Fair Attention-Based Image Captioning

Shubhang Desai
shubhang@stanford.edu

Abstract

We extend work done in debiasing image captioning models by training an attention-based captioning system using Equalizer loss. This loss function, at a high level, requires that the model be confused about gender when no gender information is present in an image, but be confident about the correct gender when it is present. Using an attention-based approach, we are able to more clearly interpret the model's decision when predicting a word at a certain timestep than when using a base model without attention. We are able to see improvements in various fairness-based metrics, and propose future steps to further improve the quality of the model.

1 Introduction

Image captioning models are being deployed in many consumer-facing applications, especially in generating captions or alt-text for images uploaded to social media sites. It is therefore critical that these models operate without learning associations which may be considered harmful or problematic. Many machine learning models tend to learn and magnify bias existing in the dataset that it is trained on, and image captioning models are certainly no exception. Due to biases in the MS-COCO dataset (including imbalance of gender representation in the context of various activities, as well as actual gender mislabels) downstream models will often inherit some amount of this bias in performing the task.

In the case of image captioning, we want the model to be, as formalized by [2], "right for the right reasons". If the model predicts the word "man" as part of its caption, it should be because it sees and focuses on a man as part of predicting that word, not on the context such as, say, a baseball bat the man is holding. [2] ensures this by masking out gender information in images and the ensuring that the model is confused about gender in this masked image. While this approach reaches state-of-the-art results on gender ratio and error rate metrics, it does not employ attention mechanism (which has been shown to outperform image captioning models without attention).

To build upon [2] by addressing some of these shortcomings, we propose the Attention-Based Equalizer. We similarly use the Equalizer model proposed by [2] to formulate our loss function, but our base is an attention-based image captioning model instead of one without attention. Because of the self-explaining nature of attention-based models, we can easily understand why our new model makes the decision that it does by examining the attention masks overlaid onto the input image.

2 Related Work

Image captioning. Producing captions for images is a task which has been explored extensively in deep learning. A seminal paper from [4] showed that it is possible to use deep representation of an image obtained from a convolutional neural network as the initial state of a recurrent network, which can be trained (in tandem with the CNN) to produce captions for the image. [5] show that one can add visual attention on top of feature maps to produce dynamic inputs at each timestep, with attention weights signifying which part of the image are important to the word prediction at that timestep. While [2] build upon [4] as their base model, we use [5] as our work's base model.

Fair image captioning. Recently, there has been some exciting work around removing gender bias from image captioning models. [2] attempt to remove gender bias by adding two additional losses to cross-entropy loss which ensure the model is only confident about gender words when gender information is present in the image (more on this below). Meanwhile, [1] train a gender classifier on images which are cropped to only include people in it.

3 Attention-Equalizer for Fair Captioning

We extend the work of [2], which proposes the Equalizer model. This model, in addition to the standard cross-entropy loss employed by [2], adds two loss values to optimize. In particular, it adds:

- Appearance Confusion Loss. This loss ensures that, at timesteps where a gendered word is in the ground-truth, the probability of man is close to the probability of woman given that the gender-information is masked out. This do so with ground-truth segmentation masks of gender information in each image.
- Confident Loss. This loss ensures that, at timesteps where a gendered word is in the ground-truth, the probability of the correct gender is greater than the probability of the incorrect gender.

We do not make any changes to these loss terms. Our modification of [2] comes in the form of using an attention-based captioning model [5] as our base model, instead of one without attention [4]. The benefit of this approach is that explainability is part of the model itself. Instead of relying of post-hoc methods to interpret the model's decision, we can directly overlay attention weights onto the image to understand why the model outputted a certain word at a given timestep.

4 Experiments

We detail our experiments by discussing the dataset and metrics used, and then describing the training process for the experiments.

4.1 Dataset

The dataset we use is the MS-COCO dataset. The dataset contains 330,000 images, segmentation masks over 80 classes of objects in the image, and 5 captions which describe the image. For our task, we work only with the images and captions.

Additionally, we collect the MS-COCO Bias dataset as defined by [2]. This dataset contains all images such that (1) at least one caption contains the word woman/man and (2) no caption contains the word man/woman. In other words, this subset of the MS-COCO dataset contains all images where the captions are certain about the gender of a person in the image. We use the captions to then label the image as "man" or "woman" depending on which word is the one present in the caption.

4.2 Metrics

We measure a handful of metrics to evaluate the performance of our models:

Gender Error Rate. We measure the number of misclassifications that happen for the "man" and "woman" classes. There are a few caveats to this metrics. Firstly, we only count a prediction of "woman" instead of "man" or vice versa is a misclassification (in other words, if the model predicts a gender-neutral word or no gender at all, we do not count this as an error). Additionally, we do not require that the predicted gender be in the same position as in the ground truth (for example, predicting "A man..." when the ground truth says "A young man..." would not be incorrect).

Gender Ratio. We measure the ratio of number predicted sentences that contain the word "woman" to the number of predicted sentences that contain the word "man". We simply count the number of predicted sentences across all images in the dataset that contain the word "woman" and not "man", and divide that by the number of sentences that contain the word "man" and not "woman". We recognize that this metric is slightly flawed (for example, what if there is *supposed* to be both "man"

Model	GER [%]	GR	GNR [%]
Baseline	9.53	0.3378	12.32
Equalizer, $\beta = 1$	41.95	2.4605	18.08
Equalizer, $\beta = 10$	1.44	0.1994	57.90
Equalizer, $\beta = 25$	0.20	0.9510	57.91

Table 1: Metrics of models from experiments.

and "woman" in the prediction as there are two important people in the image?), but the failure cases are systematic and therefore still allow us to use the gender ratio as a valid metric of comparison.

Gender-Neutral Rate. We measure the number of predicted sentences that contain a gender-neutral word. We count the number of sentences that contain "person" or "people" in the sentence and measure the rate of occurrence of this prediction. While [2] does not measure this, we decide to measure this as it gives us a good idea of how often the model decides to not guess what the gender of the person in the image is and instead decides to defer to neutrality.

4.3 Training

Our experiments are as follows. We first pretrain an image captioning model (in particular, [5]) on the full MS-COCO dataset. We do so for about 50 epochs. The pretraining stage does not employ the additional losses defined for the Equalizer model, but instead minimizes only cross-entropy loss. Following this, we fine-tune the model on the MS-COCO Bias dataset. We do so for about 5 epochs. At this stage, we minimize both the Appearance Confusion and Confidence Losses defined above, in addition to cross-entropy loss as usual.

5 Results

Table 1 shows the results of the experiments. We choose to modulate only β as part of our experiments, as the authors of [2] choose to keep α and μ equal to 1 and set β to 10. We experiment with the value of beta.

We see that the baseline model (before fine-tuning) has a fairly high gender error rate and a low ratio, indicating that there are quite a few misclassifications likely due to the model overzealously predicting "man".

When we fine-tune our model and set $\beta = 1$, we see an interesting result: we see that the model predicts "woman" almost all the time. As a result, the gender ratio is high but the error rate is abysmal. I believe the reason that this is happening is, when all terms in the loss value have equal weight, the Confidence Loss "takes over" and overcorrects the errors where "man" is predicted instead of "woman", therefore causing the model to simply output "woman" all the time.

However, as we increase β, the performance begins looking up. When $\beta = 10$, we see that we have a drastic decrease in the gender error rate. However, our ratio is still low. When we then increase to $\beta = 25$, we see that the error again decreases and the ratio becomes more balanced (i.e. closer to 1). This seems to be a great result. However, an issues still persists: the rate of gender-neutral prediction is quite high. It seems that, more than half of the time, the model simply decides to not guess what the gender of a person in the picture is an instead just say "person" or "people". While we certainly prefer a model which outputs a gender-neutral term rather than an incorrect gender, having so many gender-neutral predictions may also not be very useful. This requires more work in the future.

5.1 Qualitative Results

There are considerable improvements in the captions as well as the attention masks, indicating that the model is looking at the "right thing" when doing the prediction for a word given an image and previously predicted words. For example, in Fig. 1 we see that, where the baseline attention-based model outputted the word "man" given a picture of a woman and paid attention to her tie when making this prediction, the fine-tuned model outputted "girl" and paid attention to her face when

(a) Ground truth caption: A young woman reading a text message that appears in the air.

(b) Baseline caption: A man is a book message on says to the wall.

(c) Fine-tuned caption: A young girl is a book message on reads to the picture.

Figure 1: Captions and attention masks for an image of a girl on her phone.

making this prediction.

In 2 we see that an image of a man in a kitchen was mislabelled as "woman" by the baseline model. This misclassification is particularly troubling as there are many sexist connotations associated with this error, and we would like to not reinforce this notions. Thankfully, our model decides to output "person" instead of any gendered word for this example, which is greatly preferred over a misclassification. x We do still have some problematic examples. In 3 we see an image of a man wearing a baseball mitt, and the baseline model pays attention to the mitt as opposed to the man when predicting the word "man". Similarly, the model pays attention to the mitt when predicting the word "young". However, not that the new model does not predict a gender at all. This may indicate that, when the attention weights are centered around a part of an image that does not indicate gender, the fine-tuned model becomes genuinely confused about which gender to output and therefore completely fails. So, while the fine-tuned model does not do the "right" thing, it also does not do the "wrong" thing, which may be considered a plus.

6 Conclusion

In this paper we discuss a modification of the Equalizer model [2] which uses attention-based captioning [5] as the base model. We find that we are able to train a model which has a lower error rate for gender prediction, as well as a higher ratio to "woman" to "man" predicted in captions, but the model is still overzealous at predicting gender-neutral terms. The benefit of our model is that it has explainability baked in through its attention mechanism.

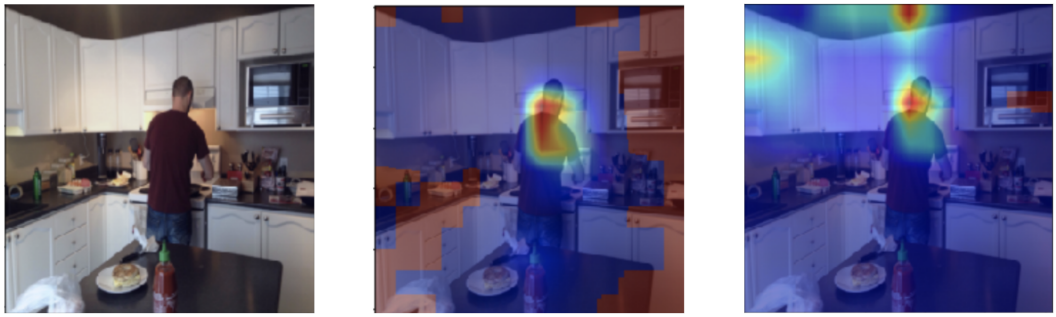
However, there is still work to be done. While the results are promising, we would like to see less gender-neutral terms being predicted and instead the model becoming more confident about its gender predictions. We may be able to do this by training the model for longer (as always predicting "person" instead of "man" or "woman" is indeed a local minimum for the problem). Additionally, we can add another loss term which further trains the attention mechanism by using the attention weights as a mask to compute the Appearance Confusion Loss.

References

[1] Bhargava, Shruti and Forsyth, David. Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models. In *arxiv*, 2019.

[2] Hendricks, Lisa Anne, Burns, Kaylee, Saenko, Kate, Darrell, Trevor, and Rohrbach, Anna. Women also snowboard: Overcoming bias in captioning models. In *European Conf. on Computer Vision (ECCV)*, 2018.

[3] Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollar, Piotr, and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In ECCV, pp. 740–755. In *Springer*, 2014.

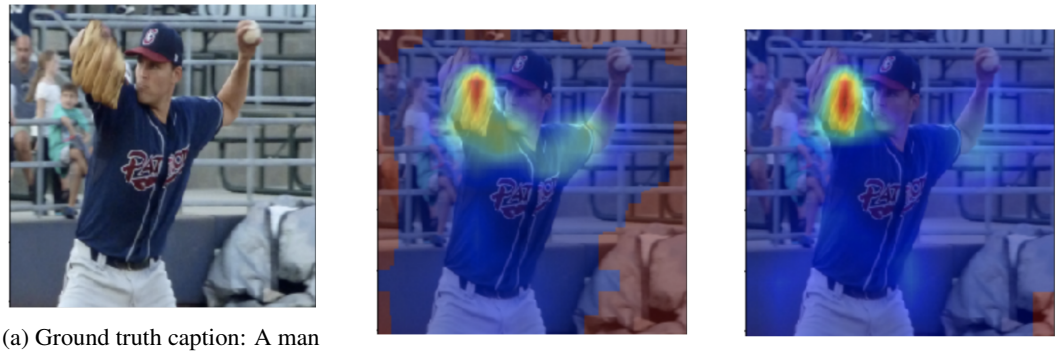


(a) Ground truth caption: A man is cooking at a stove in the kitchen.

(b) Baseline caption: A woman in preparing in a kitchen in a kitchen.

(c) Fine-tuned caption: A person is standing in a kitchen in a kitchen.

Figure 2: Captions and attention masks for an image of a man in a kitchen.



(a) Ground truth caption: A man is holding a baseball wearing a catchers mitt.

(b) Baseline caption: A man holding a baseball bat a baseball mitt.

(c) Fine-tuned caption: A young is a baseball bat a baseball mitt."

Figure 3: Captions and attention masks for an image of a man playing baseball.

[4] Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[5] Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron C., Salakhutdinov, Ruslan, Zemel, Richard S., and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.