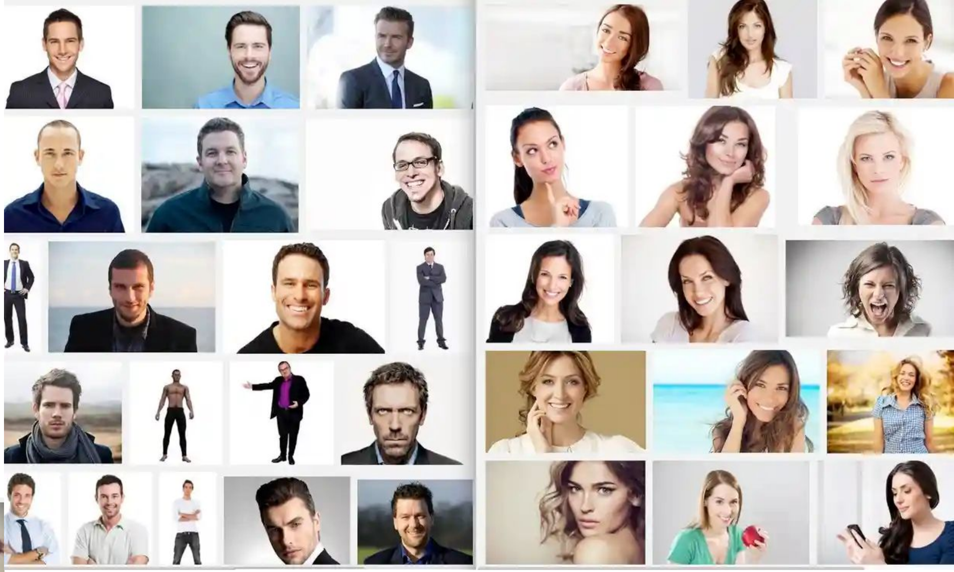
Bias and fairness

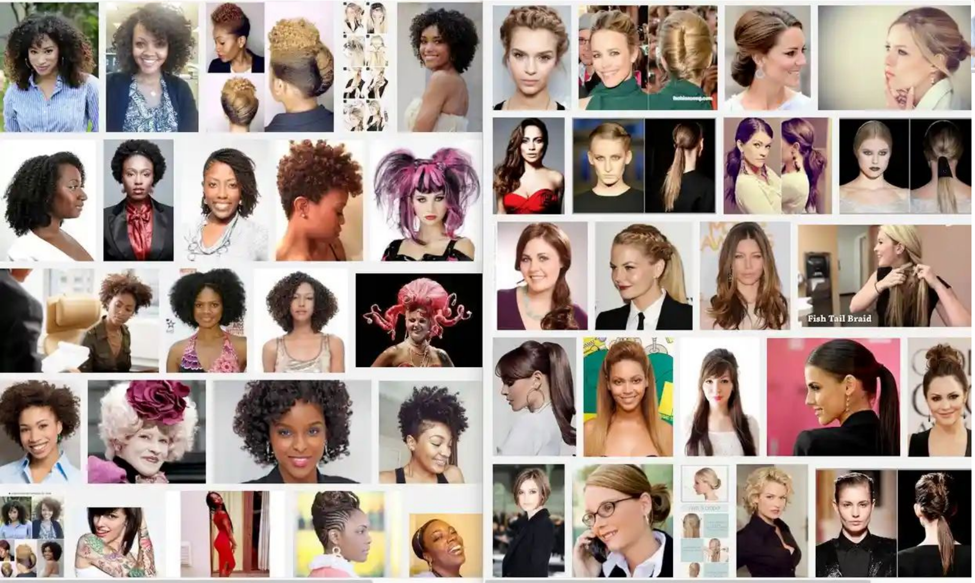
Zoe Zheng

Quantifying societal bias amplification in image captioning

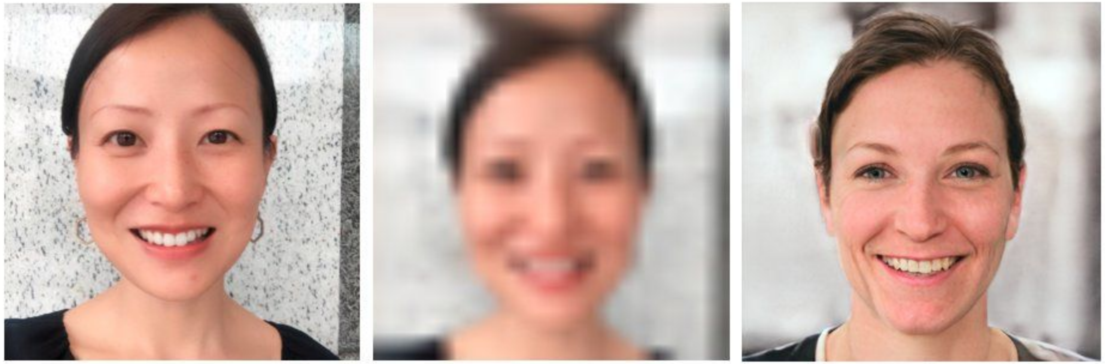
Women also snowboard: Overcoming bias in captioning models

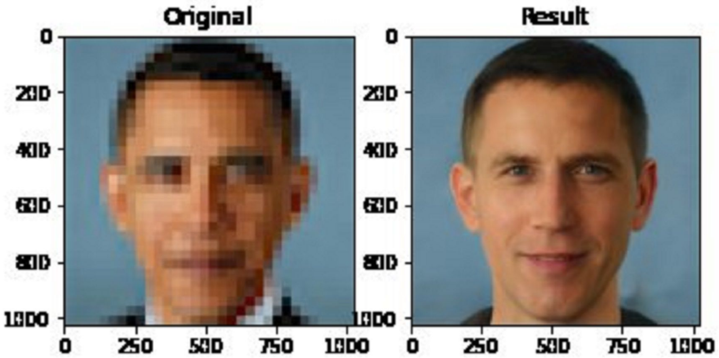
Man and Woman?

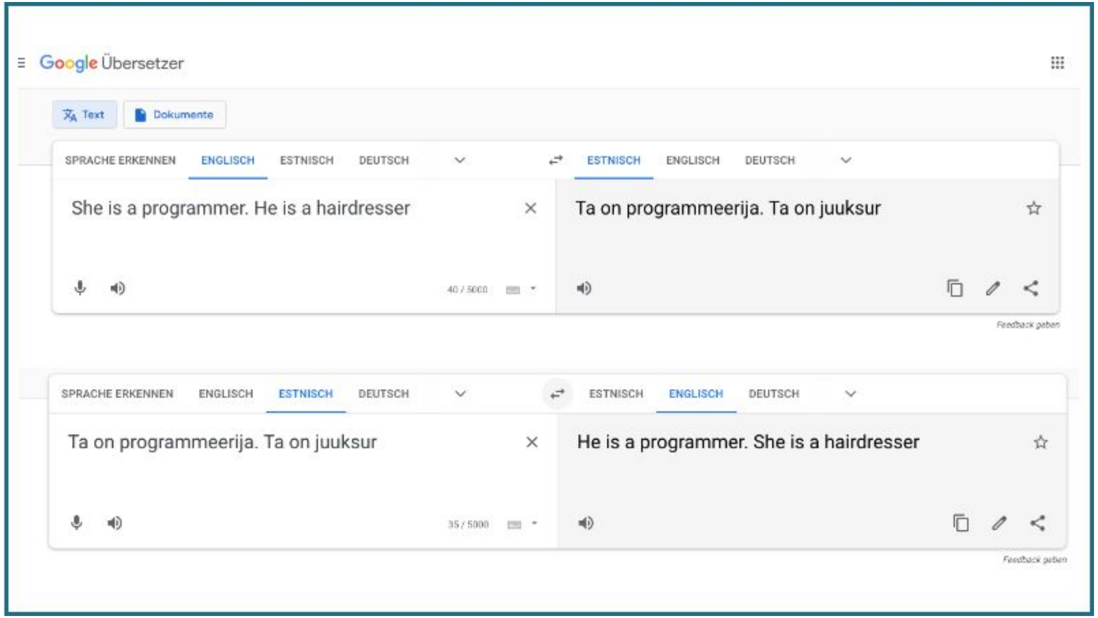




Professional and unprofessional hairstyle?







She is a programmer. He is a hairdresser

Ta on programmeerija. Ta on juuksur

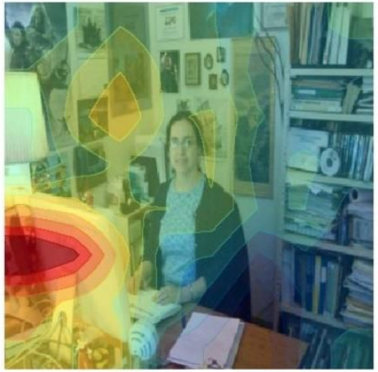
Ta on programmeerija. Ta on juuksur

He is a programmer. She is a hairdresser

Bias in CV and NLP

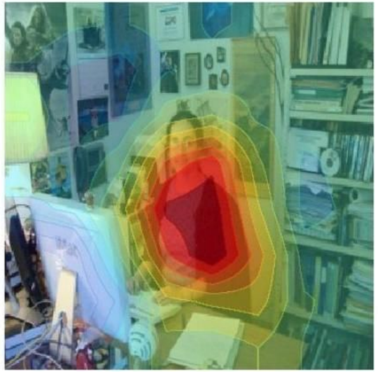
Women Also Snowboard: Overcoming Bias in Captioning Models

Wrong



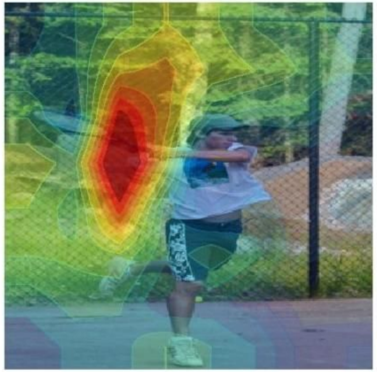
Baseline:
*A **man** sitting at a desk with a laptop computer.*

Right for the Right Reasons



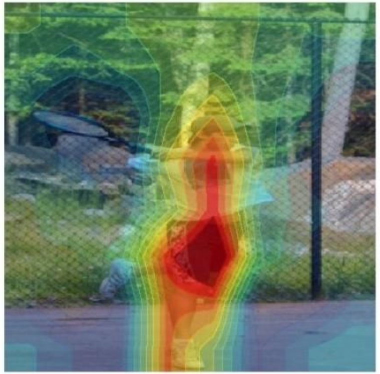
Our Model:
*A **woman** sitting in front of a laptop computer.*

Right for the Wrong Reasons



Baseline:
*A **man** holding a tennis racquet on a tennis court.*

Right for the Right Reasons



Our Model:
*A **man** holding a tennis racquet on a tennis court.*

Women Also Snowboard: Overcoming Bias in Captioning Models

- Equalizer forces models to look at a person rather than use contextual cues to make a gender-specific prediction.
- Appearance Confusion Loss and the Confident Loss

Appearance Confusion Loss

- encourages the underlying description model to be confused when making gender decisions if the input image does not contain appropriate evidence for the decision
- The Hadamard product of the mask and the original image, I ⊙ M , yields a new image, I′, with gender information that the implementer deems appropriate for classification removed

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

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

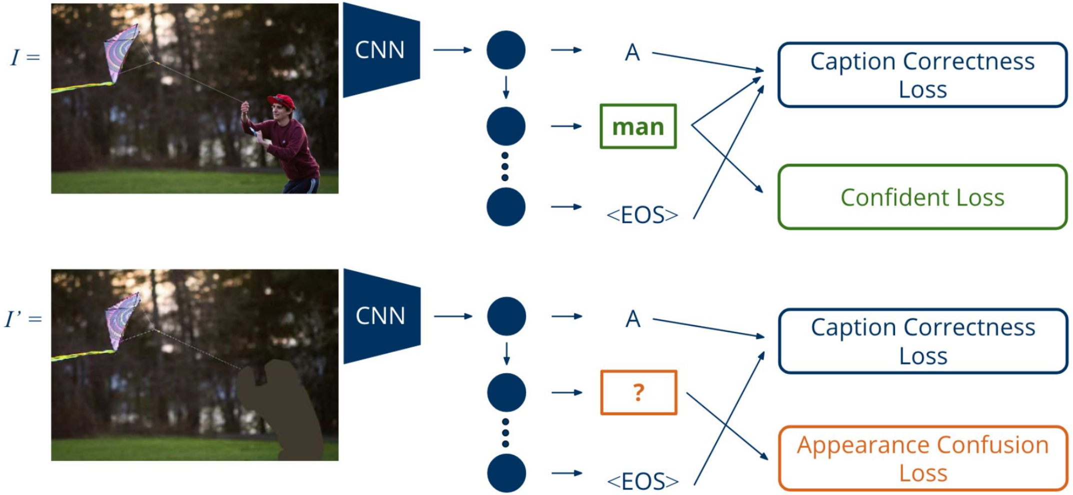
Confident Loss

- encourage our model to be confident when gender evidence is present
- When the model is confident of a gender prediction (e.g., for the word “woman”), the probability of the word “woman” should be considerably higher than the probability of the word “man”, which will result in a small value for FW and thus a small loss.

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

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

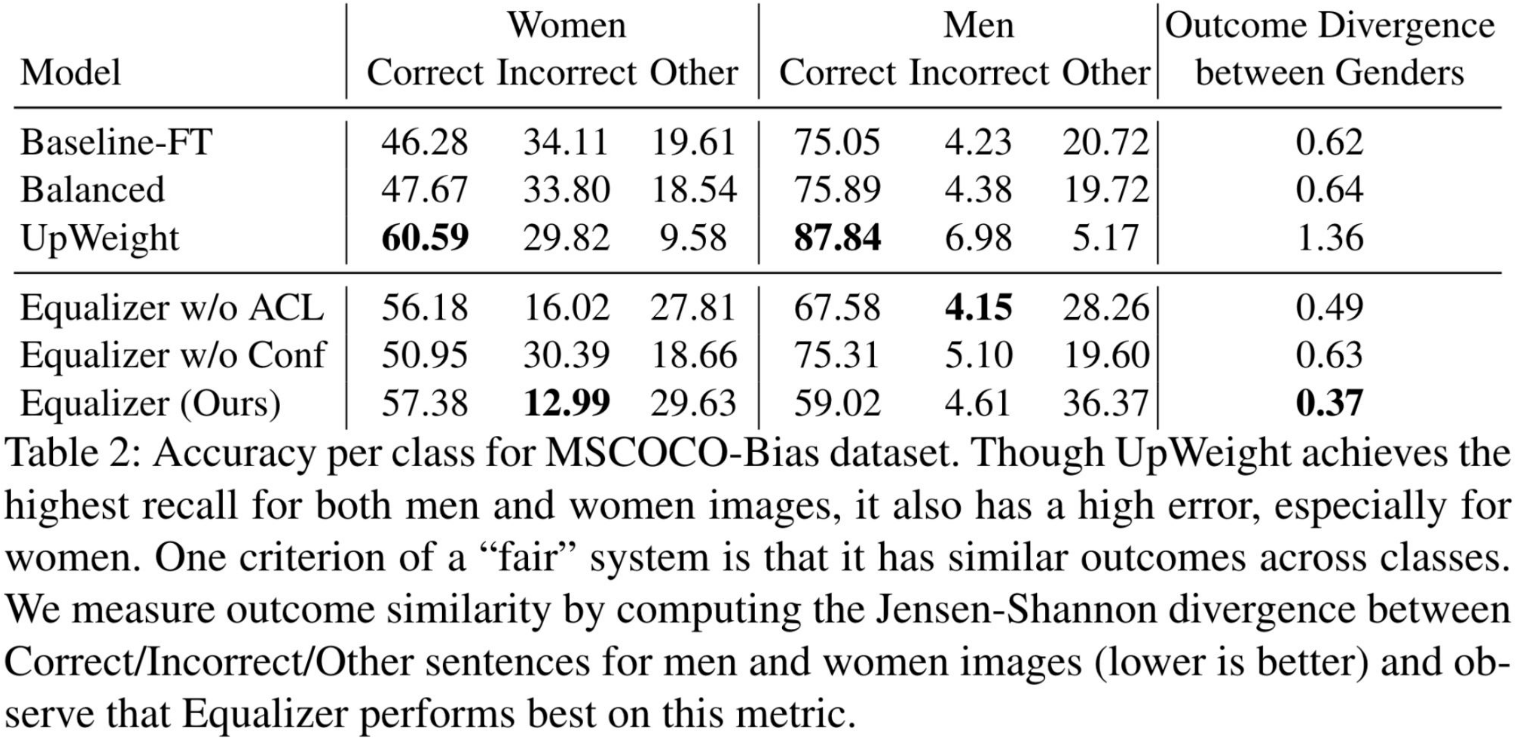
pipeline: input image → regions of interest → object classification (for each region) → captioning based on objects found



Caption Correctness Loss (cross entropy loss)

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

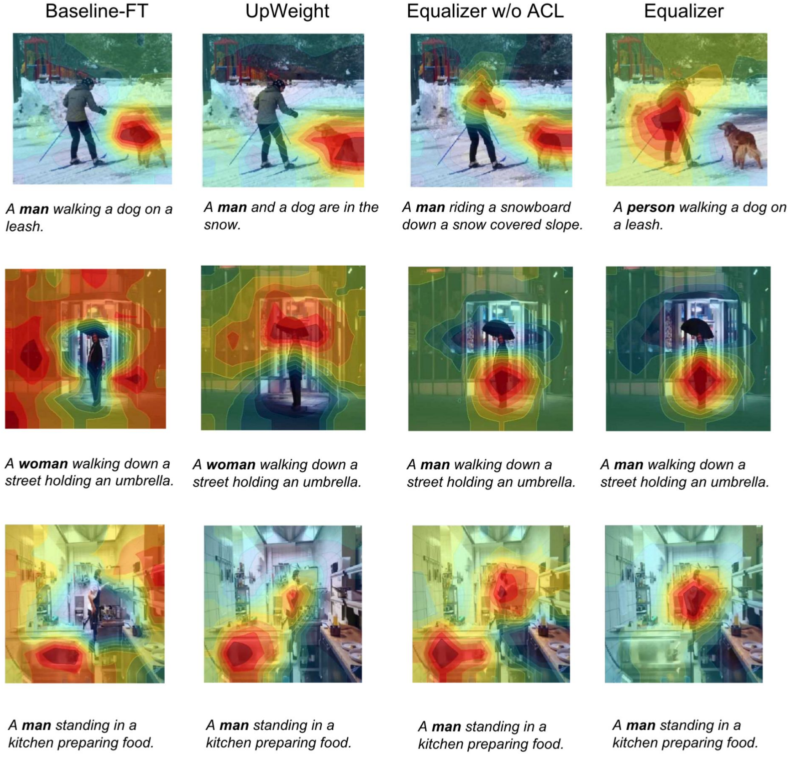
Results



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

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

What are the limitations of this paper?



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

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

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

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

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

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

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

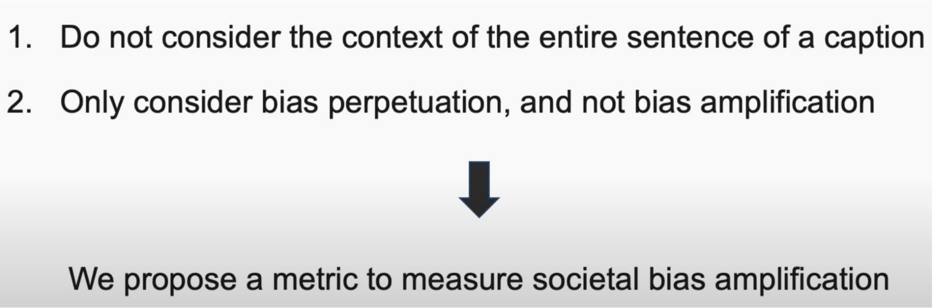
*A **man** walking down a street holding an umbrella.*

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

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

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

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

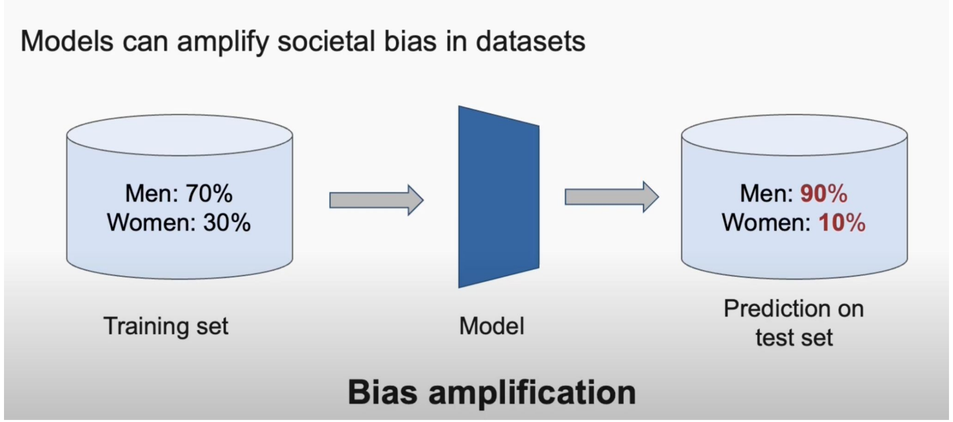


1. Do not consider the context of the entire sentence of a caption
2. Only consider bias perpetuation, and not bias amplification

⬇

We propose a metric to measure societal bias amplification

Related Bias Study



Models can amplify societal bias in datasets

Bias amplification

Even on balanced datasets, models still perpetuate bias:

— indicating that social stereotypes are occurring at the deepest levels of the image.

Previous fairness metrics in image captioning

Difference in performance

Evaluate the bias based on the difference in performance between the subgroups of a protected attribute, in terms of accuracy, ratio, or sentiment analysis.

Error: the number of demographic groups misclassifications, while neutral terms are not considered errors.

Ratio: the ratio of sentences which belong to a demographic group vs. others.

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

This check if the protected attribute has been correctly predicted in the generated caption (assumes that the attribute can be clearly identified in a sentence)

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

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

This measures whether the attention activation maps when generating a protected attribute word w in the caption

Shortcomings:

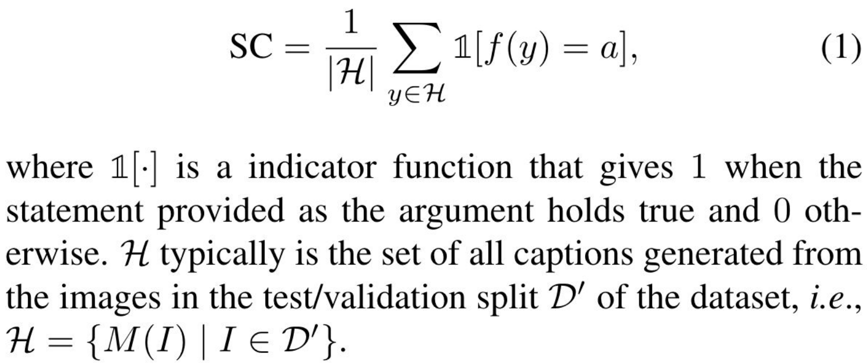
1) it needs a shortlist of protected attribute words, and a person segmentation map per image, which may not always be available

2) it assumes that visual explanations can be generated from the model, which may not always be the case

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

The reasoning is that if a classifier can distinguish between subgroups in the captions, the captions contain bias.



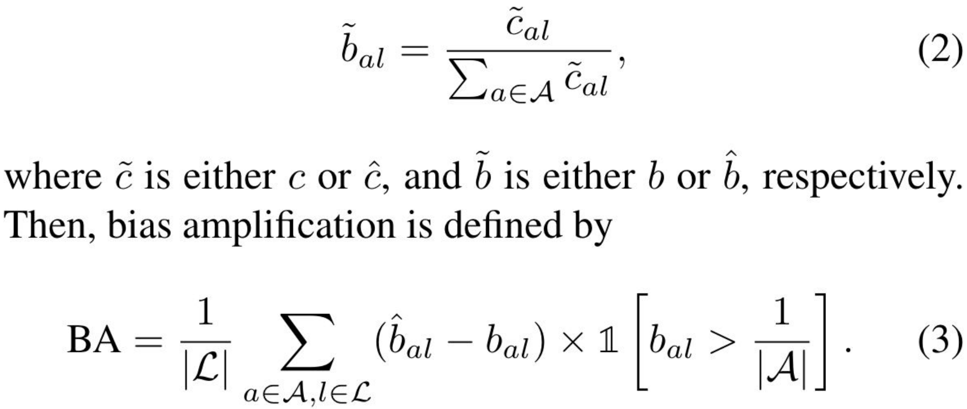
$$\mathrm{SC} = \frac{1}{|\mathcal{H}|} \sum_{y \in \mathcal{H}} \mathbb{1}[f(y) = a], \tag{1}$$

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.*, $\mathcal{H} = \{M(I) \mid I \in \mathcal{D}'\}$.

Shortcoming:
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.

Previous bias amplification metrics

Bias amplification



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

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

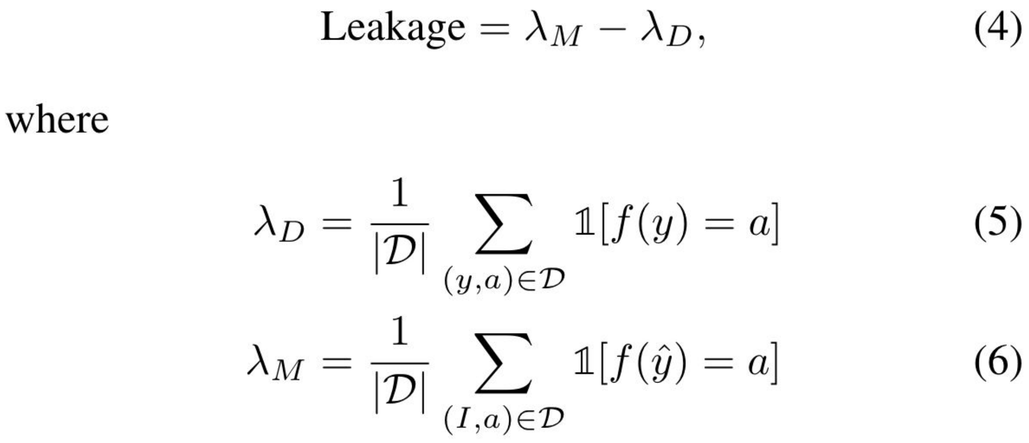
$$\mathrm{BA} = \frac{1}{|\mathcal{L}|} \sum_{a \in \mathcal{A}, l \in \mathcal{L}} (\hat{b}_{al} - b_{al}) \times \mathbb{1}\left[b_{al} > \frac{1}{|\mathcal{A}|} \right]. \tag{3}$$

Shortcomings:

- it ignores that protected attributes may be imbalanced in the dataset, e.g., in MSCOCO images, there are 2.25 more men than women, which causes most of objects to be correlated with men.

Leakage

It relies on the existence of a classifier to predict the protected attribute *a*



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

where

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

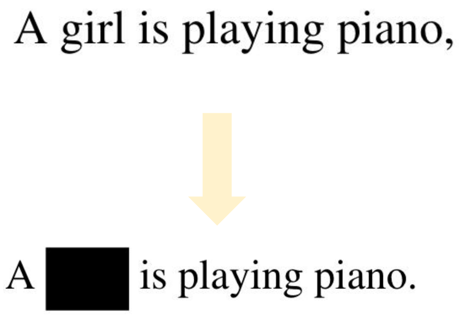
$$\lambda_M = \frac{1}{|\mathcal{D}|} \sum_{(I,a)\in\mathcal{D}} \mathbb{1}[f(\hat{y}) = a] \tag{6}$$

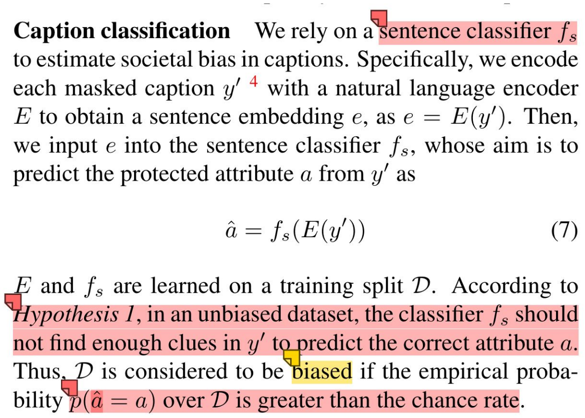
A positive leakage indicates that M amplifies the bias with respect to the training data, and mitigates it otherwise.

LIC

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

The authors preprocess captions by masking the words related to that attribute.





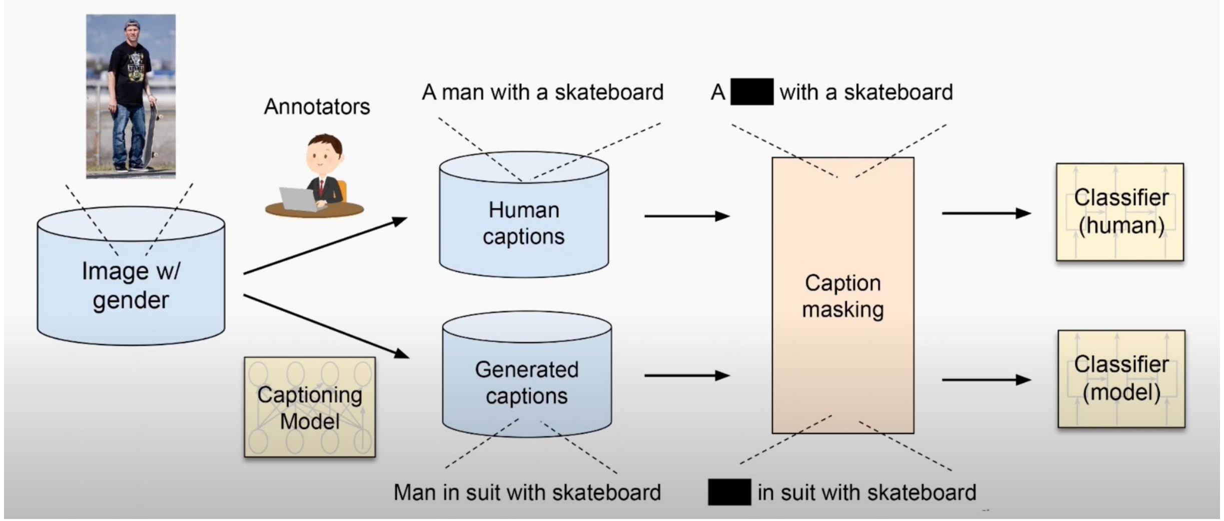
Caption classification We rely on a sentence classifier f_s to estimate societal bias in captions. Specifically, we encode each masked caption y' [4] with a natural language encoder E to obtain a sentence embedding e, as $e = E(y')$. Then, we input e into the sentence classifier f_s, whose aim is to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

To measure bias amplification:

Quantify the difference between the bias in the generated captions set (Model) with respect to the bias in the original captions (Human) in the training split D.

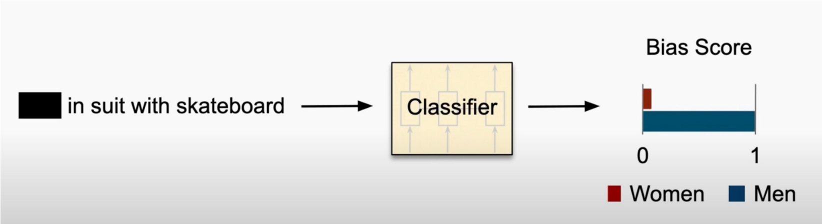


$$\mathrm{LIC}_D = \frac{1}{|\mathcal{D}|} \sum_{(y^\star, a) \in \mathcal{D}} s^\star_a(y^\star) \mathbb{1}[f^\star(y^\star) = a] \tag{9}$$

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

so that LIC is finally computed as

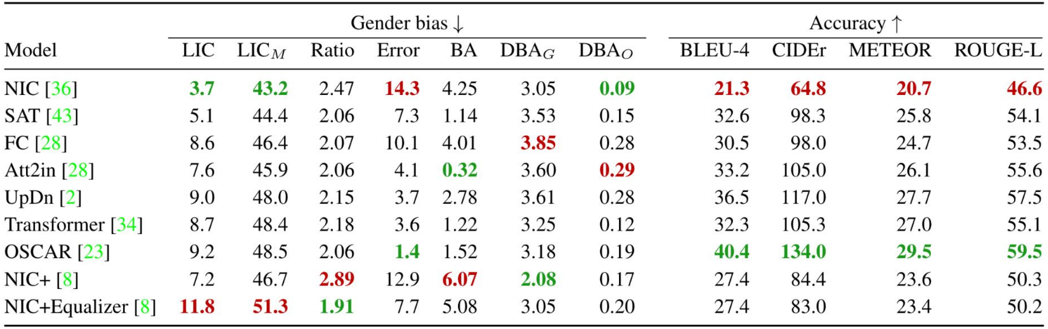
$$\mathrm{LIC} = \mathrm{LIC}_M - \mathrm{LIC}_D. \tag{11}$$



$$\hat{a} = f^\star_s(y^\star) = \mathrm{argmax}_a s^\star_a(y^\star), \tag{8}$$

Quantification of gender bias

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

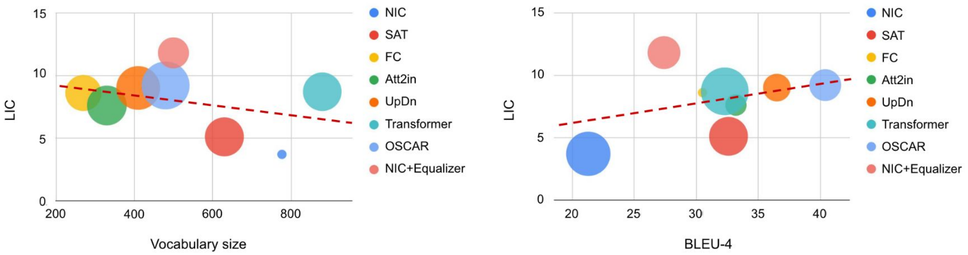


Figure 3. LIC vs. Vocabulary size (left) and BLEU-4 score (right). The size of each bubble indicates the BLEU-4 score (left) or the vocabulary size (right). Score tends to decrease with largest vocabularies, but increase with more accurate BLEU-4 models, whereas NIC+Equalizer [8] is presented as an outlier.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

All the models amplify gender bias. In Table 1, all the models have a LICM score well over the unbiased model (LICM = 25), with the lowest score being 43.2 for NIC.

Bias metrics are not consistent. LIC tends to increase with BLEU-4, and decrease with vocabulary size.

Quantification of gender bias

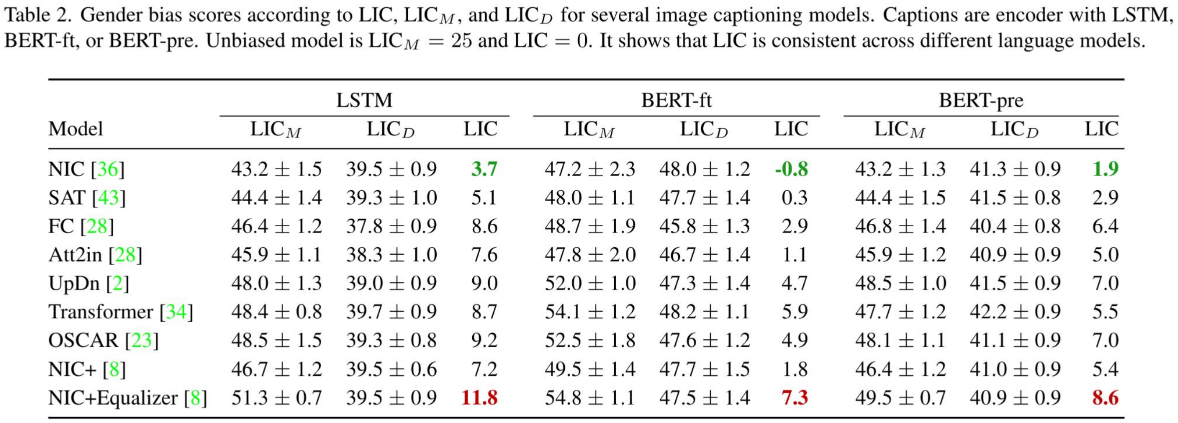
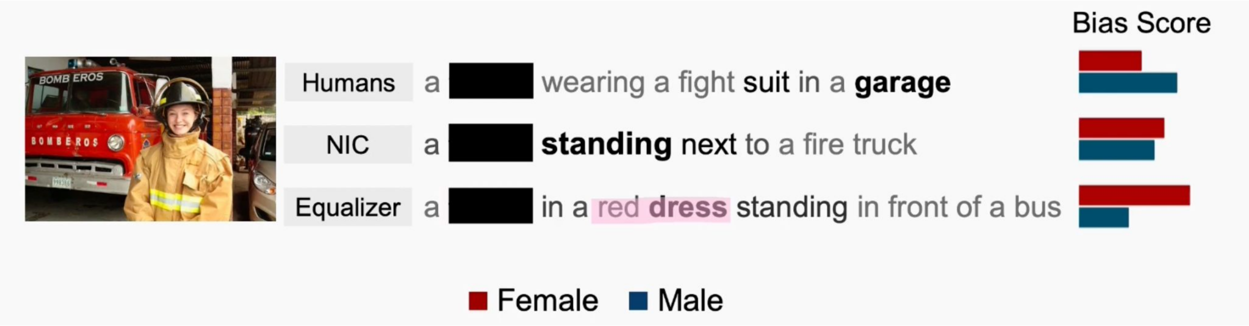


Table 2. Gender bias scores according to LIC, LIC$_M$, and LIC$_D$ for several image captioning models. Captions are encoder with LSTM, BERT-ft, or BERT-pre. Unbiased model is LIC$_M$ = 25 and LIC = 0. It shows that LIC is consistent across different language models.

	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**



LIC is robust against encoders: the tendency is maintained within the three language models: NIC shows the least bias, whereas NIC+Equalizer shows the most.

NIC+Equalizer increases gender bias with respect to the baseline.

Quantification of racial bias

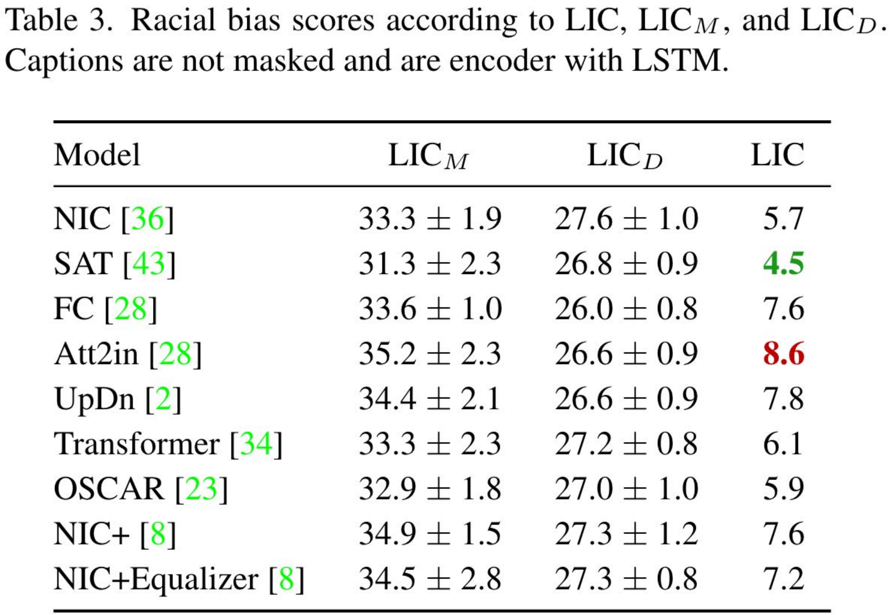


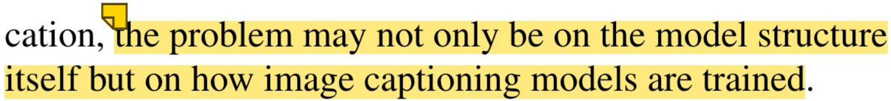
Table 3. Racial bias scores according to LIC, LIC$_M$, and LIC$_D$. Captions are not masked and are encoder with LSTM.

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

All the models amplify racial bias.

Racial bias is not as apparent as gender bias: The mean of the LICM score of all the models is 47.0 for gender and 33.7 for race.

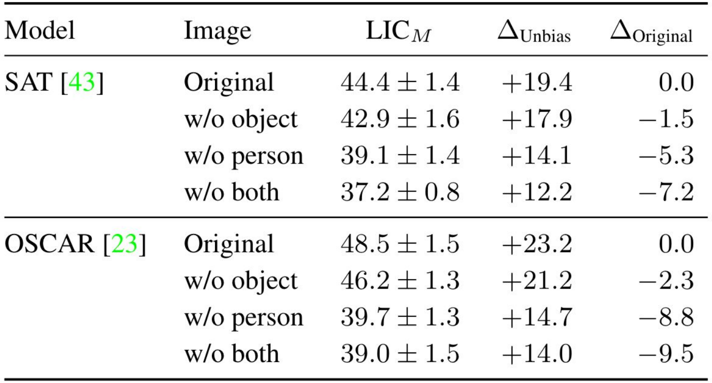
NIC+Equalizer does not increase racial bias with respect to the baseline.



cation, the problem may not only be on the model structure itself but on how image captioning models are trained.

Visual and language contribution to the bias

Table 4. Gender bias results with partially masked images. Δ_{Unbias} shows the difference with respect to a non-biased model (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

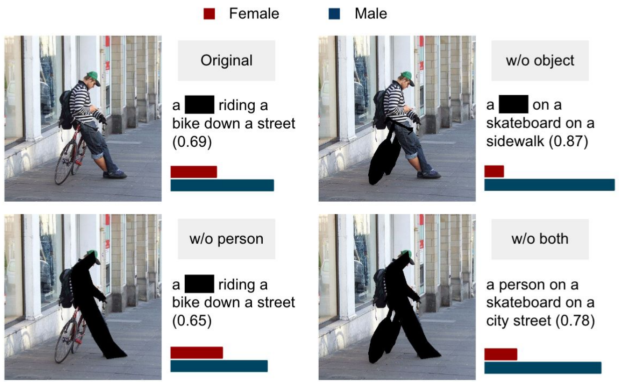

🟥 Female
🟦 Male
Original
a [redacted] riding a bike down a street (0.69)
w/o object
a [redacted] on a skateboard on a sidewalk (0.87)
w/o person
a [redacted] riding a bike down a street (0.65)
w/o both
a person on a skateboard on a city street (0.78)

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.

The contribution of objects to gender bias is minimal
The contribution of people to gender bias is higher than objects
Language models are a major source of gender bias

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

Strengths VS. Weaknesses

This paper critically points out the issues with previous paper and examined many existing models.

This paper studied deeply into both gender bias and racial bias.

By quantifying biases and examining their amplification, the paper could offer a solid methodology for evaluating and comparing different image captioning models, aiding in the design and improvement of these technologies.

It measures how much bias is introduced by the model with respect to the human captions.

It doesn’t solve the heavily rely on annotation problem.

The proposed LIC metric is simple. It can’t conclude well with pre-trained models.

Future Study

- Investigate the impact of transfer learning on bias amplification in image captioning, focusing on how biases can propagate across different domains and tasks.

- Cooperate with BLEU、ROUGE、METEOR、CIDEr matrics, the overall quality of the generated caption is also important.

Discussion

- Why is it important to study and quantify bias in machine learning models?

- Do you think LIC measures bias in image captioning systems in a meaningful way? Would you use it in your work?

- Do all machine learning models amplify bias? Why or why not?

The End

Thank you!

https://sites.google.com/view/cvpr-2022-quantify-bias/home

https://www.lri.fr/~gcharpia/deeppractice/chap_3.html

https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1204/slides/Vinodkumar_Prabhakaran_Socially_Responsible_NLP.pdf

@teenybiscuit