“I’ve Seen Things You People Wouldn’t Believe”: Hallucinating Entities in GuessWhat?!

Alberto Testoni
DISI, University of Trento
Trento, Italy
alberto.testoni@unitn.it

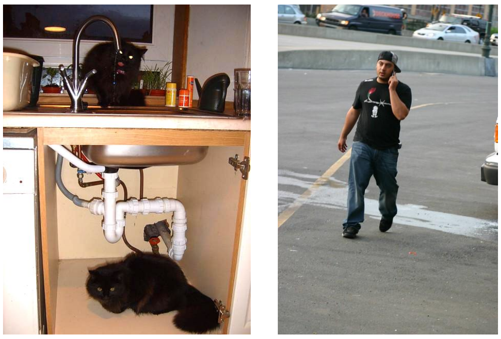
Raffaella Bernardi
CIMeC, DISI, University of Trento
Rovereto, Italy
raffaella.bernardi@unitn.it

Abstract

Natural language generation systems have witnessed important progress in the last years, but they are shown to generate tokens that are unrelated to the source input. This problem affects computational models in many NLP tasks, and it is particularly unpleasant in multimodal systems. In this work, we assess the rate of object hallucination in multimodal conversational agents playing the GuessWhat?! referential game. Better visual processing has been shown to mitigate this issue in image captioning; hence, we adapt to the GuessWhat?! task the best visual processing models at disposal, and propose two new models to play the Questioner agent. We show that the new models generate few hallucinations compared to other renowned models available in the literature. Moreover, their hallucinations are less severe (affect task-accuracy less) and are more human-like. We also analyse where hallucinations tend to occur more often through the dialogue: hallucinations are less frequent in earlier turns, cause a cascade hallucination effect, and are often preceded by negative answers, which have been shown to be harder to ground.

1 Introduction

Recent years have witnessed important progress in the quality of the output generated by deep neural network architectures. Although it is not easy to evaluate the output of natural language generation systems, some features clearly deteriorate their value, making these systems hardly employable in real-world scenarios. Crucially, state-of-the-art models are shown to generate words that are not consistent with the source inputs. This issue is generally referred to as *hallucination*.



is it a dog ? no
is it a chair ? no
is it a fridge ? no
is it a cup ? yes
on the right? yes

is it a person ? no
is it a skateboard ? no
is it a car ? yes
is it white ? no
is it green ? no

Figure 1: Hallucinations generated by the GDSE model playing GuessWhat?!. Note that the dialogue on the right also contains a question referring to an attribute (*green*) that is not related to the source image. In this paper, however, we focus only on entity hallucination.

This phenomenon applies to different NLP tasks and neural architectures. It has been explored in summarization (Kryscinski et al., 2020; Nan et al., 2021), machine translation (Koehn and Knowles, 2017; Nguyen and Chiang, 2018), and image captioning (Rohrbach et al., 2018). Hallucinating entities is particularly harmful in multimodal systems. MacLeod et al. (2017) study how blind people experience automatically generated captions describing images. The authors found that many participants in this study value more the correctness of the caption compared to a fine-grained description of the image, thus providing evidence that hallucination represents a major issue.

The problem of generating hallucinated entities is thus a relevant challenge for the community, but it is an understudied problem in multimodal conversational agents. Apart from sharing similarities with the image captioning task (e.g., generating tokens that are grounded in the image), visual dialogues have the peculiarity of being based on a complex dialogic structure. In this paper, we compare the output of neural models playing the GuessWhat?! referential visual game (de Vries et al.,

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 101–111
August 5–6, 2021. ©2021 Association for Computational Linguistics

2017). We consider different models based on the encoder-decoder framework (Sutskever et al., 2014), and we compare different architectures, with different processing of the visual input, to serve as the Encoder and Decoder modules. We adapt two multimodal models based on Transformers (Vaswani et al., 2017) to play the GuessWhat?! Questioner agent, and we highlight their strengths and weaknesses with a focus on the issue of hallucination. Examples of GuessWhat?! dialogues containing hallucinations are reported in Figure 1. We use the CHAIR metric proposed in Rohrbach et al. (2018) to quantify the number of hallucinations in the generated dialogues.

Our results confirm that hallucination heavily affects the output of generative models playing GuessWhat?!, but pre-trained Transformers (used both as Encoder and Decoder) show a consistent improvement in this respect. Moreover, our results reveal that the rate of object hallucination increases across the dialogue turns. Hallucinations frequently appear in consecutive turns and are more likely to occur after negative answers. Finally, we carry out an in-depth analysis in dialogues produced by human annotators. The main contributions of this paper can be summarized as follows:

- We investigate the issue of hallucination, an understudied problem in visual dialogue, by taking GuessWhat?! as a test-bed.
- We studied to what extent fine-grained visual representations reduce hallucinations in multimodal models.
- We show the importance of computing the CHAIR metric on models' and humans' text, and use this metric to guide a qualitative analysis to better understand the results.

2 Related Work

Hallucination in Language-only Tasks. Kryscinski et al. (2020); Nan et al. (2021) highlight the problem of factual inconsistency in abstractive summarization. This phenomenon occurs when a computational model generates a summary containing entities that do not appear in the source document. Kryscinski et al. (2020) propose a weakly-supervised, model-based approach to verify factual consistency and identify conflicts between source documents and generated summaries. Nan et al. (2021) design a set of new metrics to quantify the degree of entity hallucination in summaries. Interestingly, the authors found that ground truth summaries in the training data contain hallucinations. Similarly to these works, we focus on entity hallucination, and on inconsistencies with respect to the visual context, instead of the linguistic one.

Neural machine translation systems are also prone to such kinds of hallucinations, i.e. translations that are grammatically correct, but crucially unrelated to the source input (Koehn and Knowles, 2017; Nguyen and Chiang, 2018). A recent work (Müller et al., 2020) found that neural machine translation systems evaluated on out-of-domain test sets generate translations that are fluent but unrelated to the source sentence. These works focus on words belonging to different parts of speech, like proper nouns, adjectives, and verbs, while we only focus on entity hallucination and leave for future work the analysis of attribute hallucination.

Hallucination in Vision & Language. The generation of hallucinations affects also Multimodal Machine Translation systems. Lala and Specia (2018) highlight the issues that may arise while translating ambiguous or polysemic words given a visual context. Rohrbach et al. (2018) investigate the problem of object hallucination in image captioning, the closest task to our work. The authors propose a new metric (CHAIR) to quantify the extent to which machine-generated captions contain hallucinated entities. The authors found overreliance on language priors as a plausible cause of hallucinated tokens in the generated captions. Moreover, they found that models with a more reliable visual representation hallucinate less, suggesting that a robust processing of the visual input is important for reducing hallucination. We use the CHAIR metric to evaluate different models, and look at the role of different visual representations. A recent work (Xiao and Wang, 2021) investigates the relationship between hallucinations and predictive uncertainty in image captioning and data-to-text generation. The authors found that higher predictive uncertainty leads to a higher chance of hallucinating entities. We leave this kind of analysis for future work.

Visual Dialogues Evaluation. Among the visual dialogue datasets and tasks available (e.g., de Vries et al. 2017; Mostafazadeh et al. 2017; Das et al. 2017; Haber et al. 2019), we chose a task-oriented referential game, GuessWhat?! (de Vries et al., 2017). Task-oriented conversational agents generate dialogues to reach a goal, thus the presence of hallucinations considerably hurt the performance of such systems. We chose GuessWhat?! because of the simplicity of its dialogue structure (polar question-answer pairs). Recent work in the literature highlights the inability of the accuracy in the guessing task to serve as a good proxy of the quality of the underlying dialogues, with a particular focus on surface-level features such as the presence of repetitions (Shekhar et al., 2019; Murahari et al., 2019; Testoni et al., 2019). We extend this claim by looking at hallucination, an under-studied but crucial issue in Visual Dialogues.

3 Task and Metrics

Task The GuessWhat?! game (de Vries et al., 2017) is a cooperative two-player game in English based on a referential communication task where two players collaborate to identify a referent object in an image. This setting has been extensively used in human-human collaborative dialogue (e.g., Clark 1996; Yule 2013). GuessWhat?! is an asymmetric game involving two human participants who see a real-world image. One of the participants (the Oracle) is secretly assigned a target object within the image, and the other participant (the Questioner) has to guess it by asking binary (Yes/No) questions to the Oracle. The GuessWhat?! dataset is composed of more than 150k human-human dialogues containing an average of 5.3 questions in natural language created by annotators playing the game on MSCOCO images (Lin et al., 2014). Successful dialogues consist of around 135K dialogues grounded on about 63K unique MSCOCO images.

Metrics The first metric we consider is the raw accuracy in guessing the target object among the list of candidate objects. Secondly, to quantify the extent to which different models hallucinate entities during the dialogue, we compute the CHAIR metric *(Caption Hallucination Assessment with Image Relevance)* proposed in Rohrbach et al. (2018) for image captioning. This metric has two variants: *CHAIR-i* (per-instance), defined as the number of hallucinated objects in a sequence divided by the total number of objects mentioned, and *CHAIR-s* (per-sentence), defined as the number of sequences with at last one hallucinated entity divided by the total number of sequences. We use the same two variants of the CHAIR metric to evaluate the dialogues generated by models playing GuessWhat?!. This metric exploits the 80 MSCOCO objects which appear in the MSCOCO segmentation challenge, extended with entities mentioned in ground-truth captions, together with a list of synonyms for MSCOCO objects. We compute CHAIR for both machine-generated and human dialogues from the GuessWhat?! test set (referred to as HUMAN in the following). Computing CHAIR on human dialogues allows us to identify possible misclassification in the MSCOCO annotation and establish an upper bound for models' performance.

4 Models

To allow for a fair comparison of different Questioner models, we use the same Oracle and Guesser models in all our experiments. Following de Vries et al. (2017), we employ distinct computational models for each of the three key tasks: answering questions (Oracle), guessing the target (Guesser), and asking questions (Questioner).

4.1 Oracle

We use the baseline Oracle model proposed in de Vries et al. (2017). The model receives as input the embedding of the target object category, its spatial coordinates, and the question to be answered encoded by a dedicated Long-Short-Term Memory (LSTM) network. These three embeddings are concatenated and fed to a Multi-Layer Perceptron (MLP) that gives an answer (Yes, No, N/A).

4.2 Guesser

We use the state-of-the-art multimodal Guesser model proposed in Greco et al. (2021a) (Figure 2 bottom).[1] This Guesser is based on LXMERT (Tan and Bansal, 2019), a powerful multimodal Transformer model that is fine-tuned on the GuessWhat?! guesser task using successful human dialogues. LXMERT represents the visual input by the set of position-aware object embeddings for the 36 most salient regions detected by a Faster R-CNN network, and the text by position-aware randomly-initialized word embeddings. LXMERT has self-attention and cross-attention layers to merge and enhance the information coming from the two modalities to create a joint representation. LXMERT uses a special tokens CLS and the embedding corresponding to this token is considered a representation of the given sequence. LXMERT has been

[1] https://github.com/claudiogreco/aixia2021

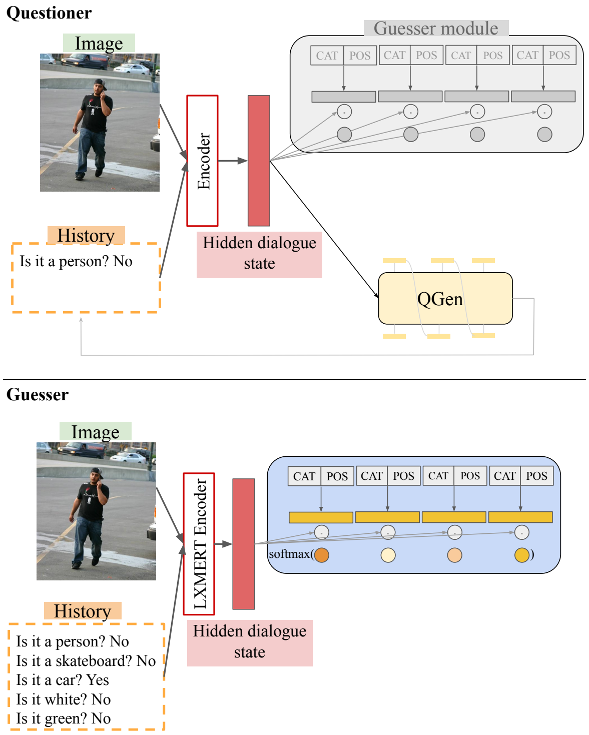


Figure 2: Skeleton architecture of the Questioner and Guesser models.

pre-trained on five tasks.[2] For the Guesser task, candidate objects are represented by the embeddings obtained via an MLP starting from the category and spatial coordinates of each candidate object. The representations so obtained are used to compute dot products with the embedding corresponding to the special token [CLS]. The scores of each candidate object are given to a softmax classifier to choose the object with the highest probability.

4.3 Questioner Models

In order to study the effect of a different (and more fine-grained) processing of the visual input, we compare two models already presented in the literature (BL and GDSE) with two Transformer-based multimodal models (LXMERT-GDSE and VLP) that we adapt to play the GuessWhat?! Questioner task. The architecture shared by the Questioner models is depicted in Figure 2. All the models discussed in the paper (except for BL) are trained to perform both the Questioner and the Guesser tasks in a multi-task fashion. For a fair comparison, we compute the accuracy in the guessing task using the same Guesser and Oracle models described above, and we use the Questioner models only to generate questions.

BL. The first model we consider is the baseline Questioner model proposed in de Vries et al. (2017). This model is implemented as a Recurrent Neural Network (RNN) with a transition function handled with LSTM, on which a probabilistic sequence model is built with a Softmax classifier. At each time step in the dialogue, the model receives as input the raw image and the dialogue history and generates the next question. The image is encoded by extracting its VGG-16 features (Simonyan and Zisserman, 2014). We consider the version of the model trained in a supervised learning fashion.

GDSE. The Visually-Grounded Dialogue State Encoder (GDSE) model was proposed in Shekhar et al. (2019). We consider the version of GDSE trained in a supervised learning fashion. The model uses a visually grounded dialogue state that takes the visual features of the input image and each question-answer pair in the dialogue history to create a shared representation used both for generating a follow-up question (QGen module) and guessing the target object (Guesser module) in a multi-task learning scenario. More specifically, the visual features are extracted with a ResNet-152 network (He et al., 2016) and the dialogue history is encoded with an LSTM network. The QGen component is optimized with the Log Likelihood of the training dialogues, and the Guesser computes a score for each candidate object by performing the dot product between a visually grounded dialogue state and each object representation. In this work, we use GDSE only to generate dialogues, since the guessing part relies on the Guesser described above.

LXMERT-GDSE. Similarly to GDSE, we implement a new Questioner model based on the LXMERT architecture described above. In this model, we take the representation corresponding to the [CLS] token as the hidden dialogue state and, similarly to GDSE, we feed this representation as input to both a QGen module (an LSTM-based decoder) and a Guesser module. We fine-tune the pre-trained LXMERT on GuessWhat?!. Again, we use this model only to generate dialogues.

VLP. Finally, we develop a Questioner model based on VLP (Zhou et al., 2020), a powerful multimodal Encoder-Decoder Transformer architecture pre-trained on image captioning. VLP is a single

[2] 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

	CHAIR-s	CHAIR-i
BL	29.53	27.32
GDSE	30.31	16.57
LXMERT-GDSE	14.98	8.83
VLP	10.78	6.60
HUMAN	7.45	4.11

Table 1: CHAIR results on human and machine-generated dialogues on the GuessWhat?! test set.

stream unified encoder-decoder architecture: its Transformer backbone is the same as BERT-base (Devlin et al., 2019). VLP represents each input image as 100 object regions extracted from a variant of Faster RCNN (Ren et al., 2016) pre-trained on Visual Genome (Krishna et al., 2017; Anderson et al., 2018), together with the class likelihood on the 1600 object categories defined in Anderson et al. (2018) as region object labels. During pre-training, the model uses a masked language modelling objective. During inference, in order to generate a sequence token-by-token, VLP masks sequentially each token by appending a special token [SEP] at the end of the sequence. VLP is trained to predict a [STOP] token at the end of the sequence, so it can stop the generation of new tokens before reaching the maximum length. We fine-tune the version of VLP pre-trained on image captioning to play the GuessWhat?! game.[3]

Implementational Details We evaluate BL, GDSE, LXMERT-GDSE, and VLP on the GuessWhat?! test set. We let the models generate 5 question-answer pairs for each game (i.e., similar to the average number of questions asked by human players in GuessWhat?!). Note that VLP is trained to predict a [STOP] token, so it can stop asking questions before reaching the 5th turn. We found that, on average, VLP asks 4 questions in a dialogue. We compare the models with respect to their accuracy in the guessing game and the quality of the generated dialogues, with a focus on the phenomenon of hallucination.

5 Experiments and Results

5.1 CHAIR Results

We compare different models against the CHAIR metric. As Table 1 shows, BL and GDSE generate many hallucinated entities, both at the sentence and instance level. On the other hand, LXMERT-GDSE and especially VLP generate less than half of the hallucinations of the previous models. Recall that LXMERT-GDSE encodes the image with 36 regions. The best model, VLP, encodes each image region together with the class likelihood on 1600 object categories, so it has access to a suitable source of information to ground the generated tokens in the image. The fine-grained visual input representation of these two models leads to a consistent reduction in hallucinations, confirming that a strong visual processing is critical for avoiding hallucination (Rohrbach et al., 2018).[4] Table 1 shows that also dialogues generated by human players contain some hallucinated entities according to the CHAIR metric, thus establishing an upper bound for models' performance. VLP is closest to the ceiling set by humans.

5.2 Performance-based Analysis

We expect the Guesser to perform better when the dialogues contain few hallucinations. In fact, as reported in Table 2, the best result is obtained with human dialogues. However, among the machine-generated dialogues, we found that the baseline model (which is shown to generate many hallucinations – Table 1) outperforms the others. We believe that this result is due to the over-reliance of the baseline model on location questions, as highlighted in Shekhar et al. (2019). These questions, though are helpful for the model to identify the target object, make its dialogues sound unnatural when asked too often. We think this confirms the failure of the overall accuracy to serve as a proxy for the quality of the generated dialogues, as recently highlighted in Shekhar et al. (2019) and Testoni and Bernardi (2021).

In order to understand this discrepancy between accuracy and hallucination, we compared dialogues that contain at least one hallucinated entity with dialogues not affected by this issue. We found that the presence of hallucinations clearly deteriorates the accuracy in the game: as shown in Table 2, dialogues containing at least one hallucinated token lead to lower accuracy in guessing the target object compared to games that do not contain

[3] Simultaneously, Suglia et al. (2021) have adapted VLP to the GuessWhat?! game; they use a different training regime, and they focus on VQA as a downstream task via transfer learning.

[4] We also computed the CHAIR metric for the model proposed in Suglia et al. (2020). We obtained from the authors the dialogues generated on a subset of the GuessWhat?! test set (corresponding to around 39% of the test set). Accuracy: 40.69%. CHAIR-s: 22.88, CHAIR-i: 12.41.

	Test Set Accuracy (5Q)	w/o hallucination	with hallucination
BL	52.36	55.39	45.15
GDSE	44.85	47.26	39.29
LXMERT-GDSE	48.53	49.62	42.38
VLP	47.55	48.18	42.34
HUMAN	69.17	69.49	64.16

Table 2: Accuracy reached by the Guesser model when receiving as input dialogues generated by different Questioner models playing with the same Oracle or full human dialogues from the GuessWhat?! test set. *'w/o hallucination'* refers to the accuracy on the subset of games that do not contain any hallucinated tokens. *'with hallucination'* refers to the accuracy on the subset of games that contain at least one hallucination.

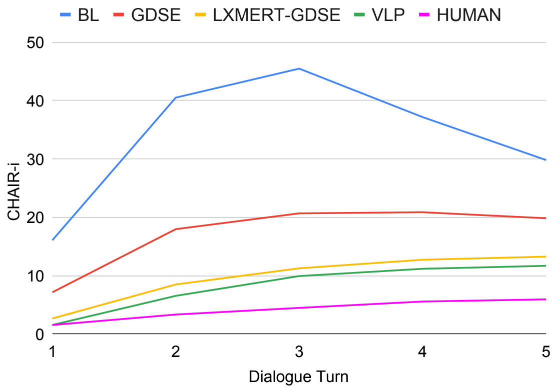


Figure 3: Per-turn CHAIR-i score for machine-generated and human dialogues. Models generate 5 questions. Hallucinated tokens tend to show up less in earlier turns.

hallucinations. Interestingly, the drop in accuracy between the two settings reveals a degree of severity from the severe hallucination encountered in BL (-10%) to the mild one in LXMERT-GDSE and GDSE (-7%) till the almost harmless one in VLP and HUMAN (-5%).[5]

5.3 Analysis of Hallucination Occurrences

In Rohrbach et al. (2018), the authors found that hallucinated entities tend to be mentioned towards the end of the sentence, and they hypothesise that some of the preceding words in the image caption may have triggered hallucination. To understand whether a similar phenomenon occurs also in visual dialogues, we run a per-turn analysis on the GuessWhat?! dialogues by computing the CHAIR-i metric after each question-answer pair. As we can see from Figure 3, hallucinations tend to show up in the latest turns of the dialogue, while the first turn contains few hallucinations.

	% consecutive halluc.
BL	24.13
GDSE	34.82
LXMERT-GDSE	38.65
VLP	25.50
HUMAN	8.09

Table 3: Percentage of hallucinated tokens appearing in consecutive turns of the dialogue.

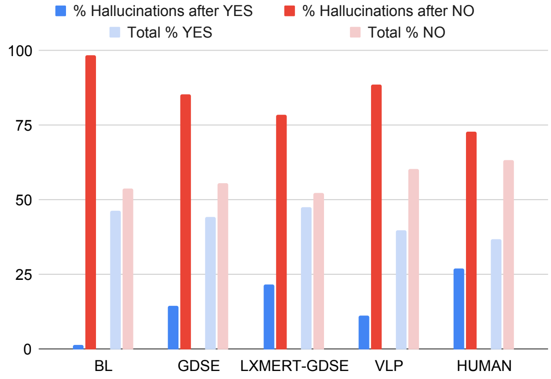


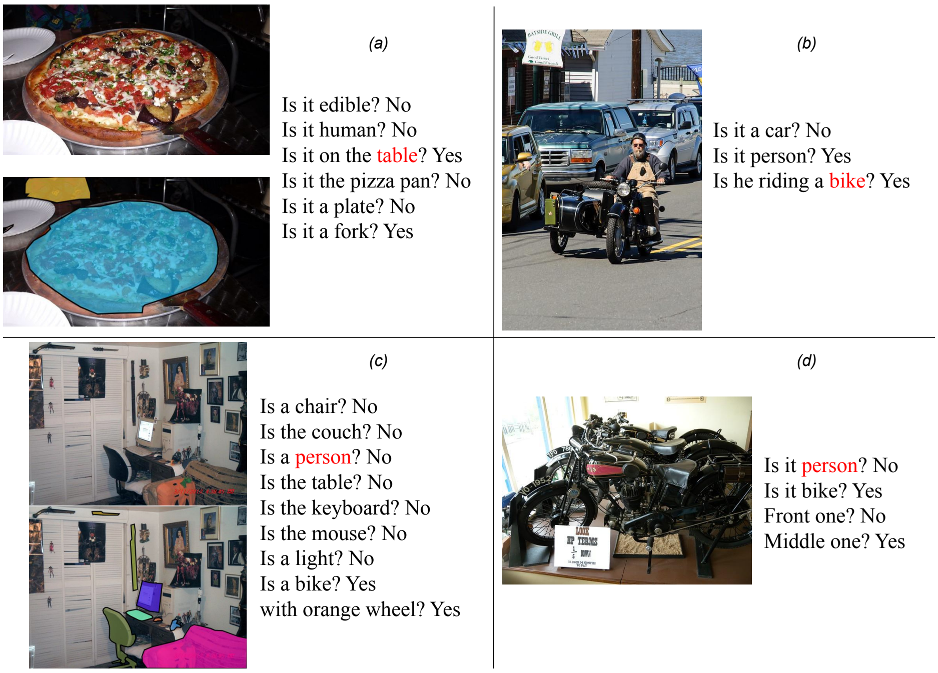
Figure 4: Percentage of hallucinated tokens appearing after a positive vs. negative answer. In light colours, we report the overall distribution of positive/negative answers in the output. The two distributions differ significantly, and this difference is particularly pronounced in machine-generated data.

To investigate the effect of hallucinations on follow-up turns, we study how the Question Generator and the Encoder modules are affected by this issue. To study the effect of hallucinations on the Question Generator, we compute how often hallucinated tokens occur in consecutive turns, i.e. the percentage of turns consisting of two consecutive questions containing at least one hallucination each, over all the turns containing at least one hallucination. As we can see from Table 3, for all the models

[5] We have also compared the accuracy in the two settings by fixing the number of candidate objects, i.e., by comparing games of the same difficulty. We found the same difference between the two settings, confirming the validity of our claim.

BL		GDSE		LXMERT-GDSE		VLP		HUMAN	
person	2803	chair	1649	bottle	716	table	480	table	389
couch	1113	person	1525	table	488	chair	462	bike	237
table	656	table	1483	bike	375	bike	352	person	211
chair	538	car	629	book	362	bottle	315	car	91
computer	404	bottle	605	cup	320	person	223	chair	88
bike	332	bench	468	bear	310	cup	220	bottle	83
car	229	book	468	chair	301	book	157	bowl	73
sink	224	phone	413	fridge	198	car	140	bear	60
dog	182	cup	376	car	195	bowl	111	cup	58
bear	171	dog	296	ball	186	ball	100	truck	54
keyboard	161	boat	255	person	163	bear	79	book	51

Table 4: Most frequent hallucinated MSCOCO categories for machine-generated and human dialogues, together with their raw frequency.



(a)
Is it edible? No
Is it human? No
Is it on the table? Yes
Is it the pizza pan? No
Is it a plate? No
Is it a fork? Yes

(b)
Is it a car? No
Is it person? Yes
Is he riding a bike? Yes

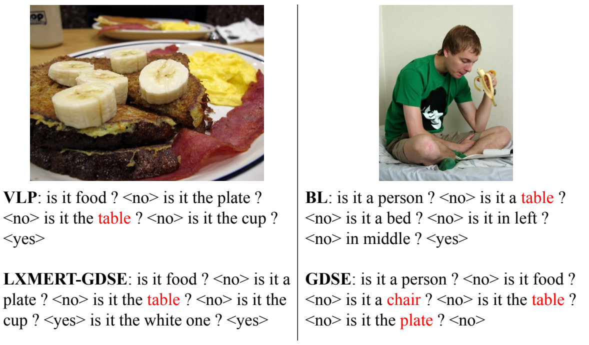
(c)
Is a chair? No
Is the couch? No
Is a person? No
Is the table? No
Is the keyboard? No
Is the mouse? No
Is a light? No
Is a bike? Yes
with orange wheel? Yes

(d)
Is it person? No
Is it bike? Yes
Front one? No
Middle one? Yes

Figure 5: Tokens counted as ‘hallucinated’ (in red) observed in human dialogues. *(a)*: the object ‘table’ is not present in MSCOCO segmentation. *(b)*: the human annotator refers to the motorcycle with ‘bike’, while they are different entities in the MSCOCO categories. *(c)*: people in paintings are not annotated. *(d)*: the dialogue contains an unrelated question.

we considered, a large part of the hallucinated tokens appear in consecutive turns, corroborating the hypothesis of Rohrbach et al. (2018) that hallucinations may cause a *cascade* effect. Crucially, in human dialogues this is not the case.

Another crucial component of the systems under analysis is the Encoder module, which plays a key role in processing the dialogue history. In Greco et al. (2021b), the authors found that computational models playing the GuessWhat?! guessing task on human dialogues struggle to profit from negatively answered questions, even when they are crucial to succeed in the game. Inspired by these findings, Figure 4 reports the percentage of hallucinations occurring *after* a positive vs. negative answer, compared with the overall distribution of answers in the generated dialogues. As we can see, hallucinations occur much more frequently after a negative answer than after a positive one, compared with the overall distribution. While in human dialogues the two answer distributions do not differ much, machine-generated dialogues have a clear tendency to generate hallucinations after a negative answer. In the baseline model, in particular, almost all hallucinated entities appear after a negative answer, while positive and negative answers are equally distributed in the generated dialogues. We conjecture that the failure in grounding negatively answered questions is behind the generation of hallucinations in the subsequent turns.



VLP: is it food ? <no> is it the plate ? <no> is it the table ? <no> is it the cup ? <yes>

LXMERT-GDSE: is it food ? <no> is it a plate ? <no> is it the table ? <no> is it the cup ? <yes> is it the white one ? <yes>

BL: is it a person ? <no> is it a table ? <no> is it a bed ? <no> is it in left ? <no> in middle ? <yes>

GDSE: is it a person ? <no> is it food ? <no> is it a chair ? <no> is it the table ? <no> is it the plate ? <no>

Figure 6: Examples of machine-generated dialogues containing hallucinations, focusing on the entity *table*. On the left, examples of *fake hallucinations* similar to those observed in human dialogues. On the right, examples of *real hallucinations*.

5.4 Qualitative Analysis

Table 1 shows that VLP is the model that is closest to humans in terms of the number of hallucinations in the output. Here, we wonder whether the hallucinations generated by VLP are human-like, i.e., whether they are similar to the ones appearing in human dialogues. The CHAIR metric relies on the MSCOCO segmentation annotation, which is not an exhaustive source for the wide variety of objects present in MSCOCO images. For this reason, Rohrbach et al. (2018) augmented the MSCOCO segmentation annotation with entities mentioned in ground truth captions. While in image captioning human annotators tend to mostly refer to salient objects in the image, in referential visual games, given the nature of the task, human annotators also refer to objects that are globally not salient, but are discriminative to perform the task. We believe that in this scenario it becomes crucial to apply the CHAIR metric both to machine-generated and human dialogues so to run a comparative analysis. Below we report what our comparison reveals.

Table 4 reports the most frequent hallucinated MSCOCO categories for each model and for humans, together with their raw frequency. We have run a manual inspection of human dialogues containing hallucinations based on the CHAIR metric, and found that in many cases they are *fake hallucinations* – they are due to missing labels in the annotation used to compute CHAIR. Figure 5-*a* reports an example with the hallucinated word "*table*": common sense would suggest the pizza is on the table, even if the latter is not visible; hence it is understandable that human players refer to it in the dialogue. The case of the word "*bike*" is illustrated by the example in 5-*b*, where rather than a hallucination, we simply have a not rigorous use of the work "*bike*" to refer to motorbikes. Finally, Figure 5-*c* illustrates why "*person*" appears in the top list of the hallucinated word: human players in their dialogues refer to entities in the paintings (in this case "*person*") which are rarely annotated in MSCOCO. Through our manual inspection of human dialogues, we have found also cases of *real hallucinations*. In most of these cases, the hallucinated entity is *person* and it occurs in the first turn – as illustrated by the example in Figure 5-*d*.

Our quantitative analysis (Table 4) suggests that entities hallucinated by VLP are similar to those appearing in human dialogues, indicating that some of them may count as *fake hallucinations*. Instead, the other models frequently hallucinate entities that are not in the human hallucination list or have low frequency; we conjecture this means that the rate of *real hallucinations* is lower for VLP than for the other models. To verify this hypothesis, we manually checked the hallucinations most frequently appearing in dialogues generated by models, and we found that, as suggested by the patterns in Table 4, VLP hallucinations are often *fake*, while BL and GDSE ones are not; LXMRT-GDSE dialogues stand in between. For instance, the example in Figure 6 illustrates a case of *fake hallucination* for VLP and LXMERT-GDSE and of *real hallucination* for the other two models.

6 Conclusion

Entity hallucination is one of the major problems that affect natural language generation systems in many NLP tasks, from machine translation to image captioning. Generating tokens that are not related to the source data compromises the possibility to use these systems in real-world scenarios. In this work, we explore to what extent this problem affects multimodal conversation agents playing the GuessWhat?! referential guessing game. We adapt two multimodal Transformer-based models to play the GuessWhat?! Questioner agent based on multimodal Transformers architectures (LXMERT-GDSE and VLP), and we compare their output with the widely used GDSE model (Shekhar et al., 2019) and the baseline model in de Vries et al. (2017). We adapt the CHAIR metric proposed in Rohrbach et al. (2018) for image captioning to assess the models' rate of object hallucination. Our analysis confirms recent findings about the inadequacy of the task success in the guessing game to serve as a good proxy of the quality of the generated dialogues. While all the models perform similarly in the GuessWhat?! game, the dialogues they generate differ dramatically. VLP and LXMERT-GDSE generate less than half of the hallucinations compared to GDSE and the baseline model, confirming the crucial role played by a strong visual processing to reduce hallucinations. The results of our in-depth analysis support the hypothesis in Rohrbach et al. (2018) that hallucinations tend to appear at the end of the sequence. Moreover, our results reveal that, in most cases, hallucinated tokens follow polar questions answered negatively. We conjecture this result is connected with our findings about the difficulties multimodal encoders have in grounding negation (Greco et al., 2021b); we believe further work is needed to understand the role of negation in visual dialogues. Finally, we highlight the importance of going beyond the simple CHAIR metric to evaluate the impact of hallucination. By running quantitative and qualitative analysis on human dialogues from the GuessWhat?! test set, we found that VLP is the model that generates less severe and more human-like hallucinations. Further work is needed to design new decoding strategies for natural language generation systems and to explore the relation between hallucination and repetitions, another major issue that heavily affects the quality of machine-generated data as recently highlighted in Testoni and Bernardi (2020). Moreover, as the example in Figure 1 (right) shows, attribute hallucination plays an important role in the quality of the generated output, and it has not received much attention from the research community.

Acknowledgments

We thank Claudio Greco for his help on the implementation side. We kindly acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research at the University of Trento. We thank the reviewers for their valuable comments.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Claudio Greco, Alberto Testoni, and Raffaella Bernardi. 2021a. Grounding dialogue history: Strengths and weaknesses of pre-trained transformers. In *AIxIA 2020 – Advances in Artificial Intelligence. Lecture Notes in Computer Science, vol 12414*, pages 263–279, Cham. Springer International Publishing.

Claudio Greco, Alberto Testoni, and Raffaella Bernardi. 2021b. "Yes" and "No": Visually grounded polar answers. *Visually Grounded Interaction and Language (ViGIL)*.

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, Florence, Italy. 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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 462–472.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

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, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1449–1454. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*.

Toan Nguyen and David Chiang. 2018. Improving lexical choice in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 334–343, New Orleans, Louisiana. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Alessandro Suglia, Yonatan Bisk, Ioannis Konstas, Antonio Vergari, Emanuele Bastianelli, Andrea Vanzo, and Oliver Lemon. 2021. An empirical study on the generalization power of neural representations learned via visual guessing games. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2135–2144, Online. Association for Computational Linguistics.

Alessandro Suglia, Antonio Vergari, Ioannis Konstas, Yonatan Bisk, Emanuele Bastianelli, Andrea Vanzo, and Oliver Lemon. 2020. Imagining grounded conceptual representations from perceptual information in situated guessing games. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1090–1102, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112.

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 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Alberto Testoni and Raffaella Bernardi. 2020. Overprotective training environments fall short at testing time: Let models contribute to their own training. In *Proceedings of the Seventh Italian Conference on*

Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Alberto Testoni and Raffaella Bernardi. 2021. The interplay of task success and dialogue quality: An in-depth evaluation in task-oriented visual dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2071–2082, Online. Association for Computational Linguistics.

Alberto Testoni, Ravi Shekhar, Raquel Fernández, and Raffaella Bernardi. 2019. The devil is in the details: 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

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.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

George Yule. 2013. *Referential communication tasks*. Routledge.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.