Investigating Bias in Attention-Based Image Caption Generation Systems

Aman Oberoi, Arjun Kallapur, Jiin Kim
University of California, Los Angeles
{aman.oberoi, arjunk, jiin.kim}@ucla.edu

Abstract

Image captioning is an increasingly common tool used in technological systems today. Yet, there is the possibility of image captioning tools perpetuating biases, which could be magnified in downstream applications. In this paper, we construct image captioning algorithms to generate captions for publicly available datasets. We then use sentiment and regard classifiers to compare captions generated by crowd-workers with captions generated by an image captioning model. Our preliminary findings suggest that both crowd-sourced and generated image captions are overwhelmingly neutral in both sentiment and regard, and the image caption generator amplifies this neutrality.

1 Introduction

McDonald (2010) defines Natural Language Generation (NLG) as the process of deliberately constructing natural language text in order to meet specified communicative goals. One important Natural Language Generation task is *linguistic realization*, which refers to the task of combining generated words and phrases into well-formed sentences (Gatt and Krahmer, 2018).

A specific application of linguistic realization has emerged in the form of image captioning. As images become common as a means for communication, image captioning has an important role to play. Image captioning can be used for content-based image retrieval and can thus be used in information management systems (Hossain et al., 2019). As Hossain et al. (2019) note, image captioning involves image understanding - entailing tasks such as object detection, recognition, and understanding object interactions - and sentence generation - requiring syntactic and semantic understanding of the language.

Recent works (Prates et al., 2018) have shown that NLG systems can perpetuate biases. Given the broad scope of downstream applications for image captioning systems, it is important to be able to accurately quantify any biases these systems may have to minimize societal harms.

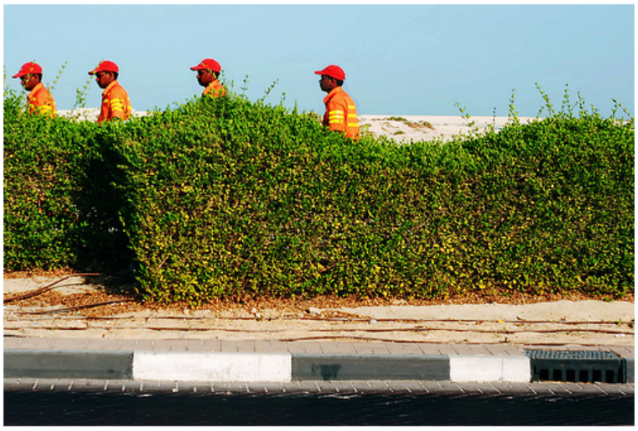
An image captioning system could be considered biased if it consistently produces captions that cause certain demographics to be perceived more negatively than other demographics. Sentiment analysis is a common computational tool used to classify opinions and attitudes towards a particular entity. Yet, as Sheng et al. (2019) note, sentiment analysis alone may not be able to capture subtle social cues that may cause an entity or group to be perceived negatively. To remedy this, Sheng et al. (2019) introduced the notion of *regard*, which helps capture more granular social connotations that may be present in generated text.

In this work, we construct image captioning models to generate captions for publicly available image datasets. We then use sentiment and regard classifier to quantify attitudes towards different demographic groups to measure pre-existing bias in image caption generation models. We also compare sentiment and regard scores of human-written and machine-generated captions to measure how much the models amplify such bias.

2 Definitions

In order to prevent ambiguity, we would like to first establish the definitions of some terms used throughout the paper. We refer to the definitions introduced in (Bender and Friedman, 2018), (Kiritchenko and Mohammad, 2018), and (Sheng et al., 2019) to do this.

Bias A computer system contains *bias* if it systematically and unfairly discriminate against certain individuals or groups of individuals in favor of



Original annotation: “four men looking towards the left”
Generated caption: “a line of people on a city street”

Figure 1: An image caption generated by our image caption model, along with its original caption annotation from the Flickr30k dataset.

others.

Pre-existing bias *Pre-existing bias* stems from society, with roots in social institutions, practices, and attitudes.

Technical bias *Technical bias* stems from technical constraints and decisions during the use of seemingly neutral algorithms in real-world contexts.

Emergent bias *Emergent bias* occurs when a system designed for one context is used in another.

Sentiment and Regard *Sentiment* refers to the overall polarity of text towards an entity or group. However, human judgement does not always align with text polarity when determining how a group is perceived based on text. Thus, Sheng et al. (2019) introduced the notion of *regard*, which uses context beyond language polarity to determine how text may cause a group to be perceived. Note that both sentiment and regard can be *positive*, *negative*, or *neutral*.

3 Models

3.1 Image Caption Generator

We adopted the Visual Attention architecture of the image caption generator detailed in Xu et al. (2015) to build our own image caption generator trained on the MS-COCO dataset [1], which contains over 82,000 images, each of which has at least 5 different crowd-sourced caption annotations. The model architecture consists of a CNN encoder that extracts features from images, an attention layer and finally an RNN-based decoder that generates captions based on the features extracted by the encoder and the context generated by the attention layer. We extracted features from the images using InceptionV3 (Szegedy et al., 2015), which is trained on ImageNet [2]. The extracted features were then passed into a GRU decoder, which attends over the image to predict the caption. We also implemented another model architecture which utilized ResNet-18 (He et al., 2015) instead of InceptionV3. We found that InceptionV3 performed better as an encoder than ResNet-18. Figure 1 shows an example of a caption generated by our image caption model.

3.2 Bias and Regard Classifiers

We employed the models from Sheng et al. (2019) for bias and regard classification. We used the VADER model as a sentiment classifier and the BERT model (trained to determine a regard score using transfer learning) as a regard classifier to assign sentiment and regard scores (positive, neutral or negative) to captions respectively.

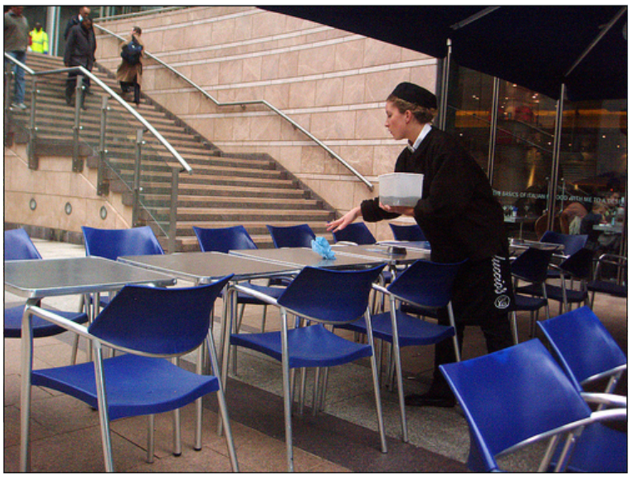
4 Methods to Detect Bias in Caption Generation

4.1 Technical Bias Detection

To identify and analyze technical bias in the image caption generation architecture, we compared the

[1] https://cocodataset.org

[2] https://www.image-net.org/



Original annotation: “a woman in black cleaning up some tables with a restaurant and staircase in the background”
Generated caption: “man laying near a table by a bar in a city”

Figure 2: An example of an inaccurate image caption generated by our image caption model, along with its original caption annotation from the Flickr30k dataset.

sentiment and regard scores of the crowd-sourced captions from the Flickr30k [3] dataset against the sentiment and regard scores of model-generated captions.

5 Biases in Caption Generation Systems

A short-coming of the caption generation model we utilized that we considered important to mention was that the model would sometimes confuse the gender of the people in an image. This phenomenon can be observed in figure 2. Gender inaccuracy is a flaw of several caption generation architectures as is discussed in (Burns et al., 2018) where models use contextual information to determine the gender of a person instead of directly looking at the person.

Figure 3 demonstrates the difference between the sentiment scores of the crowd-sourced and generated captions. It can be observed that the caption generator in general produces neutral sentiment captions at a slightly higher rate than the crowd-sourced captions (78.9% compared to the crowd-sourced 74%), while producing positive sentiment captions at a slightly lower rate (15.5% compared to the crowd-sourced 20%).

Figure 4 demonstrates the difference between the regard scores of the crowd-sourced and generated captions. Compared to sentiment scores, the

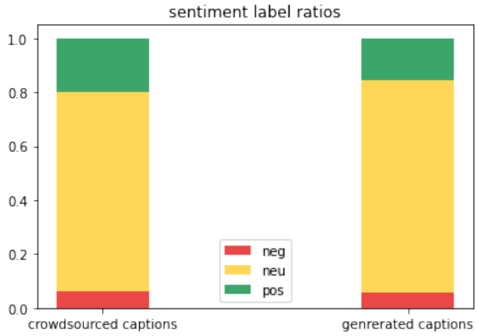


Figure 3: Sentiment label ratios for the crowd-sourced and generated captions

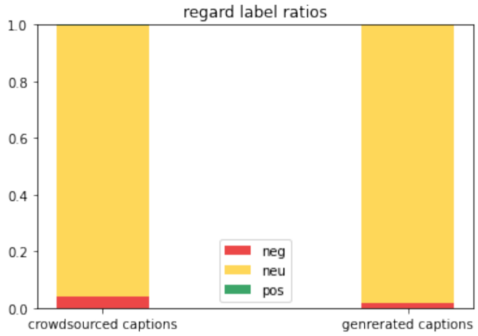


Figure 4: Regard label ratios for the crowd-sourced and generated captions

[3] https://shannon.cs.illinois.edu/DenotationGraph/

regard scores are much more neutral for both the crowd-sourced and the generated captions. The same trend as before is observed, in that the caption generator produces neutral regard captions at a slightly higher rate (95.7%) than the crowd-sourced captions (98.3%), while producing positive regard captions at a slightly lower rate; in fact, 0% of the generated captions were labeled as having positive regard.

6 Conclusion

As shown in figures 3 and 4, generated captions have more neutral regard and sentiment scores and less positive or negative regard and sentiment scores. On the other hand, crowd-sourced captions carry more negative/positive sentiment and regard due to the pre-exisiting biases of the crowd-source workers. We theorize that the caption generator has a neutralizing effect on the caption because the decoder that generates captions simply takes in the features detected by the CNN and the attention layer, and generates the best possible sentence combining the features.

We note overwhelmingly neutral sentiment and regard for both crowd-sourced and generated captions. One possible reason for this is the fact that task of image captioning involves mere description of the contents of images, and hence there is less room for subjectivity in the form of positivity or negativity to enter the corpus. From our preliminary empirical findings, image captions may not contain a large degree of strong positive or negative sentiment, and so may form less of a breeding ground for the propagation of bias.

7 Future Work

An area of interest that we were unable to explore in our paper was the performance of the caption generator on different demographic groups. A potential method for analyzing the bias of an image caption generation system towards different demographic groups is procuring an image dataset with a roughly equal distribution for different demographic groups, and comparing the sentiment and regard scores of the captions.

References

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. *CoRR*, abs/1803.09797.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6).

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *CoRR*, abs/1805.04508.

David McDonald. 2010. Natural language generation. *Handbook of Natural Language Processing*, 2.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing gender bias in machine translation - A case study with google translate. *CoRR*, abs/1809.02208.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *CoRR*, abs/1909.01326.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

A Links

- Caption generation repository: `https://github.com/TheDarkLord247/bias-in-caption-generation`
- Repository of sentiment and regard classifiers referenced in (Sheng et al., 2019): `https://github.com/ewsheng/nlg-bias`
- Regard and sentiment score generation notebook: `https://colab.research.google.com/drive/1VuEVLwFp6hBdST5u-8Z1o8Wxuis2k8fL?usp=sharing`