Women Also Snowboard: Overcoming Bias in Captioning Models

Lisa Anne Hendricks*[1][0000−0001−9340−5143], Kaylee Burns*[1][0000−0001−5713−2774],
Kate Saenko[2][0000−0002−5704−7614], Trevor Darrell[1][0000−0001−5453−8533], and
Anna Rohrbach[1][0000−0003−1161−6006]

[1]UC Berkeley, [2] Boston University
lisa_anne@eecs.berkeley.edu, kayleeburns@berkeley.edu

Abstract. Most machine learning methods are known to capture and exploit biases of the training data. While some biases are beneficial for learning, others are harmful. Specifically, image captioning models tend to exaggerate biases present in training data (e.g., if a word is present in 60% of training sentences, it might be predicted in 70% of sentences at test time). This can lead to incorrect captions in domains where unbiased captions are desired, or required, due to over-reliance on the learned prior and image context. In this work we investigate generation of gender-specific caption words (e.g. man, woman) based on the person's appearance or the image context. We introduce a new *Equalizer* model that encourages equal gender probability when gender evidence is occluded in a scene and confident predictions when gender evidence is present. The resulting model is forced to look at a person rather than use contextual cues to make a gender-specific prediction. The losses that comprise our model, the *Appearance Confusion Loss* and the *Confident Loss*, are general, and can be added to any description model in order to mitigate impacts of unwanted bias in a description dataset. Our proposed model has lower error than prior work when describing images with people and mentioning their gender and more closely matches the ground truth ratio of sentences including women to sentences including men. Finally, we show that our model more often looks at people when predicting their gender. [1]

Keywords: Image description, Caption bias, Right for the right reasons

1 Introduction

Exploiting contextual cues can frequently lead to better performance on computer vision tasks [35,34,12]. For example, in the visual description task, predicting a "mouse" might be easier given that a computer is also in the image. However, in some cases making decisions based on context can lead to incorrect, and perhaps even offensive, predictions. In this work, we consider one such scenario: generating captions about men and women. We posit that when description models predict gendered words such as "man" or "woman", they should consider visual evidence associated with the described person, and not contextual cues like location (e.g., "kitchen") or other objects

* Authors contributed equally.

[1] https://people.eecs.berkeley.edu/~lisa_anne/snowboard.html

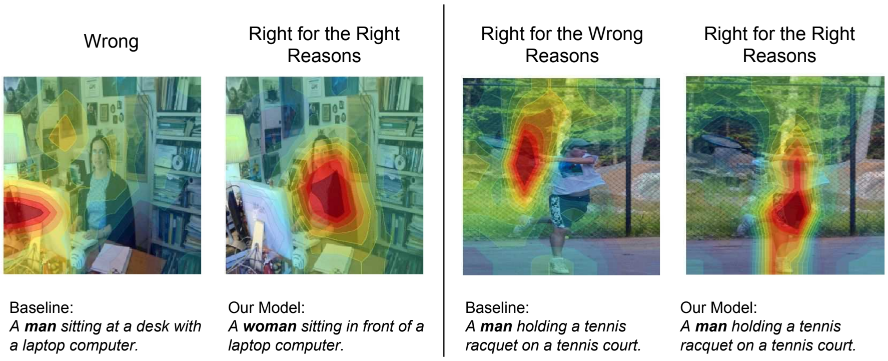

Wrong
Right for the Right Reasons
Right for the Wrong Reasons
Right for the Right Reasons
Baseline:
A **man** sitting at a desk with a laptop computer.
Our Model:
A **woman** sitting in front of a laptop computer.
Baseline:
A **man** holding a tennis racquet on a tennis court.
Our Model:
A **man** holding a tennis racquet on a tennis court.

Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

in a scene (e.g., "snowboard"). Not only is it important for description systems to avoid egregious errors (e.g., always predicting the word "man" in snowboarding scenes), but it is also important for predictions to be right for the right reason. For example, Figure 1 (left) shows a case where prior work predicts the incorrect gender, while our model accurately predicts the gender by considering the correct gender evidence. Figure 1 (right) shows an example where both models predict the correct gender, but prior work does not look at the person when describing the image (it is right for the wrong reasons).

Bias in image captioning is particularly challenging to overcome because of the multimodal nature of the task; predicted words are not only influenced by an image, but also biased by the learned language model. Though [47] studied bias for structured prediction tasks (e.g., semantic role labeling), they did not consider the task of image captioning. Furthermore, the solution proposed in [47] requires access to the entire test set in order to rebalance gender predictions to reflect the distribution in the training set. Consequently, [47] relies on the assumption that the distribution of genders is the same at training and test time. We make no such assumptions; we consider a more realistic scenario in which captions are generated for images independent of other test images.

In order to encourage description models to generate less biased captions, we introduce the *Equalizer* Model. Our model includes two complementary loss terms: the *Appearance Confusion Loss (ACL)* and the *Confident Loss (Conf)*. The Appearance Confusion Loss is based on the intuition that, given an image in which evidence of gender is absent, description models should be unable to accurately predict a gendered word. However, it is not enough to confuse the model when gender evidence is absent; we must also encourage the model to consider gender evidence when it is present. Our Confident Loss helps to increase the model's confidence when gender is in the image.

These complementary losses allow the Equalizer model to be cautious in the absence of gender information and discriminative in its presence.

Our proposed Equalizer model leads to less biased captions: not only does it lead to lower error when predicting gendered words, but it also performs well when the distribution of genders in the test set is not aligned with the training set. Additionally, we observe that Equalizer generates gender neutral words (like "person") when it is not confident of the gender. Furthermore, we demonstrate that Equalizer focuses on humans when predicting gender words, as opposed to focusing on other image context.

2 Related Work

Unwanted Dataset Bias. Unwanted dataset biases (e.g., gender, ethnic biases) have been studied across a wide variety of AI domains [29,31,4,5,3,23]. One common theme is the notion of *bias amplification*, in which bias is not only learned, but amplified [47,4,31]. For example, in the image captioning scenario, if 70% of images with umbrellas include a woman and 30% include a man, at test time the model might amplify this bias to 85% and 15%. Eliminating bias amplification is not as simple as balancing across attributes for a specific category. [31] study bias in classification and find that even though white and black people appear in "basketball" images with similar frequency, models learn to classify images as "basketball" based on the presence of a black person. One explanation is that though the data is balanced in regard to the class "basketball", there are many more white people in the dataset. Consequently, to perfectly balance a dataset, one would have to balance across all possible co-occurrences which is infeasible.

Natural language data is subject to *reporting bias* [4,13,22,21] in which people overreport less common co-occurrences, such as "male nurse" [4] or "green banana" [22]. [21] also discuss how visual descriptions reflect cultural biases (e.g., assuming a woman with a child is a mother, even though this cannot be confirmed in an image). We observe that annotators specify gender even when gender cannot be confirmed in an image (e.g., a snowboarder might be labeled as "man" even if gender evidence is occluded).

Our work is most similar to [47] who consider bias in semantic role labeling and multilabel classification (as opposed to image captioning). To avoid bias amplification, [47] rebalance the test time predictions to more accurately reflect the training time word ratios. This solution is unsatisfactory because (i) it requires access to the entire test set and (ii) it assumes that the distribution of objects at test time is the same as at training time. We consider a more realistic scenario in our experiments, and show that the ratio of woman to man in our predicted sentences closely resembles the ratio in ground truth sentences, even when the test distribution is different from the training distribution.

Fairness. Building AI systems which treat *protected attributes* (e.g., age, gender, sexual orientation) in a fair manner is increasingly important [14,9,43,25]. In the machine learning literature, "fairness" generally requires that systems do not use information such as gender or age in a way that disadvantages one group over another. We consider is different scenario as we are trying to *predict* protected attributes.

Distribution matching has been used to build fair systems [25] by encouraging the distribution of decisions to be similar across different protected classes, as well as for other applications such as domain adaption [36,46] and transduction learning [24]. Our

Appearance Confusion Loss is similar as it encourages the distribution of predictions to be similar for man and woman classes when gender information is not available.

Right for the Right Reasons. Assuring models are "right for the right reasons," or consider similar evidence as humans when making decisions, helps researchers understand how models will perform in real world applications (e.g., when predicting outcomes for pneumonia patients in [7]) or discover underlying dataset bias [33]. We hypothesize that models which look at appropriate gender evidence will perform better in new scenarios, specifically when the gender distribution at test and training time are different.

Recently, [28] develop a loss function which compares explanations for a decision to ground truth explanations. However, [28] generating explanations for visual decisions is a difficult and active area of research [26,30,11,27,48,42]. Instead of relying on our model to accurately explain itself during training, we verify that our formulation encourages models to be right for the right reason at test time.

Visual Description. Most visual description work (e.g., [37,8,15,39,1]) focuses on improving overall sentence quality, without regard to captured biases. Though we pay special attention to gender in this work, all captioning models trained on visual description data (MSCOCO [20], Flickr30k [41], MSR-VTT [38] to name a few) implicitly learn to classify gender. However current captioning models do not discuss gender the way humans do, but *amplify* gender bias; our intent is to generate descriptions which more accurately reflect human descriptions when discussing this important category.

Gender Classification. Gender classification models frequently focus on facial features [18,45,10]. In contrast, we are mainly concerned about whether contextual clues in complex scenes bias the production of gendered words during sentence generation. Gender classification has also been studied in natural language processing ([2,40], [6]).

Ethical Considerations. Frequently, gender classification is seen as a binary task: data points are labeled as either "man" or "woman". However, AI practitioners, both in industrial[2] and academic[3] settings, are increasingly concerned that gender classification systems should be inclusive. Our captioning model predicts three gender categories: male, female, and gender neutral (e.g., person) based on visual appearance. When designing gender classification systems, it is important to understand where labels are sourced from [16]. We determine gender labels using a previously collected publicly released dataset in which annotators describe images [20]. Importantly, people in the images are not asked to identify their gender. Thus, we emphasize that we are not classifying biological sex or gender identity, but rather outward gender appearance.

3 Equalizer: Overcoming Bias in Description Models

Equalizer is based on the following intuitions: if evidence to support a specific gender decision is not present in an image, the model should be *confused* about which gender to predict (enforced by an Appearance Confusion Loss term), and if evidence to support a gender decision is in an image, the model should be *confident* in its prediction (enforced by a Confident Loss term). To train our model we require not only pairs of images, I, and sentences, S, but also annotation masks M which indicate which evidence in an

[2] https://clarifai.com/blog/socially-responsible-pixels-a-look-inside-clarifais-new-demographics-recognition-model

[3] https://www.media.mit.edu/projects/gender-shades/faq

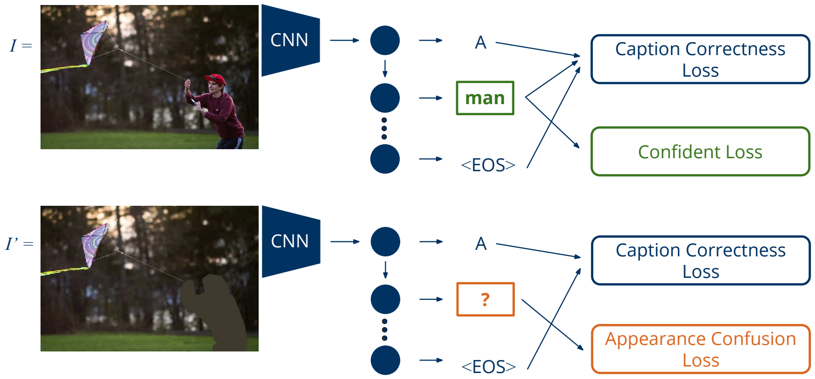


Fig. 2: Equalizer includes two novel loss terms: the Confident Loss on images with men or women (top) and the Appearance Confusion Loss on images where men and women are occluded (bottom). Together these losses encourage our model to make correct predictions when evidence of gender is present, and be cautious in its absence. We also include the Caption Correctness Loss (cross entropy loss) for both image types.

image is appropriate for determining gender. Though we use [37] as our base network, Equalizer is general and can be integrated into any deep description framework.

3.1 Background: Description Framework

To generate a description, high level image features are first extracted from the InceptionV3 [32] model. The image features are then used to initialize an LSTM hidden state. To begin sentence generation, a start of sentence token is input into the LSTM. For each subsequent time step during training, the ground truth word w_t is input into the LSTM. At test time, the previously predicted word w_{t-1} is input into the LSTM at each time step. Generation concludes when an end of sequence token is generated. Like [37], we include the standard cross entropy loss ($\mathcal{L}^{CE}$) during training:

$$\mathcal{L}^{CE} = -\frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T}\log(p(w_t|w_{0:t-1}, I)), \tag{1}$$

where N is the batch size, T is the number of words in the sentence, w_t is a ground truth word at time t, and I is an image.

3.2 Appearance Confusion Loss

Our Appearance Confusion Loss encourages the underlying description model to be *confused* when making gender decisions if the input image does not contain appropriate evidence for the decision. To optimize the Appearance Confusion Loss, we require ground truth rationales indicating which evidence is appropriate for a particular gender decision. We expect the resulting rationales to be masks, M, which are 1 for pixels which should not contribute to a gender decision and 0 for pixels which are appropriate to consider when determining gender. The Hadamard product of the mask and the original image, $I \odot M$, yields a new image, I', with gender information that the implementer deems appropriate for classification removed. Intuitively, for an image devoid of gender information, the probability of predicting man or woman should be equal. The Appearance Confusion Loss enforces a fair prior by asserting that this is the case.

To define our Appearance Confusion Loss, we first define a *confusion* function ($\mathcal{C}$) which operates over the predicted distribution of words $p(\tilde{w}_t)$, a set of woman gender words ($\mathcal{G}_w$), and a set of man gender words ($\mathcal{G}_m$):

$$\mathcal{C}(\tilde{w}_t, I') = |\sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I') - \sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I')|. \quad (2)$$

In practice, the $\mathcal{G}_w$ consists only of the word “woman” and, likewise, the $\mathcal{G}_m$ consists only of the word “man”. These are by far the most commonly used gender words in the datasets we consider and we find that using these “sets” results in similar performance as using more complete sets.

We can now define our Appearance Confusion Loss ($\mathcal{L}^{AC}$) as:

$$\mathcal{L}^{AC} = \frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} \mathbb{1}(w_t \in \mathcal{G}_w \cup \mathcal{G}_m) \mathcal{C}(\tilde{w}_t, I'), \quad (3)$$

where $\mathbb{1}$ is an indicator variable that denotes whether or not w_t is a gendered word.

For the remaining non-gendered words that correspond to images I', we apply the standard cross entropy loss to encourage the model to discuss objects which are still visible in I'. In addition to encouraging sentences to be image relevant even when the gender information has been removed, this also encourages the model to learn representations of words like “dog” and “frisbee” that are not reliant on gender information.

3.3 Confident Loss

In addition to being unsure when gender evidence is occluded, we also encourage our model to be confident when gender evidence is present. Thus, we introduce the Confident Loss term, which encourages the model to predict gender words correctly.

Our Confident Loss encourages the probabilities for predicted gender words to be high on images I in which gender information is present. Given functions $\mathcal{F}^W$ and $\mathcal{F}^M$ which measure how confidently the model predicts woman and man words respectively, we can write the Confident Loss as:

$$\mathcal{L}^{Con} = \frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} (\mathbb{1}(w_t \in \mathcal{G}_w) \mathcal{F}^W(\tilde{w}_t, I) + \mathbb{1}(w_t \in \mathcal{G}_m) \mathcal{F}^M(\tilde{w}_t, I)). \quad (4)$$

To measure the confidence of predicted gender words, we consider the quotient between predicted probabilities for man and gender words ($\mathcal{F}^M$ is of the same form):

$$\mathcal{F}^W(\tilde{w}_t, I) = \frac{\sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I)}{(\sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I)) + \epsilon} \quad (5)$$

where ϵ is a small epsilon value added for numerical stability.

When the model is confident of a gender prediction (e.g., for the word "woman"), the probability of the word "woman" should be considerably higher than the probability of the word "man", which will result in a small value for $\mathcal{F}^W$ and thus a small loss. One nice property of considering the quotient between predicted probabilities is that we encourage the model to distinguish between gendered words without forcing the model to predict a gendered word. For example, if the model predicts a probability of 0.2 for "man", 0.5 for "woman", and 0.3 for "person" on a "woman" image, our confidence loss will be low. However, the model is still able to predict gender neutral words, like "person" with relatively high probability. This is distinct from other possible losses, like placing a larger weight on gender words in the cross entropy loss, which forces the model to predict "man"/"woman" words and penalizes the gender neutral words.

3.4 The Equalizer Model

Our final model is a linear combination of all aforementioned losses:

$$\mathcal{L} = \alpha\mathcal{L}^{CE} + \beta\mathcal{L}^{AC} + \mu\mathcal{L}^{Con}, \tag{6}$$

where α, β, and μ are hyperparameters chosen on a validation set ($\alpha, \mu = 1$, $\beta = 10$ in our experiments).

Our Equalizer method is general and our base captioning framework can be substituted with any other deep captioning framework. By combining all of these terms, the Equalizer model can not only generate image relevant sentences, but also make confident gender predictions under sufficient evidence. We find that both the Appearance Confusion Loss and the Confident Loss are important in creating a confident yet cautious model. Interestingly, the Equalizer model achieves the lowest misclassification rate only when these two losses are combined, highlighting the complementary nature of these two loss terms.

4 Experiments

4.1 Datasets

MSCOCO-Bias. To evaluate our method, we consider the dataset used by [47] for evaluating bias amplification in structured prediction problems. This dataset consists of images from MSCOCO [20] which are labeled as "man" or "woman". Though "person" is an MSCOCO class, "man" and "woman" are not, so [47] employ ground truth captions to determine if images contain a man or a woman. Images are labeled as "man" if at least one description includes the word "man" and no descriptions include the word "woman". Likewise, images are labeled as "woman" if at least one description includes the word "woman" and no descriptions include the word "man". Images are discarded if both "man" and "woman" are mentioned. We refer to this dataset as MSCOCO-Bias.
MSCOCO-Balanced. We also evaluate on a set where we purposely change the gender ratio. We believe this is representative of real world scenarios in which different distributions of men and women might be present at test time. The MSCOCO-Bias set has a

roughly 1:3 woman to man ratio where as this set, called MSCOCO-Balanced, has a 1:1 woman to man ratio. We randomly select 500 images from MSCOCO-Bias set which include the word "woman" and 500 which include "man".

Person Masks. To train Equalizer, we need ground truth human rationales for why a person should be predicted as a man or a woman. We use the person segmentation masks from the MSCOCO dataset. Once the masked image is created, we fill the segmentation mask with the average pixel value in the image. We use the masks both at training time to compute Appearance Confusion Loss and during evaluation to ensure that models are predicting gender words by looking at the person. While for MSCOCO the person annotations are readily available, for other datasets e.g. a person detector could be used.

4.2 Metrics

To evaluate our methods, we rely on the following metrics.

Error. Due to the sensitive nature of prediction for protected classes (gender words in our scenario), we emphasize the importance of a low error. The error rate is the number of man/woman misclassifications, while gender neutral terms are not considered errors. We expect that the best model would rather predict gender neutral words in cases where gender is not obvious.

Gender Ratio. Second, we consider the ratio of sentences which belong to a "woman" set to sentences which belong to a "man" set. We consider a sentence to fall in a "woman" set if it predicts any word from a precompiled list of female gendered words, and respectively fall in a "man" set if it predicts any word from a precompiled list of male gendered words.

Right for the Right Reasons. Finally, to measure if a model is "right for the right reasons" we consider the pointing game [44] evaluation. We first create visual explanations for "woman"/"man" using the Grad-CAM approach [30] as well as saliency maps created by occluding image regions in a sliding window fashion. To measure if our models are right for the right reason, we verify whether the point with the highest activation in the explanation heat map falls in the person segmentation mask.

4.3 Training Details

All models are initialized from the Show and Tell model [37] pre-trained on all of MSCOCO for 1 million iterations (without fine-tuning through the visual representation). Models are trained for additional 500,000 iterations on the MSCOCO-Bias set, fine-tuning through the visual representation (Inception v3 [32]) for 500,000 iterations.

4.4 Baselines and Ablations

Baseline-FT. The simplest baseline is fine-tuning the Show and Tell model through the LSTM and convolutional networks using the standard cross-entropy loss on our target dataset, the MSCOCO-Bias dataset.

Balanced. We train a Balanced baseline in which we re-balance the data distribution at training time to account for the larger number of men instances in the training data.

	MSCOCO-Bias		MSCOCO-Balanced	
Model	Error	Ratio Δ	Error	Ratio Δ
Baseline-FT	12.83	0.14	19.30	0.51
Balanced	12.85	0.14	18.30	0.47
UpWeight	13.56	0.08	16.30	0.35
Equalizer w/o ACL	7.57	0.04	10.10	0.26
Equalizer w/o Conf	12.38	0.11	17.40	0.45
Equalizer	**7.02**	**-.03**	**8.10**	**0.13**

Table 1: Evaluation of predicted gender words based on error rate and ratio of generated sentences which include the "woman" words to sentences which include the "man" words. Equalizer achieves the lowest error rate and predicts sentences with a gender ratio most similar to the corresponding ground truth captions (Ratio Δ), even when the test set has a different distribution of gender words than the training set, as is the case for the MSCOCO-Balanced dataset.

Even though we cannot know the correct distribution of our data at test time, we can enforce our belief that predicting a woman or man should be equally likely. At training time, we re-sample the images of women so that the number of training examples of women is the same as the number of training examples of men.

UpWeight. We also experiment with upweighting the loss value for gender words in the standard cross entropy loss to increase the penalty for a misclassification. For each time step where the ground truth caption says the word "man" or "woman", we multiply that term in the loss by a constant value (10 in reported experiments). Intuitively, upweighting should encourage the models to accurately predict gender words. However, unlike our Confident Loss, upweighting drives the model to make either "man" or "woman" predictions without the opportunity to place a high probability on gender neutral words.

Ablations. To isolate the impact of the two loss terms in Equalizer, we report results with only the Appearance Confusion Loss (Equalizer w/o Conf) and only the Confidence Loss (Equalizer w/o ACL). We then report results of our full Equalizer model.

4.5 Results

Error. Table 1 reports the error rates when describing men and women on the MSCOCO-Bias and MSCOCO-Balanced test sets. Comparing to baselines, Equalizer shows consistent improvements. Importantly, our full model consistently improves upon Equalizer w/o ACL and Equalizer w/o Conf. When comparing Equalizer to baselines, we see a larger performance gain on the MSCOCO-Balanced dataset. As discussed later, this is in part because our model does a particularly good job of decreasing error on the minority class (woman). Unlike baseline models, our model has a similar error rate on each set. This indicates that the error rate of our model is not as sensitive to shifts in the gender distribution at test time.

Interestingly, the results of the Baseline-FT model and Balanced model are not substantially different. One possibility is that the co-occurrences across words are not balanced (e.g., if there is gender imbalance specifically for images with "umbrella" just bal-

	Women			Men			Outcome Divergence
Model	Correct	Incorrect	Other	Correct	Incorrect	Other	between Genders
Baseline-FT	46.28	34.11	19.61	75.05	4.23	20.72	0.62
Balanced	47.67	33.80	18.54	75.89	4.38	19.72	0.64
UpWeight	**60.59**	29.82	9.58	**87.84**	6.98	5.17	1.36
Equalizer w/o ACL	56.18	16.02	27.81	67.58	**4.15**	28.26	0.49
Equalizer w/o Conf	50.95	30.39	18.66	75.31	5.10	19.60	0.63
Equalizer (Ours)	57.38	**12.99**	29.63	59.02	4.61	36.37	**0.37**

Table 2: Accuracy per class for MSCOCO-Bias dataset. Though UpWeight achieves the highest recall for both men and women images, it also has a high error, especially for women. One criterion of a "fair" system is that it has similar outcomes across classes. We measure outcome similarity by computing the Jensen-Shannon divergence between Correct/Incorrect/Other sentences for men and women images (lower is better) and observe that Equalizer performs best on this metric.

ancing the dataset based on gender word counts is not sufficient to balance the dataset). We emphasize that balancing across all co-occurring words is difficult in large-scale settings with large vocabularies.

Gender Ratio We also consider the ratio of captions which include only female words to captions which include only male words. In Table 1 we report the *difference* between the ground truth ratio and the ratio produced by each captioning model. Impressively, Equalizer achieves the closest ratio to ground truth on both datasets. Again, the ACL and Confident losses are complementary and Equalizer has the best overall performance.

Performance for Each Gender. Images with females comprise a much smaller portion of MSCOCO than images with males. Therefore the overall performance across classes (i.e. man, woman) can be misleading because it downplays the errors in the minority class. Additionally, unlike [47] who consider a classification scenario in which the model is forced to predict a gender, our description models can also discuss gender neutral terms such as "person" or "player". In Table 2 for each gender, we report the percentage of sentences in which gender is predicted correctly or incorrectly and when no gender specific word is generated on the MSCOCO-Bias set.

Across all models, the error for Men is quite low. However, our model significantly improves the error for the minority class, Women. Interestingly, we observe that Equalizer has a similar recall (Correct), error (Incorrect), and Other rate across both genders. A caption model could be considered more "fair" if, for each gender, the possible outcomes (correct gender mentioned, incorrect gender mentioned, gender neutral) are similar. This resembles the notion of equalized odds in fairness literature [14], which requires a system to have similar false positive and false negative rates across groups. To formalize this notion of fairness in our captioning systems, we report the outcome type divergence between genders by measuring the Jensen-Shannon [19] divergence between Correct/Incorrect/Other outcomes for Men and Women. Lower divergence indicates that Women and Men classes result in a similar distribution of outcomes, and thus the model can be considered more "fair". Equalizer has the lowest divergence (0.37).

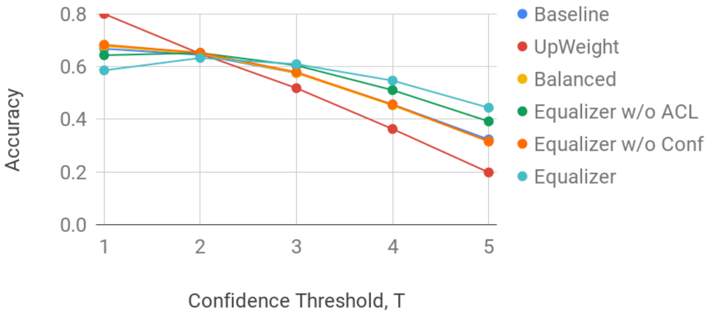


Fig. 3: Accuracy across man, woman, and gender neutral terms for different models as a function of annotator confidence. When only one annotator describes an image with a gendered word, Equalizer has a low accuracy as it more likely predicts gender neutral words but when more annotations mention gendered words, Equalizer has higher accuracy than other models.

Annotator Confidence. As described above, gender labels are mined from captions provided in the MSCOCO dataset. Each image corresponds to five captions, but not all captions for a single image include a gendered word. Counting the number of sentences which include a gendered word provides a rough estimate of how apparent gender is in an image and how important it is to mention when describing the scene.

To understand how well our model captures the way annotators describe people, instead of labeling images as either "man" or "woman", we label images as "man", "woman", or "gender neutral" based on how many annotators mentioned gender in their description. For a specific threshold value T, we consider an image to belong to the "man" or "woman" class if T or more annotators mention the gender in their description, and "gender neutral" otherwise. We can then measure accuracy over these three classes. Whereas a naive solution which restricts vocabulary to include no gender words would have low error as defined in Table 1, it would not capture the way humans use gender words when describing images. Indeed, the MSCOCO training set includes over 200,000 instances of words which describe people. Over half of all words used to describe people are gendered. By considering accuracy across three classes, we can better measure how well models capture the way humans describe gender.

Figure 3 plots the accuracy of each model with respect to the confidence threshold T. At low threshold values, Equalizer performs worse as it tends to more frequently output gender neutral terms, and the UpWeight model, which almost always predicts gendered words, performs best. However, as the threshold value increases, Equalizer performs better than other models, including at a threshold value of 3 which corresponds to classifying images based off the majority vote. This indicates that Equalizer naturally captures when humans describe images with gendered or gender neutral words.

Object Gender Co-Occurrence. We analyze how gender prediction influences prediction of other words on the MSCOCO-Bias test set. Specifically, we consider the 80 MSCOCO categories, excluding the category "person". We adopt the bias amplifica-

Accuracy	Woman	Man	All
Random	22.6	19.5	21.0
Baseline-FT	39.8	34.3	37.0
Balanced	37.6	34.1	35.8
UpWeight	43.3	36.4	39.9
Equalizer w/o ACL	48.1	39.6	43.8
Equalizer w/o Conf	43.9	36.8	40.4
Equalizer (Ours)	**49.9**	**45.2**	**47.5**

(a) Visual explanation is a *Grad-CAM* map.

Accuracy	Woman	Man	All
Random	25.1	17.5	21.3
Baseline-FT	45.3	40.4	42.8
Balanced	48.5	42.2	45.3
UpWeight	54.1	45.5	49.8
Equalizer w/o ACL	54.7	47.5	51.1
Equalizer w/o Conf	48.9	46.7	47.8
Equalizer (Ours)	**56.3**	**51.1**	**53.7**

(b) Visual explanation is a *saliency* map.

Table 3: *Pointing game* evaluation that measures whether the visual explanations for "man" / "woman" words fall in the person segmentation ground-truth. Evaluation is done for ground-truth captions on the MSCOCO-Balanced.

tion metric proposed in [47], and compute the following ratios: $\frac{count(man\&object)}{count(person\&object)}$ and $\frac{count(woman\&object)}{count(person\&object)}$, where *man* refers to all male words, *woman* refers to all female words, and *person* refers to all male, female, or gender neutral words. Ideally, these ratios should be similar for generated captions and ground truth captions. However, e.g. for *man* and *motorcycle*, the ground truth ratio is 0.40 and for the Baseline-FT and Equalizer, the ratio is 0.81 and 0.65, respectively. Though Equalizer over-predicts this pair, the ratio is closer to the ground truth than when comparing Baseline-FT to the ground truth. Likewise, for *woman* and *umbrella*, the ground truth ratio is 0.40, Baseline-FT ratio is 0.64, and Equalizer ratio is 0.56. As a more holistic metric, we average the *difference* of ratios between ground truth and generated captions across objects (lower is better). For male words, Equalizer is substantially better than the Baseline-FT (0.147 vs. 0.193) and similar for female words (0.096 vs. 0.99).

Caption Quality. Qualitatively, the sentences from all of our models are linguistically fluent (indeed, comparing sentences in Figure 4 we note that usually only the word referring to the person changes). However, we do notice a small drop in performance on standard description metrics (25.2 to 24.3 on METEOR [17] when comparing Baseline-FT to our full Equalizer) on MSCOCO-Bias. One possibility is that our model is overly cautious and is penalized for producing gender neutral terms for sentences that humans describe with gendered terms.

Right for the Right Reasons. We hypothesize that many misclassification errors occur due to the model looking at the wrong visual evidence, e.g. conditioning gender prediction on context rather than on the person's appearance. We quantitatively confirm this hypothesis and show that our proposed model improves this behavior by looking at the appropriate evidence, i.e. is being "right for the right reasons". To evaluate this we rely on two visual explanation techniques: Grad-CAM [30] and saliency maps generated by occluding image regions in a sliding window fashion.

Unlike [30] who apply Grad-CAM to an entire caption, we visualize the evidence for generating specific words, i.e. "man" and "woman". Specifically, we apply Grad-

CAM to the last convolutional layer of our image processing network, InceptionV3 [32], we obtain 8x8 weight matrices. To obtain saliency maps, we resize an input image to 299×299 and uniformly divide it into 32×32 pixel regions, obtaining a 10×10 grid (the bottom/rightmost cells being smaller). Next, for every cell in the grid, we zero out the respective pixels and feed the obtained "partially blocked out" image through the captioning network (similar to as was done in the occlusion sensitivity experiments in [42]). Then, for the ground-truth caption, we compute the "information loss", i.e. the decrease in predicting the words "man" and "woman" as $-\log(p(w_t = g_m))$ and $-\log(p(w_t = g_w))$, respectively. This is similar to the top-down saliency approach of [26], who zero-out all the intermediate feature descriptors but one.

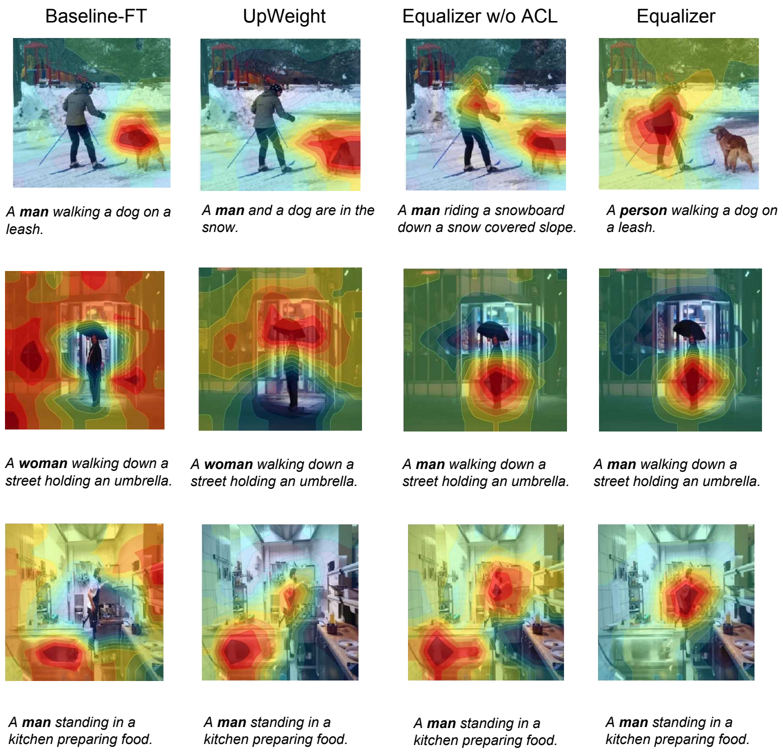
To evaluate whether the visual explanation for the predicted word is focused on a person, we rely on person masks, obtained from MSCOCO ground-truth person segmentations. We use the *pointing game* evaluation [44]. We upscale visual explanations to the original image size. We define a "hit" to be when the point with the highest weight is contained in the person mask. The accuracy is computed as $\frac{\#hits}{\#hits+\#misses}$.

Results on the MSCOCO-Balanced set are presented in Table 3 (a) and (b), for the Grad-CAM and saliency maps, respectively. For a fair comparison we provide all models with ground-truth captions. For completeness we also report the random baseline, where the point with the highest weight is selected randomly. We see that Equalizer obtains the best accuracy, significantly improving over the Baseline-FT and all model variants. A similar evaluation on the actual generated captions shows the same trends.

Looking at objects. Using our pointing technique, we can also analyze which MSCOCO objects models are "looking" at when they *do not* point at the person while predicting "man"/"woman". Specifically, we count a "hit" if the highest activation is on an object in question. We compute the following ratio for each gender: number of images where an object is "pointed at" to the true number of images with that object. We find that there are differences across genders, e.g. "umbrella", "bench", "suitcase" are more often pointed at when discussing women, while e.g. "truck", "couch", "pizza" – when discussing men. Our model reduces the overall "delta" between genders for ground truth sentences from an average 0.12 to 0.08, compared to the Baseline-FT. E.g. for "dining table" Equalizer decreases the delta from 0.07 to 0.03.

Qualitative Results. Figure 4 compares Grad-CAM visualizations for predicted gender words from our model to the Baseline-FT, UpWeight, and Equalizer w/o ACL. We consistently see that our model looks at the person when describing gendered words. In Figure 4 (top), all other models look at the dog rather than the person and predict the gender "man" (ground truth label is "woman"). In this particular example, the gender is somewhat ambiguous, and our model conservatively predicts "person" rather than misclassify the gender. In Figure 4 (middle), the Baseline-FT and UpWeight example both incorrectly predict the word "woman" and do not look at the person (women occur more frequently with umbrellas). In contrast, both the Equalizer w/o ACL and the Equalizer look at the person and predict the correct gender. Finally, in Figure 4 (bottom), all models predict the correct gender (man), but our model is the only model which looks at the person and is thus "right for the right reasons."

Discussion. We present the Equalizer model which includes an Appearance Confusion Loss to encourage predictions to be confused when predicting gender if evidence is



*A **man** walking a dog on a leash.*

*A **man** and a dog are in the snow.*

*A **man** riding a snowboard down a snow covered slope.*

*A **person** walking a dog on a leash.*

*A **woman** walking down a street holding an umbrella.*

*A **woman** walking down a street holding an umbrella.*

*A **man** walking down a street holding an umbrella.*

*A **man** walking down a street holding an umbrella.*

*A **man** standing in a kitchen preparing food.*

*A **man** standing in a kitchen preparing food.*

*A **man** standing in a kitchen preparing food.*

*A **man** standing in a kitchen preparing food.*

Fig. 4: Qualitative comparison of multiple baselines and our model. In the top example, being conservative ("person") is better than being wrong ("man") as the gender is not obvious. In the bottom example the baselines are looking at the wrong visual evidence.

obscured and the Confident Loss which encourages predictions to be confident when gender evidence is present. Our Appearance Confusion Loss, requires human rationales about what is visual evidence is appropriate to consider when predicting gender. We stress the importance of human judgment when designing models which include protected classes. For example, our model can use information about clothing type (e.g., dresses) to predict a gender which may not be appropriate for all applications. Though we concentrate on gender in this work, we believe the generality of our framework could be applied when describing other protected attributes, e.g., race/ethnicity and believe our results suggest Equalizer can be a valuable tool for overcoming bias in captioning models.

Acknowledgements. This work was partially supported by US DoD, the DARPA XAI program, and the Berkeley Artificial Intelligence Research (BAIR) Lab.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday **12**(9) (2007)
3. Barocas, S., Selbst, A.D.: Big data's disparate impact. California Law Review **104**, 671 (2016)
4. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in Neural Information Processing Systems (NIPS). pp. 4349–4357 (2016)
5. Buolamwini, J.A.: Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Ph.D. thesis, Massachusetts Institute of Technology (2017)
6. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1301–1309. Association for Computational Linguistics (2011)
7. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730. ACM (2015)
8. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2625–2634 (2015)
9. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226. ACM (2012)
10. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security **9**(12), 2170–2179 (2014)
11. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
12. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with r* cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1080–1088 (2015)
13. Gordon, J., Van Durme, B.: Reporting bias and knowledge acquisition. In: Proceedings of the 2013 workshop on Automated Knowledge Base Construction. pp. 25–30. ACM (2013)
14. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems (NIPS). pp. 3315–3323 (2016)
15. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3128–3137 (2015)
16. Larson, B.N.: Gender as a variable in natural-language processing: Ethical considerations (2017)
17. Lavie, M.D.A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). p. 376 (2014)

18. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). pp. 34–42 (2015)
19. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (1991)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)
21. van Miltenburg, E.: Stereotyping and bias in the flickr30k dataset. In: Workshop on Multimodal Corpora: Computer vision and language processing (2016)
22. Misra, I., Zitnick, C.L., Mitchell, M., Girshick, R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2930–2939. IEEE (2016)
23. of the President, U.S.E.O., Podesta, J.: Big data: Seizing opportunities, preserving values. White House, Executive Office of the President (2014)
24. Quadrianto, N., Petterson, J., Smola, A.J.: Distribution matching for transduction. In: Advances in Neural Information Processing Systems (NIPS). pp. 1500–1508 (2009)
25. Quadrianto, N., Sharmanska, V.: Recycling privileged learning and distribution matching for fairness. In: Advances in Neural Information Processing Systems (NIPS). pp. 677–688 (2017)
26. Ramanishka, V., Das, A., Zhang, J., Saenko, K.: Top-down visual saliency guided by captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, p. 7 (2017)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. ACM (2016)
28. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2017)
29. Ryu, H.J., Adam, H., Mitchell, M.: Inclusivefacenet: Improving face attribute detection with race and gender diversity. In: Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) (2018)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
31. Stock, P., Cisse, M.: Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. arXiv preprint arXiv:1711.11443 (2017)
32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
33. Tan, S., Caruana, R., Hooker, G., Lou, Y.: Detecting bias in black-box models using transparent model distillation. In: AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (2018)
34. Torralba, A.: Contextual modulation of target saliency. In: Advances in Neural Information Processing Systems (NIPS). pp. 1303–1310 (2002)
35. Torralba, A., Sinha, P.: Statistical context priming for object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). vol. 1, pp. 763–770. IEEE (2001)
36. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4068–4076. IEEE (2015)

37. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 3156–3164. IEEE (2015)
38. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5288–5296. IEEE (2016)
39. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning (ICML)
40. Yan, X., Yan, L.: Gender classification of weblog authors. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. pp. 228–230. Palo Alto, CA (2006)
41. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (TACL) **2**, 67–78 (2014)
42. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 818–833. Springer (2014)
43. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) (2018)
44. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 543–559. Springer (2016)
45. Zhang, K., Tan, L., Li, Z., Qiao, Y.: Gender and smile classification using deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). pp. 34–38 (2016)
46. Zhang, X., Yu, F.X., Chang, S.F., Wang, S.: Deep transfer network: Unsupervised domain adaptation. arXiv preprint arXiv:1503.00591 (2015)
47. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2017)
48. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. In: Proceedings of the International Conference on Learning Representations (ICLR) (2017)