Which Turn do Neural Models Exploit the Most to Solve GuessWhat? Diving into the Dialogue History Encoding in Transformers and LSTMs

Claudio Greco[1⋆], Alberto Testoni[2⋆], and Raffaella Bernardi[12]

[1] CIMeC - Center for Mind/Brain Sciences
[2] DISI - Dept. of Information Engineering and Computer Science
University of Trento
name.surname@unitn.it

Abstract. We focus on visually grounded dialogue history encoding. We show that GuessWhat?! can be used as a "diagnostic" dataset to understand whether State-of-the-Art encoders manage to capture salient information in the dialogue history. We compare models across several dimensions: the architecture (Recurrent Neural Networks vs. Transformers), the input modalities (only language vs. language and vision), and the model background knowledge (trained from scratch vs. pre-trained and then fine-tuned on the downstream task). We show that pre-trained Transformers are able to identify the most salient information independently of the order in which the dialogue history is processed whereas LSTM based models do not.

Keywords: Visual Dialogue · Language and Vision · History Encoding.

1 Introduction

Visual Dialogue tasks have a long tradition (e.g. [1]). 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 [33, 4, 8, 7, 10, 31]. Among these games, GuessWhat?! and GuessWhich [33, 4] 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.

Most Visual Dialogue systems proposed in the literature share the encoder-decoder architecture [29] and are evaluated using the task-success of the Guesser. By using this metric, multiple components are evaluated at once: the ability of the Questioner to ask informative questions, of the Oracle to answer them, of the Encoder to produce a visually grounded representation of the dialogue history and of the Guesser to select the most probable target object given the image and the dialogue history.

⋆ 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

Fig. 1: GuessWhat?! human dialogues are short and with a clear division of roles between players; most of the last questions are answered positively, are long, and contain details suitable to guess the target object.

In this paper, we disentangle the compressed task-success evaluation and focus on the ability of the Encoder to produce a dialogue hidden state representation that encodes the information necessary for the Guesser to select the target object. Therefore, we use the dialogue history generated by humans playing the referential game so to be sure of the quality of the questions and of the answers.

We run our analysis on GuessWhat?! since, as illustrated in Figure 1, its dialogues are quite simple: a sequence of rather short questions answered by Yes or No containing on average 30.1 (SD $\pm$ 17.6) tokens per dialogue. The simplicity of the dialogue structure makes the dataset suitable to be used as a diagnostic dataset.

In [23], the authors have shown that neural models are not sensitive to the order of turns in dialogues and conclude they do not use the history effectively. In GuessWhat?! dialogues the order in which questions have been asked is not crucial: we would be able to guess the target object even if the question-answer pairs in Figure 1 were provided in the reversed order. Indeed, we are able to use salient information independently of the turns where it occurs. We wonder whether the same holds for neural models trained to solve the GuessWhat?! task. As the example in the figure shows, the last question humans ask is usually quite rich in detail about the target object and is answered positively. We exploit these features of the dataset to run our in-depth analysis.

We compare encoders with respect to the architecture (Recurrent Neural Networks vs. Transformers), the input modalities (only language vs. language and vision), and the model background knowledge (trained from scratch vs. pre-trained and then fine-tuned on the downstream task). Our analysis shows that:

- the GuessWhat?! dataset can be used as a diagnostic dataset to scrutinize models' performance: dialogue length mirrors the level of difficulty of the game; most questions in the last turns are answered positively and are longer than earlier ones;
- Trasformers are less sensitive than Recurrent Neural Network based models to ther order in which QA pairs are provided;

– pre-trained Transformers detect salient information, within the dialogue history, independently of the position in which it is provided.

2 Related Work

Scrutinizing Visual Dialogues Encoding Interesting exploratory analysis has been carried out to understand Visual Question Answering (VQA) systems and highlight their weaknesses and strengths, e.g. [11, 25, 28, 12]. Less is known about how well grounded conversational models encode the dialogue history.

In [23], the authors study how neural dialogue models encode the dialogue history when generating the next utterance. They show that neither recurrent nor transformer based architectures are sensitive to perturbations in the dialogue history and that Transformers are less sensitive than recurrent models to perturbations that scramble the conversational structure; furthermore, their findings suggest that models enhanced with attention mechanisms use more information from the dialogue history than their vanilla counterpart. We take inspiration from this study to understand how State-of-the-Art (SoA) models encode the visually grounded dialogues generated by humans while playing the GuessWhat?! game.

In [13], the authors show that in many reading comprehension datasets, that presumably require the combination of both questions and passages to predict the correct answer, models can achieve quite a good accuracy by using only part of the information provided. We investigate the role of each turn in GuessWhat?! human dialogues and to what extent models encode the strategy seen during training.

SoA LSTM Based Models on GuessWhat?! After the introduction of the supervised baseline model [33], several models have been proposed. They exploit either some form of reinforcement learning [22, 36, 37, 35, 6, 34, 21] or cooperative learning [26, 21]; in both cases, the model is first trained with the supervised learning regime and then the new paradigm is applied. This two-step process has been shown to reach higher task success than the supervised approach when the Questioner and Oracle models are put to play together. Since our focus is on the Guesser and we are evaluating it on human dialogues, we will compare models that have undergone only the supervised training step. We compare these recurrent models (based on LSTMs [24]) against models based on Transformers [32].

Transformer Based Models The last years have seen an increasing popularity of transformer based models trained on several tasks to reach task-agnostic multimodal representations [14, 17, 30, 2, 27, 20]. ViLBERT [17] has been recently extended by means of multi-task training involving 12 datasets which include GuessWhat?! [18] and has been fine-tuned to play the Answerer of VisDial [19]. Among these universal multimodal models, we choose LXMERT [30]. [3] propose methods for directly analyzing the attention heads aiming to understand whether they specialize in some specific foundational aspect (like syntactic relations) functional to the overall success of the model. We take inspiration from

their work to shed light on how Transformers, that we adapt to play the GuessWhat?! game, encode the dialogues.

3 Dataset

The GuessWhat?! dataset was collected via Amazon Mechanical Turk by [33]. It is an asymmetric game involving two human participants who see a real-world image taken from the MS-COCO dataset [15]. One of the participants (the Oracle) is assigned a target object in the image and the other participant (the Questioner) has to guess it by asking Yes/No questions to the Oracle. There are no time constraints to play the game.

The dataset contains 155K English dialogues about approximately 66K different images. The answers are respectively 52.2% No, 45.6% Yes, and 2.2% N/A (not applicable); the training set contains 108K datapoints and the validation and test sets 23K each. Dialogues contain on average 5.2 question-answer (QA) pairs and the vocabulary consists of around 4900 words; each game has at least 3 and at most 20 candidates. We evaluate models using human dialogues, selecting only the games on which humans have succeed finding the target and contain at most 10 turns (total number of dialogues used: 90K in training and around 18K both in validation and testing).[3]

We run a careful analysis of the dataset aiming to find features useful to better understand the performance of models. Although the overall number of Yes/No answers is balanced, the shorter the dialogues, the higher the percentage of Yes answers is: it goes from the 75% in dialogues with 2 turns to the 50% in the 5 turn cluster to the 35% in the 10 turn cluster. Interestingly, most of the questions in the last turns obtain a positive answer and these questions are on average longer than earlier ones (see Figure 1 for an example). A model that encodes these questions well has almost all the information to guess the target object without actually using the full dialogue history. Not all games are equally difficult: in shorter dialogues the area of the target object is bigger than the one of target objects in longer dialogues, and their target object is quite often a "person" – the most common target in the dataset; moreover, the number of distractors in longer dialogues is much higher. Hence, the length of a dialogue is a good proxy of the level of difficulty of the game. Figure 2 reports the statistics of the training set; similar ones characterize the validation and the test sets.

The length of the dialogue is a good proxy of the level of difficulty of the game. Figure 3 shows that longer dialogues contain more distractors and in particular more distractors of the same category of the target object, which are supposed to be especially challenging for the models, since each candidate object is represented simply by its category and coordinates. Moreover, the area occupied by target objects is smaller in longer dialogues and the most representative category among target objects ("person") is less frequent.

We will exploit these features of the dataset to scrutinize the behaviour of models.

[3] The dataset of human dialogues is available at `https://guesswhat.ai/download`.

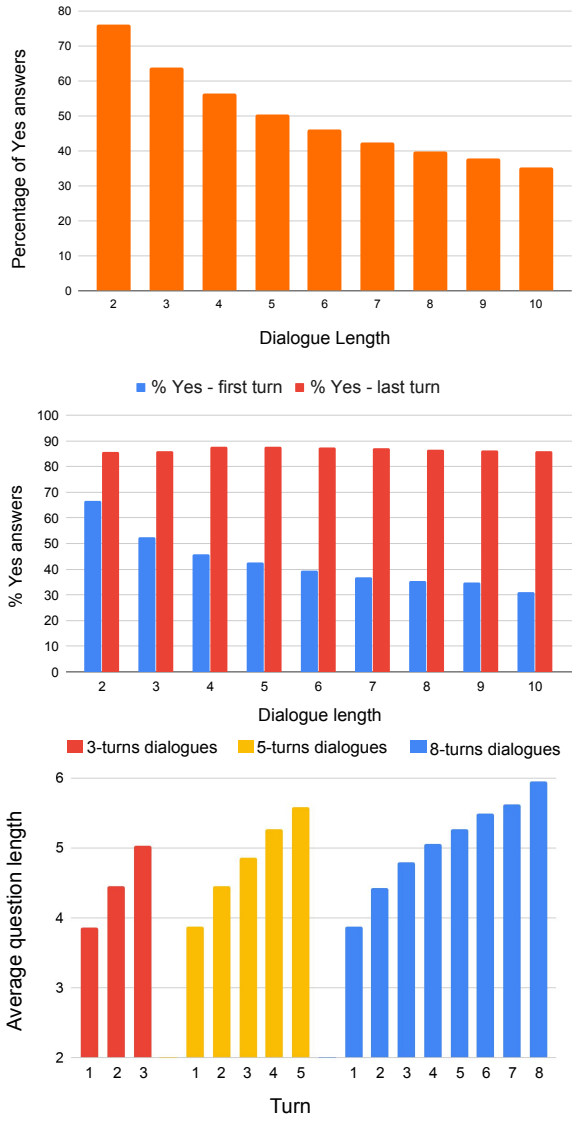

Fig. 2: Statistics of the training set (the test set has similar distributions). Dialogue length refers to the number of turns. **Up**: The distribution of Yes/No questions is very unbalanced across the clusters of games (the percentage of Yes answers is much higher in shorter dialogues); **Middle** In the large majority of games, the last question is answered positively; **Bottom**: The last questions are always longer (length of questions per turn for the clusters with dialogues having 3, 5, and 8 turns).

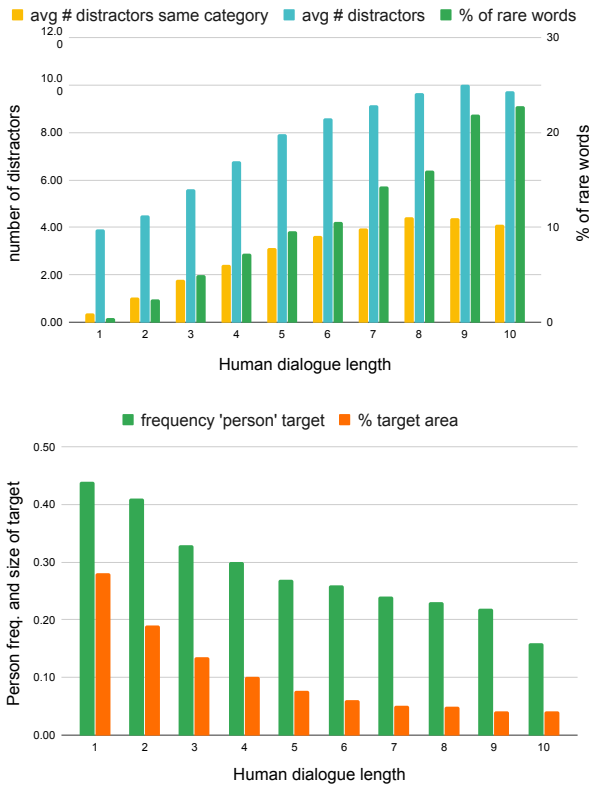

Fig. 3: **Up**: longer human dialogues contain more distractors and more distractors of the same category of the target object, and more rare words; **Down**: The distribution of target objects is unbalanced, since "person" is the most frequent target.

4 Models

All the evaluated models share the Guesser module proposed in [33]. 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 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. Figure 4 shows the shared skeleton.

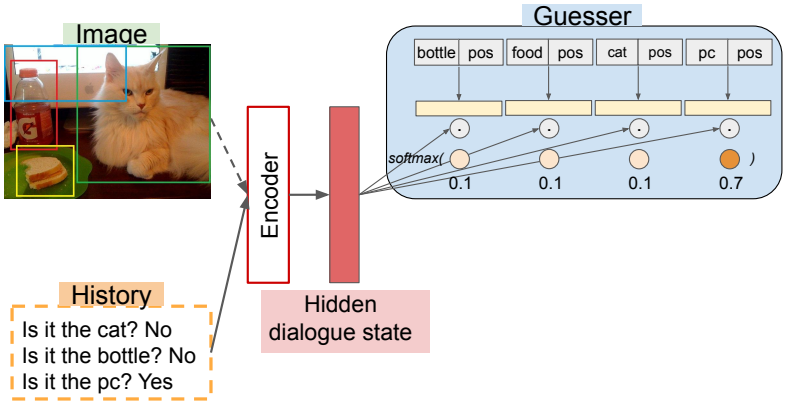

Fig. 4: Shared skeleton. Blind models do not receive the image as input.

4.1 Language Encoders

LSTM As in [33], the representations of the candidates are fused with the last hidden state obtained by an LSTM which processes only the dialogue history.

RoBERTa In the architecture of the model described above, we replace the LSTM with the robustly-optimized version of BERT [5], RoBERTa, a SoA universal transformer based encoder introduced in [16].[4] We use RoBERTa$_{BASE}$ which has been pre-trained on 16GB of English text trained for 500K steps to perform masked language modeling. It has 12 self-attention layers with 12 heads each. It uses three special tokens, namely CLS, which is taken to be the representation of the given sequence, SEP, which separates sequences, and EOS, which denotes the end of the input. We give the output corresponding to the CLS token to a linear layer and a *tanh* activation function to obtain the hidden state which is given to the Guesser. To study the impact of the pre-training phase, we have compared the publicly available pre-trained model, which we fine-tuned on GuessWhat?! (**RoBERTa**), against its counterpart trained from scratch only on the game (**RoBERTa-S**).

4.2 Multimodal Encoders

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 [9] to extract the visual vectors.

[4] We have also tried BERT, but we obtained higher accuracy with RoBERTa.

LXMERT To evaluate the performance of a universal multimodal encoder, we employ LXMERT (Learning Cross-Modality Encoder Representations from Transformers) [30]. 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 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). As RoBERTa, LXMERT uses the special tokens CLS and SEP. Differently from RoBERTa, LXMERT uses the special token SEP both to separate sequences and to denote the end of the textual input. LXMERT has been pre-trained on five tasks.[5] 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 as in RoBERTa. Similarly, we consider both the pre-trained version (**LXMERT**) and the one trained from scratch (**LXMERT-S**).

5 Experiments

We compare the models described above using human dialogues aiming to shed lights on how the encoders capture the information that is salient to guess the target object.

5.1 Task Success

Dialogues asked by human players of the GuessWhat?! games are expected to contain, together with the image they are about, the information necessary to detect the target object among the candidates. We refer to them as Ground Truth (GT) dialogues. As we can see in Table 1, the Guesser based on a blind encoder (LSTM or RoBERTa from scratch or pre-trained) obtains results higher than or comparable with V-LSTM.[6]

Table 2 reports the accuracy by clusters of games based on the dialogue length. All models reach a very high and similar accuracy in short games and differ more in longer ones. Most of the boost obtained by RoBERTa seems to come in longer dialogues where its from scratch version (RoBERTa-S) performs on a par with the other models.

[5] 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.

[6] The model proposed in [18] based on ViLBERT obtains an accuracy on GuessWhat?! with human dialogues of 65.04% when trained together with the other 11 tasks and 62.81% when trained only on it.

		GT	Reversed
BLIND	LSTM	64.7	56.0
	RoBERTa-S	64.2	57.8
	RoBERTa	**67.9**	66.5
MM	V-LSTM	64.5	51.3
	LXMERT-S	64.7	58.3
	LXMERT	64.7	60.3

Table 1: We compare the accuracy of models on the test set containing dialogues in the Ground Truth (GT) order of turns vs. the reversed order (reversed).

	LSTM	RoBERTa-S	RoBERTa	V-LSTM	LXMERT-S	LXMERT
All	64.7	64.2	67.9	64.5	64.7	64.7
3	72.5	72.7	75.3	71.9	73	73.8
5	59.3	58.3	60.1	59.3	59.2	58.7
8	47.3	45.1	51.0	47.2	46.8	43.3

Table 2: Accuracy with GT dialogues: results for all games, and for those of 3/5/8 dialogue length.

These results show that the human dialogue history alone is quite informative to accomplish the task. If we go back to the example in Figure 1, we realize it is possible to succeed in that game if we are given the dialogue only and are asked to select the target object (the scissor on the left) among candidates for which we are told the category and the coordinates – as it is the case for the Guesser.

In the following, we are running an in-depth analysis to understand whether models are able to identify salient information indipendently of the position in which they occur.

5.2 Are Models Sensitive to the Strategy Seen during Training?

In Section 3, we have seen that human dialogues tend to share a specific strategy, i.e. questions that are asked in first turns are rather short whereas those in the last turns provide relevant details about the most probable target object. We wonder whether the models under analysis become sensitive to the above-mentioned strategy and learn to focus on some turns more than others rather than on the actual salient QA pair.

Following [23], we perturb the dialogue history in the test set by reversing the order of turns from the last to the first one (reversed). Differently from them, given the nature of the GuessWhat?! dialogue history, we value positively models that are robust to this change in the dialogue history order. Our experiment (Table 1) shows that Transformers are less sensitive than LSTMs to the order in which QA pairs are provided. Interestingly, the pre-training phase seems to mitigate the effect of the change of the order even more: while RoBERTa has a drop of just -1.4, the accuracy of its from-scratch counterpart drops of -6.4.

In other words, **(pre-trained) Transformers seem to be able to identify salient information independently of the position in which it is provided within the dialogue history**.

5.3 The role of the last question

Table 3 reports the results of the models when receiving the dialogue history without the last turn. As we can see all models undergo a similar drop in accuracy. This means that all models identify the last turn as the most informative one equally well. It is worth noting that the superiority of RoBERTa compared to other models pops up even when removing the last turn, showing that RoBERTa is indeed able to better encode the full dialogue history and not only parts of it. This holds for different dialogue lengths as shown in the Table. On average, removing the last turn affects more the performance of multimodal models. For 5-turns long dialogues, the accuracy drops by -12.2 for blind models and by -14.3 for multimodal models. Similarly, or 8-Q dialogues the accuracy drops by -8 (blind) and -9.3 (multimodal).

Model	**3-Q**		**5-Q**		**8-Q**	
	All turns	**W/o last turn**	**All turns**	**W/o last turn**	**All turns**	**W/o last turn**
LSTM	72.5	53.4	59.3	46.8	47.3	38.4
RoBERTa-S	72.7	55.4	58.3	44.9	45	38.9
RoBERTa	75.3	58.2	60.1	49.3	51	42
V-LSTM	71.9	53.8	59.3	43.7	47.2	36.5
LXMERT-S	73	55.8	59.2	45	46.8	38.8
LXMERT	73.8	55.3	58.7	45.6	43.3	34.1

Table 3: Accuracy of the models when receiving all turns of the dialogue history and when removing the last turn for dialogues with 3, 5, and 8 turns.

5.4 How attention is distributed across turns

In Section 3 we have seen that the last turn is usually answered positively and it is quite informative to detect the target object. We wonder whether this is reflected on how models distribute their attention across turns within a dialogue. To this end, we analyze how much each turn contributes to the overall self-attention within a dialogue by summing the attention of each token within a turn. We run this analysis for LXMERT and RoBERTa in their various versions: **all models put more attention on the last turn** when the GT order of turns is given.

In Table 1, we have seen that Transformers are more robust than the other models when the dialogue history is presented in the reversed order (the first QA pair of the GT is presented as the last turn and the last QA pair is presented as first turn). Our analysis of the attention heads of RoBERTa and LXMERT

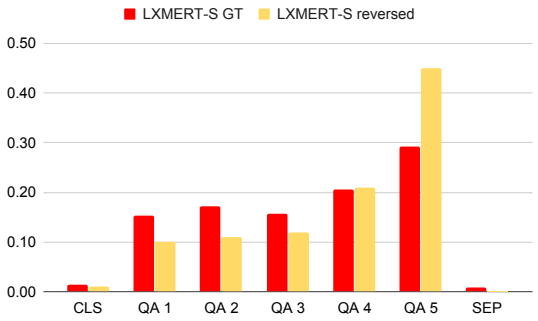

Fig. 5: Attention assigned by LXMERT-S to each turn in a dialogue when the dialogue history is given in the GT order (from QA1 to QA5) or in the reversed order (from QA5 to QA1).

shows that these models, both in their from scratch and pre-trained version, focus more on the question asked last **also in the reversed test set** where it is presented in the first position. This shows they are still able to identify the most salient information. In Figure 5, we report the attention per turn of LXMERT-S when receiving the GT and the reversed test set.

5.5 Details for reproducibility

We used the GuessWhat?! dataset in our experiments (`http://guesswhat.ai/download`). The dataset contains 155000 English dialogues about approximately 66000 different images. The Train split contains 108000 datapoints, the Validation split 23000 datapoints, and the Test split 23000 datapoints. We considered only the dialogues corresponding to the games succeeded by humans and having less or equal than 10 turns.

For training LSTM based models we adapted the source codes available at `https://github.com/shekharRavi/Beyond-Task-Success-NAACL2019` and at `https://github.com/GuessWhatGame/guesswhat/`. For training transformer based models we adapted the source code available at `https://github.com/huggingface/transformers`. The scripts for all the experiments and the modified models will be made available upon acceptance. For all models, we used the same hyperparameters of the original works. When adapting Transformers to the GuessWhat?! task, we scaled the representation of the CLS token from 768 to 512. We used PyTorch 1.0.1 for all models except for LSTM, for which we have used Tensorflow 1.3. All models are trained with Adam optimizer. For transformer based models we used a batch size equal to 16, a weight decay equal to 0.01, gradient clipping equal to 5, and a learning rate which is warmed up over the first 10% iterations to a peak value of 0.00001 and then linearly decayed.

Regarding the infrastructure, we used 1 Titan V GPU. LSTM based models took about 15 hours for completing 100 training epochs. Transformer based

models took about 4 days for completing 25 training epochs. Each experiment took about 10 minutes to evaluate the best trained models.

Details on the best epoch, the validation accuracy, and the number of parameters of each model are reported in table 4.

Model	**Best epoch**	**Validation accuracy**	**Parameters**
LSTM	18	65.6	5,030,144
RoBERTa	6	68.7	125,460,992
RoBERTa-S	13	64.7	125,460,992
V-LSTM	8	65.2	10,952,818
LXMERT-S	17	65.4	208,900,978
LXMERT	11	65.1	208,900,978

Table 4: Epoch, validation set accuracy and number of parameters for each best model.

6 Conclusion

Our detailed analysis of the GuessWhat?! dataset has revealed features of its games that we have exploited to run a diagnostic analysis of SoA models.

Our comparative analysis has shown that Trasformers are less sensitive than LSTMs to the order in which QA pairs are provided and that their pre-trained versions are even stronger in detecting salient information, within the dialogue history, independently of the position in which it is provided.

Furthermore, our results shows that RoBERTa is the encoder that provides the Guesser with the most informative representation of the dialogue history. Its advantage is particularly strong in longer dialogues. The dialogue contains already all the information necessary to guess the candidates: both with LSTM and trasformer based models the blind version obtain results higher than or comparable with their multimodal counterpart. We conjecture that this is due to the fact that the Guesser has access to the category of the target object. Important progress has been made on multimodal models since the introduction of the GuessWhat?! game. It would be interesting to see how SoA models would perform when they have to rely on visual information rather than raw category.

Acknowledgments

We kindly acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research at the University of Trento. We acknowledge SAP for sponsoring the work.

References

1. Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R.: The HCRC map task corpus. Language and Speech **34**, 351–366 (1991)
2. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Learning universal image-text representations (2019), arXiv:1909.11740
3. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? An analysis of BERT's attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 276–286 (2019)
4. Das, A., Kottur, S., Moura, J.M., Lee, S., Batra, D.: Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In: 2017 IEEE International Conference on Computer Vision. pp. 2951–2960 (2017)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, `https://www.aclweb.org/anthology/N19-1423`
6. Gan, Z., Cheng, Y., Kholy, A.E., Li, L., Liu, J., Gao, J.: Multi-step reasoning via recurrent dual attention for visual dialog. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6463–6474 (2019)
7. Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., Fernández, R.: The PhotoBook dataset: Building common ground through visually-grounded dialogue. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1895–1910 (Jul 2019). https://doi.org/10.18653/v1/P19-1184, `https://www.aclweb.org/anthology/P19-1184`
8. He, H., Balakrishnan, A., Eric, M., Liang, P.: Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 1766–1776 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Ilinykh, N., Zarrieß, S., Schlangen, D.: Tell Me More: A Dataset of Visual Scene Description Sequences. In: Proceedings of the 12th International Conference on Natural Language Generation. pp. 152–157 (2019), `https://www.aclweb.org/anthology/W19-8621`
11. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. abs/1612.06890 (2017)
12. Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1965–1973 (2017)
13. Kaushik, D., Lipton, Z.C.: How much reading does reading comprehension require? A critical investigation of popular benchmarks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 5010–5015 (2018)

14. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: A simple and performant baseline for vision and language (2019), arXiv:1908.03557
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of ECCV (European Conference on Computer Vision). pp. 740–755 (2014)
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoERTa: A robustly optimized bert pretraining approach (2019), arXiv:1907.11692
17. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems. pp. 13–23 (2019)
18. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: Proceedings of CVPR (2020)
19. Murahari, V., Batra, D., Parikh, D., Das, A.: Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. arXiv preprint arXiv:1912.02379 (2019)
20. amd Nan Duan, G.L., Fang, Y., Gong, M., Jiang, D., Zhou, M.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of AAAI (2020)
21. Pang, W., Wang, X.: Visual dialogue state tracking for question generation. In: Proceedings of 34th AAAI Conference on Artificial Intelligence (2020)
22. Sang-Woo, L., Tong, G., Sohee, Y., Jaejun, Y., Jung-Woo, H.: Large-scale answerer in questioner's mind for visual dialog question generation. In: Proceedings of International Conference on Learning Representations, ICLR (2019)
23. Sankar, C., Subramanian, S., Pal, C., Chandar, S., Bengio, Y.: Do neural dialog systems use the conversation history effectively? An empirical study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 32–37 (2019), https://www.aclweb.org/anthology/P19-1004
24. Schmidhuber, J., Hochreiter, S.: Long short-term memory. Neural Comput **9**(8), 1735–1780 (1997)
25. Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., Bernardi, R.: 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). pp. 255–265 (2017)
26. Shekhar, R., Venkatesh, A., Baumgärtner, T., Bruni, E., Plank, B., Bernardi, R., Fernández, R.: 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). pp. 2578–2587 (2019). https://doi.org/10.18653/v1/N19-1265, https://www.aclweb.org/anthology/N19-1265
27. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of generic visual-linguistic representations. In: ICLR (2020)
28. Suhr, A., Lewis, M., Yeh, J., Artzi, Y.: A corpus of natural language for visual reasoning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 217–223. Association for Computational Linguistics, Vancouver, Canada (July 2017), http://aclweb.org/anthology/P17-2034
29. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)

30. Tan, H., Bansal, M.: 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). pp. 5103–5114 (2019)
31. Udagawa, T., Aizawa, A.: A natural language corpus of common grounding under continuous and partially-observable context. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 7120–7127 (2019)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
33. de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.C.: GuessWhat?! Visual object discovery through multi-modal dialogue. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 5503–5512 (2017)
34. Yang, T., Zha, Z.J., Zhang, H.: Making history matter: History-advantage sequence training for visual dialog. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019)
35. Zhang, J., Zhao, T., Yu, Z.: Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. pp. 140–150 (2018), https://www.aclweb.org/anthology/W18-5015
36. Zhang, J., Wu, Q., Shen, C., Zhang, J., Lu, J., van den Hengel, A.: Goal-oriented visual question generation via intermediate rewards. In: Proceedings of the European Conference of Computer Vision (ECCV). pp. 186–201 (2018)
37. Zhao, R., Tresp, V.: Improving goal-oriented visual dialog agents via advanced recurrent nets with tempered policy gradient. In: Proceedings of IJCAI (2018)