Quantifying Societal Bias Amplification in Image Captioning

Yusuke Hirota
Osaka University
y-hirota@is.ids.osaka-u.ac.jp

Yuta Nakashima
Osaka University
n-yuta@ids.osaka-u.ac.jp

Noa Garcia
Osaka University
noagarcia@ids.osaka-u.ac.jp

Abstract

We study societal bias amplification in image captioning. Image captioning models have been shown to perpetuate gender and racial biases, however, metrics to measure, quantify, and evaluate the societal bias in captions are not yet standardized. We provide a comprehensive study on the strengths and limitations of each metric, and propose LIC, a metric to study captioning bias amplification. We argue that, for image captioning, it is not enough to focus on the correct prediction of the protected attribute, and the whole context should be taken into account. We conduct extensive evaluation on traditional and state-of-the-art image captioning models, and surprisingly find that, by only focusing on the protected attribute prediction, bias mitigation models are unexpectedly amplifying bias.

Figure 1. Measuring gender bias in MSCOCO captions [9]. For each caption generated by humans, NIC [36], or NIC+Equalizer [8], we show our proposed bias score for *female* and *male* attributes. This bias score indicates how much a caption is biased toward a certain protected attribute. The contribution of each word to the bias score is shown in gray-scale (bold for the word with the highest contribution). Gender revealing words are masked.

1. Introduction

The presence of undesirable biases in computer vision applications is of increasing concern. The evidence shows that large-scale datasets, and the models trained on them, present major imbalances in how different subgroups of the population are represented [7, 8, 10, 47]. Detecting and addressing these biases, often known as societal biases, has become an active research direction in our community [1, 11, 21, 30, 32, 37, 44].

Contrary to popular belief, the presence of bias in datasets is not the only cause of unfairness [16]. Model choices and how the systems are trained also have a large impact on the perpetuation of societal bias. This is supported by evidence: 1) models are not only reproducing the inequalities of the datasets but amplifying them [47], and 2) even when trained on balanced datasets, models may still be biased [40] as the depth of historical discrimination is more profound than what it can be manually annotated, *i.e.*, bias is not always evident to the human annotator eye.

The prevalence of accuracy as the single metric to optimize in most popular benchmarks [33] has made other aspects of the models, such as fairness, cost, or efficiency, not a priority (and thus, something to not look into). But societal bias is a transversal problem that affects a variety of tasks within computer vision, such as *facial recognition*, with black women having higher error rates than white men [7]; *object classification*, with kitchen objects being associated with women with higher probabilities than with men [47]; or *pedestrian detection*, with lighter skin individuals showing higher detection rates than darker skin people [42]. Although the causes of societal bias in different computer vision systems may be similar, the consequences are particular and require specific solutions for each task.

We examine and quantify societal bias in image captioning (Figure 1). Image captioning has achieved state-of-the-art accuracy on MSCOCO captions dataset [9] by means of pre-trained visual and language Transformers [23]. By leveraging very large-scale collections of data (*e.g.*, Google Conceptual Captions [29] with about 3.3 million image-caption pairs crawled from the Internet), self-attention-based models [34] have the potential to learn world representations according to the training distribution. However, these large amounts of data, often without (or with minimal) curation, conceal multiple problems, including the normalization of abuse or the encoding of discrimination [5, 10, 27]. So, once image captioning models have achieved outstanding performance on evaluation benchmarks, a ques-

tion arises: are these models safe and fair to everyone?

We are not the first to formulate this question. Image captioning has been shown to reproduce gender [8] and racial [46] bias. By demonstrating the existence of societal bias in image captioning, the pioneering work in [8] set the indispensable seed to continue to investigate this problem, which we believe is far from being solved. We argue that one of the aspects that remains open is the quantification and evaluation of societal bias in image captioning. So far, a variety of metrics have been applied to assess different aspects of societal bias in human and model-generated captions, such as whether the representation of different subgroups is balanced [8, 46] or whether the protected attributes[1] values (*e.g.*, *female*, *male*) are correctly predicted [8, 31]. However, in Section 3, we show that current metrics may be insufficient, as they only consider the effects of bias perpetuation to a degree.

With the aim to identify and correct bias in image captioning, in Section 4, we propose a simple but effective metric that measures not only how much biased a trained captioning model is, but also how much bias is introduced by the model with respect to the training dataset. This simple metric allows us to conduct a comprehensive analysis of image captioning models in terms of gender and racial bias (Section 5), with an unexpected revelation: the gender equalizer designed to reduce gender bias in [8] is actually amplifying gender bias when considering the semantics of the whole caption. This discovery highlights, even more, the necessity of a standard, unified metric to measure bias and bias amplification in image captioning, as the efforts to address societal inequalities will be ineffective without a tool to quantify how much bias a system exhibits and where this bias is coming from. We conclude the paper with an analysis of the limitation of the proposed metric in Section 6 and a summary of the main findings in Section 7.

2. Related work

Societal bias in computer vision The problem of bias in large-scale computer vision datasets was first raised by Torralba and Efros in [33], where the differences in the image domain between datasets were explored. Each dataset presented different versions of the same object (*e.g.*, cars in Caltech [17] tended to appear sidewise, whereas in ImageNet [12] were predominantly of racing type), impacting cross-dataset generalization. But it was only recently that societal bias in computer vision was formally investigated.

In the seminal work of Buolamwini and Gebru [7], commercial face recognition applications were examined across subgroups, demonstrating that performance was different according to the gender and race of each individual, especially misclassifying women with darker skin tones. Similarly, Zhao *et al*. [47] showed not only that images in MSCOCO [25] were biased towards a certain gender, but also that action and object recognition models amplified such bias in their predictions. With an increased interest in fairness, multiple methods for mitigating the effects of dataset bias have been proposed [19, 32, 39, 40, 47].

Measuring societal bias Societal bias is a problem with multiple layers of complexity. Even on balanced datasets, models still perpetuate bias [40], indicating that social stereotypes are occurring at the deepest levels of the image. This makes the manual identification and annotation of biases unfeasible. Thus, the first step towards fighting and mitigating bias is to quantify the problem.

Bias quantification metrics have been introduced for image classification. Zhao *et al*. [47] defined bias based on the co-occurrence of objects and protected attributes; Wang *et al*. [40] relied on the accuracy of a classifier when predicting the protected attributed; and Wang and Russakovsky [38] extended the definition of bias by including directionality. In addition, REVISE [37] and CIFAR-10S [41] ease the task of identifying bias on datasets and models, respectively. These solutions, however, cannot be directly applied to image captioning, so specific methods must be developed.

Societal bias in image captioning In image captioning [2, 36, 43, 45] the input to the model is an image and the output is a natural language sentence. This duality of data modalities makes identifying bias particularly challenging, as it can be encoded in the image and/or in the language. The original work by Burns *et al*. [8] showed that captions in MSCOCO [9] present gender imbalance and proposed an equalizer to force captioning models to produce gender words based on visual evidence. Recently, Zhao *et al*. [46] studied racial bias from several perspectives, including visual representation, sentiment analysis, and semantics.

Each of these studies, however, uses different evaluation protocols and definitions of bias, lacking of a standard metric. To fill this gap, we propose an evaluation metric to measure not only how biased a model is, but how much it is amplified with respect to the original (biased) dataset.

3. Analysis of fairness metrics

Bias in image captioning has been estimated using different methods: how balanced the prediction of the protected attributed is [8], the overlap of attention maps with segmentation annotations [8], or the difference in accuracy between the different protected attributes [46]. In this section, we thoroughly examine existing fairness evaluation metrics and their shortcomings when applied to image captioning.

Notation Let $\mathcal{D}$ denote the training split of a certain vision dataset with samples (I, y, a), where I is an image, y is the ground-truth annotation for a certain task, and $a \in \mathcal{A}$ is a protected attribute in set $\mathcal{A}$. The validation/test split is

[1] *Protected attribute* refers to a demographic variable (age, gender, race, *etc*.) that a model should not use to produce an output.

denoted by $\mathcal{D}'$. We assume there is a model M that makes prediction $\hat{y}$ associated with this task from the image, *i.e.*, $\hat{y} = M(I)$. For image captioning, we define a ground-truth caption $y = (y_1, y_2, \ldots, y_n)$ as a sequence of n tokens.

3.1. Fairness metrics in image captioning

Difference in performance A natural strategy to show bias in image captioning is as the difference in performance between the subgroups of a protected attribute, in terms of accuracy [8, 31, 46], ratio [8], or sentiment analysis [46]. Quantifying the existence of different behavior according to demographic groups is essential to demonstrate the existence of bias in a model, but it is insufficient for a deeper analysis, as it does not provide information on where the bias comes from, and whether bias is being amplified by the model. Thus, it is good practice to accompany difference in performance with other fairness metrics.

Attribute misclassification Another common metric is to check if the protected attribute has been correctly predicted in the generated caption [8, 31]. This assumes that the attribute can be clearly identified in a sentence, which may be the case for some attributes, *e.g.*, age (*a young person*, *a child*) or gender (*a woman*, *a man*), but not for others, *e.g.*, skin tone. This is critical for two reasons: 1) even when the attribute is not clearly mentioned in a caption, bias can occur through the use of different language to describe different demographic groups; and 2) it only considers the prediction of the protected attribute, ignoring the rest of the sentence which may also exhibit bias.

Right for the right reasons Introduced in [8], it measures whether the attention activation maps when generating a protected attribute word w in the caption, *e.g.*, *woman* or *man*, are located in the image region where the evidence about the protected attribute is found, *i.e.*, the person. This metric quantifies the important task of whether w is generated based on the person visual evidence or, on the contrary, on the visual context, which has been shown to be one of the sources of bias in image captioning models. However, it has three shortcomings: 1) it needs a shortlist of protected attribute words, and a person segmentation map per image, which may not always be available; 2) it assumes that visual explanations can be generated from the model, which may not always be the case; and 3) it does not consider the potential bias in the rest of the sentence, which (as we show in Section 5) is another critical source of bias.

Sentence classification Lastly, Zhao *et al*. [46] introduced the use of sentence classifiers for analyzing racial bias. The reasoning is that if a classifier can distinguish between subgroups in the captions, the captions contain bias. Formally, let f denote a classifier that predicts a protected attribute in $\mathcal{A}$ trained over $\mathcal{D}$, *i.e.*, $\hat{a} = f(y)$, from a caption y in an arbitrary set $\mathcal{H}$ of captions. If the accuracy is higher than the chance rate, y is considered to be biased:

$$\text{SC} = \frac{1}{|\mathcal{H}|} \sum_{y \in \mathcal{H}} \mathbb{1}[f(y) = a], \tag{1}$$

where $\mathbb{1}[\cdot]$ is a indicator function that gives 1 when the statement provided as the argument holds true and 0 otherwise. $\mathcal{H}$ typically is the set of all captions generated from the images in the test/validation split $\mathcal{D}'$ of the dataset, *i.e.*, $\mathcal{H} = \{M(I) \mid I \in \mathcal{D}'\}$.

Unlike the previous methods, this metric considers the full context of the caption. However, a major shortcoming is that, when bias exists on the generated data, the contributing source is not identified. Whether the bias comes from the model or from the training data and whether bias is being amplified or not, cannot be concluded.

3.2. Bias amplification metrics

There is a family of metrics designed to measure bias amplification on visual recognition tasks. We describe them and analyze the challenges when applied to captioning.

Bias amplification Proposed in [47], it quantifies the implicit correlations between model prediction $\hat{y} = M(I)$ and the protected attribute $a \in \mathcal{A}$ by means of co-occurrence, and whether these correlations are more prominent in the model predictions or in the training data. Let $\mathcal{L}$ denote the set of possible annotations l in the given task, *i.e.*, y and $\hat{y}$ are in $\mathcal{L}$; c_a and $\hat{c}_a$ denote the numbers of co-occurrences of a and l, counted over y and $\hat{y}$, respectively. Bias is

$$\tilde{b}_{al} = \frac{\tilde{c}_{al}}{\sum_{a \in \mathcal{A}} \tilde{c}_{al}}, \tag{2}$$

where $\tilde{c}$ is either c or $\hat{c}$, and $\tilde{b}$ is either b or $\hat{b}$, respectively. Then, bias amplification is defined by

$$\text{BA} = \frac{1}{|\mathcal{L}|} \sum_{a \in \mathcal{A}, l \in \mathcal{L}} (\hat{b}_{al} - b_{al}) \times \mathbb{1}\left[b_{al} > \frac{1}{|\mathcal{A}|}\right]. \tag{3}$$

BA > 0 means that bias is amplified by the model, and otherwise mitigated. This metric is useful for a classification task, such as action or image recognition, for which the co-occurrence can be easily counted. However, one of the major shortcomings is that it ignores that protected attributes may be imbalanced in the dataset, *e.g.*, in MSCOCO images [25] there are 2.25 more men than women, which causes most of objects to be correlated with men. To solve this and other issues, Wang and Russakovsky [38] proposed an extension called directional bias amplification.

Leakage Another way to quantify bias amplification is leakage [40], which relies on the existence of a classifier to predict the protected attribute a. For a sample (I, y, a) in $\mathcal{D}$ with a ground-truth annotation y, a classifier f predicts

the attribute $a \in \mathcal{A}$ from either y or $\hat{y} = M(I)$. Using this, the leakage can be formally defined as,

$$\text{Leakage} = \lambda_M - \lambda_D, \tag{4}$$

where

$$\lambda_D = \frac{1}{|\mathcal{D}|} \sum_{(y,a)\in\mathcal{D}} \mathbb{1}[f(y) = a] \tag{5}$$

$$\lambda_M = \frac{1}{|\mathcal{D}|} \sum_{(I,a)\in\mathcal{D}} \mathbb{1}[f(\hat{y}) = a] \tag{6}$$

A positive leakage indicates that M amplifies the bias with respect to the training data, and mitigates it otherwise.

Challenges The direct application of the above metrics to image captioning presents two major challenges. Let us first assume that, for image captioning, the set of words in the vocabulary corresponds to the set $\mathcal{L}$ of annotations in Eq. (3) under a multi-label setting. The first challenge is that these metrics do not consider the semantics of the words: *e.g.*, in the sentences *a woman is cooking* and *a woman is making dinner*, the tokens *cooking* and *making dinner* would be considered as different annotation l. The second challenge is they do not consider the context of each word/task: *e.g.*, the token *cooking* will be seen as the same task in the sentence *a man is cooking* and in *a man is not cooking*.

4. Bias amplification for image captioning

We propose a metric to specifically measure bias amplification in image captioning models, borrowing some ideas from sentence classification [46] and leakage [40]. Our metric, named LIC, is built on top of the following hypothesis:

> ***Hypothesis 1.*** In an unbiased set of captions, there should not exist differences between how demographic groups are represented.

Caption masking As discussed in Section 3, for some protected attributes (*e.g.*, age and gender), specific vocabulary may be explicitly used in the captions. For example, consider *gender* as a binary[2] protected attribute a with possible values $\{female, male\}$. The sentence

A girl is playing piano,

directly reveals the protected attribute value of the caption, *i.e.*, *female*. To avoid explicit mentions to the protected attribute value, we preprocess captions by masking the words related to that attribute.[3] The original sentence is then transformed to the masked sentence

A [redacted] is playing piano.

Note that this step is not always necessary, as some protected attributes are not explicitly revealed in the captions.

Caption classification We rely on a sentence classifier f_s to estimate societal bias in captions. Specifically, we encode each masked caption y' [4] with a natural language encoder E to obtain a sentence embedding e, as $e = E(y')$. Then, we input e into the sentence classifier f_s, whose aim is to predict the protected attribute a from y' as

$$\hat{a} = f_s(E(y')) \tag{7}$$

E and f_s are learned on a training split $\mathcal{D}$. According to *Hypothesis 1*, in an unbiased dataset, the classifier f_s should not find enough clues in y' to predict the correct attribute a. Thus, $\mathcal{D}$ is considered to be biased if the empirical probability $p(\hat{a} = a)$ over $\mathcal{D}$ is greater than the chance rate.

Bias amplification Bias amplification is defined as the bias introduced by a model with respect to the existing bias in the training set. To measure bias amplification, we quantify the difference between the bias in the generated captions set $\hat{\mathcal{D}} = \{\hat{y} = M(I) \mid I \in \mathcal{D}\}$ with respect to the bias in the original captions in the training split $\mathcal{D}$.

One concern with this definition, particular to image captioning, is the difference in the vocabularies used in the annotations and in the predictions, due to 1) the human generated captions typically come with a richer vocabulary, 2) a model's vocabulary is rather limited, and 3) the vocabulary itself can be biased. Thus, naively applying Eq. (4) to image captioning can underestimate bias amplification. To mitigate this problem, we introduce noise into the original human captions until the vocabularies in the two sets (model generated and human generated) are aligned. Formally, let $\mathcal{V}_{\text{ann}}$ and $\mathcal{V}_{\text{pre}}$ denote the vocabularies identified for all annotations and predictions in the training set, respectively. For the annotation $y = (y_1, \ldots, y_N)$, where y_n is the n-th word in y, we replace all y_n in $\mathcal{V}_{\text{ann}}$ but not in $\mathcal{V}_{\text{pre}}$ with a special out-of-vocabulary token to obtain perturbed annotations $y^\star$, and we train the classifier $f_s^\star$ over $\{y^\star\}$.

The LIC metric The confidence score $s_a^\star$ is an intermediate result of classifier $f_s^\star$, *i.e.*,

$$\hat{a} = f_s^\star(y^\star) = \text{argmax}_a s_a^\star(y^\star), \tag{8}$$

and it can be interpreted as a posterior probability $p(\hat{a} = a \mid y^\star)$ of the protected attribute a and can give an extra hint on how much $y^\star$ is biased toward a. In order words, not only the successful prediction rate is important to determine the bias, but also how confident the predictions are. The same applies to $\hat{s}_a$ and $\hat{f}_s$ trained with $\{\hat{y}\}$. We incorporate this

[2] Due to the availability of annotations in previous work, in this paper we use the binary simplification of gender, acknowledging that it is not inclusive and should be addressed in future work.

[3] A list of attribute-related words is needed for each protected attribute.

[4] If caption masking is not applied, $y' = y$.

information into the Leakage for Image Captioning metric (LIC), through

$$\text{LIC}_D = \frac{1}{|\mathcal{D}|} \sum_{(y^\star, a) \in \mathcal{D}} s_a^\star(y^\star) \mathbb{1}[f^\star(y^\star) = a] \quad (9)$$

$$\text{LIC}_M = \frac{1}{|\hat{\mathcal{D}}|} \sum_{(\hat{y}, a) \in \hat{\mathcal{D}}} \hat{s}_a(\hat{y}) \mathbb{1}[\hat{f}(\hat{y}) = a], \quad (10)$$

so that LIC is finally computed as

$$\text{LIC} = \text{LIC}_M - \text{LIC}_D. \quad (11)$$

where a model is considered to amplify bias if $\text{LIC} > 0$. We refer to $\hat{s}_a$ as the *bias score*.

5. Experiments

Data Experiments are conducted on a subset of the MSCOCO captions dataset [9]. Specifically, we use the images with binary gender and race annotations from [46]: *female* and *male* for gender, *darker* and *lighter* skin tone for race.[5] Annotations are available for images in the validation set with person instances, with a total of $10,780$ images for gender and $10,969$ for race. To train the classifiers, we use a balanced split with equal number of images per protected attribute value, resulting in $5,966$ for training and 662 for test in gender, and $1,972$ for training and 220 for test in race. Other metrics are reported on the MSCOCO val set.

Metrics We report bias using LIC, together with LIC_D in Eq. (9) and LIC_M in Eq. (10). For gender bias, we also use Ratio [8], Error [8], Bias Amplification (BA) [47], and Directional Bias Amplification [38]. Directional bias amplification is computed for object $\rightarrow$ gender direction (DBA_G) and for gender $\rightarrow$ object direction (DBA_O) using MSCOCO objects [25]. For skin tone, we only use LIC, LIC_D, and LIC_M, as there are no words we can directly associated with race in the captions to calculate the other metrics. Accuracy is reported in terms of standard metrics BLEU-4 [26], CIDEr [35], METEOR [13], and ROUGE-L [24].

Models We study bias on captions generated by the following models: NIC [36], SAT [43], FC [28], Att2in [28], UpDn [2], Transformer [34], OSCAR [23], NIC+ [8], and NIC+Equalizer [8]. NIC, SAT, FC, Att2in, and UpDn are classical CNN [22] encoder-LSTM [18] decoder models. Transformer and OSCAR are Transformer-based [34] models, which are the current state-of-the-art in image captioning. NIC+ is a re-implementation of NIC in [8] trained on the whole MSCOCO and additionally trained on MSCOCO-Bias set consisting of images of male/female. NIC+Equalizer is NIC+ with a gender bias mitigation loss, that forces the model to predict gender words only based on the region of the person. Note that most of the pre-trained captioning models provided by the authors are trained on the Karpathy split [20], which uses both train and validation splits for training. As the val set is part of our evaluation, we retrain all the models on the MSCOCO train split only.

[5] Similarly, due to the availability of annotations in previous work, we use a binary simplification for race and skin tone. We acknowledge that these attributes are much more complex in reality.

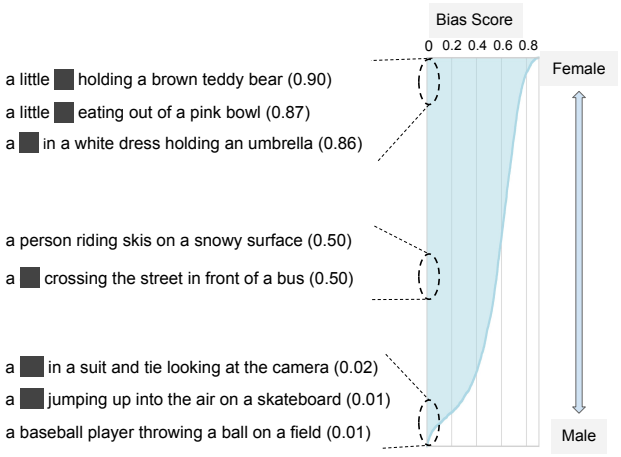

Figure 2. Gender bias score for captions generated with OSCAR. Masked captions are encoded with a LSTM and fed into a gender classifier. Bias score correlates with typical gender stereotypes.

LIC metric details For masking, we replace pre-defined gender-related words[6] with a special token `<gender>`. We do not mask any words for race prediction because race is not commonly explicitly mentioned in the captions.

The LIC classifier is based on several fully-connected layers on top of a natural language encoder. For the encoder, we use a LSTM [18] for our main results. We do not initialize the LSTM with pre-computed word embeddings, as they contain bias [6, 14]. For completeness, we also report LIC when using BERT [15], although it has also been shown to exhibit bias [3, 4] and it can affect our metric. BERT is fine-tuned (BERT-ft) or used as is (BERT-pre). The classifier is trained 10 times with random initializations, and the results are reported by the average and standard deviation.

5.1. LIC analysis

We qualitatively analyze the LIC metric to verify whether it is consistent with human intuition. We generate captions in the test set with OSCAR, mask the gender-related words, and encode the masked captions with a LSTM classifier to compute LIC bias score, $\hat{s}_a$, for the gender attribute, as formulated in Section 4. Then, we manually inspect the captions and their associated bias score.

Figure 2 shows generated captions with higher, middle, and lower bias scores. The bias score assigned to each caption matches typical gender stereotypes. For example, the

[6] The list of gender-related words can be found in the appendix.

Table 1. Gender bias and accuracy for several image captioning models. Red/green denotes the worst/best score for each metric. For bias, lower is better. For accuracy, higher is better. BA, DBA_G, and DBA_O are scaled by 100. Unbiased model is $\text{LIC}_M = 25$ and $\text{LIC} = 0$.

	Gender bias ↓							Accuracy ↑			
Model	LIC	LIC_M	Ratio	Error	BA	DBA_G	DBA_O	BLEU-4	CIDEr	METEOR	ROUGE-L
NIC [36]	**3.7**	**43.2**	2.47	**14.3**	4.25	3.05	**0.09**	**21.3**	**64.8**	**20.7**	**46.6**
SAT [43]	5.1	44.4	2.06	7.3	1.14	3.53	0.15	32.6	98.3	25.8	54.1
FC [28]	8.6	46.4	2.07	10.1	4.01	**3.85**	0.28	30.5	98.0	24.7	53.5
Att2in [28]	7.6	45.9	2.06	4.1	**0.32**	3.60	**0.29**	33.2	105.0	26.1	55.6
UpDn [2]	9.0	48.0	2.15	3.7	2.78	3.61	0.28	36.5	117.0	27.7	57.5
Transformer [34]	8.7	48.4	2.18	3.6	1.22	3.25	0.12	32.3	105.3	27.0	55.1
OSCAR [23]	9.2	48.5	2.06	**1.4**	1.52	3.18	0.19	**40.4**	**134.0**	**29.5**	**59.5**
NIC+ [8]	7.2	46.7	**2.89**	12.9	**6.07**	**2.08**	0.17	27.4	84.4	23.6	50.3
NIC+Equalizer [8]	**11.8**	**51.3**	**1.91**	7.7	5.08	3.05	0.20	27.4	83.0	23.4	50.2

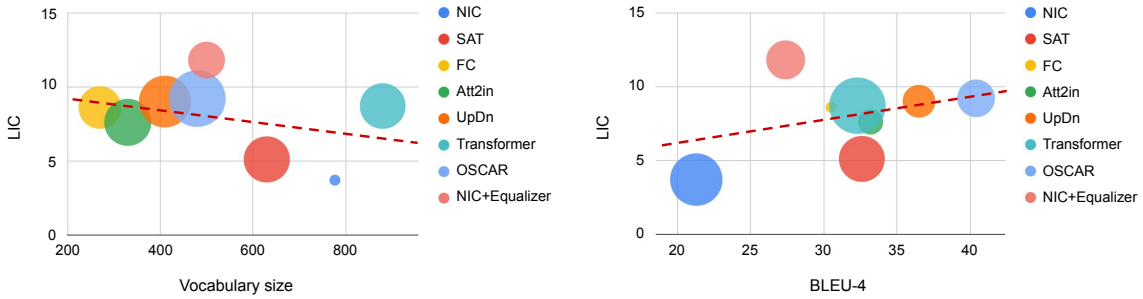

Figure 3. LIC vs. Vocabulary size (left) and BLEU-4 score (right). The size of each bubble indicates the BLEU-4 score (left) or the vocabulary size (right). Score tends to decrease with largest vocabularies, but increase with more accurate BLEU-4 models, whereas NIC+Equalizer [8] is presented as an outlier. The dotted lines indicate the tendency, $R^2 = 0.153$ (left) and $R^2 = 0.156$ (right).

third caption from the top, "a [redacted] in a white dress holding an umbrella", yields a very high bias score for *female*, probably due to the stereotype that the people who wear dresses and holds umbrellas tend to be women. On the contrary, the bottom caption, "a baseball player throwing a ball on a field", with one of the lowest scores assigned to *female*, perpetuates the stereotype that baseball players are mostly men. Additionally, when inspecting the captions with a bias score around 0.5, we see that the descriptions tend to be more neutral and without strong gender stereotypes. This support the importance of including $s_a^\star$ and $\hat{s}_a$ in the LIC computation, as in Eqs. (9) and (10).

5.2. Quantification of gender bias

We evaluate the gender bias of different captioning models in terms of LIC together with adaptations of existing bias metrics. For BA, we use the top $1,000$ common words in the captions as $\mathcal{L}$, whereas for DBA_G and DBA_O, we use MSCOCO objects [25]. More details can be found in the appendix. Results are shown in Table 1. We also show the relationship between the quality of a caption, in terms of vocabulary and BLEU-4 score, with LIC in Figure 3. Finally, we compare LIC when using different language encoders in Table 2. The main observations are summarized below.

Observation 1.1. **All the models amplify gender bias.** In Table 1, all the models have a LIC_M score well over the unbiased model ($\text{LIC}_M = 25$), with the lowest score being 43.2 for NIC. When looking at LIC, which indicates how much bias is introduced by the model with respect to the human captions, also all the models exhibit bias amplification, again with NIC having the lowest score. NIC is also the model that performs the worst in terms of accuracy, which provides some hints about the relationship between accuracy and bias amplification (*Observation 1.4*).

Observation 1.2. **Bias metrics are not consistent.** As analyzed in Section 3, different metrics measure different aspects of the bias, so it is expected to produce different results, which may lead to different conclusions. Nevertheless, all the models show bias in all the metrics except Ratio (Table 1). However, the relationship between the bias and the models presents different tendencies. For instance, NIC+Equalizer shows the largest bias in LIC (*Observation 1.3*) while Att2in has the largest bias in DBA_O.

Observation 1.3. **NIC+Equalizer increases gender bias with respect to the baseline.** One of the most surprising findings is that even though NIC+Equalizer success-

Table 2. Gender bias scores according to LIC, LIC_M, and LIC_D for several image captioning models. Captions are encoder with LSTM, BERT-ft, or BERT-pre. Unbiased model is $\text{LIC}_M = 25$ and $\text{LIC} = 0$. It shows that LIC is consistent across different language models.

	LSTM			BERT-ft			BERT-pre		
Model	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC
NIC [36]	43.2 ± 1.5	39.5 ± 0.9	**3.7**	47.2 ± 2.3	48.0 ± 1.2	**-0.8**	43.2 ± 1.3	41.3 ± 0.9	**1.9**
SAT [43]	44.4 ± 1.4	39.3 ± 1.0	5.1	48.0 ± 1.1	47.7 ± 1.4	0.3	44.4 ± 1.5	41.5 ± 0.8	2.9
FC [28]	46.4 ± 1.2	37.8 ± 0.9	8.6	48.7 ± 1.9	45.8 ± 1.3	2.9	46.8 ± 1.4	40.4 ± 0.8	6.4
Att2in [28]	45.9 ± 1.1	38.3 ± 1.0	7.6	47.8 ± 2.0	46.7 ± 1.4	1.1	45.9 ± 1.2	40.9 ± 0.9	5.0
UpDn [2]	48.0 ± 1.3	39.0 ± 0.9	9.0	52.0 ± 1.0	47.3 ± 1.4	4.7	48.5 ± 1.0	41.5 ± 0.9	7.0
Transformer [34]	48.4 ± 0.8	39.7 ± 0.9	8.7	54.1 ± 1.2	48.2 ± 1.1	5.9	47.7 ± 1.2	42.2 ± 0.9	5.5
OSCAR [23]	48.5 ± 1.5	39.3 ± 0.8	9.2	52.5 ± 1.8	47.6 ± 1.2	4.9	48.1 ± 1.1	41.1 ± 0.9	7.0
NIC+ [8]	46.7 ± 1.2	39.5 ± 0.6	7.2	49.5 ± 1.4	47.7 ± 1.5	1.8	46.4 ± 1.2	41.0 ± 0.9	5.4
NIC+Equalizer [8]	51.3 ± 0.7	39.5 ± 0.9	**11.8**	54.8 ± 1.1	47.5 ± 1.4	**7.3**	49.5 ± 0.7	40.9 ± 0.9	**8.6**

fully mitigates gender misclassification when compared to the baseline NIC+ (Error: $12.9 \rightarrow 7.7$ in Table 1), it actually increases gender amplification bias by $+4.6$ in LIC. This unwanted side-effect may be produced by the efforts of predicting gender correctly according to the image. As shown in Figure 1, NIC+Equalizer tends to produce words that, conversely, are strongly associated with that gender, even if they are not in the image. Results on DBA_O support this reasoning, revealing that given a gender, NIC+Equalizer rather produces words correlated with that gender.

Observation 1.4. **LIC tends to increase with BLEU-4, and decrease with vocabulary size.** Figure 3 shows that larger the vocabulary, the lower the LIC score. This implies that the variety of the words used in the captions is important to suppress gender bias. As per accuracy, we find that the higher the BLEU-4, the larger the bias tends to be. In other words, even though better models produce better captions, they rely on encoded clues that can identify gender.

Observation 1.5. **LIC is robust against encoders.** In Table 2, we explore how the selection of language models affects the results of LIC, LIC_M, and LIC_D when using LSTM, BERT-ft, and BERT-pre encoders. Although BERT is known to contain gender bias itself, the tendency is maintained within the three language models: NIC shows the least bias, whereas NIC+Equalizer shows the most.

5.3. Quantification of racial bias

Results for racial bias when using LSTM as encoder are reported in Table 3, leading to the following observations.

Observation 2.1. **All the models amplify racial bias.** As with gender, all models exhibits $\text{LIC} > 0$. The magnitude difference of racial bias between the models is smaller than in the case of gender (the standard deviation of LIC among the models is 2.4 for gender and 1.3 for race). This indicates that racial bias is amplified without much regard to the structure or performance of the model. In other words, as all the models exhibit similar tendencies of bias amplification, the problem may not only be on the model structure itself but on how image captioning models are trained.

Table 3. Racial bias scores according to LIC, LIC_M, and LIC_D. Captions are not masked and are encoder with LSTM.

Model	LIC_M	LIC_D	LIC
NIC [36]	33.3 ± 1.9	27.6 ± 1.0	5.7
SAT [43]	31.3 ± 2.3	26.8 ± 0.9	**4.5**
FC [28]	33.6 ± 1.0	26.0 ± 0.8	7.6
Att2in [28]	35.2 ± 2.3	26.6 ± 0.9	**8.6**
UpDn [2]	34.4 ± 2.1	26.6 ± 0.9	7.8
Transformer [34]	33.3 ± 2.3	27.2 ± 0.8	6.1
OSCAR [23]	32.9 ± 1.8	27.0 ± 1.0	5.9
NIC+ [8]	34.9 ± 1.5	27.3 ± 1.2	7.6
NIC+Equalizer [8]	34.5 ± 2.8	27.3 ± 0.8	7.2

Observation 2.2. **Racial bias is not as apparent as gender bias.** LIC_M scores in Table 3 are consistently smaller than in Table 2. The mean of the LIC_M score of all the models is 47.0 for gender and 33.7 for race, which is closer to the random chance.

Observation 2.3. **NIC+Equalizer does not increase racial bias with respect to the baseline.** Unlike for gender bias, NIC+Equalizer does not present more racial bias amplification than NIC+. This indicates that forcing the model to focus on the human area to predict the correct gender does not negatively affect other protected attributes.

5.4. Visual and language contribution to the bias

As image captioning is a multimodal task involving visual and language information, bias can be introduced by the image, the language, or both. Next, we investigate which modality contributes the most to gender bias by analyzing the behavior when using partially masked images.

We define three potential sources of bias: 1) the objects

Table 4. Gender bias results with partially masked images. Δ_{Unbias} shows the difference with respect to a non-biased model ($\text{LIC}_M = 25.0$), and Δ_{Original} with respect to the non-masked case.

Model	Image	LIC_M	Δ_{Unbias}	Δ_{Original}
SAT [43]	Original	44.4 ± 1.4	$+19.4$	0.0
	w/o object	42.9 ± 1.6	$+17.9$	-1.5
	w/o person	39.1 ± 1.4	$+14.1$	-5.3
	w/o both	37.2 ± 0.8	$+12.2$	-7.2
OSCAR [23]	Original	48.5 ± 1.5	$+23.2$	0.0
	w/o object	46.2 ± 1.3	$+21.2$	-2.3
	w/o person	39.7 ± 1.3	$+14.7$	-8.8
	w/o both	39.0 ± 1.5	$+14.0$	-9.5

being correlated with the gender [38, 40, 47], 2) the gender of the person in the image [8], and 3) the language model itself [3, 6]. To examine them, we mask different parts of the image accordingly: 1) the object that exhibits the highest correlation with gender according to the BA metric, 2) the person, 3) both of the correlated objects and the person. We analyze SAT [43] and OSCAR [23] as representative models of classical and state-of-the-art captioning, respectively. The details of the experiments can be found in the appendix. LIC_M scores are shown in Table 4.

Observation 3.1. **The contribution of objects to gender bias is minimal.** Results *w/o object* show that masking objects do not considerably mitigate gender bias in the generated captions. Compared to the original LIC_M, the score decreases only -1.5 for SAT, and -2.3 for OSCAR, concluding that objects in the image have little impact to the gender bias in the final caption.

Observation 3.2. **The contribution of people to gender bias is higher than objects.** Results *w/o person* show that by masking people in the images, we can reduce bias significantly compared to when hiding objects, indicating that regions associated with humans are the primary source of gender bias among the contents in the image.

Observation 3.3. **Language models are a major source of gender bias.** Results *w/o both* show that even when the gender-correlated objects and people are removed from the images, the generated captions have a large bias (Δ_{Unbias} is $+12.2$ for SAT, $+14.0$ for OSCAR). This indicates that the language model itself is producing a large portion of the bias. To reduce it, it may not be enough to only focus on the visual content, but efforts should also be focused on the language model. Figure 4 shows the generated captions and their bias score when images are partly masked.

6. Limitations

In Section 3, we analyzed multiple fairness metrics and their limitations when applied to image captioning. We proposed LIC with the aim to overcome these limitations and unify the evaluation of societal bias in image captioning. However, LIC also presents several limitations.

Figure 4. Generated captions and bias scores when images are partly masked. The bias score does not decrease when the object (bicycle) and the person (man) are masked.

Annotations LIC needs images to be annotated with their protected attribute. Annotations are not only costly, but may also be problematic. For example, the classification of race is controversial and strongly associated with the cultural values of each annotator [21], whereas gender is commonly classified as a binary $\{female, male\}$ attribute, lacking inclusiveness with non-binary and other-gender realities.

Training A classifier needs to be trained to make predictions about the protected attributes. The initialization of the model and the amount of training data may impact on the final results. To mitigate this stochastic effect, we recommended to report results conducted on multiple runs.

Pre-existing bias The language encoder may propagate extra bias into the metric if using pretrained biased models, *e.g.*, word embeddings or BERT. To avoid this, we recommend as much random weight initialization as possible.

7. Conclusion

This paper proposed LIC, a metric to quantify societal bias amplification in image captioning. LIC is built on top of the idea that there should not be differences between how demographic subgroups are described in captions. The existence of a classifier that predicts gender and skin tone from generated captions more accurately than from human captions, indicated that image captioning models amplify gender and racial bias. Surprisingly, the gender equalizer designed for bias mitigation presented the highest gender bias amplification, highlighting the need of a bias amplification metric for image captioning.

Acknowledgements Work partly supported by JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR216O, and JSPS KAKENHI, Japan.

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCV Workshops*, 2018. 1

[2] 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 *CVPR*, 2018. 2, 5, 6, 7

[3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *ACM FAccT*, 2021. 5, 8

[4] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating gender bias in BERT. *Cognitive Computation*, 2021. 5

[5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1

[6] 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. *NeurIPS*, 2016. 5, 8

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM FAccT*, 2018. 1, 2

[8] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 1, 2, 3, 5, 6, 7, 8

[9] 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. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2, 5

[10] Kate Crawford and Trevor Paglen. Excavating AI: The politics of training sets for machine learning. https://excavating.ai, 2019. Accessed: 2021-11-12. 1

[11] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *CVPR Workshops*, 2019. 1

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[13] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Workshop on statistical machine translation*, 2014. 5

[14] Sunipa Dev and Jeff Phillips. Attenuating bias in word vectors. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2019. 5

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019. 5

[16] Catherine D'ignazio and Lauren F Klein. *Data feminism*. MIT press, 2020. 1

[17] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004. 2

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8), 1997. 5

[19] Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. Mitigating gender bias amplification in distribution by posterior regularization. *ACL*, 2020. 2

[20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 5

[21] Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *ACM FAccT*, 2021. 1, 8

[22] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553), 2015. 5

[23] 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 *ECCV*, 2020. 1, 5, 6, 7, 8

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004. 5

[25] 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 *ECCV*, 2014. 2, 3, 5, 6

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002. 5

[27] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020. 1

[28] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 5, 6, 7

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1

[30] Pierre Stock and Moustapha Cisse. ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In *ECCV*, 2018. 1

[31] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *WWW*, 2021. 2, 3

[32] William Thong and Cees GM Snoek. Feature and label embedding spaces matter in addressing image classifier bias. *arXiv preprint arXiv:2110.14336*, 2021. 1, 2

[33] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 1, 2

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 5, 6, 7

[35] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 5

[36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2, 5, 6, 7

[37] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. In *ECCV*, 2020. 1, 2

[38] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *ICML*, 2021. 2, 3, 5, 8

[39] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, 2019. 2

[40] 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 *ICCV*, 2019. 1, 2, 3, 4, 8

[41] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020. 2

[42] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019. 1

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2, 5, 6, 7, 8

[44] 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 *ACM FAccT*, 2020. 1

[45] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016. 2

[46] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021. 2, 3, 4, 5

[47] 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 *EMNLP*, 2017. 1, 2, 3, 5, 8