Understanding and Evaluating Racial Biases in Image Captioning

Dora Zhao Angelina Wang Olga Russakovsky
Princeton University
{dorothyz,angelina.wang,olgarus}@cs.princeton.edu

Abstract

Image captioning is an important task for benchmarking visual reasoning and for enabling accessibility for people with vision impairments. However, as in many machine learning settings, social biases can influence image captioning in undesirable ways. In this work, we study bias propagation pathways within image captioning, focusing specifically on the COCO dataset. Prior work has analyzed gender bias in captions using automatically-derived gender labels; here we examine racial and intersectional biases using manual annotations. Our first contribution is in annotating the perceived gender and skin color of 28,315 of the depicted people after obtaining IRB approval. Using these annotations, we compare racial biases present in both manual and automatically-generated image captions. We demonstrate differences in caption performance, sentiment, and word choice between images of lighter versus darker-skinned people. Further, we find the magnitude of these differences to be greater in modern captioning systems compared to older ones, thus leading to concerns that without proper consideration and mitigation these differences will only become increasingly prevalent. Code and data is available at https://princetonvisualai.github.io/imagecaptioning-bias/.

1. Introduction

Computer vision applications have become ingrained in numerous aspects of everyday life, and problematically, so have the societal biases they contain. For example, gender and racial biases are prevalent in image tagging [62, 7] and image search [39, 52]; visual recognition models have disparate error rates across demographics and geographic regions [16, 24]. The perpetuation and amplification of social biases precipitate the need for a deeper exploration of these systems and of the bias propagation pathways.

We focus on the task of image captioning: the process of generating a textual description of an image [69, 50, 76, 48, 3, 33]. This task serves as an important testbed for visual reasoning and can improve accessibility of digital images for people who are blind or low vision.

In this work, we assess the pathways for bias propagation: from the images, to the manual captions, and finally to the automatically generated captions. We focus our attention on studying the Common Objects in Context (COCO) [47, 19] dataset; it is a widely used image captioning benchmark [32], thus making any biases especially problematic [22]. We collect both skin color and perceived gender annotations on 28,315 of the people in the COCO 2014 validation dataset after obtaining IRB approval. This data allows us (and future researchers) to analyze disparities in image captioning (and other visual recognition tasks) across different demographics. Concretely, we observe:

- The dataset is heavily skewed towards lighter-skinned (7.5x more common than darker-skinned) and male (2.0x more than female) individuals.[1] Further, darker-skinned females are especially underrepresented, appearing 23.1x less than lighter-skinned males.
- There are racial terms (including racial slurs) in the manual captions. The racial descriptors are not learned by the older captioning systems [59, 50], but *are* learned by the newer transformer-based models [67] – although the slurs do not yet appear to be learned.
- Image captioning systems perform slightly better (according to CIDEr [68] and BLEU [55], although not SPICE [2]) on images of lighter-skinned people. This is consistent with disparate accuracies on e.g., pedestrian detection [73] and facial recognition [16].
- There are visual differences in the depictions of lighter and darker-skinned individuals. For example, lighter-skinned people tend to be pictured more with indoor and furniture objects, whereas darker-skinned people tend to be more with outdoor and vehicle objects.
- Even after controlling for visual appearance, the captions still differ in word choices used to describe images with lighter versus darker-skinned individuals. This is particularly apparent in the manual captions and in modern transformer-based systems.

Our work lays the foundation for studying bias propagation in image captioning on the popular COCO dataset.

[1] The gender disparity was previously observed in [78] although with automatically-inferred rather than manually-annotated labels.

Data and code is freely available for research purposes at https://princetonvisualai.github.io/imagecaptioning-bias/.

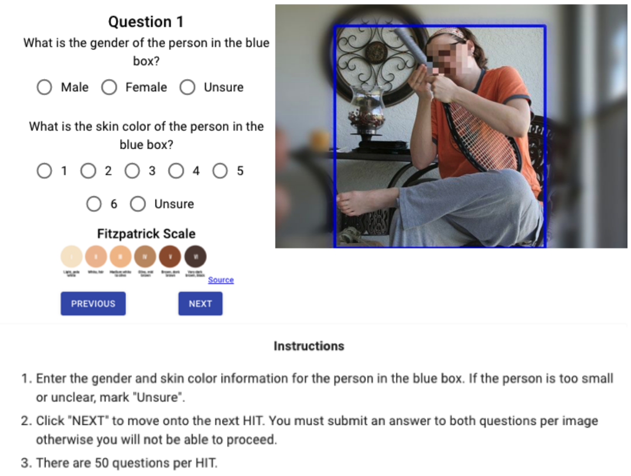
2. Related Work

Presence of dataset bias. Our work follows a long line of literature identifying, analyzing, and mitigating bias in machine learning systems. One key facet of this discussion is the bias in datasets used to train models. Under the framework of representational harms [14, 6], there is commonly a lack of representation [16, 74] and stereotyped portrayal [17, 62, 54, 65] of certain marginalized demographic groups. Along with many ethical concerns [56, 66], these dataset biases are problematic because they can propagate into models [15, 17]. In this work we analyze the biases present in a commonly-used image captioning benchmark, COCO [47, 19], using our new crowdsourced annotations.

Mitigating dataset bias. The root causes of dataset bias are complex: they stem from bias in image search engines [52], data collection practices [74, 24], and real-world disparities. Proposed solutions to dataset bias include new data collection approaches [38, 35], manual data cleanup [74, 75], synthetic data generation [58, 20, 60] – or, in extreme cases, even withdrawing the dataset after insurmountable biases have been identified [13]. Researchers have advocated for increased transparency of datasets [27, 35], including developing tools to steer researcher intervention [8, 70]. Our work does not aim to *mitigate* dataset bias but instead to articulate its impact on downstream image captioning models.

Algorithmic bias mitigation. In tandem with efforts to reform data collection, a variety of algorithmic bias mitigation techniques have been proposed; see e.g., Hutchinson and Mitchell. [34] for an overview. This work goes along with others that unveil biases present in existing algorithms [53, 49, 4]. One important theme is bias amplification [78, 71, 72], or social biases in the data getting amplified in the trained models. In this vein, we study how bias in manual image captions propagates into automated captioning systems.

Image captioning models. Image captioning models are increasingly being developed as a more complex way of labeling images [69, 50, 76, 48, 3, 33]. Recent work has discovered biases in these systems, but often with respect to gender [31, 12, 64]; the study of racial biases in captioning has been limited to analyzing bias in the *manual* captions [54, 65]. Racial bias has been identified in other automated systems [9] (e.g., speech recognition [44], facial recognition [28], pedestrian detection [73]); here we expand this work to studying racial biases in image captioning. This spurs the important question of whether race should be included in generated image captions at all. Prior works [63, 51] find that, in certain contexts, people who are blind or low vision want racial descriptors to be included. Further, this motivates the need to understand how people prefer their identities labeled by an automated captioning system, a question studied extensively by Bennett et al. [10].



Question 1

What is the gender of the person in the blue box?

○ Male ○ Female ○ Unsure

What is the skin color of the person in the blue box?

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

○ 6 ○ Unsure

Fitzpatrick Scale

Instructions

1. Enter the gender and skin color information for the person in the blue box. If the person is too small or unclear, mark "Unsure".
2. Click "NEXT" to move onto the next HIT. You must submit an answer to both questions per image otherwise you will not be able to proceed.
3. There are 50 questions per HIT.

Figure 1: The interface shown to AMT workers, who are asked to provide the inferred gender and skin color of the un-blurred person within the blue box (pixelation not seen by annotators, only to preserve privacy in figure).

3. Crowdsourcing Demographic Annotations

3.1. Annotation process

Dataset. To study bias in image captioning systems, we collect annotations on COCO [47, 19], a large-scale dataset containing images, labels, segmentations, and 5 human-annotated captions per image. COCO is a widely used image captioning benchmark. We focus on the 40,504 images of the COCO 2014 validation set, and look for person instances with sufficiently large bounding boxes (at least 5,500 pixels in area) such that there is a reasonable expectation of being able to infer gender and skin color. This results in 15,762 images and 28,315 person instances.

Annotation setup. Using Amazon's Mechanical Turk (AMT), we crowdsource race and gender labels. In our interface (Fig. 1), we present workers with a person instance in a COCO image and ask them to provide the skin color using the Fitzpatrick Skin Type scale [25], ranging from 1 (lightest) to 6 (darkest), and the binary gender expression. We also give workers the option of marking "unsure" for either. Each instance is annotated three times. We compensate the workers at a rate of $10 / hr.

Inferring race and gender. Race and gender annotations are fundamentally imperfect [29, 41, 61]. First, the annotated labels may differ from the person's identity. Second, the labels are discretized (which enables disaggregated analysis at the cost of collapsing identities). Further, the labels are for social constructs and thus subjective and influenced by the annotators' perceptions. We follow prior work [16, 73] in formulating our annotation process; we use phenotypic skin color as a proxy for race because of its visual saliency over other conceptualizations of race. However, as noted by Hanna et al. [29], we are actualizing a particular static conceptualization of observed race here. By operationalizing race this way, we miss differences that may appear in other operationalizations, such as racial identity.

Quality control. To ensure annotation quality to the extent possible, we limit the task to workers who have completed over 1,000 tasks with a 98% acceptance rate. We also construct 57 gold standard images where the gender and light-or-dark labels were agreed-upon by five independent annotators, including one of the authors. We inject 5 of these images randomly in a task with 50 images, and only allow workers who have correctly labeled these images to submit.

3.2. Gender annotations

We start by analyzing the collected gender annotations, looking at distributions at both the instance and image level.

Instance-level annotations. We analyze the gender annotations of the 28,315 `person` instances. To determine the label for a `person`, we use the majority over the three annotations. If majority is not achieved, or there are contradictory gender labels, the instance is labelled as `no consensus`. We observe that contradictory gender labels are most common when the person is a child, has obscured facial features, or possesses features that contradict social gender stereotypes (e.g. woman with short hair).

Analyzing the distribution, we see that males make up 47.4% of the instances compared to females who only comprise 23.7% (see Fig. 2). Most of the remaining instances were annotated `unsure` (26.6%), and a consensus was unable to be reached for only 2.2% of instances.

Image-level annotations. To analyze the dataset at the granularity of images, which is what the captions refer to, we map individual instance annotations to the image (as there are often multiple people per image). We use the annotations given to the largest bounding box, under the assumption that captions will mainly refer to the largest person in the image [11]. The only exception is if the second largest bounding box contains an individual of the opposite gender, and is more than half the size of the largest bounding box. In this case, we categorize the image as `both`.

The image-level distribution closely mirrors that of the instance-level (Fig. 2). Again, there are more than twice as many male images (47.4%) as female images (21.0%).

Comparing collected gender annotations with automatically derived ones [78]. Previously, works looking at gender bias in COCO have used gender labels derived from the manual captions: "[if] any of the captions mention the word man or woman we mark it, removing any images that mention both genders." [78] We compare our annotations with theirs. They label 5,413 images: our labels agree with theirs on 66.3% and disagree on 1.4%; the remaining 32.3% we determine cannot be reliably labeled with one gender, e.g., because the person is too small or there are multiple people of different genders in the image. We successfully label 10,780 images; they only label 3,591 of these correctly (details in Appendix A). This is consistent with the argument of Jacobs and Wallach [37]: gender is operationalized differently in caption-derived versus human-collected annotations.

3.3. Skin color annotations

For the skin color annotations, we follow a similar process as with our gender annotations. The only difference is that we add a method for dividing skin color into the broader categories of `lighter` and `darker`. Using these new categories, we similarly analyze the skin color distribution at both the instance and image level.

Instance-level skin color distribution. Using the same schema as in Sec. 3.2, we obtain instance-level annotations for skin color. The top two most frequently occurring Fitzpatrick Skin Types are 2 (31.5%) and 1 (15.4%). In contrast, Fitzpatrick Skin Types 5 and 6 comprise only 1.9% and 1.7% of the instances, respectively. This underrepresentation of darker-skinned individuals is an example of representational harm in and of itself.

We also include a broader skin color breakdown consisting of two categories: `lighter` and `darker`. Following previous work [16], we define the `lighter` category as all instances rated 1-3 on the Fitzpatrick scale and `darker` as containing 4-6. We also assign some of the instances that were previously uncategorized by skin color (because of conflicting labels assigned under the more granular 6-point scheme) to these broader categories. Using this skin color breakdown, 61.0% of the instances are `lighter` individuals, whereas only 8.1% are `darker` individuals. The amount of `no consensus` instances decreases from 15.4% to 13.9% when using this breakdown.

Image-level skin color distribution. At the image-level, we categorize skin color as `lighter` and `darker`, employing the same consensus method as for gender in Sec. 3.2. Of the images, 64.6% are part of the `lighter` category and 7.0% are part of `darker`, meaning there are 9.2x more lighter-skinned images than darker-skinned.

Intersectional analysis. We analyze the skin color and gender labels in tandem. Within `lighter` images, males are overrepresented at 52.8% compared to females at 25.7%. However, this difference is even starker when looking at `darker` images, where males comprise 65.1% of the images while females only make up 20.6%, reflecting the unique intersectional underrepresentation faced by darker-skinned females, as noted by Buolamwini and Gebru [16]. In fact, of the 15,762 images annotated, only 226 of them (1.4%) are of darker-skinned females.

Worker information. AMT workers were asked to optionally disclose their own race and gender identity. Of the workers asked, 97.9% provided their gender and 97.3% provided their race. As seen in Fig. 2, the annotators are pre-

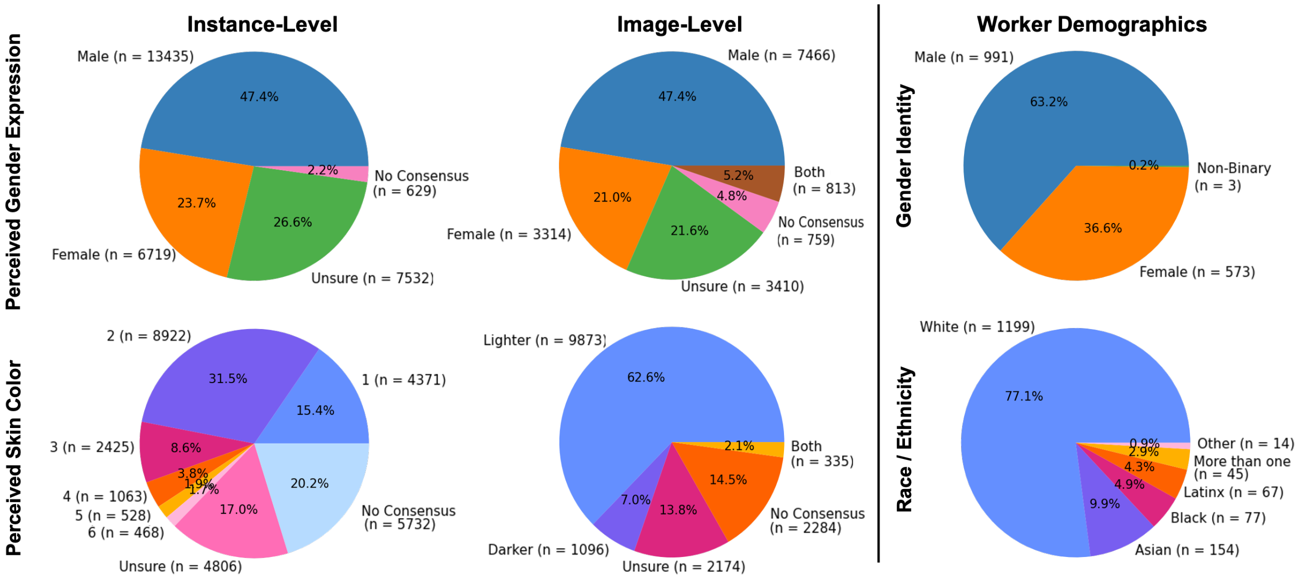


Figure 2: The results of our crowdsourced demographic annotations on the COCO 2014 validation dataset, as well as the self-disclosed demographics of the annotators. Left column: distribution of perceived skin color and gender expression of the 28,315 `people` instances. Middle column: distribution after collapsing individual annotations into image-level annotations (details in Sec. 3.2 and 3.3). Lighter-skinned people and people who are male make up the majority of their respective categories. Right column: self-reported demographics of AMT workers.

dominantly white (77.1%) and male (63.2%).

Prior work has found that annotators describe in-group versus out-group members differently [54]. Thus, there may be a concern that the skew in worker demographics could influence our collected labels. To understand whether a worker's demographics influences their selection of labels, we explore disagreements in annotations. We do so by comparing the mean difference in annotation when the pair of workers are of the same self-reported demographic group versus when they are of differing groups. If workers from different groups label images differently, we would expect pairs from distinct groups to have a greater disagreement than pairs from the same group. However, we find for skin tone there is not a substantial difference in the disagreement between pairs of the same racial group (0.870 ± 0.009) and different groups (0.857 ± 0.011). For gender, the mean difference for same gender pairs (0.109 ± 0.002) and different gender pairs (0.112 ± 0.003) is similar as well. This indicates that there is not a systematic difference between how workers of different self-reported demographic groups label images, suggesting our collected labels would be similar even if the workers came from a different demographic composition.

4. Experiments

We now discuss the findings from our experiments on understanding what kinds of biases propagate in image captioning systems. First, we examine racial terms (Sec. 4.1) and disparate performance (Sec. 4.2). We then analyze bias in terms of representation, i.e., differences between the `lighter` and `darker` images and corresponding captions. To do this we first consider the images in Sec. 4.3, before controlling for these visual differences and studying the captions in Sec. 4.4.

Models. We examine the captions generated by six image captioning models: (1) **FC** [59] is a simple sequence encoder that takes in image features encoded by a CNN; (2) **Att2in** [59] is similar but images are encoded using spatial features; (3) **DiscCap** [50] further adds a loss term to encourage discriminability; (4-6) **Transformer** [67], **AoANet** [33], and **Oscar** [45] are transformer-based models representing the current state-of-the-art. In our analysis we particularly focus on contrasting **Att2in** vs **DiscCap**, since they differ only in the added discriminability loss, and the older (1-3) vs the newer (4-6) models. We train the models on the COCO 2014 training set using proposed hyperparameters from the respective papers (e.g., the discriminability loss weight is $\lambda = 10$ for **DiscCap**). **Oscar** is further pre-trained on a public corpus of text-image pairs

Data. Our racial analysis is performed on 10,969 images of the COCO 2014 validation set which were definitively labeled as either `lighter` or `darker` (not `both` or `unsure`).

4.1. Captions contain racial descriptors

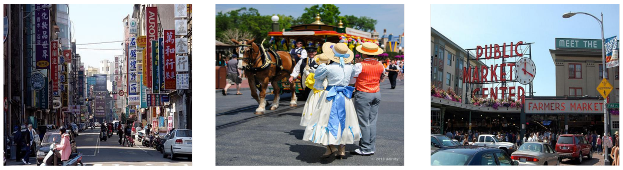
We begin by analyzing the presence of racial descriptors and offensive language in the manual as well as automatically generated captions.

Manual captions. Prior works [54, 65] show that people are more likely to use racial descriptors when describing non-white individuals. We observe this pattern in human-annotated captions by conducting a keyword search of the captions in the COCO 2014 training set using a precompiled list of racial descriptors (details in Appendix B). For ambiguous terms (e.g. "white", "black") that can be used in a non-racial context, we manually inspect the captions. Assuming the training distribution mirrors that of the validation, for the manual captions, annotators used racial descriptors to describe individuals who appear to be white 0.03% of the time versus 0.54% of the time for individuals who appear to be Black. Furthermore, in 26.9% of the instances when a racial descriptor for a white individual is used, the annotator is also mentioning an individual of a different race in the caption as well (e.g. "the white woman and Black woman"). We see this as a manifestation of the belief that "white" is the norm, and race is only salient when there is a deviation or explicit difference between multiple people.

In addition to looking for racial descriptors, we check for the presence of slurs and offensive language using a precompiled list of profane words [77]. There are 1,691 instances of profane language, occurring in 0.40% of the sentences in the COCO 2014 training set. We find alarming occurrences not only of racial slurs but also of homophobic and sexist language as well, similar to the NSFW discoveries by Prabhu and Birhane [13].

Automated captions. Racial descriptors are not found in the automated captions generated by **FC**, **Att2In**, **DiscCap**, **AoANet**, or **Oscar**. While this may be attributed to the fact that racial descriptors are uncommon in the training set, we disprove the idea that this is wholly the reason. To do so, we observe that other words which occur at similar rates (and are thus equally uncommon) are in fact still present in the model-generated captions. For example, the word "Japanese" occurs 69 times in the training set and 0 times in **AoANet**-generated captions while other descriptors, such as "uncooked" and "soaked", which appear 88 and 61 times in the training set, occur 2 and 6 times in the generated captions respectively.

While it is rare, we find that racial and cultural descriptors as well as offensive language do propagate into the captions generated by the newer transformer-based models. For **Transformer**, **AoANet**, and **Oscar**, we find instances of offensive language. In addition, there are racial descriptors in 2 of the captions generated by **Transformer** and 12 cultural descriptors. Furthermore, for 10 of the 14 images, the model uses these descriptors when the human captions do not contain any racial or cultural descriptors (Fig. 3). This leads to the worry that models may replicate offensive language or exploit spurious correlations to assign descriptors in a stereotypical and harmful way.



Human: A busy city street in an Asian country with lots of traffic.

Transformer: A city street with lots of asian businesses.

Human: People watch a horse and carriage ride by them.

Transformer: A group of indians standing around in inflatable blue.

Human: A crowded farmers market with a line of cars outside.

Transformer: A street scene with a focus on a mexican restaurant.

Figure 3: Examples of images for which the **Transformer** model [67] assigns racial or cultural descriptors to the caption. While in the first image the descriptor of "Asian" is present in the human-annotated caption, neither of the descriptors, "Indian" nor "Mexican," are applicable in the latter images.

4.2. Performance differs slightly between `lighter` and `darker` images

We next evaluate whether image captioning models produce captions of different qualities on images with lighter-skinned people than darker-skinned people. To do so, we first assess the differences in BLEU [55], CIDEr [68] and SPICE [2] scores between captions on `lighter` and `darker` images. Both BLEU and CIDEr rely on *n*-gram matching with BLEU measuring precision and CIDEr the similarity between the generated caption and the "consensus" of manual captions. SPICE, however, focuses more on semantics, capturing how accurately a generated caption describes the image's scene graph (e.g. objects, attributes).

From these results (Tbl. 1), we make two key observations. First, according to both BLEU and CIDEr, the models **Att2in**, **Transformer**, **AoANet**, and **Oscar** perform somewhat better on `lighter` images than `darker` images: e.g., they achieve 2.7 ± 0.7, 3.2 ± 1.2, 1.9 ± 1.6, and 3.0 ± 1.1 higher CIDEr scores respectively on `lighter` than `darker` images. We observe that these differences in BLEU and CIDEr are not significant for the **FC** and **DiscCap** — likely because their overall CIDEr scores are worse, at only 87.2 and 71.1 respectively, whereas the other four models attain CIDEr scores above 90.0 (see Appendix C). This suggests that the *way* models are choosing to describe the images may be better-suited for the majority group. In fact, we see there is a slight positive correlation between the performance of the model (as measured by CIDEr) and the differences in performance between the two groups with an R^2 of 0.343 (Fig. 4). Second, there are no noticeable differences with SPICE, indicating that the captions identify key visual concepts equally accurately across both groups. Nonetheless, it is important to note that negative results do not indicate something is bias-free, but merely that our particular experiment did not uncover strong biases.

Table 1: The differences in captioning performance (score on `lighter` - score on `darker`) as measured by BLEU [55], CIDEr [68], and SPICE [2] multiplied by 100 on the COCO 2014 validation dataset. Error bars represent 95% confidence intervals across random seeds used to train 5 models per architecture.

	BLEU	CIDEr	SPICE
FC [59]	0.5 ± 0.5	-0.8 ± 1.8	0.2 ± 0.3
Att2in [59]	2.4 ± 0.4	2.7 ± 0.7	0.0 ± 0.1
DiscCap [50]	0.3 ± 0.5	0.6 ± 0.7	0.0 ± 0.2
Transformer [67]	2.5 ± 0.9	3.2 ± 1.2	-0.1 ± 0.3
AoANet [33]	1.8 ± 0.8	1.9 ± 1.6	0.0 ± 0.2
Oscar [45]	2.1 ± 0.7	3.0 ± 1.1	0.1 ± 0.3

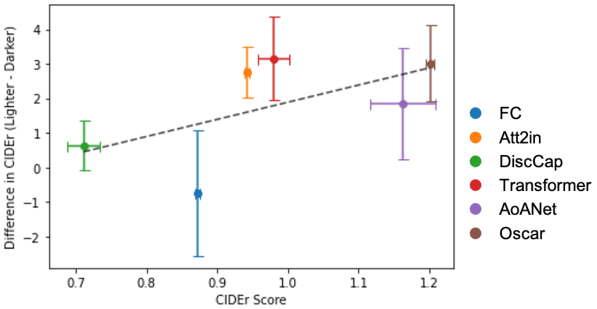


Figure 4: Regressing model performance, as measured by CIDEr [68], against difference in performance (CIDEr on `lighter` - CIDEr on `darker`) suggests that as performance increases, the difference may correspondingly increase as well ($R^2 = 0.343$). The horizontal and vertical error bars in the graph represent the 95% confidence intervals for the performance and differences, respectively

4.3. Visual appearance differs between `lighter` and `darker` images

The analyses so far only consider issues in the captions themselves, irrespective of the image. We now explore how the visual depictions of people of different groups differ. We analyze simple image layout statistics, apply the REVISE [70] tool for discovering bias in datasets, and consider differences in visual appearance of the image content.

We split our skin-tone-labeled image dataset of 10,969 images into 9,609 images for training and 1,360 for testing.[2] We use area under the ROC curve (AUC) as our metric on a balanced (through re-weighting) test set, so random guessing would have an AUC of 50%. We bootstrap over 1,000 resamples and report a 95% confidence interval.

Image layout statistics. We consider the following simple image layout statistics as our features: number of people in the image, largest person bounding box size, distance of the largest bounding box from the center of the image, and gender (`male`, `female`, `unsure`, or `no consensus`) one-hot coded. We train logistic regression models using LBFGS through the `sklearn` package [57] to predict whether the input corresponds to the `lighter` or `darker` label. An ability to classify serves as a signal for how distinguishable the input features of the two groups are. We use a balanced class weight and run five-fold cross-validation to tune the L2 regularization hyperparameter (1e−4 to 1e4).

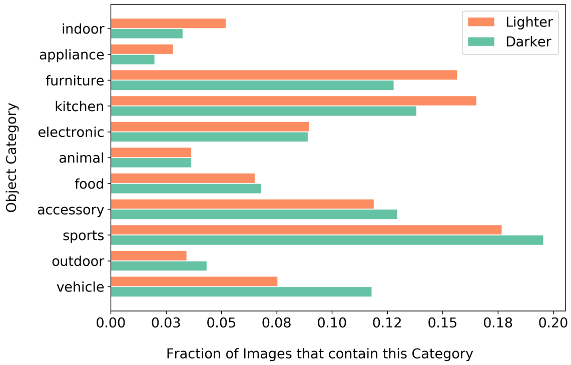


Figure 5: Images with people of lighter and darker skin tones co-occur with object categories at different frequencies. Whereas the former tend to be pictured with object categories that are indoor, the latter tend to be pictured with object categories that are more likely to be outdoors.

Our two best performing models are trained on the distance from center and the distance plus the gender. Distance alone achieves an AUC of 56.6 ± 5.2; adding gender increases the AUC to 57.8 ± 4.9. Distance is predictive because darker-skinned individuals tend to be further from the image center than lighter-skinned individuals; this is troubling since the "important" parts of an image tend to be more centered [11]. Gender is a useful feature since from Sec. 3.3 we know that the gender distribution differs between the two groups.

REVISE [70] bias discovery. We next apply the REvealing VIsual biaSEs (REVISE) tool.[3] Using REVISE we discovered that darker-skinned people appear more frequently with outdoor objects, and lighter-skinned people appear more frequently with indoor objects (Fig. 5). Specifically, objects like sink, potted plant, and toothbrush all appear with lighter-skinned people over 13x as much as with darker-skinned people, despite lighter-skinned people only appearing in 7x as many images as darker-skinned people. Although at the moment the differences in object co-occurrences do not appear to have noticeable downstream effects (Sec. 4.2), these differences may lead to discrepancies in performance as certain objects become more easily identifiable for different skin tone groups.

Visual appearance. Finally, we use image classification models for a detailed examination of how the content of the images differs between different skin tones. To ensure that the skin color of the pictured individual does not affect the model's prediction, we use COCO's object-level segmentations to mask all the `people` objects. We fill in these

[2] These images belong to the COCO 2017 training and validation set respectively; recall that all belong to the COCO 2014 validation set.

[3] We additionally include the 813 images labeled `both` in both groups.

Table 2: Three bias analyses on manual and automated captions of images for which visual content has been controlled. For the first column the VADER sentiment score [36] is multiplied by 100. For the last two columns, the number is AUC×100 for classification ability, where higher numbers indicate a greater ability to distinguish between the two groups. Error bars represent 95% confidence intervals across random seeds used to train 5 models per architecture.

	Sentiment	Embedding	Vocab.
	(Δ)	(AUC)	(AUC)
Human	1.5	68.9	61.8
FC [59]	1.0 ± 0.4	55.8 ± 5.7	65.9 ± 4.2
Att2in [59]	0.3 ± 0.2	55.3 ± 2.5	62.8 ± 1.5
DiscCap [50]	0.9 ± 0.7	52.2 ± 2.8	63.0 ± 3.2
Transformer [67]	0.6 ± 0.8	54.2 ± 3.1	66.0 ± 1.4
AoANet [33]	1.0 ± 0.4	56.3 ± 3.0	68.0 ± 1.8
Oscar [45]	0.6 ± 0.6	54.0 ± 2.6	64.4 ± 2.8

masks with the average color pixel in the image. Using the masked images, we fine-tune a pre-trained ResNet-101 [30] over five epochs using the Adam optimizer [42] and a batch size of 64. We oversample the `darker` images to account for the imbalanced class sizes. During training, the learning rate is initialized to be 0.01 and decays by a factor of 0.1 after three epochs. The model achieves an AUC of 55.4 ± 4.9, indicating that there is a slight learnable difference between the scenes of `lighter` and `darker` images.

4.4. Captions describe people differently based on skin tone

Finally, we consider how both manual and automatic captions differ when describing `lighter` versus `darker` images. To do so, we first control for the visual differences, in order to disentangle the issues coming from the image content versus from the words used in the caption. We do so by finding images that are as similar as possible in content, and differ only by the skin color of the people pictured, i.e., constructing counterfactuals within the realm of our existing dataset. Concretely, for each `darker` image, we find the corresponding `lighter` image that minimizes the Euclidean distance between the extracted ResNet-34 features [30] of the masked images using the Gale-Shapley algorithm [26] for stable matching (Fig. 6). After examining the results, we select the top 40% most similar image pairs.

The resulting dataset has 876 images. When needed, we use 700 for training (80%) and 176 for testing (20%); otherwise we compute statistics over the whole dataset. As expected, a visual classifier trained on these images (with the people masked) achieves an AUC of only 44.7 ± 9.3, failing to differentiate between the two groups.

In the following analyses, we use the same six models and training setup as in previous experiments. However, we use the dataset, introduced above, which consists of 876 unmasked images for evaluation. This data thus allows us to examine whether human-annotated and model-generated captions diverge even when visual differences (except skin color) are controlled.

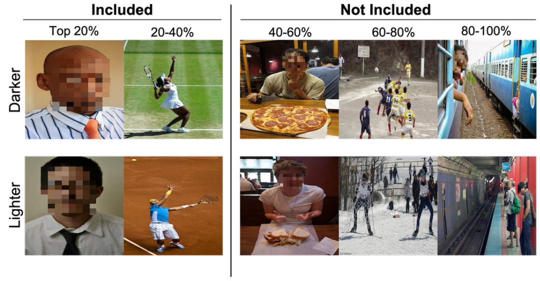


Figure 6: Examples of paired images along 20 percentile increments of similarity. The leftmost images represent an example pair from the most similar top 20% of pairs, and the rightmost represent the bottom 20%. We pick 40% as our threshold for controlled images to include.

4.4.1 Sentiment Analysis

For our first line of inquiry, we use the Valence Aware Dictionary and Sentiment Reasoner (VADER) [36] to perform sentiment analysis on the human-annotated captions. Limitations include that sentiment analysis tools have been shown to encode societal biases themselves [43, 23], and may not generalize well to out-of-distribution machine-generated text. VADER returns a compound polarity score from -1 (strongly negative) to 1 (strongly positive). Scores less than -0.05 are considered negative; scores greater than 0.05 positive. We find that human-annotated captions describing `lighter` images have a mean compound score of 0.073 ± 0.01 whereas those describing `darker` images have a mean compound score of 0.059 ± 0.01. The difference in compound scores is statistically significant ($p = 0.005$), with captions describing `lighter` images being more positive.

We find that automated captioning systems do not appear to amplify the difference in sentiment scores between the two groups (Tbl. 2). The lack of difference is largely due to the fact that automated captions tend to be more neutral than the human-annotated ones, thus removing most of the sentiment. In fact, the compound scores were all less than 0.03, excluding scores for captions generated by **Transformer** (0.046 for `lighter` and 0.042 for `darker`).

4.4.2 Sentence embedding differences

For our next analysis, we use sentence embeddings from the Universal Sentence Encoder [18] to compare how the semantic content of captions differs between `lighter` and `darker` images. To note, racial descriptors in the captions are not removed for this experiment. We train a multilayer perceptron classifier (MLP) on the embeddings and run five-fold cross validation to tune the learning rate (1e$-$5 to 1) and number of epochs (1 to 150). We find that the classifier can differentiate between the captions with an AUC of 68.9 ± 3.5, indicating a learnable difference in the resulting caption content despite the visual content (with skin tone masked) being indistinguishable.

We see in Tbl. 2 that the ability to differentiate based on embeddings drops in the generated captions, especially for the more advanced **Transformer** model to 54.2±3.1, which is almost random. Although humans appear to be assigning different content to similar images with people of different skin tones, automated captioning models do not appear to uphold this trend, at least with respect to the particular sentence embeddings we use.

4.4.3 Vocabulary differences

Finally, we consider word choice in the captions. We use a logistic regression model and a vocabulary of the 100 most commonly used words (filtering out articles, prepositions, and racial descriptors, e.g. "white") in the COCO 2014 training set. Our features are size 100 binary indicators of whether a particular word is present in a caption. The classifier achieves an AUC of 61.8 ± 3.8 on human captions. Beyond the differential use of racial descriptors we already observed in Sec. 4.1, this suggests annotators use different vocabularies to describe images even with similar visual content (other than skin tone).

The ability to distinguish between `lighter` and `darker` images further increases when automated captions are used. Particularly, in Tbl. 2 we see from **Att2in** to **DiscCap** and **FC** to **Transformer**, the AUCs slightly increases from 62.8±1.5 to 63.0±3.2 and 65.9±4.2 to 66.0±1.4, respectively. From **FC** to **AoANet**, there is a greater increase in AUC from 65.9 ± 4.2 to 68.0 ± 1.8. We do note that, for **Oscar**, the ability to differentiate based on vocabulary decreases compared to **FC** as the AUC drops from 65.9 ± 4.2 to 64.4 ± 2.8. This may be due to the fact that **Oscar** is pre-trained on a larger corpus of data; the greater dataset diversity may help diminish the differences between the vocabularies used. Overall, this leads us to believe that more advanced models are more likely to employ different word choices when describing different groups of people.

Interpreting these results relative to that of the previous section in which we found that the semantic content of generated captions did not differ much between different groups, we consider whether different words are being used despite caption content being similar. As an example, the sentences "Apples are good." and "Apples are great." may map to similar sentence embeddings, but the specific word choice employed is different. In this vein, we find, for instance, that on **AoANet**'s captions, the average coefficent of the word "road" is 0.226 higher than that of the word "street" (where higher coefficients are predictive of `darker`), even though upon manual inspection the images being described are similar (see Appendix D). While differences in the usage of words, such as "road" and "street," are relatively innocuous, these subtle differences in vocabulary may become more problematic when we consider how certain words like "articulate" have developed a different meaning when applied to Black people [21, 1]. Thus, in future work, it is important to consider not only the semantic differences captured in the sentence embeddings but also the specific words being employed.

5. Discussion and Conclusion

In this work, we seek to understand not only what racial biases are present in the COCO image captioning dataset, but also how these biases propagates into models trained on them. We annotate skin color and gender expression of people in the images, and consider various forms of bias such as those in the form of differentiability between different groups. We find instances of bias in the dataset and the automated image captioning models. However, we are careful to note that cases in which we did not find bias do not mean there are not any, merely that our particular experiments did not uncover them. By looking at the models that seem to be most indicative of where the image captioning space is progressing, we can see that the bias appears to be increasing. For researchers, this serves as a reminder to be cognizant that these biases already exist and a warning to be careful about the increasing bias that is likely to come with advancements in image captioning technology.

Based on these analyses, we propose directions for mitigating the biases found in captioning systems. First, from our findings in Sec. 3.2 and 4.1, we see that human annotators make assumptions about the demographics of people pictured or use different language when describing people of different skin tone groups. To mitigate this, dataset collectors can provide more explicit instructions for annotators (e.g. do not label gender or include racial descriptors to people). In addition, we also find that ground-truth captions contain profane language (Sec. 4.1). In line with existing mitigation efforts [74, 13], manual captions containing slurs or other offensive concepts should be removed from the dataset. Additionally, in Fig. 2 we see that only 7.0% of the dataset contained images of people with darker skin tones, i.e., 1096 images. We need to collect more diverse datasets such that we can measure disaggregated statistics and compare metrics such as the difference in SPICE scores with the knowledge that our measurements do not suffer from a high sampling bias. Finally, from our analysis of generated captions (Sec. 4.4), we note that **Oscar** exhibits less bias compared to the other transformer-based models. This suggests the greater dataset diversity from pre-training the model may help reduce the amount of bias that propagates into the automated captions.

Acknowledgements. This work is supported by the National Science Foundation under Grant No. 1763642 and the Friedland Independent Work Fund from Princeton University's School of Engineering and Applied Sciences. We thank Arvind Narayanan, Karthik Narasimhan, Sunnie S. Y. Kim, Vikram V. Ramaswamy, and Zeyu Wang for their helpful comments and suggestions, as well as the Amazon Mechanical Turk workers for the annotations.

References

[1] H. Samy Alim and Geneva Smitherman. Articulate while black: Barack Obama, language, and race in the U.S. *Oxford University Press*, 2012. 8

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 5, 6

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016. 2

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[6] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. In *Special Interest Group in Computing, Information, and Society (SIGCIS)*, 2017. 2

[7] Alistair Barr. Google mistakenly tags Black people as 'gorillas,' showing limits of algorithms. *The Wall Street Journal*, 2015. 1

[8] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019. 2

[9] Ruha Benjamin. *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity, 2019. 2

[10] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. "It's complicated": Negotiating accessibility and (mis)representation in image descriptions of race, gender, and disability. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021. 2

[11] Alexander C. Berg, Tamara L. Berg, Hal Daumé, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, and Kota Yamaguchi. Understanding and predicting importance in images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3, 6

[12] Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. Master's thesis, University of Illinois at Urbana-Champaign, 2019. 2

[13] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2, 5, 8

[14] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Association for Computational Linguistics (ACL)*, 2020. 2

[15] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Neural Information Processing Systems (NeurIPS)*, 2016. 2

[16] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2018. 1, 2, 3

[17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 2

[18] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder for english. In *Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, 2018. 7

[19] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 1, 2

[20] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning (ICML)*, 2020. 2

[21] Lynette Clemetson. The racial politics of speaking well. *The New York Times*, 2007. 8

[22] Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision making systems. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. 1

[23] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Workshop on Abusive Language Online (ALW)*, 2019. 7

[24] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *CVPR Workshop on Fairness, Accountability Transparency, and Ethics in Computer Vision*, 2019. 1, 2

[25] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. In *Archives of Dermatology*, 1988. 2

[26] David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962. 7

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. In *Fairness, Accountability, Transparency in Machine Learning (FAT/ML)*, 2018. 2

[28] Patrick J Grother, Patrick J Grother, and Mei Ngan. *Face Recognition Vendor Test (FRVT)*. US Department of Commerce, National Institute of Standards and Technology, 2014. 2

[29] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020. 2, 3

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016. 7

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

[32] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6), 2019. 1

[33] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 4, 6, 7

[34] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019. 2

[35] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. 2

[36] Clayton Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *International AAAI Conference on Web and Social Media (ICWSM)*, 2014. 7

[37] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. 3

[38] Eun So Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020. 2

[39] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2015. 1

[40] Os Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. *ACM Conference on Computer Supported Cooperative Work (CSCW)*, 2018.

[41] Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. 2

[42] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*, 2015. 7

[43] Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. 7

[44] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. 2

[45] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision (ECCV)*, 2020. 4, 6, 7

[46] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 1, 2

[48] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[49] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016. 2

[50] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 4, 6, 7

[51] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2017. 2

[52] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. 1, 2

[53] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. 2

[54] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. How do we talk about other people? group (un)fairness in natural language image descriptions. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2019. 2, 4, 5

[55] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*, 2002. 1, 5, 6

[56] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *NeurIPS Machine Learning Retrospectives Workshop*, 2020. 2

[57] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 12:2825–2830, 2011. 6

[58] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[59] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for

image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 4, 6, 7

[60] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. Fairness GAN. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019. 2

[61] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. In *ACM Conference on Computer Supported Cooperative Work (CSCW)*, 2020. 2

[62] Carsten Schwemmer, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6, 2020. 1, 2

[63] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. ”Person, shoes, tree. is the person naked?” What people with vision impairments want in image descriptions. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2020. 2

[64] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *The Web Conference (WWW)*, 2021. 2

[65] Emiel van Miltenburg. Stereotyping and bias in the Flickr30K dataset. In *Multimodal Corpora: Computer Vision and Language Processing (MMC)*, 2016. 2, 5

[66] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. Talking about other people: An endless range of possibilities. In *International Conference on Natural Language Generation (INLG)*, 2018. 2

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NeurIPS)*, 2017. 1, 4, 5, 6, 7

[68] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 5, 6

[69] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[70] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 6

[71] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *International Conference on Machine Learning (ICML)*, 2021. 2

[72] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[73] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. In *CVPR Workshop on Fairness, Accountability Transparency, and Ethics in Computer Vision*, 2019. 1, 2

[74] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2020. 2, 8

[75] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. *CoRR*, abs/2103.06191, 2021. 2

[76] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[77] zacanger. profane-words. `https://github.com/zacanger/profane-words/blob/master/words.json`, 2021. 5

[78] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 1, 2, 3