

Stereotyping and Bias in the Flickr30K Dataset

Emiel van Miltenburg

Vrije Universiteit Amsterdam

emiel.van.miltenburg@vu.nl

Abstract

An untested assumption behind the crowdsourced descriptions of the images in the Flickr30K dataset (Young et al., 2014) is that they “focus only on the information that can be obtained from the image alone” (Hodosh et al., 2013, p. 859). This paper presents some evidence against this assumption, and provides a list of biases and unwarranted inferences that can be found in the Flickr30K dataset. Finally, it considers methods to find examples of these, and discusses how we should deal with stereotype-driven descriptions in future applications.

Keywords: image annotation, stereotypes, bias, Flickr30K

1. Introduction

The Flickr30K dataset (Young et al., 2014) is a collection of over 30,000 images with 5 crowdsourced descriptions each. It is commonly used to train and evaluate neural network models that generate image descriptions (e.g. (Vinyals et al., 2015)). An untested assumption behind the dataset is that the descriptions are based on the images, and nothing else. Here are the authors (about the Flickr8K dataset, a subset of Flickr30K):

“By asking people to describe the people, objects, scenes and activities that are shown in a picture without giving them any further information about the context in which the picture was taken, we were able to obtain conceptual descriptions that focus only on the information that can be obtained from the image alone.” (Hodosh et al., 2013, p. 859)

What this assumption overlooks is the amount of *interpretation* or *recontextualization* carried out by the annotators. Let us take a concrete example. Figure 1 shows an image from the Flickr30K dataset.



Figure 1: Image 8063007 from the Flickr30K dataset.

This image comes with the five descriptions below. All but the first one contain information that cannot come from the image alone. Relevant parts are highlighted in **bold**:

1. A blond girl and a bald man with his arms crossed are standing inside looking at each other.
2. A **worker** is **being scolded** by her **boss** in a **stern lecture**.

3. A **manager** talks to an employee about job performance.
4. A hot, blond girl **getting criticized by her boss**.
5. Sonic employees **talking about work**.

We need to understand that the descriptions in the Flickr30K dataset are subjective descriptions of events. This can be a good thing: the descriptions tell us what are the salient parts of each image to the average human annotator. So the two humans in Figure 1 are relevant, but the two soap dispensers are not. But subjectivity can also result in *stereotypical* descriptions, in this case suggesting that the male is more likely to be the manager, and the female is more likely to be the subordinate. Rashtchian et al. (2010) do note that some descriptions are speculative in nature, which they say hurts the accuracy and the consistency of the descriptions. But the problem is not with the lack of consistency here. Quite the contrary: the problem is that stereotypes may be pervasive enough for the data to be consistently biased. And so language models trained on this data may propagate harmful stereotypes, such as the idea that women are less suited for leadership positions. This paper aims to give an overview of linguistic bias and unwarranted inferences resulting from stereotypes and prejudices. I will build on earlier work on linguistic bias in general (Beukeboom, 2014), providing examples from the Flickr30K data, and present a taxonomy of unwarranted inferences. Finally, I will discuss several methods to analyze the data in order to detect biases.¹

2. Stereotype-driven descriptions

Stereotypes are ideas about how other (groups of) people commonly behave and what they are likely to do. These ideas guide the way we talk about the world. I distinguish two kinds of verbal behavior that result from stereotypes: (i) linguistic bias, and (ii) unwarranted inferences. The former is discussed in more detail by Beukeboom (2014), who defines linguistic bias as “a systematic asymmetry in word choice as a function of the social category to which the target belongs.” So this bias becomes visible through the *distribution* of terms used to describe entities in a particular

¹The Flickr30K data also contains examples where annotators judge the subjects of the images on their looks. E.g. description #4 above calling the girl in the image *hot*. Analyzing this judgmental language goes beyond the scope of this paper.

category. Unwarranted inferences are the result of speculation about the image; here, the annotator goes beyond what can be glanced from the image and makes use of their knowledge and expectations about the world to provide an overly specific description. Such descriptions are directly identifiable as such, and in fact we have already seen four of them (descriptions 2–5) discussed earlier.

2.1. Linguistic bias

Generally speaking, people tend to use more concrete or specific language when they have to describe a person that does not meet their expectations. Beukeboom (2014) lists several linguistic ‘tools’ that people use to mark individuals who deviate from the norm. I will mention two of them.²

Adjectives One well-studied example (Stahlberg et al., 2007; Romaine, 2001) is sexist language, where the sex of a person tends to be mentioned more frequently if their role or occupation is inconsistent with ‘traditional’ gender roles (e.g. *female surgeon*, *male nurse*). Beukeboom also notes that adjectives are used to create “more narrow labels [or subtypes] for individuals who do not fit with general social category expectations” (p. 3). E.g. *tough woman* makes an exception to the ‘rule’ that women aren’t considered to be tough.

Negation can be used when prior beliefs about a particular social category are violated, e.g. *The garbage man was not stupid*. See also (Beukeboom et al., 2010).

These examples are similar in that the speaker has to put in additional effort to mark the subject for being unusual. But they differ in what we can conclude about the speaker, especially in the context of the Flickr30K data. Negations are much more overtly displaying the annotator’s prior beliefs. When one annotator writes that *A little boy is eating pie without utensils* (image 2659046789), this immediately reveals the annotator’s normative beliefs about the world: pie should be eaten *with* utensils. But when another annotator talks about *a girls basketball game* (image 8245366095), this cannot be taken as an indication that the annotator is biased about the gender of basketball players; they might just be helpful by providing a detailed description. In section 3 I will discuss how to establish whether or not there is any bias in the data regarding the use of adjectives.

2.2. Unwarranted inferences

Unwarranted inferences are statements about the subject(s) of an image that go beyond what the visual data alone can tell us. They are based on additional assumptions about the world. After inspecting a subset of the Flickr30K data, I have grouped these inferences into six categories (image examples between parentheses):

Activity We’ve seen an example of this in the introduction, where the ‘manager’ was said to be *talking about job performance* and *scolding [a worker] in a stern lecture* (8063007).

Ethnicity Many dark-skinned individuals are called *African-American* regardless of whether the picture has been taken in the USA or not (4280272). And people who



Figure 2: Image 4183120 from the Flickr30K dataset.

look Asian are called Chinese (1434151732) or Japanese (4834664666).

Event In image 4183120 (Figure 2), people sitting at a gym are said to be watching a game, even though there could be any sort of event going on. But since the location is so strongly associated with sports, crowdworkers readily make the assumption.

Goal Quite a few annotations focus on explaining the *why* of the situation. For example, in image 3963038375 a man is fastening his climbing harness *in order to have some fun*. And in an extreme case, one annotator writes about a picture of a dancing woman that *the school is having a special event in order to show the american culture on how other cultures are dealt with in parties* (3636329461). This is reminiscent of the Stereotypic Explanatory Bias (Sekaquaptewa et al., 2003, SEB), which refers to “the tendency to provide relatively more explanations in descriptions of stereotype inconsistent, compared to consistent behavior” (Beukeboom et al., 2010, p. 5). So in theory, odd or surprising situations should receive more explanations, since a description alone may not make enough sense in those cases, but it is beyond the scope of this paper to test whether or not the Flickr30K data suffers from the SEB.

Relation Older people with children around them are commonly seen as parents (5287405), small children as siblings (205842), men and women as lovers (4429660), groups of young people as friends (36979).

Status/occupation Annotators will often guess the status or occupation of people in an image. Sometimes these guesses are relatively general (e.g. college-aged people being called *students* in image 36979), but other times these are very specific (e.g. a man in a workshop being called a *graphics designer*, 5867606).

3. Detecting stereotype-driven descriptions

In order to get an idea of the kinds of stereotype-driven descriptions that are in the Flickr30K dataset, I made a browser-based annotation tool that shows both the images and their associated descriptions.³ You can simply leaf through the images by clicking ‘Next’ or ‘Random’ until you find an interesting pattern.

²Examples given are also due to (Beukeboom, 2014).

³Code and data is available on GitHub: <https://github.com/evanmiltenburg/Flickr30K-Image-Viewer>

Asian		Average 60%
2339632913	Asian child/baby	2
3208987435	Asian baby, Asian/oriental woman	3
7327356514	Asian girl/baby, Asian/oriental woman	4
Black		Average 40%
1319788022	African-American (AA)/black baby	3
149057633	African/AA child, black baby	3
3217909454	Dark-skinned baby	1
3614582606	AA baby	1
White		Average 20%
11034843	White baby boy	1
176230509	White baby boy	1
2058947638	White baby	1
3991342877	White baby	1
4592281294	White baby stroller	FP
661546153	White baby stroller	FP
442983801	Fair-skinned baby	1

Table 1: Number of times ethnicity/race was mentioned per category, per image. The average is expressed as a percentage of the number of descriptions. Counts in the last column correspond to the number of descriptions containing an ethnic/racial marker. Images were found by looking for descriptions matching (asian|white|black|African-American|skinned) baby. I found two false positives, indicated with FP.

3.1. Ethnicity/race

One interesting pattern is that the ethnicity/race of babies doesn’t seem to be mentioned *unless* the baby is black or asian. In other words: white seems to be the default, and others seem to be marked. How can we tell whether or not the data is actually biased?

We don’t know whether or not an entity belongs to a particular social class (in this case: ethnic group) until it is marked as such. But we can approximate the proportion by looking at all the images where the annotators have used a marker (in this case: adjectives like *black*, *white*, *asian*), and for those images count how many descriptions (out of five) contain a marker. This gives us an *upper bound* that tells us how often ethnicity is indicated by the annotators. Note that this upper bound lies somewhere between 20% (one description) and 100% (5 descriptions). Figure 1 presents count data for the ethnic marking of babies. It includes two false positives (talking about a *white baby stroller* rather than a *white baby*). In the Asian group there is an additional complication: sometimes the mother gets marked rather than the baby. E.g. *An Asian woman holds a baby girl*. I have counted these occurrences as well.

The numbers in Table 1 are striking: there seems to be a real, systematic difference in ethnicity marking between the groups. We can take one step further and look at all the 697 pictures with the word ‘baby’ in it. If there turn out to be disproportionately many white babies, this strengthens the conclusion that the dataset is biased.⁴

⁴Of course this extra step does constitute an additional annotation effort, and it is fairly difficult to automate; one would have to train a classifier for each group that needs to be checked.

I have manually categorized each of the baby images. There are 504 white, 66 asian, and 36 black babies. 73 images do not contain a baby, and 18 images do not fall into any of the other categories. While this does bring down the average number of times each category was marked, it also increases the contrast between white babies (who get marked in less than 1% of the images) and asian/black babies (who get marked much more often). A next step would be to see whether these observations also hold for other age groups, i.e. children and adults.³

3.2. Other methods

It may be difficult to spot patterns by just looking at a collection of images. Another method is to tag all descriptions with part-of-speech information, so that it becomes possible to see e.g. which adjectives are most commonly used for particular nouns. One method readers may find particularly useful is to leverage the structure of Flickr30K Entities (Plummer et al., 2015). This dataset enriches Flickr30K by adding coreference annotations, i.e. which phrase in each description refers to the same entity in the corresponding image. I have used this data to create a coreference graph by linking all phrases that refer to the same entity. Following this, I applied Louvain clustering (Blondel et al., 2008) to the coreference graph, resulting in clusters of expressions that refer to similar entities. Looking at those clusters helps to get a sense of the enormous variation in referring expressions. To get an idea of the richness of this data, here is a small sample of the phrases used to describe beards (cluster 268): *a scruffy beard*; *a thick beard*; *large white beard*; *a bubble beard*; *red facial hair*; *a braided beard*; *a flaming red beard*. In this case, ‘red facial hair’ really stands out as a description; why not choose the simpler ‘beard’ instead?⁵

4. Discussion

In the previous section, I have outlined several methods to manually detect stereotypes, biases, and odd phrases. Because there are many ways in which a phrase can be biased, it is difficult to automatically detect bias from the data. So how should we deal with stereotype-driven descriptions?

Neutralizing stereotypes for production One way to move forward might be to work with multilingual data. Elliott et al. (2015) propose a model that generates image descriptions given data from multiple languages, in their case German and English. Multilingual, or better: multicultural data might force models to put less emphasis on features that are only salient to annotators from one particular country.

Stereotypes and interpretation While stereotypes might be a problem for production, further study of cultural stereotyping might be beneficial to systems that have to interpret human descriptions and determine likely referents of those descriptions. E.g. knowing that *baseball player* probably refers to a *male* baseball player is very useful.

Levels of describing an image There is a large body of work in art, information science, library science and related fields dedicated to the description and categorization

⁵Code and data is available on GitHub: <https://github.com/evanmilteneburg/Flickr30k-clusters>

of images (Shatford, 1986; Jaimes and Chang, 1999). A common thread is that we can divide image description into multiple levels or stages, starting from concrete physical attributes up to abstract contextual information. These levels build on each other; we first have to recognize separate entities before we can reason about their relation. But recent neural network models like (Vinyals et al., 2015) do not match this procedure. Rather, they are trained to create a direct mapping between images and their descriptions. With this paper, I hope to have shown that the Flickr30K dataset is *layered*, reflecting not only the physical contents of the images, but also whether the images match the everyday expectations of the crowd. An interesting challenge would be for image description models to learn separate representations for both layers: the perceptual and the contextual.

Representativeness My argument here is not that we should explicitly remove bias from crowdsourced descriptions of images. This may result in normalising the data into a form that is less representative of actual human descriptions. I do, however, contend that we should accept that crowdsourced descriptions of images *are* biased. Acknowledging this fact is an important step towards designing models that can accommodate data based on a mixture of facts *and stereotypes* about the world.

5. Conclusion

This paper provided a taxonomy of stereotype-driven descriptions in the Flickr30K dataset. I have divided these descriptions into two classes: linguistic bias and unwarranted inferences. The former corresponds to the annotators' choice of words when confronted with an image that may or may not match their stereotypical expectancies. The latter corresponds to the tendency of annotators to go beyond what the physical data can tell us, and expand their descriptions based on their past experiences and knowledge of the world. Acknowledging these phenomena is important, because on the one hand it helps us think about what is learnable from the data, and on the other hand it serves as a warning: if we train and evaluate language models on this data, we are effectively teaching them to be biased.

I have also looked at methods to detect stereotype-driven descriptions, but due to the richness of language it is difficult to find an automated measure. Depending on whether your goal is production or interpretation, it may either be useful to suppress or to emphasize biases in human language. Finally, I have discussed stereotyping behavior as the addition of a contextual layer on top of a more basic description. This raises the question what kind of descriptions we would like our models to produce.

6. Acknowledgments

Thanks to Piek Vossen and Antske Fokkens for discussion, and to Desmond Elliott and an anonymous reviewer for comments on an earlier version of this paper. This research was supported by the Netherlands Organization for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen (SPI 30-673, 2014-2019).

7. Bibliographical references

- Beukeboom, C. J., Finkenauer, C., and Wigboldus, D. H. (2010). The negation bias: when negations signal stereotypic expectancies. *Journal of personality and social psychology*, 99(6):978.
- Beukeboom, C. J. (2014). Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In J. Laszlo, et al., editors, *Social cognition and communication*, volume 31, pages 313–330. Psychology Press. Author's pdf: <http://dare.ubvu.vu.nl/handle/1871/47698>.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.
- Elliott, D., Frank, S., and Hasler, E. (2015). Multilingual image description with neural sequence models. *CoRR*, abs/1510.04709.
- Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, pages 853–899.
- Jaimes, A. and Chang, S.-F. (1999). Conceptual framework for indexing visual information at multiple levels. In *Electronic Imaging*, pages 2–15. International Society for Optics and Photonics.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE ICCV*, pages 2641–2649.
- Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (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*, pages 139–147. Association for Computational Linguistics.
- Romaine, S. (2001). A corpus-based view of gender in british and american english. *Gender Across Languages: The linguistic representation of women and men*, 1:153–175.
- Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P., and von Hippel, W. (2003). Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology*, 39(1):75–82.
- Shatford, S. (1986). Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly*, 6(3):39–62.
- Stahlberg, D., Braun, F., Irmen, L., and Sczesny, S. (2007). Representation of the sexes in language. *Social communication*, pages 163–187.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of CVPR*, pages 3156–3164.
- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.