The Task Matters: Comparing Image Captioning and Task-Based Dialogical Image Description

Nikolai Ilinykh, Sina Zarrieß, David Schlangen
Dialogue Systems Group
University of Bielefeld
Germany
{first.last}@uni-bielefeld.de

Abstract

Image captioning models are typically trained on data that is collected from people who are asked to describe an image, without being given any further task context. As we argue here, this context independence is likely to cause problems for transferring to task settings in which image description is bound by task demands. We demonstrate that careful design of data collection is required to obtain image descriptions which are contextually bounded to a particular meta-level task. As a task, we use MeetUp!, a text-based communication game where two players have the goal of finding each other in a visual environment. To reach this goal, the players need to describe images representing their current location. We analyse a dataset from this domain and show that the nature of image descriptions found in MeetUp! is diverse, dynamic and rich with phenomena that are not present in descriptions obtained through a simple image captioning task, which we ran for comparison.

1 Introduction

Automatic description generation from real-world images has emerged as a key task in vision & language in recent years (Fang et al., 2015; Devlin et al., 2015; Vinyals et al., 2015; Bernardi et al., 2016), and datasets like Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014) or Microsoft CoCo (Lin et al., 2014; Chen et al., 2015) are typically considered to be general benchmarks for visual and linguistic image understanding. By exploiting these sizeable data collections and recent advances in computer vision (e.g. ConvNets, attention, etc.), image description models have achieved impressive performance, at least for in-domain training and testing on existing benchmarks.

Nevertheless, the actual linguistic definition and foundation of image description as a task remains unclear and is a matter of ongoing debate, e.g. see (van Miltenburg et al., 2017) for a conceptual discussion of the task from a cross-lingual perspective. According to (Bernardi et al., 2016), image description generation involves *generating a textual description (typically a sentence) that verbalizes the most salient aspects of the image*. In practice, however, researchers have observed that eliciting descriptions from naive subjects (i.e. mostly crowd-workers) at a consistent level of quality is a non-trivial task (Rashtchian et al., 2010), as workers seem to interpret the task in different ways. Thus, previous works have developed relatively elaborate instructions and quality checking conventions for being able to systematically collect image descriptions.

In this paper, we argue that problems result from the fact that the task is typically put to the workers without providing any further context. This entirely monological setting essentially suggests that determining the salient aspects of an image (like highly important objects, object properties, scene properties) can be solved in a general, "neutral" way, by humans and systems. We present ongoing work on collecting image descriptions in task-oriented dialogue where descriptions are generated collaboratively by two players. Importantly, in our setting (which we call the MeetUp! environment), image descriptions serve the purpose of solving a higher-level task (meeting in a room, which in the game translates to determining whether an image that is seen is the same as the one that the partner sees). Hence, our participants need not be instructed explicitly to produce image descriptions. In this collaborative setting,

we observe that the notion of saliency is non-static throughout a dialogue. Depending on the history of the interaction, and the current state, speakers seem to flexibly adjust their descriptions (ranging from short scene descriptions to specific object descriptions) to achieve their common goal. Moreover, the descriptions are more factual than those collected in a monological setting. We believe that this opens up new perspectives for image captioning models, which can be trained on data that is bounded to its contextual use.

2 Related Work

As described above, the fact that the seemingly simple task of image captioning can be interpreted differently by crowd-workers has already been recognised in the original publications describing the datasets (Hodosh et al., 2013; Young et al., 2014; Chen et al., 2015). However, it has been treated as a problem that can be addressed through the design of instructions (e.g., "do not give people names", "do not describe unimportant details", (Chen et al., 2015)). (van Miltenburg et al., 2016; van Miltenburg, 2017) later investigated the range of pragmatic phenomena to be found in such caption corpora, with the conclusion that the instructions do not sufficiently control for them and leave it to the labellers to make their own decisions. It is one contribution of the present paper to show that providing a task context results in more constrained descriptions.

Schlangen et al. (2016) similarly noted that referring expressions in a corpus that was collected in a (pseudo-)interactive setting (Kazemzadeh et al., 2014), where the describers were provided with immediate feedback about whether their expression was understood, were more concise than those collected in a monological setting (Mao et al., 2016).

Similar to MeetUp, the use of various dialogue game set-ups has lately been established for dialogue data collection. Das et al. (2017) designed the "Visual Dialog" task where a human asks an agent about the content of an image. De Vries et al. (2017) similarly collected the GuessWhat? corpus of dialogues in which one player has to ask polar questions in order to identify the correct referent in the pool of images. de Vries et al. (2018) also develop a new navigation task, where a "tourist" has to reach a target location via communication with a "guide" given 2D images of various map locations. While similar in some respects, MeetUp is distinguished by being a symmetrical task (no instruction giver/follower) and broader concerning language data (more phenomena such as repairs, strategy negotiation).

3 Data collection

3.1 MeetUp image descriptions

The MeetUp game is a two-player text-based communication game set in a visual 2D environment.[1] The game starts with two players being placed in different 'rooms'. Rooms are represented to the players through images.[2] Each player only sees their own location. The objective of the game is to find each other; that is, to be in the same room. To solve this task, players can communicate via text messages and move freely (but unnoticed by the other player) to adjacent rooms. In the process of the game, the players naturally produce descriptions of what they currently see—and, interestingly, sometimes of what they have previously seen—to determine whether they have reached their goal or not. When they think that they have indeed achieved their goal, they indicate this via a particular command, and the dialogue ends.

The corpus we use here consists of 25 MeetUp (MU) games, collected via crowd-sourcing with Amazon Mechanical Turk. Workers were required to be native speakers of English. The dialogues all end with *a matching phase* where the players try to establish whether they are in the same room, by exchanging descriptions and come to the conclusion that they are (correctly in fact, in all but one dialogue). In some games, the players earlier already suspected to be in the same room and had such a "matching phase", but concluded that they weren't.

The complexity of the game board is likely to have an influence on the shape of the dialogue. For this data collection, we handcrafted a set of game boards to contain a certain degree of room type redundancy (e.g., more than one bed room per game board) and varying levels of overall complexity, as indicated in Table 2.

For our investigations here, we take these "matching phase" sub-dialogues and the images

[1] See https://github.com/dsg-bielefeld/meetup for more details.

[2] The images were taken from the ADE20k dataset (Zhou et al., 2016), which is a collection of images of indoor and outdoor scenes.

*1. Modern kitchen with **grey marble accents** featuring The popular **stainless steel appliances**.*

*2. Modern kitchen with **stainless steel appliances** well decorated*

*3. This kitchen looks very beautiful I can eat off **the floors** that's how clean it looks.*

4. A very clean looking kitchen, black and silver are the color theme. Looks like it is in an expensive place.

Figure 1: Example of a scene and corresponding monological captions

	Time	Private to A	Public	Private to B
31	(01:45)		A: I am now in a kitchen with **wood floors** and **a poster** that says **CONTRATTO**	
			....	
59	(02:50)		B: Wait– I found the kitchen!	
			....	
60	(02:55)	$\xrightarrow{N}$ kitchen		
61	(02:55)	You can go [/n]orth [/e]ast [/s]outh [/w]est		
62	(03:13)		A: I am back in kitchen. It has **a white marble dining table** in center	
63	(03:29)		B: Yes. There are **four chairs** on **the island**.	
64	(03:35)		A: Exactly	
65	(03:37)		B: And **the big Contratto poster**.	
66	(03:48)		B: **Three lights** above **the island**?	
67	(03:53)		A: yep	
71	(04:05)			B: /done
72	(04:07)	A: /done		
73	(04:10)		Well done! You are all indeed in the same room!	

Table 1: Excerpt from a dialogue involving image description (of the image shown in Figure 1)

Gameboard	Rooms	Types	R/T Ratio
House	11	9	1.2
Airport	22	15	1.5
Hospital	13	12	1.08
Shopping Mall	15	14	1.07
School	17	14	1.2

Table 2: Description of gameboards

that they are about (note that for the non-matching situations, there are two images for one subdialogue), to give us a set of 33 images together with corresponding utterances. We will call these utterances *dialogical image descriptions (DDs)*, in contrast to the *monological image descriptions (MDs)* described in the next subsection.

An example of such a description is shown in Table 1. From left to right columns represent line number in a dialogue, timestamp of a message, messages private to player A, messages seen by both, private messages of player B. Lines 60-72 in the transcript contain part of the dialogue where players act on suspicion that they might be in the same location and start describing images presented to them individually. In earlier stages of the dialogue (lines 31 and 59), this room had already been referenced to. It indicates that the player keeps a memory of what has already been mentioned and can refer back to that.

3.2 Monological image descriptions

In order to compare dialogical descriptions with data produced in a typical non-context caption environment, we also collected MDs on Amazon Mechanical Turk (AMT). We presented workers with the 33 images and instructed them to produce captions for them. We adopted the instructions from the MS COCO collection (Chen et al., 2015), which ask workers to "describe the important parts" of the image, and, importantly, to provide at least eight (8) words per image description. We collected four captions per image; and thus 132 captions overall. An example of four monological image descriptions for one image is shown in Figure 1.

4 Analysis

An important task is to determine what types of referring expressions are present in the datasets. In order to identify and analyse referring expressions, we used the brat annotation tool (Stenetorp et al., 2012) to tokenise and annotate both DDs and MDs. The first author of the paper annotated whether an utterance contains descriptions of scene with objects (*a kitchen with wood floors*), or objects only (*a white marble dining table*), expresses players' actions (*moving north now*) or is

	MDs	DDs		
		DD_{roam}	DD_{match}	DD_{all}
Number of descr.	132	94	174	268
Number of tokens	1655	915	1400	2315
Average length	12.5	9.73	8.05	8.64
Type / Token ratio	0.30	0.38	0.34	0.29
Number of REXs	138	184	344	528
REX per description	1.0	1.96	1.98	1.97
Average length of REXs	1.82	2.26	2.23	2.24

Table 3: Analysis of image descriptions

related to players' beliefs about their current state (*I think we are in the same room*). For our analysis we define referring expressions (REXs) as nominal phrases that refer to the objects in the scene (*four chairs*) or to the scene itself (*a kitchen*).

Additionally, we identified parts of speech in both DDs and MDs using Stanford Log-linear Part-Of-Speech Tagger (Toutanova et al., 2003). Examples of REXs according to this definition are displayed in bold in Figure 1 for MDs and in Table 1 for DDs.

Table 3 gives some basic statistics about the two data sets. The goal is to look at the task dependence of image descriptions. Each MU dialogue can be divided into phases, two of which are exemplified in Table 1. The *roaming phase* (part of it are lines 31, 59) is typically filled with movements and players informing each other about their location. The *matching phase* (lines 60-72) ends the dialogue with the determination that the two players are present in the same place. In order to demonstrate dynamics of interactions in MeetUp, we look at all DDs as well as at their statistical characteristics in two phases. There were almost two times more DDs than MDs overall, with matching phase requiring a high number of descriptions as well. At the same time, MDs tend to be longer than all DDs, though both sets have nearly identical type/token ratio.

4.1 Referring expressions in MDs and DDs

When looking at the number of REXs in Table 3, in the MeetUp set-up players produced almost four times more overall referring expressions than the workers that produced the MD set. The majority of these occurred in the matching phases, which indicates that the different subgoals between phases have an influence. There were also nearly two times as many REXs per individual description in the MeetUp setting than in the monological descriptions. Additionally, given the fact that MeetUp descriptions are generally more condensed than the MDs (8.64 vs. 12.5), it appears that MDs contain much material not directly relevant for reference to the scene or its objects. In particular we observed that there are on average 11 words in MDs (88%) which are not REXs and thus not related to an image, while there are nearly only 6 (70%) non-REX words in DDs. MeetUp players also produce longer REXs and this parameter is stable for all MeetUp phases. These observations show that the MeetUp descriptions are more focused on the task, less broad, contain much more referring expressions, which are longer then the ones in the non-task-driven set-up.

4.2 Adjectives in MD and DD

Table 4 displays the most frequent adjectives in both datasets in the spirit of (Baltaretu and Ferreira, 2016), who compared type and frequency of adjectives in a similar task design. It clearly shows a trend that seems to be present in the overall data: MDs cover a broader range of object properties or image attributes than the DDs.

Adj	Num
clean	15
small	11
large	9
empty	9
beautiful	8
white	7
nice	7
dark	5
old	5
many	5

Adj	Num
white	18
blue	11
red	11
left	7
small	6
same	5
open	4
yellow	4
black	4
right	3

Table 4: First 10 most frequent adjectives in both MDs (left) and DDs (right)

For example, evaluative adjectives (*beautiful*, *nice*) appear very often in MDs, while none of them is observed for the DDs. The latter ones seem to concentrate on attributes like colour, size, position, qualities of objects, while monological captions additionally have adjectives which refer

to age, feelings, a number of objects in the scene. Furthermore, 78 adjectives occur only once among all words in MDs, while this number is almost half that for the DDs (38). It additionally supports the idea that absence of the task makes humans to produce broad image descriptions, which are not necessarily grounded in scene objects.

5 Conclusion

The task of collecting appropriate training data for image caption generation systems, and language & vision in general, is not a trivial one. We found that in a standard crowdsourcing-based collection procedure, annotators tend to produce interpretative, non-factual descriptions, leading to potentially unsystematic or noisy data. We have presented a task-oriented interactive set-up for data collection where image descriptions are naturally used by speakers to solve a higher level task. Our data collected in a small-scale pilot study indicates that dialogical image descriptions consistently lead to factual descriptions containing many more reasonable referring expressions than monological descriptions. The analysis presented here will be used to further control MeetUp! data collection in order to avoid data that is similar to non-task-drive monological captions.

References

Adriana Baltaretu and Thiago Castro Ferreira. 2016. Task demands and individual variation in referring expressions. In *INLG 2016 - Proceedings of the Ninth International Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK*, pages 89–93. The Association for Computer Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China. Association for Computational Linguistics.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of CVPR*, Boston, MA, USA. IEEE.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of CVPR 2016*, Las Vegas, USA.

Emiel van Miltenburg. 2017. Pragmatic descriptions of perceptual stimuli. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10. Association for Computational Linguistics.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. *arXiv preprint arXiv:1707.01736*.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*, pages 54–59. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of ACL 2016*, Berlin, Germany.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *CoRR*, abs/1807.03367.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2016. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*.