On the use of human reference data for evaluating automatic image descriptions

Emiel van Miltenburg
Tilburg center for Cognition and Communication (TiCC), Tilburg University
Warandelaan 2, 5037 AB Tilburg, The Netherlands
C.W.J.vanMiltenburg@tilburguniversity.edu

1. Introduction

Automatic image description systems are commonly trained and evaluated using crowdsourced, human-generated image descriptions [1]. The best-performing system is then determined using some measure of similarity to the reference data (BLEU [8], Meteor [2], CIDER [16], etc). Thus, both the quality of the systems as well as the quality of the evaluation depends on the quality of the descriptions. As Section 2 will show, the quality of current image description datasets is insufficient. I argue that there is a need for more detailed guidelines that take into account the needs of visually impaired users, but also the feasibility of generating suitable descriptions. With high-quality data, evaluation of image description systems could use reference descriptions, but we should also look for alternatives.

2. The language of image descriptions

By virtue of their size, current image description datasets such as Flickr30K [17] or MS COCO [7] provide a unique opportunity to study how people talk about images. Doing so has revealed that:

1. Despite the fact that the guidelines tell them not to, crowdworkers often speculate about the contents of the images [11, 12]. The presence of these *unwarranted inference* means that the data does not constitute a reliable basis for evaluation. This work also argues that speculation occurs because of the decontextualized nature of the task. It is hard to describe an image without interpreting it, and interpretating an image often means filling in any missing details.

2. There is a high degree of variation in the descriptions [11, 14, 13]. This means that it unclear what descriptions should look like. The diversity in image description datasets is partly due to differences between the annotators [12], but also due to the undefined nature of the crowdsourcing task, which does not specify what the descriptions will be used for. Desmond Elliott (p.c.) notes that crowdworkers for the German portion of the Multi30K corpus [3] actively discussed the purpose of the task on crowdworker forums online. Successful communication requires interlocutors to be aware of the purpose of the exchange, so that they can adjust their contributions accordingly [5].

3. The descriptions contain linguistic constructions, such as negations (e.g. *A man **not** wearing a shirt playing tennis.*) that require high-level reasoning (e.g. knowing that people usually wear shirts, and signaling that this behavior is unusual), and as such are impossible to generate for current systems [15].

These observations raise the question of what image description systems should look like. Although the development of image description guidelines is an active area of research (see [9] for an overview), there is a disconnect between the human-computer interaction (HCI) literature and the image captioning (IC) literature. Specifically: (1) the HCI literature does not look at variation in image description data; and (2) on the IC-side, current evaluation practices and image description datasets do not take existing guidelines into account. For the latter, crowdworkers receive fairly generic instructions, which leave the datasets open to (a subset of) the flaws discussed in this section.

3. Captions for visually impaired people

To my knowledge, there are currently two image description datasets for blind or visually impaired people. The first is developed by Gella & Mitchell [4], who interviewed potential users, and developed the Expressive Captions Dataset to address their needs. For this dataset, crowdworkers were explicitly asked to talk about emotional content of the images. Few other details are known, as the dataset remains unreleased.

The second dataset was developed by Gurari and colleagues [6], and is currently the target of the VizWiz 2020 image captioning task. A notable improvement from earlier published work, is that the crowdsourcing procedure clearly notes that the description should be useful to someone who

is blind. This hopefully lowers the amount of variation in the descriptions, and gives rise to a more uniform dataset. But otherwise, the instructions mostly tell users what *not* to say. Future work should also provide more positive support.

4. Conclusion

Given the above, we should be careful when evaluating image description systems using human reference data. A further complicating factor is that textual similarity-based evaluation metrics have been shown to be unreliable indicators of output quality for natural language generation systems [10]. So what are we to do? I offer two suggestions:

1. Complement user studies with the assessment of human-generated descriptions, to develop more detailed guidelines. This will enable us to more precisely specify the needs of visually impaired users. E.g. the question 'which features of human entities are/should be described?' has been looked at from both the user [9] and data [14] side. Combining these perspectives hopefully means that we do not miss any relevant features.

2. Develop automatic checks to see whether human- or computer-generated descriptions conform to those guidelines. This might also lead to a "description-checking interface" which could provide real-time feedback to users writing image descriptions.

This proposal hopefully brings us closer together, to tailor image description systems to their users' needs.

References

[1] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016. 1

[2] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014. 1

[3] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics. 1

[4] Spandana Gella and Margaret Mitchell. Residual multiple instance learning for visually impaired image descriptions. In *11th Women in Machine Learning Workshop*, 2016. 1

[5] Herbert Paul Grice. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. New York: Academic Press, 1975. 1

[6] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. *arXiv preprint arXiv:2002.08565*, 2020. 1

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 1

[9] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. "person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. 1, 2

[10] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan, Oct.–Nov. 2019. Association for Computational Linguistics. 2

[11] Emiel van Miltenburg. Stereotyping and bias in the flickr30k dataset. In Jens Edlund, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of Multimodal Corpora: Computer vision and language processing*, pages 1–4, 2016. 1

[12] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain, September 2017. Association for Computational Linguistics. 1

[13] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. Measuring the diversity of automatic image descriptions. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*, 2018. 1

[14] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. Talking about other people: an endless range of possibilities. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 415–420. Association for Computational Linguistics, 2018. 1, 2

[15] Emiel van Miltenburg, Roser Morante, and Desmond Elliott. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*, pages 54–59, Berlin, Germany, August 2016. Association for Computational Linguistics. 1

[16] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[17] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014. 1