Photobook dataset: investigating the importance of history in a reference resolution model

Bob Borsboom (10802975), Emma Hokken (10572090), Linda Wouters (11000139)
Group 3

Abstract

In this report the Photobook dataset (Haber et al., 2019) with task-oriented visually grounded dialogues and corresponding reference chains, containing utterances (segments) from multiple game rounds referring to a single image by the same interlocutors, were analysed. Additionally, their pre-trained reference resolution models, that either did or did not use history, that classified target images given these reference chains were investigated. This was done by analysing the cosine similarities in the reference chains and the part-of-speech (POS) tag changes for segments that were correctly or incorrectly classified by their models. Furthermore, perturbations were performed on the reference chains to investigate how the models use grounding. It was found that nouns were important in the grounding and for the models, as well as the order of the history. Additionally, for the History models the history alone was not sufficient to deal with changed segments. Finally, the performance of the History model was shown to somewhat depend on the interlocutors of the segments and on the grounding process.

1 Introduction

When communicating, humans create a common ground which they use to communicate more efficiently (Stalnaker, 1978). Because this common ground differs between conversations and interlocutors, more information about the process of grounding can aid artificial dialogue agents to improve their communication.

Haber et al. (2019) investigated grounding in humans in a task-oriented, visually grounded dialogue. This was created by letting participant pairs play an online multi-round game where each was shown 6 images and they had to identify which image(s) the other player was shown as well by open written chat communication.

Additionally, Haber et al. (2019) used this dataset to create reference chains containing utterances that refer to a specific image over multiple rounds, forming a segment per round, from two interlocutors, creating a history of references to a single image. A simple model was then trained to predict the image given a segment and the image features, either with (History model) or without the game history (No History model) for that image.

In this research project, the reference chains and the pre-trained History and No History model from Haber et al. (2019) are used to investigate the grounding in the data and the importance of history on the models by separating the segments in conditions based on them being correctly or incorrectly classified by either model and analysing their differences.

Firstly, to investigate whether differences in grounding in the conditions could cause the models to classify them differently, the similarities of segments on an image between the first referring expression and later ones are determined. To investigate the difference in syntactic information between the condition, their part-of-speech tag distributions are analysed. Following this, four types of perturbations are performed on the reference chains. (1) The importance of the development of the grounding in the history is tested by changing the order of the segments from different rounds in a game (Sankar et al., 2019). (2) To investigate the dependence of the grounding development on the interlocutors, segments are exchanged between different participant pairs. (3) To test the effect of changing the text of a segment on its own, to control for the previous perturbation, segments referring to different images within a game are exchanged. Finally, (4) to find more information about how utterances become more effective through grounding, half or all of specific POS tags (nouns, verbs, adjectives and adverbs) are removed.

These experiments will lead to more insight in both the models proposed by Haber et al. (2019), and on how grounding develops and leads to more effective communication. It was found that nouns were important in the grounding and for the models, as well as the order of the history. Additionally, for the History models the history alone was not sufficient to deal with changed segments. Finally, the performance of the History model was shown to somewhat depend on the interlocutors of the segments and on the grounding process.

2 Related Work

Over the past few years, a couple large datasets for visually grounded dialogues have been created (Das et al., 2017b; De Vries et al., 2017). These datasets are both factored towards Visual Question Answering (Antol et al., 2015), where a user asks questions about one specific image, and another user answers these questions. The roles of Questioner and Answerer are set and there is not much interaction between users. Ilinykh et al. (2019) created a dataset where users are (virtually) located in the same room and are asked to find each other. This dataset takes navigational aspects into account and allows users to converse freely. The Photobook dataset (Haber et al., 2019) is similar in that way: users are allowed to converse freely, resulting in a much more natural conversation. Furthermore, users are shown 6 images at a time, resulting in (possible) rapid changes in topics.

To study whether neural generative models use conversation history effectively, Sankar et al. (2019) introduced perturbations in their utterances and found that the model outcome was not affected by this in any way. When shuffling all words in a sentence, for example, the model was still able to generate a coherent sentence. This showed that models, or at least the model they used, do not seem to use the conversation history. Furthermore, Das et al. (2017a) found that, during Visual Question Answering, humans look at (or pay attention to) very different regions than a Visual Question Answering model does. This suggests that models look at text in a different manner than humans do. These findings invoke a spark of interest: how does the model used for the Photobook dataset use history? What would happen to model performance if perturbations were introduced?

3 Approach

3.1 Dataset

The Photobook dataset (Haber et al., 2019), which was used as the basis for this report, consists of multiple games where two interlocutors both separately see 6 images per round. The participants' goal is to find whether the highlighted images are similar or different. They can reach their goal by having a written dialogue in an online chat. Each game consists of 5 rounds and each round contains images on the same subject (e.g. where a dog has a prominent place in the image), which were taken from the COCO dataset (Lin et al., 2014). The Photobook dataset can be found at `https://dmg-photobook.github.io/analysis.html`.

The data set contains all the utterances, game ids and game rounds which were made during all the games. From this dataset segments are created which refer to one or multiple images (targets; 75% refers to one image, 25% refers to 2 or more images). This is done by a heuristic model. A segment contains utterances for a single target image in a single game round, including the questions and answers. This resulted in 6801 segments in the test set from 2811 chains. In this research project, only the segments with one image target are used.

A reference chain is built when the interlocutors talk again about the same image in later game rounds. For example, if the same image appears multiple times and is discussed in rounds 1, 3, and 4, the chain consist of these utterances. The rank of an image is the amount of times an image appears in a game (in the game rounds). For example, rank 3 means it is the third time that a specific image occurs in a game.

In this research project, focus is laid on two different aspects in the prediction of the model given a segment. Namely, if the model returns a correct or incorrect prediction with the history and if the model returns a correct or incorrect prediction without the history. Whereby the history is defined as all the previous segments in a reference chain.

3.2 Models

The general goal of the models from Haber et al. (2019) is to predict the corresponding image given a segment or a history of segments using two different models. First, the No-History model converts the segments to features with a Long Short-Term Memory (LSTM), (Hochreiter and Schmidhuber, 1997). The visual features from the candidate images are extracted from the ResNet-152 which was pre-trained on ImageNet (Deng et al., 2009). After this, a dot product between segment and image features is calculated. Finally, the sigmoid function is applied to the segment and image features. This results in a prediction for all candidate images.

Secondly, the History model works similar to the No-History model, except that it also makes use of previous segments in a reference chain. These previous segments contain linguistic common ground, as images re-appear during a game. An image that has been seen before by both users might yield a description which is shorter than the initial description. Thus, this additional history should ideally help the model predict an image. The history is fed into an LSTM, which converts it to features that can be added to the segment features. A dot product between segment and image features is once again computed, followed by a sigmoid.

3.3 Analyses

The previously described data and models were taken from Haber et al. (2019). Using this the importance of history in the data for the pre-trained models (history and no history) is analysed by separating the segments and images into four conditions. The first condition contains segments and images that were correctly classified by both the history model and the no history model (hT-nhT, n=1702). The second condition contains segments and images that were correctly classified by the history model and incorrectly classified by the no history model (hT-nhF, n=112). The third condition contains segments and images that were incorrectly classified by the history model and correctly classified by the no history model (hF-nhT, n=121). The last condition contains segments and images that were incorrectly classified by both models (hF-nhF, n=248). The segments in these four conditions are first analysed on their similarity to segments in previous game rounds as in Haber et al. (2019). Secondly, they are analysed by their POS tags between first and current segments.

First, the cosine similarity is calculated between the segment embeddings, using the average of each segment. The similarity is calculated between the first segment with all other segments from the same chain, categorised by rank. Here, it is assumed that a larger distance between the segment embeddings indicates a more advanced grounding. If the distance remains similar, this indicates that the language of the players has not changed much, suggesting not much grounding took place. This cosine similarity is compared over the four conditions. In order to compare these segments word2vec embeddings, trained on Google news articles Mikolov et al. (2013), were used. It is expected that, when for example comparing a segment from rank 4 with rank 1 and a segment from rank 2 with rank 1, the cosine similarity will be smaller from the first segment and larger for the latter.

To investigate the difference in syntactic information in the segments between the conditions, the usage of certain part-of-speech (POS) tag is investigated. This is done by calculating the differences in POS ratios for each segment with the first segment in its chain. The words in the segments are all converted to lower case and are checked for misspellings by the oov dictionary (as created by Haber et al. (2019)), which consists of 182 common misspelled words in the games with their correctly spelled words. The POS tags per segment are then retrieved by using the natural language toolkit (nltk) POS tag package (Bird et al., 2009).

3.4 Perturbations

To investigate in what way grounding is present in the data and used by the History model, various perturbations are applied to the input segments.

The first perturbation randomly shuffles the order of the rounds within one chain of rounds that refer to a single image, which can cause later segments to appear earlier in a chain and vice versa. The segments are then given as input to the History model with their altered history. Grounding between players is expected to increase with the rounds. Therefore, changing the order of the ranks is expected to decrease accuracy, since if the segment of a later rank is placed earlier in a chain some development of grounding is missing. Investigating this will allow for insight into the amount of grounding present in the data and into how the model uses history. If the accuracy does not decrease, this can indicate that there is not much grounding present or that the model is not using the history effectively. This is expected for the segments that were correctly predicted by the No History model.

In the second perturbation experiment the utterances of a segment in one game is changed with those of a segment from the same rank of another game, but which refers to the same image. For example, the utterance in segment 12 from game A that refers to image X can be changed with the text of segment 23 from game B that refers to image X. This perturbation is performed by creating a list of all the segments in a certain rank referring to a certain image and randomly shuffle the texts of the corresponding utterances. Only the utterances of segments that only occur in one chain are changed, leading to unique utterances for each perturbed segment. As the image that the segments refer to is the same for the games of the initial and new utterances, only the speakers change. If the History model uses the history effectively, this perturbation should decrease the accuracy, as other players would have created a different grounding than the current players. This effect is expected to be stronger for rounds later in a chain, where more grounding has been established, as well as in segments that were only correctly predicted by the History condition, since grounding is assumed to be stronger here. However, because the utterances of the new segment of the other players refers to the same image as the original players, the model could still predict the correct image well.

A third perturbation is performed to investigate whether the effect of the previous perturbation is caused by simply changing the segment or because of the different interlocutors. Here the utterances of a segment are changed with those of another segment from the same game and rank, but which refers to a different image. For example, the utterance in segment 12 from game A that refers to image X can be changed with the text of segment 23 from game A that refers to image Y. Again this perturbation is performed only on segments that occur in a single chain, by shuffling the utterances of the segments of a certain rank and game. It is expected that the History model will perform close to chance on this perturbation, since the segments can now refer to any image, even outside of the image subject. This would indicate that the results from the previous perturbation are caused by the change in interlocutors. It is expected that the effect of this perturbation is the smallest on the segments only correctly predicted by the History model (hT-nhF), since these are thought to have a strong grounding or informative history.

For the last perturbation, the importance of certain part of speech (POS) tags and in which ways grounding can be understood by the model is investigated. Either all or half of the instances of a specific tag is removed in every segment. This is only done for the content words (i.e. nouns, verbs, adjectives, and adverbs). It is hypothesised that grounding becomes more effective by using relatively more nouns and adjectives, since these will be used to refer to objects, and by using less verbs and adverbs. It is therefore expected that model performance decreases more when nouns and adjectives are removed than when verbs and adverbs are removed. Additionally, it should perform worse when more instances of a tag are removed, because there will be less information in the segments.

4 Result and Analyses

4.1 Analyses

4.1.1 Cosine Similarity

The mean and standard deviation of the cosine similarity between each segment with the first segment in its chain was calculated over the four different conditions: hT-nhF, hT-nhT, hF-nhT and hF-nhF. Results can be found in Table 1. For three segments, there was no embedding available, so these segments were not included in this analyses. A Kruskal–Wallis one-way analysis of variance was performed to investigate whether there was a significant difference between the four conditions. The Kruskal-Wallis test returned a p-value < 0.001 and a statistic of 15.071. A t-test showed that there was a significant difference between the cosine similarity in the hT-nhT and the hT-nhF conditions (p-value = 0.001, statistic = 3.123). It was expected that when there is more grounding, the cosine similarity with the first segment in the chain would be larger. Together this indicates that there was more grounding in the segments that were only correctly predicted with their history.

The cosine similarity for different ranks was also calculated. As mentioned before, it was expected that the cosine similarity would decrease as the rank increased. The results in Figure 1 indeed show that the cosine similarity decreases over time for almost all conditions, but not for the condition where both the History and the No History models predict the images correctly. This could be a sign of these segments being correctly predicted because there was no grounding.

4.1.2 Model performances per rank

In order to investigate the difference between the History and the No History model, accuracies were compared between each segment and the first seg-

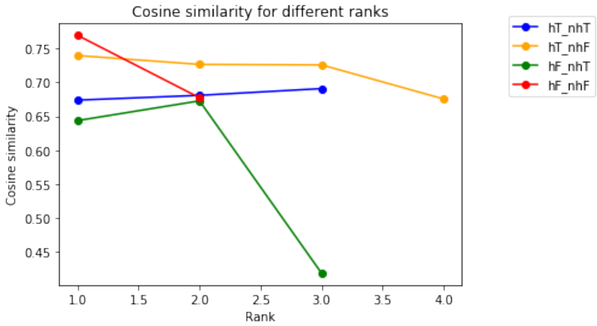

Figure 1: Cosine similarity between segments in the first and various last ranks.

Condition	Mean	STD
hT-nhT	0.747	0.141
hT-nhF	0.695	0.161
hF-nhT	0.736	0.160
hF-nhF	0.666	0.193

Table 1: Cosine similarity

ment of its chain. Results can be found in Table 2, where on the right the accuracy of the segments in the 2nd, 3rd, 4th, 5th and the overall last rank (combining all ranks) are shown. On the left the accuracy of the first segments in the chains corresponding to the segments on the right are shown. As expected, the accuracy for the No history model drops when it predicts an image based on the segment of a higher rank. This is likely because more grounding has taken place between interlocutors, as a lot of previously established information is missing. The accuracy for the History model seems to increase when predicting based on a higher rank, likely due to the larger size of the history.

4.1.3 POS tag distribution

To investigate the difference in syntactic information in the segments between the conditions, the POS tag ratio differences for each segment with the first segment in its chain were calculated.

	First rank		y-axis rank	
	H	NH	H	NH
2nd	0.752	0.818	0.839	0.801
3rd	0.741	0.816	0.841	0.834
4th	0.693	0.808	0.830	0.784
5th	0.619	0.728	0.883	0.666
last	0.752	0.818	0.863	0.825

Table 2: The accuracies per rank on the right side and accuracy of the corresponding first segment in their chains on the left side, for the History and No History model.

For the results see Figure 2. In all four conditions, the ratio of nouns versus all other words increases slightly (by 0-3%) as the rounds progress. Thus, in round 5 there are relatively more nouns used than in round 1. This is supported by the finding in Haber et al. (2019) where overall the ratio of the noun use increases. The ratio of verbs used decreases in almost all conditions (1 to 2 percent). They seem somewhat less important in later rounds. The ratio of adjectives used increases in all conditions in later rounds. This is because adjectives describe something about the objects in images (like ”golden” and ”oval”). This is an important property for the interlocutors to determine which image is common or different (game goal). There is almost no difference in the adverb use ratio in later rounds for the different conditions.

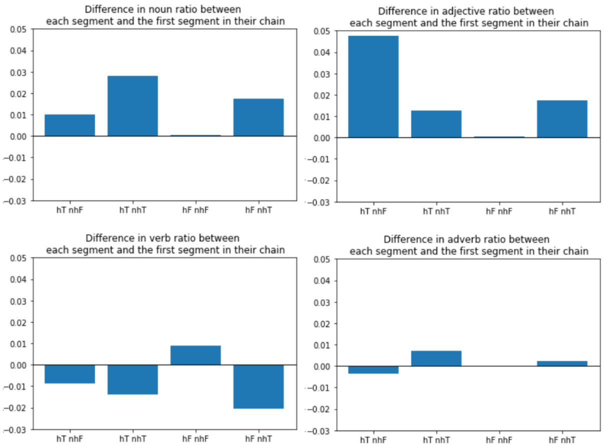

Figure 2: Pos tag distribution per tag and group

4.2 Perturbations

To investigate the grounding in the different conditions, various perturbations are performed. In the first perturbation experiment, the order of the segments in a chain were shuffled. Table 3 shows the mean and standard deviation on three simulations of the accuracies of the History model per condition after shuffling. The History model scored a slightly lower, but still high accuracy on the segments it classified correctly before shuffling. Of these segments those that were incorrectly classified without history showed a stronger decrease, indicating that in these segments the grounding was more important. Additionally, the History model performed better on the segments that were previously incorrectly classified. This could be caused by some of

Condition	Mean	STD
hT-nhT	0.954	0.00375
hT-nhF	0.862	0.0345
hF-nhT	0.286	0.0407
hF-nhF	0.224	0.0189
all	0.834	0.00317

Table 3: Accuracy of the History model after shuffle perturbation per condition and all segments together (mean and std, n=3).

these segments now containing more information about the target than before, because their histories can now contain more segments.

4.2.1 Exchange games and images

In the game perturbations the utterances of a segment were changed for the utterances of another segment of the same rank, referring to the same image, but from another game and thus other interlocuters. Additionally, in the image perturbation the utterances were changed for the utterances of another segment of the same rank and the same game, but referring to a different image.

The accuracies of the History model following these perturbations are displayed in Figure 3 for the conditions where the segments were originally correctly classified by the History model, and a group of all segments that were evaluated by the models together. When the segments are changed for one from another game, but the same image (the continuous line), the accuracy decreases in any rank to an accuracy around 0.8, indicating that some grounding was present between the interlocuters, but often enough information from another game has been transferred to still enable a correct prediciton. However, when the segments are changed for one from the same game, but referring to another image (the dashed line), the performance decreases drastically. This can be seen as a control task, which indicates that the decrease in performance of the game perturbation is indeed due to the different interlocuters, since otherwise the decrease in accuracy would be similar for the game and image perturbation. Additionally, for the image perturbation the History models performance decreases less for later ranks, which can mean that the model relies more on the history for these segments, which contains information from more rounds.

Figure 4 shows the performance of the History model on these perturbations for the segments originally incorrectly classified and a group of all segments that were evaluated. Changing the segments for those referring to another image has no effect on the lack of performance. However, changing them to a segment referring to the same image, but from a different game increases the performance. This indicates that these segments themselves were not informative enough, since changing their utterances for those of another segment increased the performance. Additionally, since the performance is increased by changing the utterances of another segment, this can mean that the original segments did not contain much information in their history, possibly caused by grounding. However, although the accuracy is increased, it does not reach the mean score of all segments together, which can signal that the history is still important. Nonetheless, the effect is the same for all ranks, making this idea more uncertain.

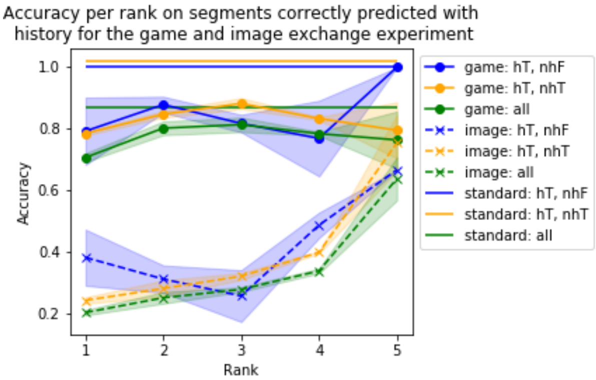

Figure 3: Mean accuracy of the History model on the segments originally correctly classified and all evaluated segments together following the game and image perturbations. Shaded area shows the standard deviation.

4.2.2 Removing POS tags

In investigating how the model uses POS tags, 50% or 100 % of the nouns, verbs, adjectives or adverbs were removed from the segments. The results per condition can be seen in Table 4 for the History model and Table 5 for the No History model.

Removing the nouns from the segments leads to the largest decrease in accuracy in both the History and No History model. Removing all nouns resulted in an almost 50% accuracy drop in the History model for the segments that were correctly predicted by this model before (hT-nhF and hT-nhT). Removing half of the nouns resulted in a 20% to 30% accuracy drop for the same group. In the No History model the performance also decreased for

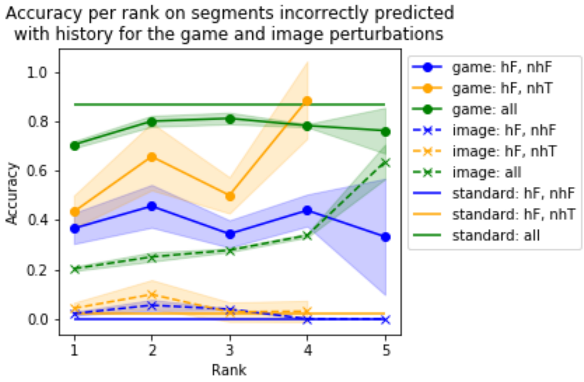

Figure 4: Mean accuracy of the History model on the segments originally incorrectly classified and all evaluated segments together following the game and image perturbations. Shaded area shows the standard deviation.

the segments previously correctly classified, with the largest decrease in the hF-nhT condition. Nouns are thus important for the model as they directly describe the objects. This indicates that the model depends on the presence of nouns in the segments.

For the segments that were incorrectly predicted before, removing POS tags increased the performance slightly, but around or below chance (i.e. 1/6th, as there are 6 target images), which therefore does not hold much significance. One exception is that when all the adjectives are removed, models perform above chance on the segments that were first wrongly categorised by the models.

Removing the verbs and adjectives from the segments has a much smaller effect on the accuracy of the model compared to removing the nouns. For the History model, however, the decrease in accuracy is more pronounced in the segments that were only correctly classified by the History model (hT-nhF). For the No History model, the segments that were only correctly classified by the No History model (hF-nhT) also had a more pronounced decrease in accuracy. This could indicate that the segments that were only correctly predicted by the History model are fundamentally different from the segments only correctly predicted by the No History model. Additionally, removing the adverbs has almost no effect on the performance, which can be a result from the task, since these words are not important for describing images. Finally, there is almost no difference in removing all or half of these tags.

5 Conclusion and Discussion

In this report the data and models created by (Haber et al., 2019) were analysed. It was found that more grounding appeared in later rounds, as shown by an increased ratio of nouns and adjectives and that the nouns were the most important words for the models to predict well. Additionally, the History models used history, but history alone was not sufficient. Furthermore, in segments that were initially correctly classified, the grounding was somewhat dependent on the interlocutors. For segments initially incorrectly classified, the history did not contain enough grounding. And finally, for the segments that were only correctly predicted with their history, their grounding process was more important.

First, the analyses show that the accuracy does not drop for the History model when predicting an image for a higher rank, as opposed to the No History model, indicating the usage of the history in the chains. Additionally, analyses on the cosine similarities between the different conditions indicated more grounding in segments that were only predicted correctly with their history. Other metrics to quantify grounding can give more insight in this. This grounding is assumed to be present, as shown by the increase in the ratio of nouns and adjectives in later ranks, while the use of verbs decreases.

Next, various experiments were performed. Shuffling the order of the history in a reference chain showed that for a segment that was only correctly predicted with its history, the order was more important than for segments that were always correctly classified, indicating more grounding here. However, since the shuffling resulted in changes in history sizes, future work should investigate this effect while keeping the history size the same, or by removing only certain ranks to control for this.

When changing a segment in a game that was correctly predicted by the History model with a segment from another game but pertaining to the same image, the model was still able to predict the correct image. This suggests that, when including history, changing the interlocutors only has a small effect on the performance. Additionally, the history itself is not enough to combat the misinformation from using a segment from the same game but pertaining a different image, as performance dropped drastically.

Additionally, for the segments that were initially incorrectly predicted, changing them to another segment from the same interlocutors but for a different

History				
	hT-nhT	hT-nhF	hF-nhT	hF-nhF
Noun 100	0.52	0.54	0.09	0.09
Verb 100	0.97	0.88	0.16	0.06
Adj 100	0.94	0.77	0.25	0.07
Adv 100	0.99	0.99	0.08	0.02
Noun 50	0.83 ± 0.0023	0.73 ± 0.029	0.14 ± 0.027	0.086 ± 0.015
Verb 50	0.98 ± 0.0011	0.94 ± 0.016	0.096 ± 0.018	0.046 ± 0.0022
Adj 50	0.97 ± 0.0023	0.87 ± 0.018	0.17 ± 0.038	0.057 ± 0.0059
Adv 50	0.997 ± 0.00028	0.99 ± 0.0086	0.059 ± 0.012	0.013 ± 0.0035

Table 4: Accuracy of the No History model when removing 100% or 50% of POS-tag words per condition. For 50% the mean and std are shown of 3 simulations.

No history				
	hT-nhT	hT-nhF	hF-nhT	hF-nhF
Noun 100	0.47	0.11	0.29	0.07
Verb 100	0.96	0.13	0.88	0.04
Adj 100	0.92	0.23	0.69	0.05
Adv 100	0.99	0.06	0.97	0.03
Noun 50	0.80 ± 0.0095	0.15 ±0.018	0.59 ± 0.019	0.066 ± 0.0096
Verb 50	0.98 ± 0.0019	0.059 ± 0.0037	0.92 ± 0.0054	0.025 ± 0.0018
Adj 50	0.97 ± 0.0017	0.11 ± 0.0078	0.81± 0.010	0.033 ± 0.012
Adv 50	0.996 ± 0.00074	0.025 ± 0.0086	0.99 ± 0.0039	0.014 ± 0.0020

Table 5: Accuracy of History model when removing 100% or 50% of POS-tag words per condition. For 50% the mean and std are shown of 3 simulations.

image did not increase performance, but changing to a segment from a different game but where interlocutors were still talking about the same image did increase performance. This suggests that these initially incorrectly classified segments were not informative enough for the model to make a correct prediction. This result does however indicate that the model history for these segments does not contain much information about grounding. With these experiments, the conditions (hT-nhT, hT-nhF, hF-nhT, hF-nhF) where the new text of the exchanged segment came from was not taken into account. Future work could further experiment with this, in particular investigating whether exchanging segments from certain conditions to another has any specific effect on model performance. For exchanging images, future research could look into whether swapping a segment regarding a different image but from the same category as the current image changes model predictions. If the History model is still able to predict the correct image, this could indicate that the history has a strong influence on the prediction.

Removing some or all words with a content-word POS tags almost always resulted in an accuracy drop, except for removing adverbs. Removing (a percentage of) nouns resulted in the largest performance drop, indicating that nouns are important to determine which image an interlocutor is referring to. Here, future work could look at removing words with certain POS tags in segments from specific ranks. Furthermore, the reason adverbs were not as important could be caused by the images being static. If in the task set by Haber et al. (2019), videos were used the importance of adverbs might change as actions in videos might result in more adverb-heavy language (e.g. ”man who jumps high”).

As a note on the dataset, Haber et al. (2019) created the segments on which this project depends by using certain heuristics. However, some segments ended up occurring in multiple chains, indicating some sub-optimal categorisations. Additionally, the models used in this report from Haber et al. (2019) were created as simple baseline models. As the original authors mentioned themselves, more sophisticated models should be created to continue the investigation of the development of grounding.

6 Acknowledgements

The first sections of the report were divided as follows. Together we wrote the abstract, Linda wrote the introduction, Emma the related work section, and Bob wrote the section about the models. Together we wrote the approach section (dataset, models, and analysis). The experiments were divided as follows. Emma focused on the cosine similarity, Linda on the perturbations and Bob on the POS tags. Together we wrote the code, ran the experiments, and wrote the results in the report. After that we together wrote the conclusion and discussion sections.

The github of the code can be found at: https://github.com/lindalw/CDM/tree/master/photobook_dataset.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017a. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017b. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. 2017. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The photobook dataset: Building common ground through visually-grounded dialogue.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment. *arXiv preprint arXiv:1907.05084*.

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 *European conference on computer vision*, pages 740–755. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. *arXiv preprint arXiv:1906.01603*.

Robert C. Stalnaker. 1978. Assertion. In P. Cole, editor, *Syntax and Semantics*, volume 9, pages 315–332. New York Academic Press.