Measuring Representational Harms in Image Captioning

Angelina Wang
Princeton University
Princeton, NJ, USA
angelina.wang@princeton.edu

Solon Barocas
Microsoft
New York, NY, USA
solon.barocas@microsoft.com

Kristen Laird
Microsoft
New York, NY, USA
kristen.laird@microsoft.com

Hanna Wallach
Microsoft
New York, NY, USA
wallach@microsoft.com

ABSTRACT

Previous work has largely considered the fairness of image captioning systems through the underspecified lens of "bias." In contrast, we present a set of techniques for measuring five types of representational harms, as well as the resulting measurements obtained for two of the most popular image captioning datasets using a state-of-the-art image captioning system. Our goal was not to audit this image captioning system, but rather to develop normatively grounded measurement techniques, in turn providing an opportunity to reflect on the many challenges involved. We propose multiple measurement techniques for each type of harm. We argue that by doing so, we are better able to capture the multi-faceted nature of each type of harm, in turn improving the (collective) validity of the resulting measurements. Throughout, we discuss the assumptions underlying our measurement approach and point out when they do not hold.

CCS CONCEPTS

• **Social and professional topics → User characteristics**; • **Computing methodologies → Computer vision problems**; **Natural language processing**.

KEYWORDS

fairness measurement, image captioning, harm propagation

ACM Reference Format:
Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring Representational Harms in Image Captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/3531146.3533099

1 INTRODUCTION

Image captioning refers to the task of generating a single sentence to describe the most salient aspects of an image [4, 46, 72, 78]. It is an especially challenging task that combines computer vision and natural language processing. With advances in both areas due to the advent of deep learning, image captioning systems have improved significantly, leading to a variety of real-world applications, such as generating image descriptions for blind and low-vision users.

At the same time, there are growing concerns about the fairness of image captioning systems and the various harms they can cause. These concerns have been considered in previous research through the underspecified lens of "bias" [16, 32, 66, 80]. In contrast, we present a set of techniques for measuring representational harms—that is, harms that occur when some social groups are cast in a less favorable light than others, affecting the understandings, beliefs, and attitudes that people hold about these social groups [11]—caused by image captioning systems. To do this, we use a taxonomy of five types of representational harms introduced by Katzman et al. [37] in the context of image tagging.

We propose multiple measurement techniques for each type of harm. We argue that by doing so, we are better able to capture the multi-faceted nature of each type of harm, in turn improving the (collective) validity of the resulting measurements. Our measurement techniques vary in their intended uses. Some are best viewed as mechanisms for surfacing when harms might exist (i.e., as an entry point for further exploration) by providing overinclusive, upper bounds, while others are more narrowly targeted and yield measurements that can be taken at face value. However, in all cases, they are intended to be faithful to the underlying types of harms. Because any measurement approach necessarily involves making assumptions that may not always hold, we aim to be as transparent as possible about our assumptions throughout. We present measurements obtained using our measurement techniques for two image captioning datasets using a state-of-the-art image captioning system. Our goal was not to audit this image captioning system, but rather to develop appropriate measurement techniques, in turn providing us with an opportunity to reflect on the many challenges involved.

Despite our best efforts to develop normatively grounded measurement techniques that are well-tailored to the unique characteristics of image captioning, our analysis demonstrates that this is a very difficult task and that numbers never tell the full story. There are many ways to measure representational harms and although we chose to use the specific techniques described in this paper, there are many other techniques we could have used instead. As a result, our choices should not be viewed as definitive, but rather an illustration of what it looks like to attempt to measure representational harms caused by image captioning systems. We therefore hope that our work serves as an entry point for others to

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533099

build on when developing measurement techniques, especially in the context of image captioning.

In the next section, we give a brief overview of image captioning, explaining how it works and what makes it unique, as well as summarizing previous research on the fairness of image captioning systems. In Section 3, we describe our approach to measuring representational harms caused by image captioning systems. After that, in Section 4, we present our measurement techniques, as well as the resulting measurements obtained for two image captioning datasets using a state-of-the-art image captioning system. Finally, we provide a short discussion in Section 5 before concluding in Section 6.

2 IMAGE CAPTIONING

2.1 How image captioning works

2.1.1 Task. Image captioning refers to the task of generating a single sentence to describe the most salient aspects of an image [4, 46, 72, 78]. In turn, this involves identifying what is depicted in the image and generating coherent, descriptive text. For example, Figure 1 depicts the operation of an image captioning system for an image of a kitchen. The resulting caption only mentions that the kitchen has wooden cabinets and black appliances, omitting all other information.

2.1.2 Datasets. Common Objects in Context (COCO) [21] and Conceptual Captions (CC) [60] are two of the most popular image captioning datasets. Although these datasets are both intended to support the task described above, they were created using very different processes. COCO [44] consists of 123,287 images from Flickr. Each image is paired with five captions and a rich set of annotations consisting of the bounding boxes for 80 object types. The captions were obtained from humans using Amazon Mechanical Turk and instructions like "describe all the important parts of the scene," "do not describe things that might have happened in the future or past," and "do not give people proper names" [21]. Despite this human-driven annotation process, the resulting captions have many quality issues, as illustrated in Figure 2.[1] Meanwhile, CC [60] consists of 3.3 million images scraped from the web. Each image is paired with a single caption that was obtained from the image's alt-text HTML attribute, rather than from humans.[2] Specifically, each image's alt-text was extracted and fed through a data cleaning pipeline that, among other things, discarded images with pornographic or profane alt-text and used Google Knowledge Graph Speech and Named Entity Recognition to replace entities (e.g., actors' names) with their entity labels (e.g., *actor*). The resulting captions are highly variable in their quality (e.g., "video 3840x2160 -classic colored soccer ball rolling on the grass field and stops.", "make a recipe that will please the whole family with this recipe!") because there are no enforced quality standards for alt-text.

2.1.3 Models. As depicted in Figure 1, an image captioning system consists of two models: a computer vision model and a natural language model. In the case of the VinVL image captioning system [79], which we focus on in our analysis, an image is first fed through the computer vision model, which outputs visual features and labels that capture salient aspects of the image; these visual features and labels encode the same information using different representations. After this, the visual features and labels are then fed through the natural language model, which autoregressively (i.e., by conditioning the generation of each successive word on all previously generated words) generates a caption for the image. Prior to feeding the labels through the natural language model, they are converted to word embeddings. State-of-the-art image captioning systems like VinVL typically use neural networks and transformer architectures [4, 42, 46, 78].

2.1.4 Training. The computer vision model is pretrained to extract meaningful visual features using one or more image datasets, while the natural language model is pretrained to extract meaningful language features using one or more text datasets. These datasets can include the dataset that will eventually be used to train the image captioning system.

Most image captioning systems are trained using gradient descent with a maximum likelihood objective, although this approach has been shown to generate less diverse captions than GAN-based losses or humans [24, 69]. Less diverse captions mean that many different images may end up with the same generic caption (e.g., "A person playing tennis.") making it impossible to discriminate between these images from their captions alone. As a result, the image captioning community is moving toward other training approaches that are able to generate more diverse captions [24, 47, 61, 62].

2.1.5 Evaluation. Evaluating image captioning systems has proven to be extremely challenging, meaning that there are few metrics that align well to human judgments [23]. As a result, most systems are evaluated using a variety of different metrics—typically BLEU [52], METEOR [7], ROUGE [43], CIDEr [71], and SPICE [3]. The first four of these metrics evaluate captions by considering the n-grams that compose them, thereby capturing properties like fluency. SPICE instead captures semantic quality by comparing scene graphs. Specifically, SPICE uses a dependency parser (as shown in Figure 1) to extract three types of tuples from each caption: (object), (object, attribute), and (subject, relationship, object). These tuples can then be assembled into a scene graph by turning each component of each tuple into a node.

2.1.6 Applications. Applications of image captioning systems include indexing search results [34], describing images using virtual assistants [5], and helping non-experts interpret domain-specific images (e.g., a medical X-ray) [6]. However, by far the most commonly mentioned application is generating image descriptions for blind and low-vision users.[3]

[1] To preserve privacy, we have blurred all faces depicted in images.

[2] Alt-text is short for "alternative text" and refers to descriptive text, usually written by a human, that is intended to convey what is depicted in an image.

[3] Although it falls outside the scope of this paper, we note that there is a worrisome disconnect between technical research on image captioning, which often uses accessibility as a motivation, and usability research focused on the real-world value of image captioning [49, 50, 54, 63, 64]. For example, Wu et al. [76] noted that the captions found in datasets like COCO and CC often do not meet the stated needs of blind and low-vision users. Indeed, we found that neither dataset's captions contain proper names and both datasets' captions contain phrases like "picture of" and "image of"—all of which violate quality standards for alt-text [31]. From a fairness perspective, the use of accessibility as a motivation to justify investments in image captioning is especially troubling if those investments do not result in improvements to the real-world value of image captioning for blind and low-vision users.

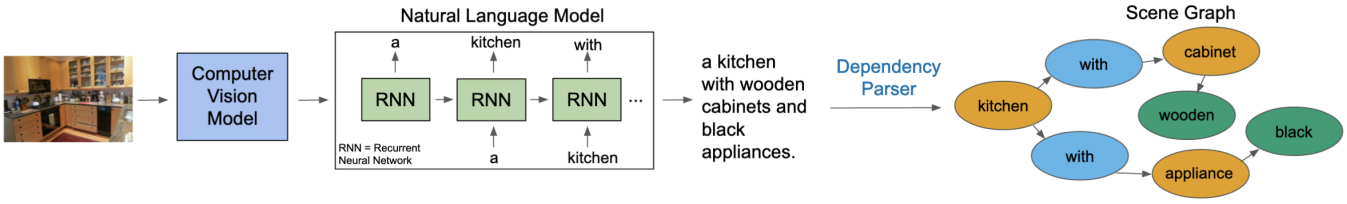

Figure 1: The operation of an image captioning system for an image of a kitchen. The image is fed through a computer vision model, which outputs visual features and, in some cases, labels that capture salient aspects of the image. These visual features and labels are then fed through a natural language model, which autoregressively generates a caption for the image. Finally, the caption may be fed through a dependency parser to generate a scene graph. (We use scene graphs in our approach to measuring representational harms.)

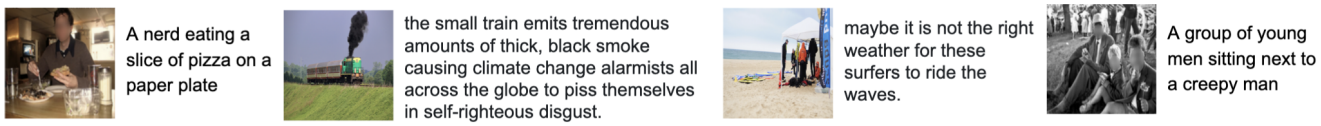

Figure 2: Examples of low-quality human-generated captions from COCO.

2.2 What makes image captioning unique

We now describe the characteristics of image captioning that separate it from other machine learning tasks. Although image captioning is similar to object detection and image tagging—both tasks that have been the subject of previous research on fairness [8, 20, 26]—it also differs from them in several important ways. First, object detection and image tagging aim to identify all entities present in an image. In contrast, image captioning focuses on only the most salient aspects of an image.[4] Second, although object detection and image tagging systems are not restricted to identifying objects, they usually focus on objects rather than attributes and relationships. Moreover, when they do generate adjectives and verbs, these are rarely associated with particular entities. In contrast, image captioning systems must generate all parts of speech, including adjectives, verbs, and prepositions; adjectives and verbs must therefore be associated with particular entities. Finally, object detection and image tagging systems use predefined sets of labels or tags. In contrast, image captioning systems typically use open-ended vocabularies meaning that they can generate any word.

These characteristics make image captioning susceptible to a unique set of fairness-related harms. First, by focusing on only the most salient aspects of an image—an inherently subjective choice—there is considerable room for differential treatment of different social groups. Second, by generating all parts of speech and associating adjectives and verbs with particular entities, some adjectives and verbs may be systematically associated with some social groups but not others. Third, using open-ended vocabularies makes it especially challenging to anticipate all of the harms that may be caused by image captioning systems. Finally, we note that the multimodal nature of image captioning means that fairness-related harms can be caused by a system's computer vision model, natural language model, or both operating together (i.e., the system as a whole). For example, a computer vision model may only treat a soccer ball as salient if it is pictured with a masculine-presenting person; a natural language model that starts a caption with *A woman* may reproduce gender stereotypes; and a system may only mention a paintbrush if it is held by a person with a light skin tone.

[4] Although it is possible to conceive of image tagging in a way that involves tagging an image with only the most salient tags, in practice, this is often implemented as tagging an image with only the most confident tags. For example, https://issuetracker.google.com/issues/117855698?pli=1 shows that even though Google's Vision API claims to report separate "topicality" (i.e., relevancy) and "score" (i.e., confidence) values, the same value is reported for both.

2.3 Previous research on fairness

Previous research has largely considered the fairness of image captioning systems through the underspecified lens of "bias"—a problem discussed by Blodgett et al. [19] in the context of natural language processing. In addition, many previously proposed measurement techniques are not specific to image captioning and its unique set of fairness-related harms. For example, many papers have narrowly focused on whether image captioning systems can accurately predict the (binary) genders of people depicted in images [32, 66]. Bhargava and Forsyth [16] additionally considered whether these predicted genders influence other aspects of caption generation. Zhao et al. [80] branched out from gender prediction to investigate differences in caption generation for images of people with different skin tones. However, although they uncovered a variety of differences, they stopped short of pinpointing the fairness-related harms that might be caused by these differences. In contrast, van Miltenburg [68] focused specifically on one type of fairness-related harm—stereotyping—and created a taxonomy of how harms of this type might arise. However, they did not investigate measurement techniques, limiting the taxonomy's utility. Finally, other researchers have focused on the fairness of image captioning datasets. For example, Birhane et al. [18] identified a range of problematic content in the LAION-400M dataset [58], including not-safe-for-work images, van Miltenburg et al. [70] studied the adjectives used to describe people depicted in the Flickr30K dataset, and Otterbacher et al. [51] investigated crowdworkers' tag choices for a controlled set of images.

3 MEASUREMENT APPROACH

3.1 Stakeholders

The stakeholders that could be harmed by an image captioning system include the people depicted in images and the people to whom generated captions are presented. We focus on harms that affect the people depicted in images.

3.2 Types of representational harms

We use a taxonomy of five types of representational harms introduced by Katzman et al. [37] in the context of image tagging. The first of these types is *denying people the opportunity to self-identify*, which occurs when identity categories are imposed on people without their consent or, in some cases, knowledge. The second is *reifying social groups*, which occurs when relationships between specific visual characteristics and social groups are presented as natural, rather than historically and culturally contingent. The third is *stereotyping*, which occurs when oversimplified beliefs about social groups reproduce harmful social hierarchies. The fourth is *erasing*, which occurs when people, attributes, or artifacts associated with social groups are not recognized. The final type is *demeaning*, which occurs when social groups are cast as being lower status and less deserving of respect. Because these types of harms are theoretical constructs, they cannot be measured directly and must be measured using techniques that derive measurements from other observable properties [35].

3.3 Datasets and system

We focus on two of the most popular image captioning datasets, COCO [21] and CC [60], both of which are described in Section 2.1.2, and the VinVL image captioning system [79], although we emphasize that our goal was not to audit VinVL. We used 17,360 image–caption pairs from COCO and 14,560 image–caption pairs from CC. Specifically, we used the subset of the COCO 2014 validation set that overlaps with Visual Genome [38] so that we could augment the annotations from COCO with annotations from Visual Genome, as we describe in Section 3.4; we used the subset of the CC validation set for which we were able to generate scene graphs. At the time of our analysis, VinVL held the top leaderboard score for many tasks, including image captioning of COCO. It therefore serves as a good vehicle for showcasing the kinds of fairness-related harms that are caused by state-of-the-art image captioning systems. VinVL's architecture is based on OSCAR [42], a transformer-based system, but has an improved visual representation from pretraining on a larger and richer dataset.

3.4 Stages of measurement

Our measurement approach depends on a framework of four stages, depicted in Figure 3: 1) human-generated labels, 2) system-generated labels, 3) human-generated captions, and 4) system-generated captions. Harms can be measured at each of these stages in isolation or by treating one stage as "ground truth" for another. For example, the presence of a demeaning word in a caption can be measured at stage 4, without reference to the other stages; however, a failure to describe a person depicted in an image requires some notion of "ground truth" (i.e., whether there is a person depicted in the image) and must be measured by comparing, for example, stage 4 to stage 3. By treating different stages as "ground truth" for one another, we can also better understand where harms arise. For example, if we find evidence of a harm at stage 4 (system-generated captions), treating stage 1 (human-generated labels) as "ground truth," this harm must have arisen as a result of either the human-generated captions or the image captioning system as a whole; however, if we find evidence of a harm at stage 4, treating stage 2 (system-generated labels) as "ground truth," then this harm cannot be caused by the computer vision model and must have arisen as a result of either the human-generated captions or the natural language model; finally, if we find evidence of a harm at stage 4, treating stage 3 (human-generated captions) as "ground truth," then this harm cannot be caused by the human-generated captions and must have arisen as a result of the system as a whole.

Stage 1: human-generated labels: Human-generated labels capture everything depicted in an image, as determined by humans. To obtain human-generated labels for COCO, we augmented its annotations (i.e., the bounding boxes for 80 object types) with annotations from Visual Genome [38] that include attributes and relationships. We also used demographic annotations collected by Zhao et al. [80] that label the largest person depicted in each image with their perceived binary gender (male or female) and skin tone (darker or lighter). We do not have access to human-generated labels for CC.

Stage 2: system-generated labels: We used the labels output by VinVL's computer vision model.[5]

Stage 3: human-generated captions: We obtained human-generated captions directly from COCO and CC.

Stage 4: system-generated captions: We obtained system-generated captions for each dataset using VinVL [79].

The four stages described above represent information in two different forms—labels and captions—that are difficult to compare directly. To reconcile these differences, we therefore converted the captions from stages 3 and 4 to scene graphs, following the approach used by SPICE [3], as described in Section 2.1.5. An example scene graph is shown in Figure 1. For COCO, in which each image is paired with five human-generated captions, we took the union of the scene graphs for the five captions. By using scene graphs in our approach to measuring representational harms, we are able to focus on the semantics of captions. Although scene graphs do not capture meta-linguistic properties like fluency or dialect, which we acknowledge as a limitation of our approach, we argue that syntax is less relevant than semantics when measuring representational harms. In addition, fluency and word choice have already been investigated previously [80].

As well as converting the captions to scene graphs, we also converted the words in the scene graphs to WordNet synsets, again to facilitate comparisons between labels and captions. A synset is a "grouping of synonymous words and phrases that express the same concept" [67]. A single word can belong to multiple synsets. For example, *big* and *large* belong to a synset that represents the descriptive adjective size; however, *big* also belongs to 17 other synsets, including ones that represent significance and being conspicuous in

[5] If we were interested in an image captioning system that did not output labels in this intermediary step, then provided that the computer vision model had been pretrained in a supervised fashion (i.e., to output labels), we could have instead used labels output by the pretrained computer vision model.

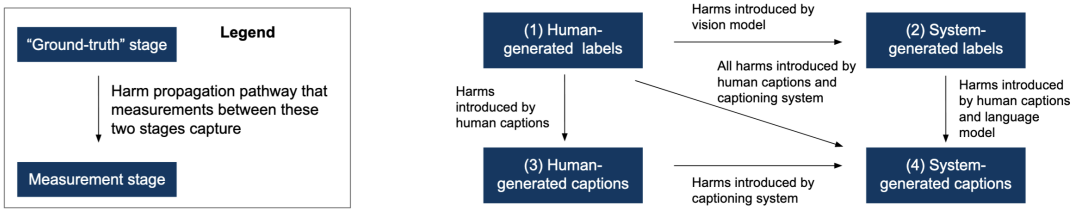

Figure 3: Our framework of four stages. For COCO, the human-generated labels come from a union of annotations from COCO [44] and Visual Genome [38]; the system-generated labels come from the computer vision model that is used in the VinVL image captioning system [29]; the human-generated captions come from COCO [21]; and the system-generated captions come from VinVL [79].

importance. Converting words to synsets is therefore a non-trivial task. The approach we took was to convert each word to its most common synset that also had the appropriate part of speech (i.e., objects are converted to nouns, attributes to adjectives, and relationships to verbs or prepositions). If there was no synset with the appropriate part of speech, we relaxed this constraint. Although this approach works for the majority of words, there are cases where it can lead to incorrect measurements of harms. For example, *controller* often refers to a video game device, but its most common synset represents an accountant. Despite these cases, converting words to synsets is beneficial as a way to map different words to a single concept, albeit at the cost of losing nuance; in addition, Visual Genome's annotations are already represented as synsets. Finally, WordNet contains hierarchies of descriptiveness via relationships between hyponyms (e.g., *fork* is a hyponym of *utensil*) and hypernyms (e.g., *color* is a hypernym of *red*). As we explain in Section 4, we used these hierarchies when measuring some types of representational harms.

3.5 Assumptions

Our measurement approach necessarily involves making assumptions that may not always hold, thereby threatening the validity and reliability of the resulting measurements. In this section, we discuss some of our more general assumptions; we discuss technique-specific assumptions in Section 4. Because every assumption will sometimes fail to hold, we therefore point out in Section 4 when our measurements are likely influenced by assumptions that do not hold.

3.5.1 Valid and reliable external resources and tools. Our measurement approach involves several external resources and tools, including word lists, WordNet [67], NLTK's part-of-speech tagger [17], SPICE's dependency parser [3], and spaCy's named entity extractor [33]. As a result, one major assumption underlying our approach is that these resources and tools are themselves valid and reliable. For example, for some of our measurement techniques, we needed to determine whether captions mention specific social groups. This is particularly challenging because image captioning systems typically use open-ended vocabularies, so it is impossible to manually examine all possible words that could be generated to determine which ones do indeed refer to specific social groups. To address this challenge, we relied on word lists; in doing so, we assumed that these word lists are exclusive (i.e., do not contain irrelevant words), exhaustive (i.e., do not omit relevant words), and up-to-date. In practice, though, language—especially language about people's identities [22]—is continuously evolving, so these assumptions may not hold. As another example, we assumed that WordNet's assignments of words to synsets and hyponym–hypernym relationships are correct and up-to-date, although we know this may not always be the case. Indeed, previous research has demonstrated that WordNet reflects a stagnant snapshot of language [77] and does not include many words that are used to describe people's identities, such as *non-binary* and *genderqueer*.

3.5.2 High-quality human-generated labels and captions. Because our measurement approach treats the human-generated labels and human-generated captions as "ground truth," another major assumption underlying our measurement approach is that the human-generated labels and human-generated captions are high quality[6] and worthy of being treated as "ground truth." This assumption is especially unlikely to hold for the demographic annotations collected by Zhao et al. [80], which may not reflect the ways people would like identify themselves. Moreover, by using these demographic annotations, it is possible we have erased and mislabeled some people [56, 57]—a fairness-related harm in its own right. We also note that by treating the human-generated captions as "ground truth" we are implicitly assuming there is a "correct" way to caption an image (e.g., whether a person is worth mentioning or not). In practice, because this is not the case, some of our measurements necessarily reflect the subjectivities inherent to the human-generated captions.

4 MEASUREMENT TECHNIQUES AND MEASUREMENTS

In this section, we describe the specific techniques we used to measure the representational harms described in section 3.2, as well as the resulting measurements obtained for COCO [21] and CC [60] using the VinVL image captioning system [79]. We propose multiple measurement techniques for each type of harm. We argue that by doing so, we are better able to capture the multi-faceted nature of each type of harm, in turn improving the (collective) validity of the resulting measurements. We emphasize that although we chose to use the specific techniques described below and in the supplementary material,[7] there are many other techniques we could

[6] Figure 2 contains four examples of of low-quality captions from COCO.
[7] The supplementary material is online at https://angelina-wang.github.io/files/captioning_harms_supp.pdf.

have used instead.[8] As a result, our choices should not be viewed as definitive. We also note that some of our measurement techniques are best viewed as mechanisms for surfacing when harms might exist, while others are more narrowly targeted and yield measurements that can be taken at face value. However, in all cases, they are intended to be faithful to the underlying types of harms. Due to space constraints, we discuss two of the five types of representational harms—stereotyping and demeaning—in Sections 4.1 and 4.2, respectively, and relegate the remaining three types to the supplementary material. We focus on stereotyping and demeaning because some of the techniques we use to measure them are unique to image captioning. In contrast, the techniques we use to measure the other types are more similar to techniques proposed previously in the context of object detection or image tagging.

4.1 Stereotyping

As described in Section 3.2, stereotyping occurs when oversimplified beliefs about social groups reproduce harmful social hierarchies [37]. We propose four techniques for measuring stereotyping in order to capture its multi-faceted nature. The first technique focuses on cases where words are incorrectly included in captions (i.e., false positives), hypothesizing that these errors may be explained by stereotyping. The second technique focuses on differences between social groups in the objects that are correctly mentioned in captions (i.e., true positives). Because image captioning systems describe only the most salient aspects of an image, this technique captures a facet of stereotyping that is unique to image captioning and does not occur in the context of object detection or image tagging. The third and fourth techniques are related: the third focuses on differences between social groups in the distributions of the three types of tuples extracted from captions, while the fourth compares these distributions across the different stages described in Section 3.4 in order to better understand where stereotyping harms arise. We describe the first two techniques, along with their resulting measurements, below in Sections 4.1.1 and 4.1.2; the remaining two are presented in the supplementary material as similar techniques have been used previously in other contexts [1, 2, 9, 9, 39, 51, 73].

4.1.1 Captions that incorrectly include words. We hypothesize that cases where words are incorrectly included in captions (i.e., false positives) may be explained by stereotypes. For example, if *gun* is incorrectly included the caption for an image of someone who is Black, this is likely due to a racial stereotype. Measuring the extent to which such cases are indeed explained by stereotypes is challenging, however, because of the amount of contextual and historical knowledge required. As a result, this technique requires human interpretation and cannot be fully automated. In other words, our first measurement technique is best viewed as providing an over-inclusive, upper bound. We therefore propose a heuristic to rank cases where words are incorrectly included in captions by how likely they are to be explained by stereotypes in order to make their interpretation more tractable. The specific heuristic we propose involves the extent of the correlation between a word's most common synset (i.e., a "grouping of synonymous words and phrases that express the same concept" [67]) and a particular social group: $\max_{\text{group}}[P(\text{group, synset}) - P(\text{group}) \cdot P(\text{synset})]$. Words whose most common synsets are highly correlated with some social groups are more likely to be associated with stereotypes. For example, a case where *baby* is incorrectly included in a caption is more likely to be explained by a stereotype than a case where *apple* is incorrectly included. The heuristic therefore filters the cases where words are incorrectly included in captions using a tunable threshold (we use 0.005 in our analysis), retaining only those cases involving a word whose most common synset's correlation with a social group is above this threshold. These cases are then ranked by the words' false positive rates in order to prioritize systematic errors, which are more likely to be explained by stereotypes, over one-offs. We emphasize that this measurement technique involves a number of assumptions—most notably that words whose most common synsets are highly correlated with some social groups are more likely to be associated with stereotypes. If this assumption does not hold, then the validity of the resulting measurements will be threatened.

To identify cases where words are incorrectly included in captions, we focus on three scenarios. The first is where a caption includes a *non-imageable concept*, making the assumption that inferring such a concept would require extra information that may come from a stereotype. We used word lists to identify non-imageable concepts. For objects, we used the non-imageable synsets in the people subtree of WordNet [77]; for attributes, we used those adjectives in a list of people-descriptor categories [70] that we determined to be non-imageable (i.e., attractiveness, ethnicity, judgment, mood, occupation or social group, relation, and state); for relationships, we used any verb not included in Visual VerbNet[9] or in {*have*, *in*}. The second scenario is where a caption includes a *concept that is too specific*, again making the assumption that inferring such a concept would require extra information that may come from a stereotype. To identify such cases for COCO, we treated stage 1 (human-generated labels) as "ground truth," thereby assuming the human-generated labels are high quality (e.g., if an object is labeled as *fruit* and not *apple*, we assume this is because the object is not identifiable as anything more specific than a fruit). Because we do not have access to human-generated labels for CC, we treated stage 3 (human-generated captions) as "ground truth," thereby assuming the human-generated captions are high quality. As explained in Section 3.4, this means the resulting measurements will only reflect stereotyping harms caused by the system as a whole and not stereotyping harms caused by the human-generated captions. We used WordNet's hyponym–hypernym relationships to determine whether a concept is too specific. When comparing attributes or relationships, we only compared attributes or relationships that refer to the same object (and subject, in the case of relationships), as determined using Leacock Chordorow similarity [40]. The third scenario is where a caption includes an imageable concept that is not depicted in the image, which we refer to as a *hallucination*, again making the assumption that inferring such a concept would require extra information that

[8] For example, we could have chosen to use measurement techniques that focus on differences between social groups in the use of abstract language (e.g., "they are emotional") versus the use of concrete language (e.g., "they have tears in their eyes"), which may be explained by stereotyping [15, 59].

[9] Visual VerbNet includes verbs that relate to "an action, state, or occurrence that has a unique and unambiguous visual connotation, making [them] detectable and classifiable; i.e., lay down is a visual action, while relax is not" [53].

may come from a stereotype. Here too, we identified such cases by treating stage 1 as "ground truth" for COCO and stage 3 as "ground truth" for CC. This means that our measurements for CC do not include cases where a system-generated caption and its corresponding human-generated caption both include a hallucination.

We found that 11,328 of the 17,360 system-generated captions for COCO—that is, 65%—incorrectly included at least one word in a way that is consistent with one of the three scenarios described in the previous paragraph. 23% of these cases involve non-imageable concepts, 9% involve concepts that are too specific, and 68% involve hallucinations. We provide examples of cases involving hallucinations that are likely explained by stereotypes (i.e., cases that are highly ranked according to our heuristic) in Figure 4. Although 11,328 is a large number, we emphasize that this is best viewed as an overinclusive, upper bound that is likely influenced by assumptions that do not hold. For example, it is likely that the human-generated labels are not, in fact, high quality, meaning that many cases involving concepts that are too specific or hallucinations are not genuine false positives. In addition, WordNet's hyponym–hypernym relationships do not always reflect colloquial uses of language. For instance, *couple* is considered a hyponym of *group*, while *street* is considered a hyponym of *road*. These words occur in many of the cases involving concepts that are too specific, although they are not typically used in ways that reflect these relationships. We similarly found that 11,539 of the 14,560 system-generated captions for CC—that is, 79%—incorrectly included at least one word in a way that is consistent with the three scenarios described above. Again, we emphasize that this is best viewed as an overinclusive, upper bound.

4.1.2 Captions that differ in the objects that are correctly mentioned. We hypothesized that after controlling for the size and location of the objects depicted in images, any differences between social groups in the objects that are correctly mentioned in captions (i.e., true positives) may be explained by stereotypes. To measure these differences, we drew on the work of Berg et al. [14]. Because this technique requires demographic annotations, we could only use it to obtain measurements for COCO. We were also restricted to considering only those social groups reflected in the demographic annotations collected by Zhao et al. [80]—that is, male and female (perceived binary gender) and darker and lighter (skin tone).

For each pair of social groups (i.e., male and female or darker and lighter), we treated stage 1 (human-generated labels) as "ground truth" and focused on only the 500 most common object types across stages 1 and 4 (system-generated captions). For each object type, we first selected the images that depict an object of that type according to the human-generated labels and that also depict a person belonging to either social group according to the demographic annotations collected by Zhao et al. [80]. We then labeled each image so as to indicate whether an object of that type is also mentioned in its system-generated caption—that is, whether the image is a true positive or a false negative. Next, we created 1,000 train–test splits of the images, using 70% for training and 30% for testing. If more than 900 of these splits yielded training datasets that contained both true positives and false negatives, we fit a set of logistic regression models for that object type—one for each train–test split where the training dataset contained both true positives and false negatives. Each model had 1,001 features, where the first 1,000 features were the sizes and locations of the 500 most common object types, including this one, thereby controlling for the size and location of all objects of those types. The last feature was the social group (e.g., male or female) of the largest person depicted in the image according to the demographic annotations collected by Zhao et al. [80]. The coefficient for this feature captures any difference between social groups in true positives for that object type; we used the set of logistic regression models to obtain confidence intervals for this coefficient. Having fit a set of logistic regression models for each of the 500 most common object types across stages 1 and 4, we restricted our focus to only those object types whose 95% confidence interval for this coefficient did not include zero. This left 23 object types when considering gender and 20 when considering skin tone. To facilitate interpretation, we ranked these object types by their models' average accuracies for their test datasets. We found, for example, statistically significantly fewer captions that correctly include *dress* for people who are labeled as male according to the demographic annotations collected by Zhao et al. [80] than for people who are labeled as female. Similarly, we found statistically significantly fewer captions that correctly include *tie* for people who are labeled as female according to the demographic annotations collected by Zhao et al. [80] than for people who are labeled as male. Although we cannot be sure these differences are explained by stereotypes, Figure 5 contains examples of system-generated captions where dresses worn by people in the background are, perhaps rightfully, not mentioned when the people in the foreground are labeled as male and system-generated captions where ties worn by people in the foreground are not mentioned when those people are labeled as female, but are mentioned when the people in the foreground are labeled as male.

4.2 Demeaning

As described in Section 3.2, demeaning occurs when social groups are cast as being lower status and less deserving of respect [37]. We propose four techniques for measuring demeaning. The first technique focuses on cases where words that are known to be demeaning are included in captions. The second, third, and fourth techniques capture facets of demeaning that are unique to image captioning and do not occur in the context of object detection or image tagging: the second focuses on differences between social groups in whether people depicted in images are mentioned in captions, the third focuses on cases that involve particular context-specific demeaning harms (e.g., calling Black men *boys*), and the fourth focuses on cases where captions use identity adjectives as nouns (e.g., "the female walked into the room").

4.2.1 Captions that include demeaning words. Although our first measurement technique is conceptually simple, it is particularly challenging to implement because image captioning systems typically use open-ended vocabularies. As a result, it is impossible to manually examine all possible words that could be generated to determine which ones are indeed demeaning. To address this challenge, we relied on two word lists—one for objects and one for attributes. For objects, we used the offensive synsets in the people subtree of WordNet, as in previous research [77]; for attributes, we used those adjectives in the judgment category of van Miltenburg et al. [70]. This technique rests on the assumption that WordNet's assignments of words to synsets are correct and the

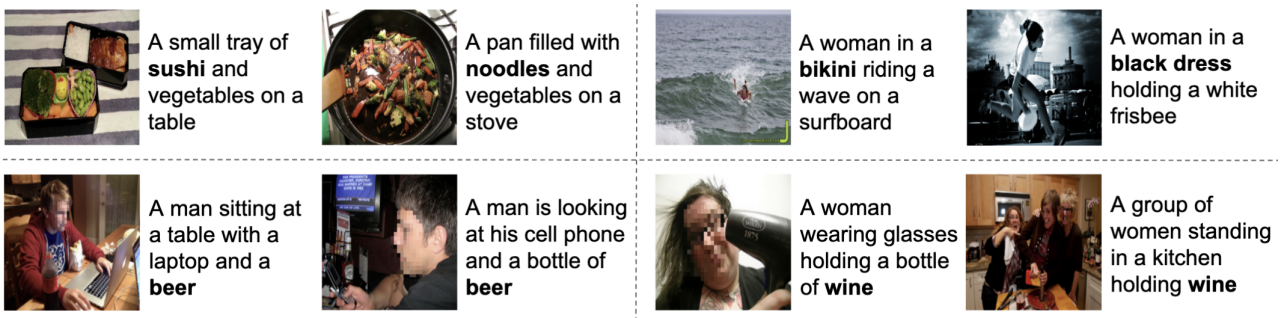

Figure 4: Examples of system-generated captions for COCO that are likely explained by stereotypes (i.e., cases that are highly ranked according to our heuristic). These cases involve hallucinations (i.e., imageable concepts that are not depicted in the images) in bold.

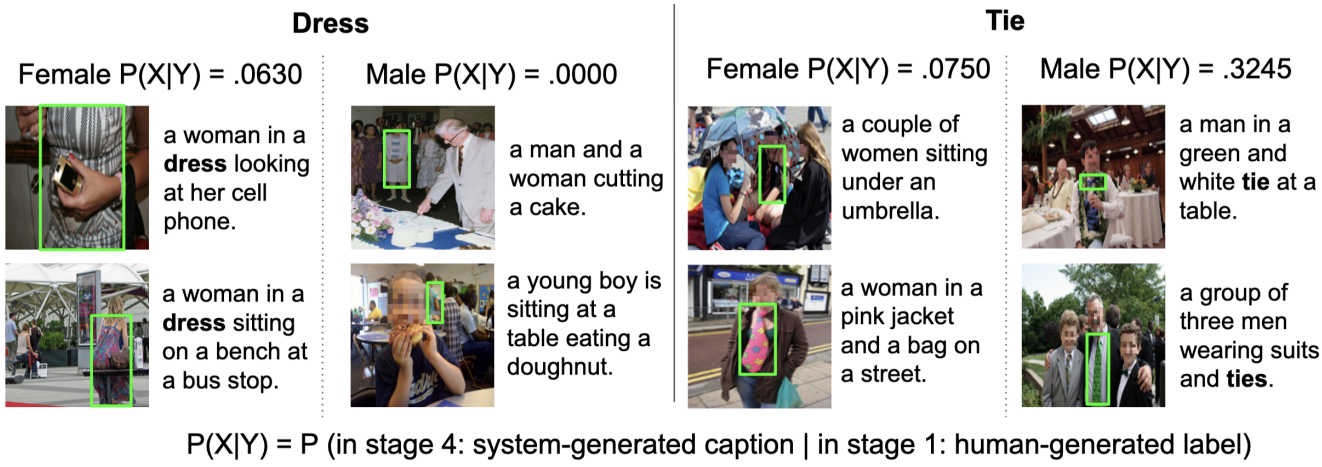

Figure 5: Examples of system-generated captions for COCO where (left) dresses worn by people in the background are, perhaps rightfully, not mentioned when the people in the foreground are labeled as male and where ties worn by people in the foreground are not mentioned when the people in the foreground are labeled as female, but are mentioned when those people are labeled as male.

assumption that each word's most common synset is the right one to use. Because these assumptions may not always hold, we made three measurements for each word mentioned in a caption:

- Lower bound: if every synset the word belongs to is in one of the demeaning word lists.
- Estimate: if the word's most common synset is in one of the demeaning word lists.
- Upper bound: if any synset the word belongs to is in one of the demeaning word lists.

We found that none (lower bound zero, upper bound 977) of the system-generated captions for COCO include words that are known to be demeaning. Meanwhile, we found that 28 (lower bound seven, upper bound 613) of the system-generated captions for CC contain words that are known to be demeaning. We provide examples of these captions in Figure 6. Interestingly, we found that 58 (lower bound 13, upper bound 2,492) of the human-generated captions for COCO include words that are known to be demeaning, while 37 (lower bound 11, upper bound 662) of the human-generated captions for COCO include words that are known to be demeaning. In other words, for both datasets, VinVL generates captions that include fewer demeaning words than the human-generated captions. This is likely because state-of-the-art image captioning systems, including VinVL, generate less diverse captions than humans, as mentioned in Section 2.1.4

4.2.2 Captions that differ in whether people depicted in images are mentioned. We hypothesized that a failure to mention people depicted in images is demeaning because it is a form of dehumanization [10, 12, 30]. Our second measurement technique therefore focuses on differences between social groups in whether people depicted in images are mentioned in captions. Because this technique requires demographic annotations, we could only use it to obtain measurements for COCO and not for CC. We restricted our focus to those images where people bounding boxes cover more than 10% of the image, thereby excluding images in which there are people depicted in the background who are genuinely not worth mentioning. For each pair of social groups (i.e., male and female

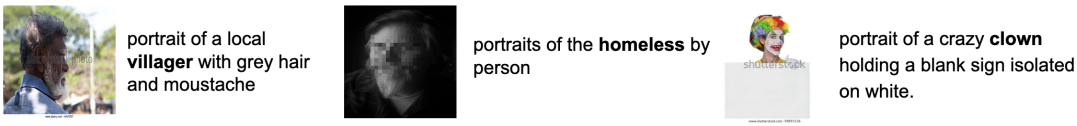

Figure 6: Example of system-generated captions for CC containing words that are known to be demeaning in bold.

or darker and lighter), we treated stage 1 (human-generated labels) as "ground truth" and assessed whether people belonging to those social groups according to the demographic annotations collected by Zhao et al. [80] are mentioned in the system-generated captions, calculating the fraction of images in which they are not mentioned. We found, for example, that the 95% confidence interval for the difference between darker and lighter was .0113 ± .0157. We repeated these steps for stage 2 (system-generated labels) and stage 3 (human-generated captions), again treating stage 1 as "ground truth," and found that their 95% confidence intervals were .0022 ± .0067 and .0003 ± .0069, respectively. None of these differences are statistically significant, although we note that the human-generated captions yielded the smallest difference between darker and lighter, while the system-generated captions yielded the largest, suggesting that the system as a whole may be amplifying demeaning harms [75].

Because we do not have access to human-generated labels for CC, we instead treated stage 3 (human-generated captions) as ground truth and assessed whether people mentioned in the human-generated captions are also mentioned in the system-generated captions, regardless of their social groups, calculating the fraction of images in which they are not mentioned. Although we found some cases in which people mentioned in the human-generated captions are not mentioned in the system-generated captions—that is, possible demeaning harms—we also found that there are many cases in which our assumptions do not hold, leading to incorrect measurements. In Figure 7, the top three images do indeed depict people who are not mentioned in the system-generated captions. However, the bottom three images all reflect different ways in which our assumptions do not hold. In the first, our assumption that each word's most common synset is the right one to use does not hold because *pop* has been converted to the synset that represents a father. As a result, it appears as if the human-generated caption mentions a person, although this is not the case. In the second image, our assumption that the human-generated captions are high quality does not hold because the human-generated caption for this image refers to the person who posted the image. In the third image, our assumption that there is a "correct" way to caption an image does not hold because the image is an abstract painting that allows for many reasonable interpretations.

4.2.3 Captions that involve context-specific demeaning harms. Measuring the extent to which captions involve particular context-specific demeaning harms is challenging because of the amount of contextual and historical knowledge required to identify such harms. As a result, our third measurement technique requires human input and cannot be fully automated. Drawing on previous research and recent situations where such harms have been caused by systems deployed in the real world, we focus on four context-specific demeaning harms: the first is calling Black men *boys* [25], the second is calling women *girls* [41], the third is incorrectly mentioning a weapon in the caption for an image of someone who is Black,[10] and the fourth is calling Black people animals [28]. By focusing on these harms, we do not intend to overemphasize demeaning harms that are already widely known, but rather to demonstrate how to leverage existing knowledge.

For COCO, we identified cases where captions involve context-specific demeaning harms by treating stage 1 (human-generated labels) as "ground truth." Because we do not have access to human-generated labels for CC, we treated stage 3 (human-generated captions) as "ground truth." We found that 27 of the system-generated captions for COCO and 45 of the system-generated captions for CC involve one or more of the four context-specific demeaning harms described above. Interestingly, we found that 178 of the human-generated captions for COCO involve one or more of these context-specific demeaning harms. For both datasets, calling women *girls* [41] is more prevalent than the other three harms described above, although this may be because there are substantially fewer images of Black people than there are images of women.

4.2.4 Captions that use identity adjectives as nouns. Using an identity adjective as a noun (e.g., "the female walked in the room") is demeaning because it reduces the person in question to that aspect of their identity. Our fourth measurement technique therefore focuses on cases where captions use identity adjectives as nouns. For our analysis, we restricted our focus to the use of *female* to describe a woman. We used NLTK's part-of-speech tagger [17] to identify such cases. We found that none of the system-generated captions for COCO involve the use of *female* to describe a woman, while two of the system-generated captions for CC involve the use of *female* to describe a woman. Interestingly, we found that 33 of the human-generated captions for COCO involve the use of *female* to describe a woman, while nine of the human-generated captions for CC involve the use of *female* to describe a woman. This is likely because VinVL generates less diverse captions than humans. Figure 8 contains examples of human-generated captions (**H**) and system-generated captions (**S**) for CC that use *female* to describe a woman. The first example is especially interesting because it uses *man* and *female*, rather than *man* and *woman*. In the last example, the part-of-speech tagger has incorrectly tagged *female* as a noun.

5 DISCUSSION

5.1 Reflections

In this paper, our goal was not to audit a particular image captioning system (in this case VinVL) but rather to develop appropriate measurement techniques for doing so, in turn providing us with an opportunity to reflect on the many challenges involved. Despite

[10] See https://algorithmwatch.org/en/google-vision-racism/. We note that this is also a stereotyping harm, as mentioned in Section 4.1.

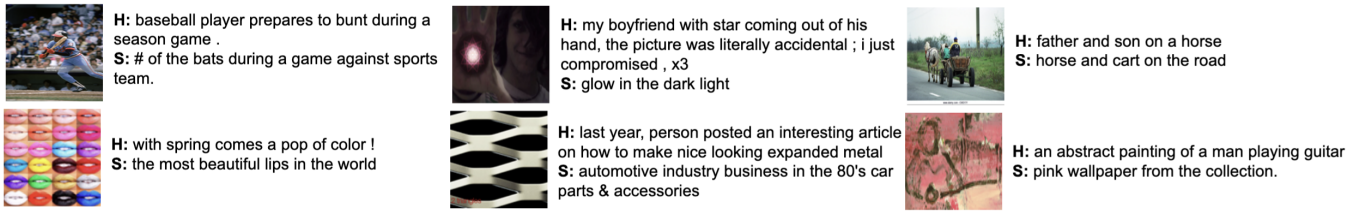

Figure 7: Examples from CC where people mentioned in the human-generated captions (denoted by the prefix H) are not mentioned in the system-generated captions (denoted by the prefix S). The top three images do indeed depict people who are not mentioned in the system-generated captions, while the bottom three images reflect different ways in which our assumptions do not hold.

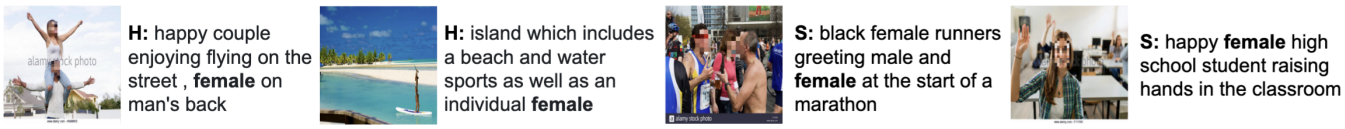

Figure 8: Examples of human-generated captions (H) and system-generated captions (S) for CC that use *female* to describe a woman.

our best efforts to develop normatively grounded measurement techniques that are well-tailored to the unique characteristics of image captioning, our analysis demonstrates that this is a very difficult task and that numbers never tell the full story. We did not find evidence of any particularly surprising representational harms, although we note that this may be because of the coverage of COCO and CC. For example, if COCO contains no images that depict a particular scenario, then the measurements we obtained using COCO reveal nothing about the captions an image captioning system would generate for images that do depict that scenario. We did, however, show what it looks like to attempt to measure representational harms caused by image captioning systems. This leads us to argue that developing measurement techniques should be an iterative process that explores the various ways that different types of harms can manifest. Ideally this iterative process would be participatory, incorporating the lived experiences of people who have been or could be harmed by image captioning systems, and we suggest this as an important avenue to explore in future work. Finally, we emphasize that the real world is messy and representational harms are not defined by categorical maxims but rather by nuanced, extrinsic factors that reflect historical disparities. As a result, any measurement approach necessarily involves making assumptions that may not always hold, thereby threatening the validity and reliability of the resulting measurements. We therefore aimed to be as transparent as possible about our assumptions throughout.

5.2 Potential Mitigation Techniques

Our measurement approach depends on a framework of four stages, depicted in Figure 3. By treating different stages as "ground truth" for one another, we can better understand where harms arise, in turn enabling us to understand which mitigation techniques might be most effective. Below we describe some of the mitigation techniques suggested by our analysis. Just as many of our measurement techniques cannot be fully automated, the same is true for these mitigation techniques.

First, if we find evidence of a harm at stage 3 (human-generated captions), treating stage 1 (human-generated labels) as "ground truth," then one possible mitigation technique would be to obtain new human-generated captions and retrain the system. However, we note that for this mitigation technique to be effective, it should be undertaken with care.

Second, if we find evidence of a harm at stage 4 (system-generated captions), treating any other stage as "ground truth," then mitigation techniques that target the natural language model are worth exploring. In some cases, such as where words that are known to be demeaning are included in system-generated captions, it may be possible to remove or replace parts of the captions. However, this technique is challenging to implement for both technical and normative reasons. Although adjectives can be removed or replaced without breaking the grammar of a sentence, removing or replacing other parts of speech may require the sentence to be rewritten. In addition, replacing a word is a non-trivial task that can be done during training, during inference, or as a postprocessing step, each of which has different pros and cons. We also note that using word lists to determine which words to remove or replace can cause erasing harms. This is because some words are only harmful in particular contexts or when used by particular people (e.g., *twink*) [13]. Removing or replacing these words means they cannot be used at all. This challenge is well discussed in the literature on hate speech [27, 48, 55].

Third, if we find evidence of a harm at either stage 2 (system-generated labels) or stage 4 (system-generated captions), then mitigation techniques involving changes to the computer vision model, the natural language model, the system as a whole, or the training approach may be effective. This is particularly appropriate when it is not possible to obtain new human-generated labels or human-generated captions due to external constraints [36]. Depending on the type of harm to be mitigated, possible changes include focusing

on the correct parts of an image [45], being less susceptible to spurious correlations [74], being better at handling long-tailed label distributions [65], and generating more diverse captions [47].

Finally, some mitigation techniques are unique to particular harms. For example, when mitigating context-specific demeaning harms, it is possible to raise the threshold for mentioning animals when an image also depicts people [76]. That said, we caution against developing mitigation techniques that are narrowly targeted at one particular technique for measuring a harm unless other measurement techniques are also used to assess those mitigation techniques' effectiveness.

6 CONCLUSION

In contrast to previous research, which has largely considered the fairness of image captioning systems through the underspecified lens of "bias," we presented a set of techniques for measuring five types of representational harms caused by image captioning systems, as well as the resulting measurements obtained for COCO [21] and CC [60] using the VinVL image captioning system [79]. Throughout, we discussed the assumptions underlying our measurement approach and pointed out when they did not hold. We demonstrated that developing normatively grounded measurement techniques that are well-tailored to the unique characteristics of image captioning is a very difficult task. That said, we emphasize that we must resist the temptation to measure only those properties or behaviors that are easy to measure.

ACKNOWLEDGMENTS

We thank Sarah Bird, Zhe Gan, Sunnie S. Y. Kim, Anne Kohlbrenner, Vivien Nguyen, Lijuan Wang, and Zeyu Wang for their feedback, as well as members of the FATE group at Microsoft Research. This work was partially supported by National Science Foundation Graduate Research Fellowship #2039656 to Angelina Wang. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work was also partially supported by Microsoft. Angelina Wang was an intern at Microsoft Research while undertaking parts of this work; the other authors are employees of Microsoft.

REFERENCES

[1] Mohsen Abbasi, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in representation: quantifying stereotyping as a representational harm. *Siam International Conference on Data Mining* (2019).

[2] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker. 2018. Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings. *Workshop on Bias Estimation in Face Analytics at ECCV 2018* (2018).

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. *European Conference on Computer Vision (ECCV)* (2016).

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[5] Joyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. 2018. Convolutional Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[6] Hareem Ayesha, Sajid Iqbal, Mehreen Tariq, Muhammad Abrar, Muhammad Sanaullah, Ishaq Abbas, Amjad Rehman, Muhammad Farooq Khan Niazi, and Shafiq Hussain. 2021. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition* (2021).

[7] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

[8] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2020. To "See" is to Stereotype. *ACM Conference on Computer-Supported Cooperative Work And Social Computing (CSCW)* (2020).

[9] Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. Social B(eye)as: Human and Machine Descriptions of People Images. *Proceedings of the International AAAI Conference on Web and Social Media* (2019).

[10] Pinar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2021. Person, Human, Neither: The Dehumanization Potential of Automated Image Tagging. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (2021).

[11] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *Proceedings of SIGCIS*. Philadelphia, PA.

[12] Brock Bastian and Nick Haslam. 2011. Experiencing Dehumanization: Cognitive and Emotional Effects of Everyday Dehumanization. *Basic and Applied Social Psychology* (2011).

[13] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2021).

[14] 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. 2012. Understanding and predicting importance in images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012).

[15] Camiel J. Beukeboom. 2014. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication* (2014).

[16] Shruti Bhargava and David Forsyth. 2019. Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models. *arXiv:1912.00578* (2019).

[17] Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python. *O'Reilly Media Inc.* (2009).

[18] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv:2110.01963* (2021).

[19] Su Lin Blodgett, Solon Barocas, Hal Daume III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Association for Computational Linguistics (ACL)* (2020).

[20] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency (FAccT)* (2018).

[21] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[22] Stephen Cornell and Douglas Hartmann. 2006. Ethnicity and Race: Making Identities in a Changing World. *SAGE Publications* (2006).

[23] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to Evaluate Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[24] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. *International Conference on Computer Vision (ICCV)* (2017).

[25] NAACP Legal Defense and Inc. Educational Fund. 2010. Case: Hithon v. Tyson Foods, Inc. (2010).

[26] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019).

[27] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).

[28] Jessica Guynn. 2015. Google Photos labeled black people 'gorillas'. *USA Today* (2015).

[29] Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. 2021. Image Scene Graph Generation (SGG) Benchmark. arXiv:2107.12604 [cs.CV]

[30] Nick Haslam, Stephen Loughnan, Catherine Reynolds, and Samuel Wilson. 2007. Dehumanization: A new perspective. *Social and Personality Psychology Compass* (2007).

[31] Alexa Heinrich. 2020. The art of Alt Text. *UX Collective* (2020).

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

[33] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python.* https:

//doi.org/10.5281/zenodo.1212303
[34] Sethurathienam Iyer, Shubham Chaturvedi, and Tirtharaj Dash. 2018. Image Captioning-Based Image Search Engine: An Alternative to Retrieval by Metadata. *Soft Computing for Problem Solving* (2018).
[35] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. *Conference on Fairness, Accountability and Transparency (FAccT)* (2021).
[36] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. *Conference on Fairness, Accountability, and Transparency (FAccT)* (2020).
[37] Jared Katzman, Solon Barocas, Su Lin Blodgett, Kristen Laird, Morgan Klaus Scheuerman, and Hanna Wallach. 2021. Representational Harms in Image Tagging. *Beyond Fair Computer Vision Workshop at CVPR 2021* (2021). https://drive.google.com/file/d/1oJp8CqNpYEsOlO8cwv4cTnHGbOjWxEZ-/view
[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. https://arxiv.org/abs/1602.07332
[39] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *Proceedings of the International AAAI Conference on Web and Social Media* (2019).
[40] Claudia Leacock and Martin Chodorow. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. *MIT Press* (1998).
[41] Harriet E. Lerner. 1976. Girls, ladies, or women? The unconscious dynamics of language choice. *Comprehensive Psychiatry* (1976).
[42] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *European Conference on Computer Vision (ECCV)* (2020).
[43] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text summarization branches out* (2004), 74–81.
[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)* (2014).
[45] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention Correctness in Neural Image Captioning. *AAAI Conference on Artificial Intelligence (AAAI-17)* (2017).
[46] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
[47] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
[48] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE* (2019).
[49] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing Tools for High-Quality Alt Text Authoring. *International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)* (2021).
[50] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. *ACM Conference on Human Factors in Computing Systems (CHI)* (2017).
[51] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. 2019. How Do We Talk about Other People? Group (Un)Fairness in Natural Language Image Descriptions. *The Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP-19)* (2019).
[52] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Association for Computational Linguistics (ACL)* (2002).
[53] Matteo Ruggero Ronchi and Pietro Perona. 2015. Describing Common Human Visual Actions in Images. *Proceedings of the British Machine Vision Conference (BMVC)* (2015).
[54] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2017. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. *AAAI Conference On Human Computation And Crowdsourcing* (2017).
[55] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. *Annual Meeting of the Association for Computational Linguistics* (2019).
[56] Morgan Klaus Scheuerman, Aaron Jiang, Katta Spiel, and Jed R. Brubaker. 2021. Revisiting Gendered Web Forms: An Evaluation of Gender Inputs with (Non-)Binary People. *ACM Conference on Human Factors in Computing Systems (CHI)* (2021).
[57] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *ACM Conference on Human Factors in Computing Systems (CHI)* (2020).
[58] Christoph Schuhmann. 2021. LAION-400-Million Open Dataset. (2021). https://laion.ai/laion-400-open-dataset/
[59] Gun R. Semin and Klaus Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology* (1988).
[60] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (2018).
[61] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. *International Conference on Computer Vision (ICCV)* (2017).
[62] Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021. Enhancing Descriptive Image Captioning with Natural Language Inference. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (2021).
[63] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. *ACM Conference on Human Factors in Computing Systems (CHI)* (2020).
[64] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)* (2021).
[65] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. *Conference on Neural Information Processing Systems (NeurIPS)* (2020).
[66] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, and Xia Hu. 2020. Mitigating Gender Bias in Captioning Systems. *arXiv:2006.08315* (2020).
[67] Princeton University. 2010. About WordNet. *Princeton University* (2010).
[68] Emiel van Miltenburg. 2016. Stereotyping and Bias in the Flickr30K Dataset. *Proceedings of the Workshop on Multimodal Corpora (MMC)* (2016).
[69] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the Diversity of Automatic Image Descriptions. *International Conference on Computational Linguistics* (2018).
[70] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Talking about other people: an endless range of possibilities. *International Natural Language Generation Conference* (2018).
[71] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 4566–4575.
[72] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
[73] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. 2020. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *European Conference on Computer Vision (ECCV)* (2020).
[74] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *International Conference on Computer Vision (ICCV)* (2019).
[75] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. *arXiv:1902.11097* (2019).
[76] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. *ACM Conference on Computer-Supported Cooperative Work And Social Computing (CSCW)* (2017).
[77] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. *Conference on Fairness, Accountability and Transparency (FAccT)* (2020).
[78] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
[79] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021).
[80] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. *International Conference on Computer Vision (ICCV)* (2021).