a Faster R-CNN and the text by position-aware randomly-initialized word embeddings. Both the visual and linguistic representations are processed by a specialized 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 vision output) enhanced with the other modality as well as their joint representation (cross-modality output). LXMERT uses the special tokens CLS and SEP; CLS is taken to be the representation of the given sequence, whereas SEP is used both 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 take CLS as hidden dialogue state. To isolate the effect of the pre-training phase, we consider both the pre-trained version (**LXMERT**) and the one trained from scratch (**LXMERT-S**).

3 Experiments

We divide the games of the test set on which humans have been successful into two subsets: those in which the human dialogue ends with a positively answered question (Yes-set) and those in which it ends with a negatively answered question (No-set) – illustrated by the examples in Figure 1, on the left and right, respectively. The former consists of 16366 games, whereas the latter of 2350.[2] The two subsets are rather similar in terms of length and number of candidates (Table 1).

As shown in Table 2, LXMERT obtains a much higher overall accuracy than the LSTM based models (+4.5 than LSTM); this difference in performance seems to come mostly from its pre-training phase, since LXMERT-S performs on a par with the LSTM based models. Interestingly, when we zoom into the accuracy models reach on the Yes- vs. No-set, we see that all models obtain around +20% accuracy on the Yes-set and that LXMERT's advantage over the other simpler models comes exclusively from the Yes-set (Yes-set: +4.9 than LSTM vs. No-set: +1.9). This seems to suggest that both LSTM and transformer based models have a hard time interpreting negatively answered questions. In the following, we aim to understand this result better by studying the role played by the last turns and how the probability assigned by the Guesser to the candidate objects changes after a Yes- vs. No-turn.

	Nr. Games	# Turns	# Candidates
Full test set	18840	4.5	8
Yes-set	16366	4.5	8
No-set	2350	4.5	7.8

Table 1: Statistics on the full test set and on the Yes- (resp. No-) subsets obtained by selecting only dialogues with a positively (resp. negatively) answered question in the last turn.

	All games	Yes-set	No-set
Random	12.5	16.4	16.4
LSTM	64.7	67.0	49.0
V-LSTM	64.5	67.0	48.3
LXMERT-S	64.4	66.6	49.5
LXMERT	**69.2**	**71.9**	50.9

Table 2: Task Accuracy obtained by models when receiving the human dialogues of all the games in the test-set or only those in the Yes-set vs. No-set.

Last turn: Yes vs. No As commented above, the last turn in the Yes-set vs. No-set is expected to play a rather different role. In particular, we conjecture that already alone a positively answered question in the last turn is rather informative whereas a last turn answered negatively is not. On the other hand, last turns containing a negative answer are expected to enrich the dialogue history

	Yes-set		No-set	
	W/o last	last	W/o last	last
LSTM	48.3	51.8	39.9	24.5
V-LSTM	48.6	47.3	37.8	20.7
LXMERT-S	48.4	51.7	41.0	22.2
LXMERT	49.9	**61.2**	41.9	26.6

Table 3: Accuracy comparison obtained when giving the dialogue without the last turn (W/o last) or with only the last turn (last).

[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] The full dataset of human dialogues is availabe at https://guesswhat.ai/download. We exclude games of the test set in which humans have not succeed (2502) or have not completed the game (1289); we keep only games with maximum length 10. In the remaining games, 124 ends with an underspecified answer (NA).

	All dialogue history			Last turn	
	$T_i : Yes$	$T_i : No$	$T_i : N/A$	$T_i : Yes$	$T_i : No$
LSTM	14.5	2.9	2.3	16.4	5.1
V-LSTM	14.0	3.1	2.9	13.9	2.9
LXMERT-S	12.3	4.4	2.1	14.6	7.2
LXMERT	16.4	4.1	1.4	19.7	**8.8**

Table 4: Change across consecutive turns in the probability assigned to the target after Yes- vs. No- vs. N/A-turns, i.e., $P(o)_{T_{i+1}} - P(o)_{T_i}$ (all dialogue history) and before/after the last turn (Last turn).

and help to guess the target. Hence, they are an interesting test-bed for our research question. First of all, as shown in Table 3, when evaluating models on the dialogue without the last turn, the difference between the accuracy they reach in the Yes- vs. No-set is lower than what we have observed when considering the full dialogue; for LXMERT it is 21% when considering the full dialogue (71.9 vs. 50.9 in Table 2) and 9% when removing the last turn (49.9 vs. 41.9 in Table 3). Moreover, when removing the last turn the drop in accuracy is around 20% in the Yes-set and of only 10% in the No-set. This clearly shows the important role played by the last turn in solving the game and suggests that LXMERT does much better than the other models in encoding and grounding the last turn in the Yes-set, but its advantage over the other models is moderate when grounding negatively answered questions and it is not better than the other models in encoding the dialogue history. Finally, as expected, when receiving only the last turn, models obtain a high accuracy when the answer is positive (Yes-set) and are near to chance level when it is negative (No-set).

The Probability computed by the Guesser We compute how the probability assigned by the Guesser to the target object $P(o)$ changes after each turn ($P(o)_{T_{i+i}} - P(o)_{T_i}$) and compare turns T_i with a "Yes", "No", or "N/A" answer. We expect that throughout the dialogue, it is easier to use the Yes-turns than the No ones (viz. the probability might have a higher boost after a positive turn), but we hope models are able to benefit from the questions answered negatively better that those answered by N/A. Moreover, given the role played by the Yes- vs. No-turns at the end of the dialogue, we would expect that the latter brings an even higher boost in probability than the former, if the model properly interprets negative answers. In Table 4 we report the results for all the dialogues when considering the average change across all turns or of only those at end of the dialogues.[3] As we can see, questions answered with "Yes" bring a much higher increase of probability than questions answered with "No" – which for LSTM have on average the same impact as those answered by N/A (2.9 vs. 2.3).[4] However, after the last turn the probability change is still higher in the Yes- than in the No-turns, and LXMERT does slightly better than the baselines: the change in probability assigned after a No-turn is still rather low (Yes: 19.7 vs. No: 8.8).

4 Related Work

Interesting exploratory analysis has been carried out to understand Visual Question Answering (VQA) systems and highlight their strengths and weaknesses, e.g., (Johnson et al., 2017; Shekhar et al., 2017; Suhr et al., 2017; Kafle and Kanan, 2017). Less is known about how well grounded conversational models encode the dialogue history and, in particular, negatively answered questions. (Greco et al., 2020) show that pre-trained transformers detect salient information in the dialogue history independently of its position.

5 Conclusion

In this paper, we study to what extent a state-of-the-art multimodal universal encoder, LXMERT, is able to encode negation when playing GuessWhat?!. Our results show that while LXMERT greatly outperforms a simple multimodal LSTM based model in grounding positive answers, its boost is moderate on grounding negatively answered questions and it is none when encoding a full dialogue. These results call for further studies on how to improve models' understanding of negation in visually grounded referential games.

[3] We have obtained similar patterns when comparing the models on games with a given number of candidate objects.

[4] On average the probability before a Yes-turn and a No-turn are similar for all models the difference is lower than 10%.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34:351–366.

P. Carpenter and M. Just. 1975. Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, 82:45–73.

H. Clark and W. Chase. 1972. On the process of comparing sentences against pictures. *Cognitive Psychology*, 3:472–517.

R. Dale and N. Duran. 2011. The cognitive dynamics of negated sentence verification. *Cognitive Science*, 35:983–996.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In *2017 IEEE International Conference on Computer Vision*, pages 2951–2960.

Claudio Greco, Alberto Testoni, and Raffaella Bernardi. 2020. Grounding dialogue history: Strengths and weaknesses of pre-trained transformers. In *XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 24-27, 2020, Revised and Selected papers*, volume 12414 of LNAI of *AIxIA 2020: Advances in Artificial Intelligence*. Springer Nature Switzerland AG, 2021.

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*, pages 1895–1910.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1766–1776.

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*, pages 152–157.

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.

Germán Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2016. There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, 42(4):637–660.

Ann Nordmeyer and Michael C. Frank. 2014. A pragmatic account of the processing of negative sentences. *Cognitive Science*, 36.

Mike Oaksford. 2002. Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Thinking & reasoning*, 8(2):135–151.

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.

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.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

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 5503–5512.