SEEING DOUBLE: EVALUATING BIAS IN IMAGE CAPTIONING TECHNIQUES WITH SIMILAR IMAGES

Brandy Chen
brandyc@

Dora Zhao
dorothyz@

December 7, 2020

1 Introduction

Image captioning is the task of processing an input image and outputting a textual description of the image. In recent years, this has been done using deep learning techniques which have been made possible with the creation of large-scale datasets, like Microsoft Common Objects in Context (MSCOCO) [1], which provide both images and corresponding captions. Given the multimodal nature of the image captioning, including both a visual and language component, it has been a technically interesting task for computer vision researchers. However, beyond just the academic value, automated image captioning also has commercial and societal value. For one, image captioning has the potential for improving the accessibility of online images for visually impaired individuals [2].

Nonetheless, like many other computer vision techniques, image captioning models must confront the question of bias, especially with regards to protected attributes, such as gender or age. While gender classification is not an explicit task in image captioning, models have been found to implicitly learn to predict gender [3][4]. In doing so, they can produce results that not only reproduce but also augment harmful social and cultural biases. Therefore, being able to evaluate and then mitigate bias within image captioning systems is a critical task.

Given that image captioning models are predicting gender, it becomes important that they predict gender based on the visual features associated with the person being described rather than relying on contextual clues, such as the environment or the objects with which the person cooccurs. For example, ideally, a model will classify the gender in an image of a man snowboarding based on the man's features, not the snowboarding environment. However, it has instead been shown that many models tend to exploit contextual cues when generating captions, leading to biased models and potentially harmful gender biases [3] (see Figure 1).

Consequently, there has been a shift in focus towards bias mitigation. Multiple models and techniques that claim to address gender bias have been created. In this project, we focus on one specific bias mitigation technique: Hendricks and Burns et al.'s [3] Equalizer model presented in their 2018 paper "Women Also Snowboard: Overcoming Bias in Captioning Models." We introduce a new method, adapted from Stock and Cisse's [5] work on uncovering biases in ImageNet, that uses pairs of similar images differing only by the protected attribute of interest (e.g. gender) to critique bias mitigation techniques. Furthermore, we evaluate two methods– training on a more representative dataset and removing gendered assumptions for human annotators– for altering ground-truth data as a means of overcoming bias in captioning models.

2 Background and Related Work

2.1 Bias in Computer Vision

Our study on gender biases in image captioning is grounded in the recent increase in interest among the computer vision and artificial intelligence communities to evaluate and mitigate biases in automated decision making models. While biases can take many forms, there has been a particular focus on bias towards protected attributes, such as gender, given the sensitive nature and cultural implications of these biases. Within the field of computer vision, there have been works

Right for Wrong Reasons **Wrong**

A **woman** standing in a kitchen preparing food. A **man** in a suit and tie holding a book.

Figure 1: Two examples of the image captioning model exploiting contextual clues to predict gender. The image on the left predicts gender correctly but focuses on the contextual cue of the baking tray rather than the person. The image on the right predicts the wrong gender and focuses on the laptop in the background rather than the person.

focused on fairness spanning across many subdomains including facial recognition [6], pedestrian detection [7], image labelling [8], and image captioning [3][4].

Efforts to combat biases in computer vision address the issue at different stages within the pipeline. One method has been to address the bias present in the ground-truth or training data itself, which often underrepresent groups in certain contexts or contains stereotyped depictions. For example, in Buolamwini and Gebru's [6] seminal study "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," they introduce the Pilot Parliaments Benchmark, a more representative dataset both in terms of the skin color and gender of the subjects. Furthermore, in REVISE (REvealing VIsual biaSEs), Wang et al. [9] propose a tool that automatically detects forms of bias in a given visual dataset along three dimensions: object-based, gender-based and geography-based.

Another means of addressing bias is through creating models specifically focused on fairness and mitigating biases in their predicted results. More specifically in the field of image captioning, there have been different approaches to addresses gender biases in image captioning models. For instance, Tang et al.'s [4] Guided Attention Image Captioning (GAIC) model provides self guidance on visual attention to encourage models to capture correct gender visual evidence. At a high level, this is done with two complementary streams, the caption generation stream (Scg) and gender evidence mining stream (Sgm), where Scg helps generate high-quality descriptions and synthesize attention maps of gender words while Sgm forces the attention to be focused on the correct gender evidence. The GAIC model has been shown to reduce gender prediction errors with a competitive caption quality. Another image captioning model that attempts to mitigate gender bias is Hendricks and Burns et al.'s [3] Equalizer model, which will be further discussed below.

2.2 Similar Examples for Model Criticism

In Stock and Cisse's [5] 2018 paper "ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases," they propose using an adversarial example approach to model criticism for uncovering undesired biases, such as gender bias. Stock and Cisse sample pairs of images that are identical save for the race of the subject pictured. Using these pairs of images, Stock and Cisse then compare the image tags produced by a ResNet-101. In their analysis of the subcategory `basketball`, they found that images containing Black individuals are classified to be `basketball` whereas similar images with non-Black individuals are labeled differently. While Stock and Cisse do not conduct further analysis as to why this disparity occurs, their proposed adversarial example approach does provide a method for uncovering bias present in models.

However, in presenting their methodology for the adversarial example approach, Stock and Cisse do not provide rationale as to how or why certain images are chosen as similar images. They state that the similar images are sampled from the Internet but do not provide further details. In this study, we adopt Stock and Cisse's approach of finding similar image pairs for model criticism and apply a more rigorous approach by quantitatively measuring the similarity between the images. Through this more detailed sampling approach, we are able to propose a replicable method that can be applied to uncovering gender biases in computer vision tasks more broadly.

2.3 Show and Tell Model

Hendricks and Burns et al.'s [3] Equalizer is built on top of Vinyals et al.'s [10] Show and Tell model pre-trained on MSCOCO. The Show and Tell model is a neural image caption generator that consists of two parts, an encoder and a decoder. The encoder is a deep convolutional neural network that takes an image and returns a fixed-length vectorized representation of the input image. The decoder than takes the vectorized representation of the image and decodes it to produce a caption.

For the encoder, the Show and Tell model specifically uses the Inception v3 image recognition model pretrained on ImageNet. The layers of the Inception v3 consist of convolutions, average pooling, max pooling, concats, dropouts and fully connected layers. The loss is computed using softmax.

For the decoder, the Show and Tell model specifically uses a long short-term memory (LSTM) network. More generally, LSTM networks are a type of recurrent neural networks that are typically used for sequence modeling tasks, such as machine translation or speech recognition. In Show and Tell, the LSTM network is trained as a language model conditioned on the image encodings from the encoder.

2.4 Equalizer

Hendricks and Burns et al.'s [3] Equalizer model overcomes bias in captioning models by forcing models to look at a person rather than contextual cues in the surrounding environment to make gender-specific predictions. While within fairness in machine learning literature the focus tends to be more on not utilizing protected attributed (e.g. gender or sexual orientation) when making automated decision, the Equalizer model is trying to predict the protected attribute of gender albeit using the correct evidence in the image. To do so, the model is built on top of the Show and Tell Model with a novel loss function. This loss function utilizes three terms, the Appearance Confusion Loss ($\mathcal{L}^{AC}$), the Confident Loss ($\mathcal{L}^{\text{Con}}$) and standard cross entropy loss ($\mathcal{L}^{CE}$) to mitigate impacts of unwanted bias in a dataset. The model itself is a linear combination of the three losses:

$$\mathcal{L} = \alpha\mathcal{L}^{CE} + \beta\mathcal{L}^{AC} + \mu\mathcal{L}^{Con} \tag{1}$$

with hyperparamters α, $\beta = 1$ and $\mu = 10$.

2.4.1 Appearance Confusion Loss (ACL)

Appearance Confusion Loss encourages models to be confused when predicting gender in an image that does not contain appropriate evidence. It uses masks, which are 1 for pixels which do not contribute to a gender decision and 0 for pixels which are appropriate to be considered in gender decisions, to provide the ground truth rationales that indicate which evidence is appropriate for a particular gender decision. Thus, given an image I and a mask M, the Hadamard product of I and M, or $I \odot M$, yields a new image I' that contains everything except the gender information that the implementer deems appropriate for gender decisions. Examples of such images are shown in Figure 2.

The ACL ($\mathcal{L}^{AC}$) is defined as follows:

$$\mathcal{L}^{AC} = \frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T} \mathbb{1}\!\!\!/ \left(w_t \in \mathcal{G}_w \cup \mathcal{G}_m\right) \mathcal{C}\left(\tilde{w}_t, I'\right) \tag{2}$$

$$\mathcal{C}\left(\bar{w}_t, I'\right) = \mid \sum_{g_w \in \mathcal{G}_w} p\left(\tilde{w}_t = g_w \mid w_{0:t-1}, I'\right) - \sum_{g_m \in \mathcal{G}_m} p\left(\bar{w}_t = g_m \mid w_{0:t-1}, I'\right) \mid \tag{3}$$

$\mathcal{C}\left(\bar{w}_t, I'\right)$ is the confusion function that operates over the predicted distribution of words $p(\bar{w}_t)$, a set of woman gender words ($\mathcal{G}_1$) and a set of man gender words ($\mathcal{G}_m$). $\mathbb{1}\!\!\!/$ is an indicator variable that denotes whether $\bar{w}_t$ is a gendered word. In this equation, N is the batch size, T is the number of words in the sentence, w_t is the ground-truth word at time t, and $\bar{w}_t$ is the predicted word at time t.

2.4.2 Confident Loss (Conf)

Confident Loss encourages models to be confident on images in which gender information is present. Given images I and functions $\mathcal{F}^W$ and $\mathcal{F}^M$, which measure how confidently the model predicts woman and man words respectively, the Confident Loss ($\mathcal{L}^{\text{Con}}$) is defined as follows:

$$\mathcal{L}^{\text{Con}} = \frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T} \left(\mathbb{1}\!\!\!/(w_t \in \mathcal{G}_w)\, \mathcal{F}^W\left(\bar{w}_t, I\right) + \mathbb{1}\!\!\!/\left(w_t \in \mathcal{G}_m\right) \mathcal{F}^M\left(\bar{w}_t, I\right)\right) \tag{4}$$

Figure 2: Examples of person masks needed to train the Appearance Confusion Loss term. Original images are on the left and masked images are on the right. Here, the masked images contain everything except with people blocked out, as the people represent gender evidence deemed appropriate for gender predictions.

Again, N is the batch size, T is the number of words in the sentence, w_t is the ground-truth word at time t, and $\bar{w}_t$ is the predicted word at time t. $\mathcal{F}^W(\bar{w}_t, I)$ and $\mathcal{F}^M(\bar{w}_t, I)$ are the quotients between the predicted probability for man and woman gender words and for woman and man gender words, respectively. Thus, when the model is confident of a gendered prediction, such as for the word "man," the probability of the word "man" should be higher than the probability for the word "woman." This results in a small value for $\mathcal{F}^M$ and therefore a small loss. Also, by considering the quotients between predicted probabilities, models are encouraged to distinguish between gendered words without being forced to predict a gendered word.

2.4.3 Metrics and Evaluations

The Equalizer Model has a lower error rate, or the number of man/woman misclassifications, and is able to more closely match the ground truth ratio of sentences containing "woman" to sentences containing "man." Using pointing game evaluation and the Gradient-weight Class Activation Mapping (Grad-CAM) [11] system to measure if the model is right for the right reasons, the Equalizer model also performs better than others models in consideration. More specifically, pointing game evaluation measures whether the visual explanations created using the Grad-CAM approach for gender words fall in the person segmentation ground-truth.

3 Approach

3.1 Dataset Creation

We start by creating a dataset of pairs of similar images that we will then use to evaluate the Equalizer model's performance. The intuition behind creating this dataset is that, if given two images that differ only by the gender of the subject, an unbiased model would return identical captions except with the appropriate gendered language. Ideally a

researcher would be able to take photos to create these pairs; however, given the resources that may go into undertaking such a task, we propose a method that utilizes existing images readily available on the Internet.

Given its prevalence within image captioning research, we use Microsoft Common Objects in Context (MSCOCO) dataset [1] as the basis for our experimentation. To start creating our similar images validation set, we first determine what categories of images to include. This is done by determining the top 25 most commonly co-occurring MSCOCO object categories with male or female images. To determine if an image is male or female, we look at the captions for uses of male indicator (e.g. "man", "boy", "men") or female indicator (e.g. "woman", "girl", "women") language. Since MSCOCO includes 5 captions per image, if the majority of the captions contain a use of a certain indicator and none of the captions for that image contain an indicator for the opposite gender, we consider the image to be of that particular gender. To illustrate, if three of the captions use a male indicator word and none of the captions use a female indicator word, we consider this image to be depicting a male.

After selecting our gendered images, we evaluate the object co-occurrences as the ratio between the number of times an object appears with male / female images over the total number of male / female images that we have.

$$\text{Co-occurrence}(x, p) = \frac{\text{\# of images } x \text{ appears in with } p}{\text{\# of images with } p} \tag{5}$$

Here x refers to an object category in MSCOCO (e.g. skateboard, tie) and p refers to a person category (e.g. male or female).

For example, we find that "skateboard" and "tie" often commonly co-occur with males while "handbag" and "hair drier" commonly co-occur with images containing females. From the top 25 most commonly co-occurring categories, we hand-select categories that contain more images featuring only one person given that this was easier for us to find matching similar images.

The goal here is to have a small number of categories where each contain a fair number of images instead of the opposite. Considering that and challenges of finding similar images of males in categories that are more female biased, we end up with six categories of objects that males are commonly found to be pictured with and four categories for females. The exact categories chosen are shown in Table 1.

Male Categories	Female Categories
frisbee	hairdrier
racket	handbag
skateboard	refrigerator
sports ball	toothbrush
surfboard	
tie	

Table 1: The categories of objects from MSCOCO chosen for each gender that we find similar images for

Having decided on the categories for each gender, at least two images of the associated gender are selected from the MSCOCO validation set for each category. These selected images from the validation set serve as one of the two images in every similar image pair to be created. With that, each chosen validation set image is then used to find five similar potential pair-forming images that differ only in gender. These images are gathered from the Internet, largely from Flickr, and found by querying pairwise combinations for words (e.g. "woman + tie"). Here, we attempt to replicate Lin et al.'s [1] process for creating the MSCOCO dataset as closely as possible. Figure 3 shows the MSCOCO validation image and its five potential similar pair-forming images.

After finding five similar images for each of the selected MSCOCO validation set images, we determine the most similar image out of the five. Whereas previous attempts to find similar images [5] do not delineate a clear and rigorous method for determining similar images, we want to provide a quantitative rationale for selecting the similar image. Thus, we determine the most similar image for each by extracting the image features using a ResNet 50 and calculate the Euclidean loss. We also conduct an experiment in which we use a histogram of gradients to extract global features paired with Euclidean loss. However, the losses are much higher than those with the ResNet and upon human validation of the matches, are less visually similar than those matches found using ResNet.

Using the flattened image features extracted using ResNet, we visualize this information using the t-SNE dimensionality reduction technique. The images are now plotted in two dimensions with the image features that are more similar being plotted closer together in the plot. As shown in Figure 4, the images for the same object categories tend to cluster

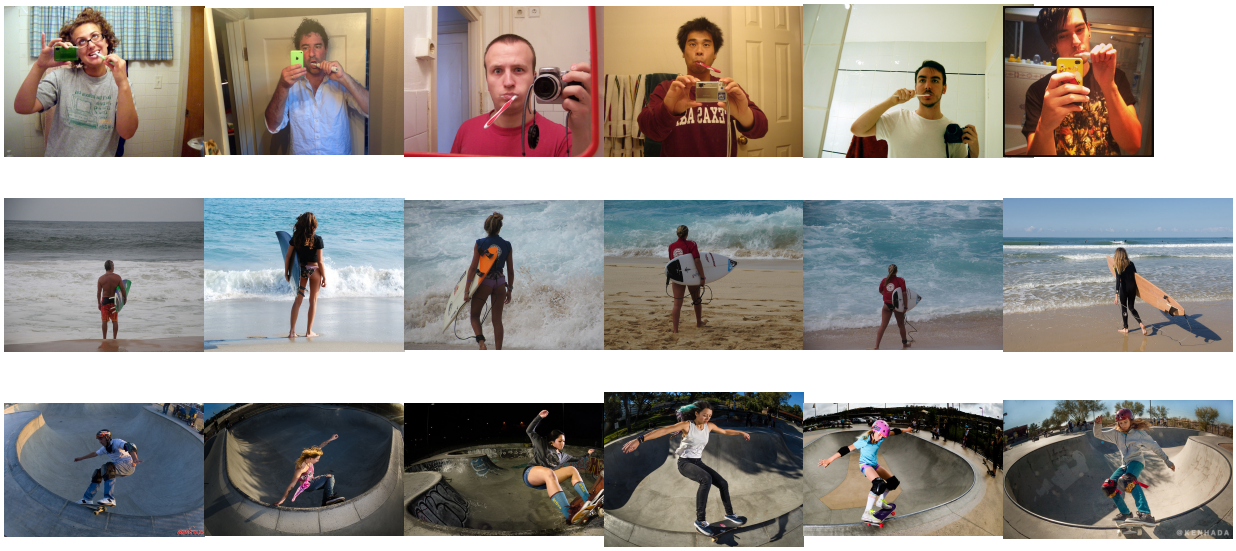


Figure 3: From left to right, the MSCOCO validation set image and its five potential similar pair-forming images.

together. For example on the lower right-hand side, there is a noteable cluster of surboard images. We use this technique as a check to ensure that there is not a noticeable divide between the MSCOCO and Flickr images.

Thus, using the image features extracted using ResNet, the image with the lowest Euclidean loss is considered to be the most similar to its counterpart image from the MSCOCO validation set. Finally, we ensure that the selected most similar image and its counterpart MSCOCO validation set image are both from the same object category. With this procedure, we end up with 51 pairs of similar images, some of which are shown in Figure 5.

To create the ground truth captions for the Flickr similar images, we utilize the ground truth captions from the MSCOCO validation set images. Since the pairs of images are all fairly similar, the captions did not change significantly during the conversion. Specifically, during conversion, gender specific language is updated to reflect the correct gender and objects that appear in the similar Flickr image (see Figure 5). Captions containing gender neutral language are not changed.

3.2 Analysis

Using our dataset of similar images, we now have a new means for analyzing the results of the Equalizer model proposed by Hendricks and Burns et al [3]. Our guiding intuition is that, given our pairs of images are similar in composition minus the gender of the target individual, the captions produced should be identical barring the predicted gender.

3.2.1 Models

We use the six models that Hendricks and Burns et al. describe in "Women Also Snowboard". The baseline model, **Baseline-FT**, is the Show and Tell model that has been fine-tuned on the target dataset. **Balanced** and **Upweight** are two other baseline models in which the gender distribution has been balanced to take into account the higher proportion of men and the loss value for gendered words has been increased, respectively. The other three models are **Equalizer w/o ACL** or Equalizer with only Confidence Loss, **Equalizer w/o Conf** or Equalizer with only Appearance Confusion Loss, and **Equalizer** or the full Equalizer model. These three versions of the Equalizer model allow for more isolated analysis of the impact of the two loss terms in Equalizer.

3.2.2 Quantitative Metrics

The first set of metrics that we use are **error rate** and **gender ratio**.

- **Error Rate**: The number of images where gender is incorrectly predicted over the total number of images. We do not penalize the model for predicting a gender neutral term.
- **Gender Ratio**: The ratio of "female" sentences to "male" sentences. A sentence is considered "female" if it contains a word that belongs to a precompiled list of female indicator words and "male" if it contains a word that belongs to a precompiled list of male indicator words.

Figure 4: t-SNE visualization of validation images from MSCOCO and the five potential similar pair-forming images selected from Flickr

We report gender ratio using Ratio Δ which is the difference between the ground-truth gender ratio and the predicted gender ratio. These two metrics are related to the gender classification and are the original metrics used in the "Women Also Snowboard" paper [3].

The second type of metric that we use is **sentence similarity**, which is calculated as the average cosine similarity between the predicted captions for the male versus the female images in a pair. This is done using Cer et al.'s [12] Universal Sentence Encoder, which creates sentence level embeddings that are 512 dimensional vectors. The semantic similarity between two sentences is simply the inner product between its two encodings. While we are focused on evaluating the models and how accurate their gender predictions are, we are also interested in the sentence similarity score to see the accuracy of the caption content excluding the gender and whether or not there is a trade off. Thus, prior to inputting the generated captions from the six models into the Universal Sentence Encoder, we convert all captions to be more gender neutral by replacing gender words like "woman" or "his" to more neutral words like "person" or "their." In doing so, we prevent the differences in gender words in the captions from skewing the similarity scores.

Finally, we include an evaluation on four standard image captioning evaluate metrics:

- **BLEU (Bilingual Evaluation Understudy)** [13]: Analyzes the co-occurrences of n-grams, or a contiguous sequence of n items in a given sample of text, between the candidate sentence and reference sentences. The overall BLEU score is computed using a weighted geometric mean of the individual n-gram precision for n = 1, 2, 3 and 4.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** [14]: Generates an alignment between the words in the candidate and reference sentences, with an aim of 1:1 correspondence. Given a set of alignments, the METEOR score is the harmonic mean of precision and recall between the best scoring reference and candidate.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** [15]: A set of three evaluation metrics designed to evaluate text summarization algorithms. In this paper, ROGUE_L is used. ROGUE_L uses a measure based on the Longest Common Subsequence (LCS), or a set of words shared by two sentences which

MSCOCO Image

A girl is taking a selfie while brushing her teeth.
In bathroom taking selfie of herself brushing teeth.
A woman taking a picture of herself while brushing her teeth.
A woman brushing her teeth while taking a selfie.
A woman brushing her teeth while taking a picture of herself.

A man and his surfboard survey a stormy sea.
A man with a surfboard is watching the waves.
A man holding a surfboard on top of a sandy beach.
A man stands on the beach holding his surfboard.
A surfer holding a surf board at the edge of the beach watching the surf.

Flickr Image

A man is taking a selfie while brushing his teeth.
In bathroom taking selfie of himself brushing teeth.
A man taking a picture of himself while brushing his teeth.
A man brushing his teeth while taking a selfie.
A man brushing his teeth while taking a picture of himself.

A woman and her surfboard survey a stormy sea.
A woman with a surfboard is watching the waves.
A woman holding a surfboard on top of a sandy beach.
A woman stands on the beach holding her surfboard.
A surfer holding a surf board at the edge of the beach watching the surf.

Figure 5: Examples of the similar image pairs consisting of a MSCOCO image and a Flickr image. Each image has five ground truth captions that we manually converted to fit the Flickr image and replace any instances of gendered language. If gendered language is not present (e.g. surfer), we keep the caption as is.

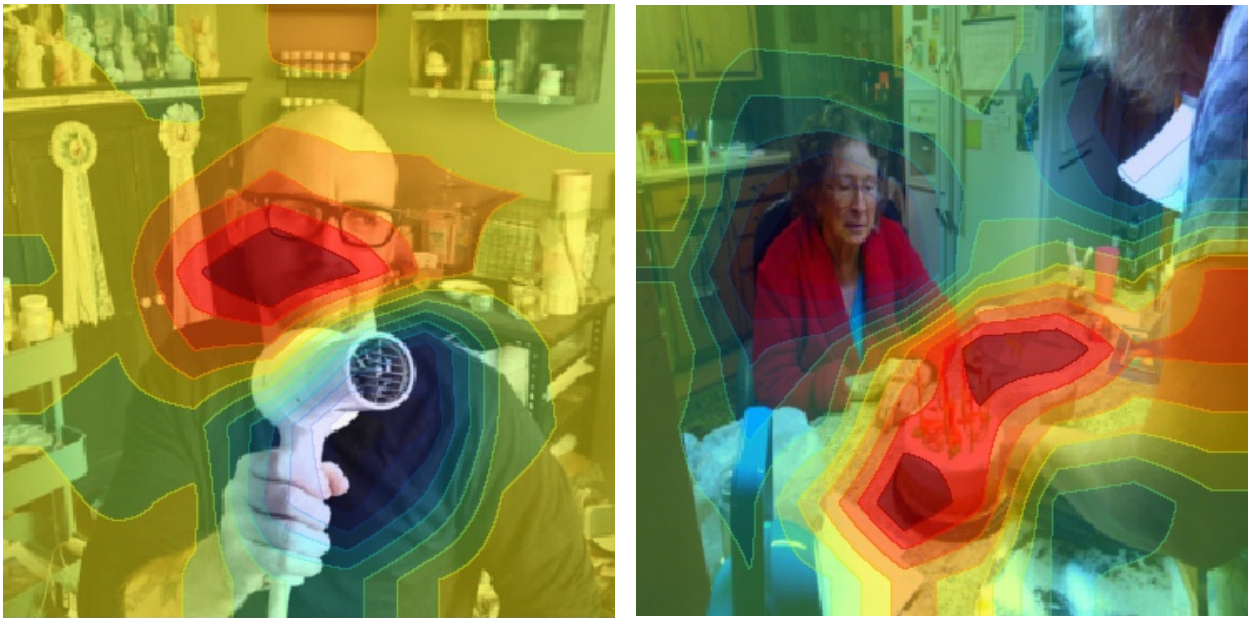
occur in the same order. Given the length of the LCS between a pair of sentences, the ROGUE_L score computes an F-measure.

- **CIDEr (Consensus-based Image Description Evaluation)** [16]: Measures consensus in captions by performing a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram. The CIDEr score for n-grams of length n is computed using the average cosine similarity between the candidate sentence and the reference sentences.

While not originally included in "Women Also Snowboard," we feel it is valuable to include metrics that take into consideration the predicted objects in the caption in addition to the set of metrics that look at gender classification. We chose these four metrics since they are the original metrics used by the MSCOCO evaluation server [17].

3.2.3 Qualitative Metrics

In addition to the quantitative metrics, we also look at heat maps produced using Grad-CAM [11] for applying the pointing game evaluation on the models. In particular, we are interested in seeing whether the models are "right for the right reasons" [3]. As illustrated in Figure 6, this means that when a model predicts gendered words, we want to ensure that the person picture is most important for determining the gender rather than contextual clues or objects, such as a surfboard or handbag. Thus with this qualitative evaluation, we are mainly focused on the images for which gender is predicted.



*A **man** is holding a glass of wine.*
Right for the Right Reasons

*A **woman** sitting at a table with a cake.*
Right for the Wrong Reasons

Figure 6: Equalizer Grad-CAM Results illustrating right for the right reasons (L) and right for the wrong reasons (R)

3.2.4 Model Performance

The Equalizer model performs as expected when evaluating based on the gender ratio and error rate metrics that were originally proposed by Hendricks and Burns et al.'s [3]. From Table 3, we see that the Equalizer model has the lowest error rate at 13.8%, which is 8.5% lower than the best performing baseline model, Upweight. Also consistent with the findings from Hendricks and Burns et al. [3], we see that the ACL and Conf loss terms are complementary in that the combined loss in Equalizer performs better than either loss term separately. Similar to the error rates, Equalizer also predicts a gender ratio closest to the that of the ground truth. However, the positive Ratio Δ indicates that Equalizer is still overpredicting "male" sentences compared to the ground truth. Compared to the Ratio Δs of the Equalizer in the original paper, which were -0.03 and 0.13, the Ratio Δ of 0.344 appears relatively high. We believe this is due to the fact that our similar images datasets consists of images from categories that are highly skewed towards one gender in the MSCOCO dataset. This may make the task of correctly predicting the gender harder than for images more generally in MSCOCO.

Figure 7: Poor Grad-CAM images of Equalizer. From left to right, Equalizer focuses on the background corners for the first two images and focuses on the soccer ball for the last image, despite there being clear gender evidence.

The Equalizer does not have a boost in performance when we compare the sentence similarity and MSCOCO image captioning metrics. In Table 3, we see that Equalizer has the lowest sentence similarity score of 0.767 out of the six models. In particular, Equalizer is very successful in generating captions with similar content for iconic images focused on isolated individuals in categories like skateboard and surfboard. But it performs poorly when it comes to images with higher variability, such as categories like tie or hair dryer. While this is a common trend among all six models, Equalizer is observed to especially struggle and more often perform the worst in terms of similarity score among the six models when it comes to generating captions for a given pair of similar images.

Furthermore, as seen in Table 4, Equalizer performs worse than the Baseline-FT model on the CIDEr metric with a score of 1.103 versus 1.130. Because of the fluctuating values for the MSCOCO image captioning metrics, we also provide qualitative observations on the prediction captions. Upon closer examination of the predicted captions for Equalizer, we find that, while the predicted gender is correct more often than for other models, the objects predicted are often incorrect. For example, we notice that Equalizer incorrectly predicts a "nintendo wii game controller"– an object that does not appear in our image set– for two pictures for which the gender is predicted correctly. Incorrectly predicting objects associated with a gendered individual can also reinforce gender stereotypes, which we are looking to prevent. More generally, there is a trade off between accuracy of gender predictions and caption content in Equalizer, where Equalizer tends to get the gender right at the expense of the rest of the caption content.

When it comes to the qualitative metrics obtained from analyzing the Grad-CAM images, Equalizer has the highest ratio classified right for the right reasons to right for the wrong reasons. Specifically, it performs second best behind the Upweight model in determining the most number of images right for the right reasons. Even though Upweight predicts the correct gender in more instances, it has a higher error rate compared to Equalizer since it rarely predicts gender neutral options and thus has a higher number of images for which gender is predicted incorrectly as well. Equalizer also has the least number of images that are right for the wrong reasons, as expected. But despite the promising performance, we find that the Grad-CAM images are not as correct or focused on the right people as we would have expected them to be, even for Equalizer. For example, as shown in Figure 7, when predicting some images, Equalizer is focused on the corners of the image, an irrelevant part of the image with no gender evidence, or an object that should be avoided for gender evidence.

One last weakness of the Equalizer model we noticed is that it requires high quality image segmentation masks for each photo. While we chose to make these masks manually, we also know that another option could be to use a computer vision method, such as a Mask R-CNN [18] for this purpose. However, given that the original Mask R-CNN model [18] was also trained on MSCOCO, we want to note that this could be another source for injecting bias. Regardless of the method used, we believe making these masks is a timely task. Thus, we propose a method in our experiments section that circumvents the need to make these masks.

4 Experiments

Based on our analysis of the Equalizer Model, we have found that while it has a gender ratio closer to the ground truth and a lower error rate than the other models that Hendricks and Burns et al. evaluated, it still has several shortcomings. In particular, when looking at the image captioning metrics that evaluate the accuracy of the overall caption via sentence similarity, including the objects predicted, rather than just the gender ratio, Equalizer does not do better and in fact performs worse than the baseline on several metrics.

Model	Right for the Right Reasons	Right for the Wrong Reasons	Ratio
Baseline-FT	38	22	1.73
Balanced	32	32	1.00
Upweight	**46**	26	1.77
Equalizer w/o ACL	35	29	1.21
Equalizer w/o Conf	33	25	1.32
Equalizer	40	**21**	**1.90**

Table 2: The number of images models got right for the right or wrong reasons. Upweight had the most images classified as right for the right reasons as well as the most images classified. However, Equalizer has the highest ratio of images classified right for the right reasons to right for the wrong reasons.

Model	Error Rate	Ratio Δ	Sentence Similarity
Baseline-FT	0.287	0.515	0.818
Balanced	0.266	0.528	0.825
Upweight	0.223	0.465	**0.838**
Equalizer w/o ACL	0.213	0.470	0.773
Equalizer w/o Conf	0.277	0.541	0.802
Equalizer	**0.138**	**0.344**	0.767

Table 3: Error rate, gender ratio and sentence similarity for the six models. Equalizer performs significantly better for error rate and ratio Δ. The three baseline models perform in line for sentence similarity and better than all three Equalizer variations.

4.1 Training on a More Representative Dataset

It has been suggested that training on a more representative dataset, such as Buolamwini and Gebru's [6] Pilot Parliaments Benchmark, can be used as a technique for combating biases. We acknowledge that Hendricks and Burns et al. utilize this technique with the Balanced baseline model, which is trained on a dataset where the number of training examples of women is the same as the number of training examples of men. However, simply balancing the number of women and men training examples in MSCOCO does not address the imbalances within categories of objects that predominantly co-occur with men and women. For example, object categories such as surfboard and skateboard contain more images of men than women. Moreover, even when a dataset or category does contain a balanced number of men and women images, there may still exist appearance differences and stereotyped representations that create bias. For example, Wang et al. [9] found that in OpenImages, images of men with flowers are often in official and formal settings while images of women with flowers are often in paintings and other staged settings.

With that in mind, we decide to use the remaining images not used for the creation of the similar image pairs to create a more representative dataset containing more balanced categories. In other words, we hope that we can create more balanced categories in which categories of objects predominantly containing images of one gender will be more balanced out with the addition of our images of the opposite gender. After removing any duplicates and conflicting images with the images used in the similar pairs, there were 200 remaining unique images, which we plan to use to fine-tune the Baseline-FT model to see if it helps with bias mitigation.

We also implement the following data augmentation techniques to expand our small set of 200 images to 1000 images:

- **Shear**: Uniformly sample shear parameters within $[-0.3, 0.3]$ and shear all images along the x and y axes. The shear parameters are the numbers added to the 2 x 3 shear matrix.
- **Scale**: Scale all images by 150%.
- **Crop**: Crop all images so that the main person or people take up most of pixels in the cropped images. Essentially, the cropped images remove unnecessary background elements and features and force the models to focus more on the individuals for gender evidence by limiting the number of available features. The cropped images also stayed true to the ground truth captions of the non-cropped images.
- **Flip**: Flip all images randomly, either vertically, horizontally or both vertically and horizontally.

Similar to how the ground truth captions are generated for the Flickr images in the similar image pairs, the ground truth captions for the 1000 images are also the captions of the corresponding MSCOCO images but updated and edited appropriately. Specifically, the ground truth captions for the data augmented images are the same as the captions for

Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Baseline-FT	0.723	0.557	0.414	0.319	0.285	0.572	1.130
Balanced	0.719	0.555	0.412	0.318	0.278	0.566	1.137
Upweight	0.720	0.554	0.413	0.320	0.278	0.577	1.103
Equalizer w/o ACL	0.687	0.511	0.360	0.266	0.258	0.537	0.964
Equalizer w/o Conf	**0.734**	0.564	**0.425**	0.334	**0.287**	0.577	**1.145**
Equalizer	0.725	**0.565**	0.431	**0.339**	**0.287**	**0.578**	1.103

Table 4: MSCOCO evaluation metrics for the six models. Equalizer w/o Conf and Equalizer appear to perform the best across the four metrics. However, there is not one model that performs the best compared to the other five across all four image captioning metrics.

the original 200 non-augmented images, as the content of the images does not change during the data augmentations mentioned above.

Given the time and computing constraints, we choose to fine-tune the existing models provided by Hendricks and Burns et al. [3] as opposed to training the Show and Tell model from scratch. We freeze the Inception v3 submodel variables and fine-tune only the LSTM model. To start, we fine-tune the Baseline-FT model using the non-augmented 200 images and all four of the augmentations over ten epochs. However, after evaluating the model on our created validation set and the MSCOCO 2017 validation set, we found that the model was overfitting to the data even when we trained on only one epoch. The model was only predicting keywords, such as "skateboard" and "tie," even for images in which these objects did not occur. Thus, we then decided to remove one of the augmentations and train on only the original 200 non-augmented images and three data augmentation techniques– shear, crop, and flip. We choose to remove scale since we felt that it was not contributing anything "new" and only providing a larger duplicate of the non-augmented images. Again, we start to fine-tune the baseline model over 10 epochs but found that the model was overfitting and predicting only "a" in the captions. Therefore, for the final model, which we refer to as **Baseline-Aug**, we train over 5 epochs with 800 images– non-augmented, sheared, cropped, and flipped images.

Model	Error Rate	Ratio Δ	Sentence Similarity
Baseline-FT	0.287	0.515	0.818
Equalizer	0.138	0.344	0.767
Baseline-Aug	0.255	0.534	0.756

Table 5: Error rate, Ratio Δ, and sentence similarity for the augmented baseline model, Equalizer, and Baseline-FT. Baseline-Aug performs worse in gender ratio and sentence similarity than both Baseline-FT and Ratio Δ. The error rate is in line with Baseline-FT.

Based on results shown in Table 5, Baseline-Aug does not improve much compared to Baseline-FT. Its error rate decreased marginally but its sentence similarity and Ratio Δ worsened. Overall, Equalizer still performs the best out of the three. We hypothesize that the lack of performance improvement is due to the small size of the dataset we are using to fine-tune, despite our data augmentations to expand the dataset, and being only able to fine-tune rather than train from scratch. Finally, even with our augmented dataset, we still do not perfectly balance all the co-occurrences, which is in part what makes the concept of having a perfectly balanced or representative dataset difficult to attain.

4.2 Correcting for Gendered Assumptions in Ground-Truth Captions

One of the shortcomings of the Equalizer model is that it requires high-quality person segmentation masks for each image in the training set in order to calculate the ACL loss term. Creating these masks for a large-scale data is extremely time-consuming [4]. Thus, while image segmentation annotations are provided for MSCOCO, the cost of creating masks for other large-scale datasets is most likely infeasible. To address this, we propose a new method that corrects for gendered assumptions made in the ground-truth captions and circumvents the need for making image masks.

Our intuition behind proposing this method comes from the fact that human annotators tend to assume male as the default in the ground-truth data even when gender information is not present. Thus, whereas the ACL loss term is attempting to correct for gender assumptions being made at the model-level, we are attempting to correct this within the ground-truth data itself.

We correct for the gendered assumptions made by human annotators by running a face detector through the MSCOCO 2017 validation set. If no faces are detected in the image, we take this to indicate that there is not enough relevant

information to infer gender. While we recognize that there are instances in which gender may be inferred even if the face is not detectable, we self-validated the first 50 photos and found the face detector to be a good proxy. We found that in most instances, the individuals were either turned away from the camera or too small for a human to be able to detect gender (see Figure 8). Moreover, given that we are trying to create a method that is more time efficient compared to the ACL loss term, we posit that a simplistic yet efficient method such as face detection is best suited for our goal. In total, we found that faces were not detected in 2264 of the images.

(a) Image 532481 from MSCOCO 2017 Val

(b) Image 458755 from MSCOCO 2017 Val

Figure 8: The first two images from the MSCOCO validation set in which faces could not be detected. The image on the left is an example for which the person is too small to infer gender. The image on the right is an example for which the person's face is occluded / turned away from the camera so gender again cannot be inferred.

After finding the images in which faces cannot be detected we then check the corresponding ground-truth captions using a pre-defined list of terms to see if there are any gendered terms present. If so, we then replace these terms with gender neutral terms, such as "person" versus "man" / "woman" or "their" versus "his" / "her." In total, we found that of the 2264 images that faces were not detected, 1475 (65.2 %) of the images used gendered language. Furthermore, 1128 of these images or 76.4% of those containing gendered languages used masculine words in the caption. This confirms our intuition that human annotators tend to default to male when gender information may not present.

Using the gender neutral captions and their corresponding images, we fine-tune the LSTM models for both the Baseline-FT and the Equalizer w/o ACL model over three epochs. The fine-tuned models will be referred to as **Baseline-Neutral** and **Equalizer w/o ACL-Neutral**, respectively. Given that our gender-neutral corrections to the ground-truth captions are meant to serve in the place of the ACL loss term, we expected that combining the corrected data with the Equalizer w/o ACL model should perform in-line with the Equalizer model.

After observing the disproportionate number of images assumed to be male over images assumed to be female even when adequate gender information was not present, we proposed a variation in which we correct all captions assumed as male to be female. For example, captions that used "man" were now changed to "woman". Similar to the gender neutral captions, we fine-tuned both the Baseline-FT and the Equalizer w/o ACL models on this female-assumed data, which will be referred to as **Baseline-Women** and **Equalizer w/o ACL-Women** respectively.

Training on the corrected ground-truth captions helps improve the models' ability to correctly predict the gender of individuals even for the models trained on the baseline. As shown in Table 6, for both Baseline-Neutral and Baseline-Women, the error rate and Ratio Δ have decreased compared to the Baseline-FT model. This means that these models are not only more likely to correctly identify the gender of the individual pictured but also predict a gender ratio closer to the ground truth. Despite the fine-tuned models' improvement for error rate and Ratio Δ, they still perform worse compared even to the baseline Upweight model as well as the Equalizer model. Furthermore, there appears to be a trade-off between the two aforementioned metrics and sentence similarity– a general theme that we saw when analyzing the different models proposed in "Women Also Snowboard."

Turning to look at the models trained on Equalizer w/o ACL, we find that Equalizer w/o ACL-Neutral and Equalizer w/o ACL-Women perform the same across all metrics except for sentence similarity (see Table 6). In our further analysis between the Equalizer and our fine-tuned models, we will be comparing to Equalizer w/o ACL-Neutral given its slightly better performance. However, we wanted to note that, contradictory to our hypothesis that Equalizer w/o

Model	Error Rate	Ratio Δ	Sentence Similarity
Baseline-FT	0.287	0.515	**0.818**
Equalizer w/o ACL	0.213	0.470	0.773
Equalizer	**0.138**	0.344	0.767
Baseline-Neutral	0.245	0.429	0.773
Baseline-Women	0.223	0.481	0.777
Equalizer w/o ACL-Neutral	0.213	**0.033**	0.764
Equalizer w/o ACL-Women	0.213	**0.033**	0.611

Table 6: Error rate, Ratio Δ, and sentence similarity. Though Equalizer performs better in terms of error rate, Equalizer w/o ACL-Neutral and Equalizer w/o ACL-Women have gender ratios much closer to the ground-truth gender ratio.

Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Baseline-FT	0.723	0.557	0.414	0.319	0.285	0.572	**1.130**
Equalizer w/o ACL	0.687	0.511	0.360	0.266	0.258	0.537	0.964
Equalizer	**0.725**	**0.565**	**0.431**	**0.339**	0.287	**0.578**	1.103
Baseline-Neutral	**0.725**	0.552	0.418	0.324	0.275	0.561	1.126
Baseline-Women	0.722	0.551	0.413	0.316	0.274	0.558	1.099
Equalizer w/o ACL-Neutral	**0.725**	0.557	0.417	0.318	**0.288**	0.570	1.104
Equalizer w/o ACL-Women	**0.725**	0.557	0.417	0.318	**0.288**	0.570	1.104

Table 7: MSCOCO image captioning metrics. Equalizer w/o ACL-Neutral performs in-line with Equalizer on these metrics but slightly better than the Equalizer w/o ACL model.

ACL-Women would overpredict women, the model actually performs around the same as Equalizer w/o ACL-Neutral. We hypothesize that this may be due to the limited training examples that we are fine-tuning the model on. Alternatively, if we were able to train Show and Tell from scratch with the female-corrected data for ambiguous images, we believe that the model would default to predicting women over men.

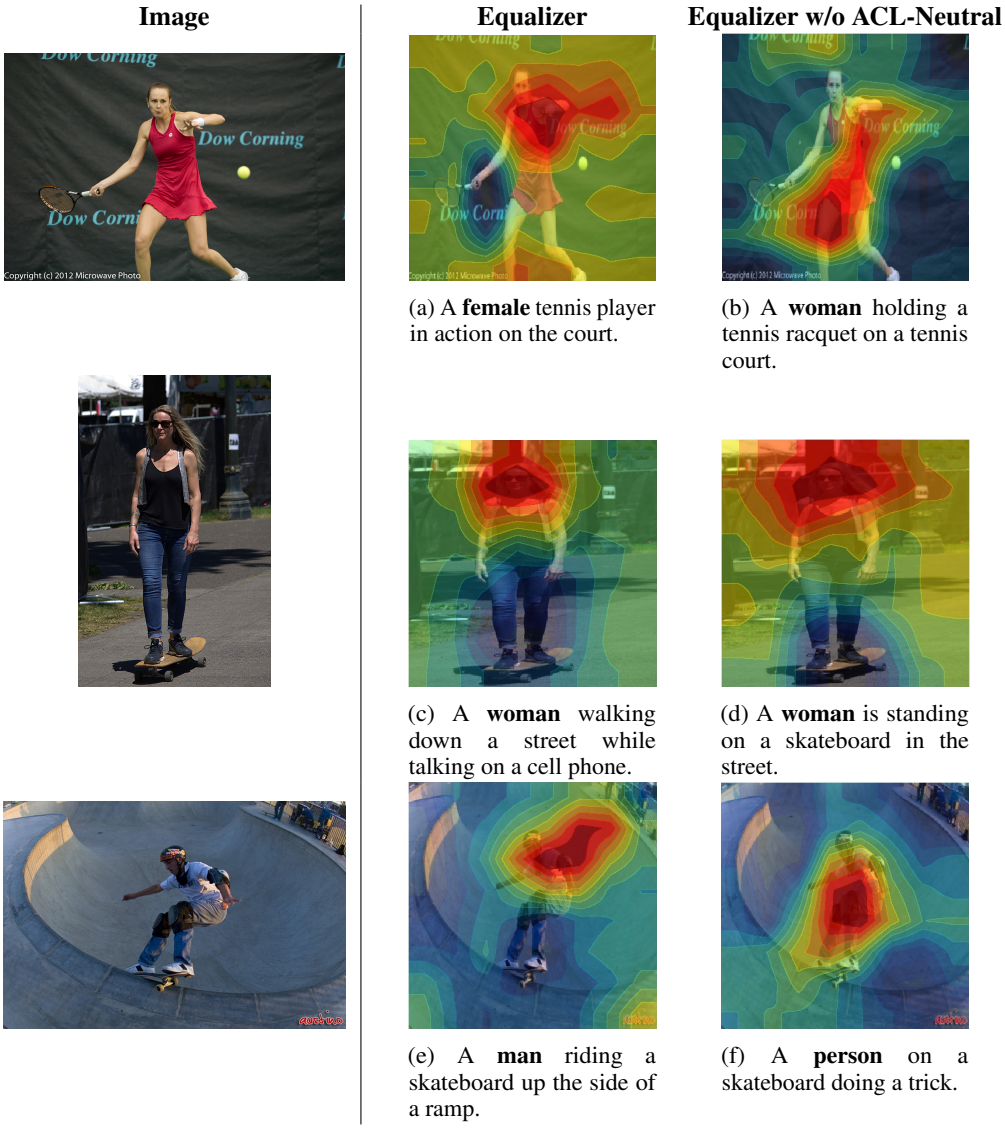
Finally, we compare the performance of Equalizer w/o ACL-Neutral to two of the original models proposed in "Women Also Snowboard." We find that Equalizer w/o ACL-Neutral outperforms not the only Baseline-FT model but also the Equalizer model in terms of Ratio Δ. In fact, with a Ratio Δ of only 0.033, Equalizer w/o ACL-Neutral's gender predictions closely mirror the ratio of the ground-truth. One potential concern is that the improved performance comes from the fact that the model has defaulted to predicting gender neutral terms. Upon further analysis, for the 94 images, Equalizer w/o ACL-Neutral predicts 35 males and 31 females whereas Equalizer predicts 47 males and 27 females. While Equalizer does detect gender at a higher rate, the difference in detection between the two is only 8 images, and, even if Equalizer w/o ACL-Neutral had predicted all 8 of these images to be male, their new Ratio Δ of 0.197 would still be significantly lower than that of Equalizer.

While Equalizer w/o ACL-Neutral performs better in terms of the Ratio Δ, its error rate is higher than than of the Equalizer model. However, upon examination of the errors, we find that the Equalizer model incorrectly classifies 11 images of females and 2 images of males. On the other hand, Equalizer w/o ACL-Neutral incorrectly classifies 11 images of females and 9 of males. Thus, even though the error rate is higher for the Equalizer w/o ACL-Neutral, it provides parity in performance across genders whereas Equalizer still performs disproportionately better on male images.

When comparing the Grad-CAM heat maps for the Equalizer model versus the Equalizer w/o ACL-Neutral, we observe that one shortcoming of the proposed model is that it is less likely to focus on the person pictured to make a prediction. As pictured in Figure 9, we can see, however, that when the model does not focus on the person in the image, it is more likely to default to gender neutral terms rather than try to assume gender. Nonetheless, while we would prefer the model to focus on the person and be able to accurately predict gender, being more conservative and predicting a gender neutral term (e.g. "person") is a better alternative than defaulting to male or inaccurately predicting the gender of the individual.

5 Conclusion and Future Work

We started by presenting similar image pairs as a new means of uncovering biases in image captioning techniques. While we varied the pairs used in the study along the axes of gender, this methodology can be extended to any attribute



(a) A **female** tennis player in action on the court.

(b) A **woman** holding a tennis racquet on a tennis court.

(c) A **woman** walking down a street while talking on a cell phone.

(d) A **woman** is standing on a skateboard in the street.

(e) A **man** riding a skateboard up the side of a ramp.

(f) A **person** on a skateboard doing a trick.

Figure 9: Qualitative analysis of the Grad-CAM heatmaps for Equalizer and Equalizer w/o ACL-Neutral. First row and second row shows an instance where both models are right for the right reasons (e.g. focused on the woman). Third row shows both are focused on the person but Equalizer w/o ACL-Neutral makes the more conservative guess of person. This may be preferred given that it is debatable whether enough gender information is present to make an inference.

of interest, such as race or age. In addition to the similar image pairs, we also introduced a new metric, sentence similarity, which can be used to measure the difference in predicted captions for similar image pairs. For an unbiased model, the sentences should be exactly similar since the only variable changing in the input image is the attribute of interest (e.g. gender).

Using our similar images pairs created on the MSCOCO dataset, we then evaluated Hendricks and Burns et al.'s [3] Equalizer model. From our analysis of the Equalizer model, we found that while it performed better on the fairness metrics outlined in the paper, there was still room for improvement. Furthermore, we found the need for segmented people masks to be too costly, especially when considering the scale of datasets used for the deep learning methods for which Equalizer is designed.

We propose a new method for addressing bias by correcting for human annotator bias found in the ground-truth data. This approach not only produced better results compared to Equalizer but also was more time efficient. However, beyond just using this method for fine-tuning models, we also would urge researchers to apply this method and correct for known human biases present in the ground-truth data as part of the dataset creation process.

Finally, we recognize that our work was limited by the scale of our similar images dataset and computational resources available for training. We believe that extending work in the following ways could not only bolster the analysis we have done but also provide meaningful ways for mitigating biases in image captioning:

- **Larger Similar Image Pairs:** Given the scope of this project, we only used 51 pairs of similar images. Collecting a large set of similar image pairs would potentially provide deeper insight into the biases that may exist. Despite the limited number of pairs, we believe that our results indicate that the similar images approach for model criticism is a valuable and replicable tool for evaluating biases.
- **Different Models:** While we chose to focus on the Equalizer model, there are many others, such as Tang et al.'s [4] GAIC model we described in our related works sections. The similar pairs approach can be applied to evaluate the effectiveness of bias mitigation models but also to measure the biases that may be present in general image captioning models as well. Furthermore, we also believe our method for correcting gendered assumptions in ground-truth data can also be applied to other models.
- **Other Protected Attributes:** As mentioned already, our similar pairs approach for model criticism can be extended to analyze other protected attributes, such as race or age. Given that most of the work on biases in image captioning have been focused on gender, extending this method can provide valuable insight on other biases that may exist.

6 Code and References

- **Our code and dataset**: `https://github.com/dorazhao99/cos429-fina..`
- **Model Checkpoints**: `https://drive.google.com/file/d/1RhGDa56g6hUQJVN4m7O6zy2RI9nbCV45/view?usp=sharing`
- **Our fork of the Equalizer code**: `https://github.com/dorazhao99/women-snowboard`
- **Equalizer Model**: `https://github.com/kayburns/women-snowboard`
- **MSCOCO Evaluation Code**: `https://github.com/dorazhao99/pycocoevalcap`
- **Universal Sentence Encoder**: `https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder`
- **Image Similarity**: `https://www.oreilly.com/library/view/practical-deep-learning/9781492034858/ch04.html`
- **t-SNE Embedding**: `https://www.learnopencv.com/t-sne-for-feature-visualization/`

7 Acknowledgements

We would like to acknowledge our TA, Angelina Wang, for providing feedback and giving advice throughout the project. In addition, we would like to acknowledge Hendricks and Burns et al. for authoring and providing code for "Women Also Snowboard: Overcoming Bias in Captioning Models", Chen et al. for developing the MSCOCO Evaluation code (and Github user salaniz for converting the code from Python 2 to Python 3), Lin et al. for creating the MSCOCO dataset, and Piotr Skalski for creating makesense.ai and allowing us to create our image segmentations easily, and Cer et al. for creating the Universal Sentence Encoder. Finally, we would like to thank Professors Russakovsky and Deng for teaching us about computer vision this semester.

8 Note on Independent Work

I (Dora) am conducting my senior thesis, advised by Professor Russakovsky, on racial biases in image captioning. In that project, I am collecting demographic annotations on the racial and gender composition of the MSCOCO dataset. I am then using these annotations to evaluate racial biases present in the ground-truth captions and iamges of MSCOCO.

There are several components of this project that are separate from my senior thesis. First, we create a novel dataset of similar image pairs using image features extracted from a ResNet-50. Second, this work is focused mainly on Equalizer, although we do claim that our results could be applied to other systems. On the other hand, my independent work focuses more on doing analysis on MSCOCO rather than zeroing in on one particular model.

9 Honor Code

I pledge my honor that this paper represents my own work in accordance with University regulations.

/s/ Brandy Chen and Dora Zhao

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[2] A. Stangl, M. R. Morris, and D. Gurari, ""' person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[3] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *European Conference on Computer Vision*, pp. 793–811, Springer, 2018.

[4] R. Tang, M. Du, Y. Li, Z. Liu, and X. Hu, "Mitigating gender bias in captioning systems," *arXiv preprint arXiv:2006.08315*, 2020.

[5] P. Stock and M. Cisse, "Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 498–512, 2018.

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (S. A. Friedler and C. Wilson, eds.), vol. 81 of *Proceedings of Machine Learning Research*, (New York, NY, USA), pp. 77–91, PMLR, 23–24 Feb 2018.

[7] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," *arXiv preprint arXiv:1902.11097*, 2019.

[8] C. Schwemmer, C. Knight, E. D. Bello-Pardo, S. Oklobdzija, M. Schoonvelde, and J. W. Lockhart, "Diagnosing gender bias in image recognition systems," *Socius*, vol. 6, p. 2378023120967171, 2020.

[9] A. Wang, A. Narayanan, and O. Russakovsky, "Revise: A tool for measuring and mitigating bias in visual datasets," 2020.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," 2015.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[12] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," 2018.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[14] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, pp. 376–380, 2014.

[15] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[16] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

[17] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

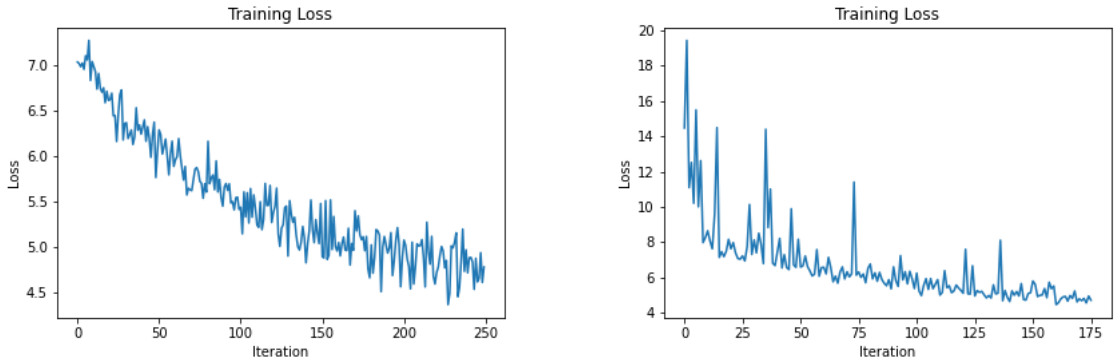
10 Appendix

10.1 Gender Indicative Words

Pre-compiled list of gender indicative words we used for females and males respectively:

- **Female**: Woman, women, female, girl, girls, lady
- **Male**: Man, men, male, boy, boys, guy, dude

10.2 Loss Curves



Baseline-Aug Equalizer w/o ACL-Neutral

Figure 10: Training loss curves over 10 epochs for Baseline-Aug (L) and 5 epochs for Equalizer w/o ACL-Neutral (R)