RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] Reproducibility Study of "Quantifying Societal Bias Amplification in Image Captioning"

Farrukh Baratov[1,2], Göksenin Yüksel[1,2], Darie Petcu[1,2], and Jan Bakker[1,2]
[1]University of Amsterdam, The Netherlands – [2]Equal contributions

Edited by
Koustuv Sinha,
Maurits Bleeker,
Samarth Bhargav

Received
04 February 2023

Published
20 July 2023

DOI
10.5281/zenodo.8173715

Reproducibility Summary

Scope of Reproducibility – We study the reproducibility of the paper *Quantifying Societal Bias Amplification in Image Captioning* [1] by Hirota, Nakashima, and Garcia. In this paper, the authors propose a new metric to measure bias amplification, called LIC, and evaluate it on multiple image captioning models. Based on this evaluation, they make the following main claims which we aim to verify: (1) all models amplify gender bias, (2) all models amplify racial bias, (3) LIC is robust against encoders, and (4) the NIC+Equalizer model increases gender bias with respect to the baseline. We also extend upon the original work by evaluating LIC for age bias.

Methodology – For our reproduction, we were able to run the code provided by the authors without any modifications. For our extension, we automatically labelled the images in the dataset with age annotations and adjusted the code to work with this dataset. In total, 38 GPU hours were needed to perform all experiments.

Results – The reproduced results are close to the original results and support all four main claims. Furthermore, our additional results show that only a subset of the models amplifies age bias, while they strengthen the claim that LIC is robust against encoders. However, we acknowledge that our extension to age bias has its limitations.

What was easy – The author's code and the data needed to run it are publicly available. The code required no modification to run and the scripts were provided with an extensive argument parser, allowing us to quickly set up our experiments. Moreover, the details of the original experiments were clearly stated in the appendix.

What was difficult – We found that it was difficult to interpret the author's code as the provided documentation contained room for improvement. Also, the scripts contained repetitive code. While the authors retrained all image captioning models, they did not share the model weights, making it difficult to extend upon their work.

Communication with original authors – No (attempt at) communication with the original authors was performed.

Correspondence should be addressed to Farrukh Baratov (farrukh.baratov@student.uva.nl)
The authors have declared that no competing interests exist.
Code is available at https://github.com/fbaratov/mlrc-lic-plus-age. – SWH swh:1:dir:42a11dd6a0c47c1ad4d57649d08238230c570fb3;origin=https://github.com/fbaratov/mlrc-lic-plus-age;visit=swh:1:snp:222b530d4c368546db5fe7ae6ef7d18e4f949429;anchor=swh:1:rev:32f722b1d60dfd9412a3c98b7e874f679536ac77.
Data is available at https://drive.google.com/drive/folders/1PI03BqcnhdXZi2QY9PUHzWn4cxgdonT- – DOI 10.48550/arXiv.2106.08503.
Open peer review is available at https://openreview.net/forum?id=eJmQJT0Dtt.

1 Introduction

Image captioning models take advantage of correlations between co-occurring captions and images in the training data. Therefore, the presence of societal biases in the large-scale datasets on which these models are trained can cause these models to not only reproduce the inequalities in the datasets but also amplify them [2].
Hirota, Nakashima, and Garcia investigate this problem in the paper "Quantifying Societal Bias Amplification in Image Captioning." They argue that existing bias amplification metrics, including Bias Amplification (BA) [2] and Leakage [3], are insufficient when applied to image captioning. Therefore, they propose a new metric to specifically study bias amplification in image captioning, called Leakage for Image Captioning (LIC). It is built on top of the idea that a model amplifies societal bias if a classifier can predict the protected attributes[1] values (e.g. female, male) more accurately from the generated captions than from human captions.
The authors use this metric to conduct extensive evaluation on both traditional and state-of-the-art image captioning models. In this work, we aim to verify their main claims by reproducing their main findings. Furthermore, we extend upon their work by evaluating LIC for age as a protected attribute. This subsequent experiment allows us to assess the generalizability of LIC for other protected attributes, as well as its overall robustness.
The remainder of this work will be structured as follows: first, we identify the main claims from the original paper in Section 2. Section 3 outlines our approach towards verifying these claims and extending upon the original work through reproduction and additional experimentation respectively. We give an overview of our experimental results in Section 4. Finally, we discuss these results and our experience with reproducing and extending the paper in Section 5.

2 Scope of reproducibility

Hirota, Nakashima, and Garcia study bias amplification on captions generated by various image captioning models. The LIC score for these models is evaluated on a captions dataset with binary gender and race annotations [4]. Furthermore, three language encoders [5, 6] are used for encoding captions in order to evaluate LIC's robustness against changes in sentence representation. The reader is invited to study Section 3 for implementation-specific details. We use the provided methodology and source code in order to verify the authors' main claims:

- *Claim 1*: All the models amplify gender bias.
- *Claim 2*: All the models amplify racial bias.
- *Claim 3*: LIC is robust against encoders.
- *Claim 4*: NIC+Equalizer increases gender bias with respect to the baseline NIC+.

Upon verifying these claims, we take a step further and evaluate the LIC metric in the context of another type of bias, namely age bias. We specifically use age as additional protected attribute because mentions of age are likely to appear in the available captions dataset. This opens up a new possibility to assess the LIC metric.

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

3 Methodology

3.1 Calculating the LIC score

Hirota et al. [1] use Leakage for Image Captioning (LIC) as a metric to measure bias amplification in image captioning models. In order to evaluate the LIC score for a model M, one needs to have access to a training set D and test set D' with samples $(I, y, \hat{y}, a)$, where I is an image, y is a human caption, $\hat{y} = M(I)$ is a generated caption and a is the protected attributes value. Then the LIC score is calculated as follows.
First, all captions are preprocessed by masking words related to the protected attribute. Details about the preprocessing are found in Section 3.3.
Second, two classifiers are trained to predict the protected attribute from a masked caption. The first classifier f is trained on the human captions in the train set D, while the second classifier $\hat{f}$ is trained on the generated captions. Both classifiers rely on a natural language encoder, which also needs to be learned on the training data. Let us denote the confidence score of the first classifier as s_a. Then this classifier predicts the protected attribute a from a caption y as

$$\hat{a} = f(y) = \arg\max_a s_a(E(y)) \tag{1}$$

where E is the natural language encoder. The same applies for $\hat{f}$, $\hat{E}$, $\hat{s}_a$ and $\hat{y}$ in case of the second classifier.
After the classifiers are trained, they are evaluated on the corresponding test sets of human and generated captions. LIC is built on top of the hypothesis that in an unbiased set of captions, there should not exist differences between how demographic groups are represented. Thus, if a classifier was trained on an unbiased dataset, it could not have learnt how to correctly predict the protected attribute of a masked caption. The higher the accuracy and confidence scores of the predictions, the more biased the captions are. From the evaluation of the first classifier on the test set, the bias in the human captions can then be quantified as

$$LIC_D = \frac{1}{|\mathcal{D}'|} \sum_{(y,a)\in\mathcal{D}'} s_a(y)\mathbb{1}[f(y) = a] \tag{2}$$

where $\mathbb{1}[\cdot]$ is the indicator function. Similarly, from the evaluation of the second classifier on the test set, the bias in the generated captions can be quantified as

$$LIC_M = \frac{1}{|\mathcal{D}'|} \sum_{(\hat{y},a)\in\mathcal{D}'} \hat{s}_a(\hat{y})\mathbb{1}[\hat{f}(\hat{y}) = a]. \tag{3}$$

Finally, the LIC score is calculated as the amount of bias introduced by the model with respect to the human captions:

$$LIC = LIC_M - LIC_D. \tag{4}$$

A model amplifies bias if LIC > 0 and mitigates it otherwise.

3.2 Model descriptions

We describe the model types (captioning models, sentence classifiers, and language encoders) used for reproducing, as well as their relevance, in the following section.

Captioning models — Image captions are generated using the following models: NIC [7], SAT [8], FC [9], Att2in [9], UpDn [10], Transformer [11], OSCAR [12], NIC+ and NIC+Equalizer [13]. NIC+Equalizer is NIC+ with a gender bias mitigation loss, which forces the model to focus on the image region with a person when predicting gender words.

Language encoders — Different sentence encoders are experimented with in order to assess the robustness of the LIC metric. These are a bidirectional LSTM [6] and BERT [5]. The weights of the LSTM are randomly initialized and learned via training, whereas BERT is initialized with pretrained weights. Either all its weights are fine-tuned as a part of training (BERT-ft) or only the final fully connected layers are fine-tuned (BERT-pre).

Sentence classifiers — Sentence classifiers are implemented as encoder classification heads. The LSTM uses a fully connected final layer trained with Adam [14] as a classification head. BERT uses a 2-layer MLP with Leaky ReLU activations and one-dimensional batch normalization, trained with AdamW [15]. The weights for both classifiers are randomly initialized.

3.3 Datasets

The dataset used for the experiments (available here) is a subset of the MSCOCO 2014 captions dataset [16] with binary gender and race annotations [2]. Gender annotations are either *male* or *female*, and race annotations are *lighter* or *darker*. Age annotations are either *young* or *old*.

Dataset splits — The dataset contains 10,780 images with gender annotations, and 10,969 images with race annotations. A 90% train-test split is utilized for both race and gender in order to train the classifiers. The number of images for each protected attribute value is made equal, thereby assuring that the model does not learn additional biases due to inequalities in the data. This results in 5,966 training and 662 testing images for gender, and 1,972 training and 220 testing images for race.
We implement an additional age dataset that consists of the gender dataset with age-related sentence tags. Each image contains five unique captions, and age tags are assigned based on the human captions' contents. If the set of captions contains only *young* or *old* words, then it is labeled accordingly. If it contains both, then it is labeled *young*. If it contains neither, the image is given the special *unknown* tag and excluded from the dataset. The same tag is then given to the corresponding captions in the generated dataset. We keep 3,130 training and 403 testing images for each age value in order to ensure that the number of images for each age value is equal. This results in an age dataset that has 6,260 training and 806 testing images total.

Preprocessing — Preprocessing is done by modifying the contents of the captions. Words related to the protected attribute are masked by neutral tags. For example, if the protected attribute is gender, words such as *boy* and *girl* would be replaced with a neutral tag "[MASK]". A list of masked words for gender is available in Appendix A.1. No words are masked for race. Furthermore, noise is added to human captions by replacing all words that do not occur within the generated vocabulary with a "[UNK]" token. This is done in order to simplify the generally more complex vocabulary found within human captions and align it with the simpler generated caption vocabulary.

3.4 Hyperparameters

We use the hyperparameters provided by the authors in the original paper. For the LSTM and BERT-pre classifiers, that means training is conducted for 20 epochs with a learning rate of $5 \cdot 10^{-5}$, while BERT-ft is trained for 5 epochs with a learning rate of $1 \cdot 10^{-5}$.

3.5 Experimental setup and code

In the original paper, the classifiers were trained 10 times per model. Due to limited computational resources, in this reproducibility study we only train the classifiers with

3 different seeds per model. We make an exception for the evaluation of NIC+ and NIC+Equalizer with gender as protected attribute, since claim 4 relies heavily on these results. The mean and standard deviation of the LIC scores are computed for each combination of captioning model, language encoder and protected attribute.

3.6 Computational requirements

We utilized two computational resources for training our models: a computational cluster with a TITAN X GPU, and a Google Cloud Virtual Machine (VM) instance with a K80 GPU. We used the VM instances to train all the LSTM models, and BERT models with generated captions. BERT models with human-annotated captions, which require more space, were trained on the cluster, as the GPU is more powerful.

Model	LSTM	BERT-ft	BERT-pre
Human	1 hour	4 hours 30 minutes	5 hours
Generated	15 minutes	1 hour	1 hour

Table 1. Average training times (in GPU time) for each model-dataset pair.

The GPU hours spent on training each model configuration are presented in Table 1. The LSTM with generated captions has the shortest training time at 15 minutes, while BERT-pre with human captions took 5 hours, marking the longest average training time.

4 Results

4.1 Results reproducing original paper

We set out to reproduce the four claims listed in Section 2 with the experiments. Having obtained results consistent with Hirota, Nakashima, and Garcia on all performed experiments, we were able to reproduce all four claims. These observations were inferred from Tables 2 and 3. Table 2 illustrates the bias score for all the models (LIC_M) in the gender experiment, as well as the human captions bias (LIC_D) and the bias amplification for each model (LIC). Table 3 shows these scores for the race experiments. An unbiased model should score $LIC_M = 25$ and $LIC = 0$. [1]. This value is obtained from a classification accuracy of 50% (random guess), multiplied with the confidence of 0.5.

	LSTM			BERT-ft			BERT-pre		
Model	LIC$_M$	LIC$_D$	LIC	LIC$_M$	LIC$_D$	LIC	LIC$_M$	LIC$_D$	LIC
NIC	43.0 ± 1.7	40.6 ± 0.9	2.4	46.7 ± 1.5	48.2 ± 0.9	-1.5	43.7 ± 0.5	41.0 ± 0.8	2.6
SAT	42.9 ± 1.0	39.8 ± 0.8	3.1	47.7 ± 1.0	48.0 ± 1.1	-0.3	42.6 ± 0.6	41.6 ± 1.0	1.0
FC	47.0 ± 1.2	38.5 ± 1.0	8.5	50.4 ± 1.6	46.1 ± 0.9	4.2	47.3 ± 1.3	40.7 ± 0.6	6.6
Att2in	46.2 ± 1.1	38.7 ± 0.4	7.5	48.2 ± 1.2	46.8 ± 0.9	1.3	46.1 ± 0.9	41.1 ± 0.8	5.0
UpDn	48.6 ± 0.5	39.6 ± 0.5	9.1	53.0 ± 0.3	47.6 ± 0.8	5.4	48.5 ± 0.9	41.7 ± 0.8	6.9
Transformer	48.6 ± 0.3	40.4 ± 1.0	8.2	55.4 ± 0.2	48.6 ± 0.8	6.9	48.3 ± 1.1	42.3 ± 0.8	6.0
OSCAR	48.5 ± 2.1	39.9 ± 0.5	8.7	52.9 ± 2.2	47.7 ± 0.7	5.1	47.7 ± 1.5	41.0 ± 0.8	6.7
NIC+	46.7 ± 1.1	39.4 ± 0.7	7.3	50.8 ± 1.4	48.2 ± 0.9	2.6	47.7 ± 0.5	40.7 ± 0.8	7.0
NIC+Equalizer	51.3 ± 0.7	39.5 ± 0.8	11.8	54.9 ± 1.7	47.8 ± 0.5	7.1	50.0 ± 0.8	40.8 ± 0.6	9.2

Table 2. Gender bias amplification LIC scores for all used image captioning models. Results are grouped by the used encoder: LSTM, BERT-ft, or BERT-pre.

Claim 1: All the models amplify gender bias. — The first observation to make about the gender experiment results is that all the LIC_M scores are well above the unbiased value of 25, sitting anywhere between 42.6 and 54.9. Bias amplification is exhibited by all models under at least two of the three used encoders, which can be seen by examining the LIC

	LSTM			BERT-ft			BERT-pre		
Model	LIC$_M$	LIC$_D$	LIC	LIC$_M$	LIC$_D$	LIC	LIC$_M$	LIC$_D$	LIC
NIC	33.1 ± 2.7	27.3 ± 1.0	5.8	33.0 ± 1.9	36.5 ± 0.7	-3.5	32.5 ± 1.0	32.8 ± 0.7	-0.3
SAT	31.2 ± 2.0	26.7 ± 0.6	4.6	38.2 ± 2.1	36.3 ± 0.8	1.9	34.9 ± 1.6	33.0 ± 0.4	1.9
FC	33.6 ± 0.9	26.4 ± 0.6	7.2	41.5 ± 0.4	36.0 ± 0.7	5.5	39.4 ± 2.7	32.2 ± 0.5	7.2
Att2in	35.3 ± 2.3	26.2 ± 0.3	9.1	41.4 ± 1.0	36.2 ± 0.7	5.2	38.3 ± 1.4	32.1 ± 0.2	6.2
UpDn	36.7 ± 0.4	26.3 ± 0.4	10.4	42.3 ± 0.8	36.6 ± 0.2	5.7	39.4 ± 0.9	33.1 ± 0.3	3.2
Transformer	34.7 ± 0.6	27.3 ± 0.3	7.5	40.8 ± 1.5	37.4 ± 0.8	3.4	36.3 ± 1.3	33.8 ± 1.0	5.6
OSCAR	34.0 ± 2.5	27.1 ± 0.7	6.9	40.1 ± 1.8	36.7 ± 0.6	3.4	36.3 ± 2.7	32.9 ± 0.4	3.4
NIC+	34.9 ± 1.4	27.4 ± 0.9	7.5	40.2 ± 1.5	36.4 ± 1.0	3.8	37.9 ± 2.5	33.3 ± 1.1	4.6
NIC+Equalizer	34.6 ± 2.5	27.3 ± 0.9	7.3	39.0 ± 1.5	36.8 ± 0.8	2.2	37.0 ± 3.1	33.1 ± 0.8	3.9

Table 3. Race bias amplification LIC scores for all used image captioning models. Results are grouped by the used encoder: LSTM, BERT-ft, or BERT-pre.

scores. The only two exceptions, where $LIC < 0$, are under the finetuned BERT encoder for the NIC ($LIC = -1.5$, lowest value in the table) and SAT ($LIC = -0.3$) caption models. It should be noted, however, that these values are still close to 0 and therefore do not show a strong ability to mitigate bias.

Claim 4: NIC+Equalizer increases gender bias with respect to the baseline NIC+ — Looking at the bottom two rows of Table 2, we can see that NIC+Equalizer achieves the highest LIC score regardless of the used encoder. While the ”ground truth” LIC_D scores remain consistent between NIC+ and NIC+Equalizer, the increase in LIC_M is substantial across all columns of the table. This is associated with an increased gender bias introduced by the model, which supports the claim that the Equalizer increases gender bias compared to the baseline NIC+.

Claim 2: All the models amplify racial bias. — The race experiments, depicted in Table 3, show that the captioning models exhibit bias not only towards the gender attribute, but also towards the race. None of the captioning models manage to score negative LIC values with all encoders. Instead, almost all the captioning models only achieve positive LIC scores, showing they amplify racial bias. Once again, the NIC model is the only exception. For any of the BERT encoders, the LIC score is below 0. Looking at the LSTM column, however, NIC does not score the lowest LIC, which therefore shows that the merit of negative LIC does not primarily correspond to the NIC captioning model.

Claim 3: LIC is robust against encoders. — Despite minor variations in some parts of the results tables, the general trends in the scores remain consistent across all used encoders. The LIC score remains at comparable values for all models when comparing between different used encoders, which goes to show that the encoder choice does not disrupt the LIC result. We can therefore say LIC is indeed robust against encoders and complete the four claims by Hirota, Nakashima, and Garcia that we reproduced and confirmed.

4.2 Results beyond original paper

Age experiment — As described in Section 2, we extended the original experiments by computing LIC scores for the protected attribute of age. The results can be observed in Table 4. The first observation to make is that the LIC scores are overall lower than for the other attributes: we now see at least half the models achieve sub-zero (bias mitigation) LIC scores under all encoders. The LIC_M scores are comparable to the other protected attributes, meaning the models are not biased towards age more than towards the other tested attributes. Moreover, the LIC_D scores are higher, meaning that the observed bias in the dataset is larger. This is especially noticeable in the finetuned BERT, compared to both results in our extension and LIC_D scores in the original work. The spread of the

	LSTM			BERT-ft			BERT-pre		
Model	LIC$_M$	LIC$_D$	LIC	LIC$_M$	LIC$_D$	LIC	LIC$_M$	LIC$_D$	LIC
NIC	37.7 ± 0.4	46.5 ± 0.4	-8.8	40.5 ± 0.8	51.6 ± 1.0	-11.1	41.0 ± 0.8	41.8 ± 0.5	-0.8
SAT	45.7 ± 1.9	46.1 ± 0.7	-0.5	47.9 ± 1.7	51.8 ± 0.3	-3.9	41.6 ± 1.0	42.0 ± 0.2	-0.4
FC	44.6 ± 0.8	45.2 ± 0.8	-0.5	47.7 ± 0.8	50.5 ± 0.4	-2.8	40.7 ± 0.6	41.1 ± 0.4	-0.4
Att2in	47.0 ± 1.5	45.5 ± 0.7	1.5	49.4 ± 1.3	50.5 ± 0.8	-1.1	41.1 ± 0.8	40.9 ± 0.3	0.2
UpDn	48.9 ± 0.7	46.5 ± 0.6	2.5	51.8 ± 1.2	51.8 ± 0.7	0.0	41.7 ± 0.8	41.6 ± 0.5	0.1
Transformer	49.3 ± 1.2	46.7 ± 0.6	2.6	52.8 ± 1.9	52.6 ± 0.1	0.2	42.3 ± 0.8	42.5 ± 0.3	-0.2
Oscar	50.8 ± 0.9	46.4 ± 0.8	4.4	51.0 ± 1.6	51.5 ± 0.6	-0.4	47.9 ± 1.8	41.9 ± 0.5	5.9
NIC+	41.5 ± 1.4	46.4 ± 0.4	-4.9	43.4 ± 1.8	51.6 ± 1.1	-8.2	39.3 ± 0.5	41.6 ± 0.2	-2.3
NIC+Equalizer	40.8 ± 1.2	46.6 ± 0.4	-5.8	43.2 ± 2.0	51.5 ± 0.6	-8.3	39.3 ± 1.1	41.7 ± 0.2	-2.4

Table 4. Age bias amplification LIC scores for all used image captioning models. Results are grouped by the used encoder: LSTM, BERT-ft, or BERT-pre.

LIC scores, between the smallest and largest values for each encoder, appears to have increased from the protected attributes observed in the reproduction results: while those exhibit differences of 6-8 units, this difference for age reaches values around 8-12 units.

Encoder analysis using attention visualizer — In a method similar to Hendricks et al. [17], we implemented an attention visualizer that we used on the two BERT encoders: finetuned (BERT-ft) and pretrained (BERT-pre). Upon inspecting multiple layers and attention heads, we found head 2 in layer 4 controls correlations relevant to our research. Figure 1 shows the class token is linked by the attention mechanism to the word *motorcycle* for the old class, and to the word *skate* for the young class. Figure 2 further illustrates the difference in attention tuning between the finetuned and pretrained BERT model. In finetuned BERT, attention links the class token to the word *skate*. This link is also present in pretrained BERT, but with a smaller weight (weaker connection).

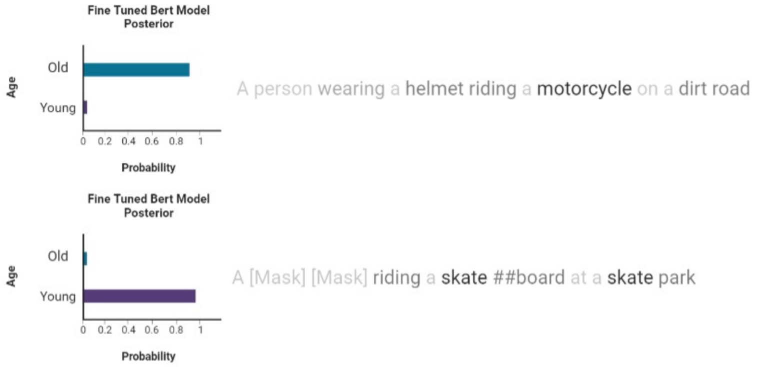

Figure 1. Captions and confidence levels predicted as "old" and "young" by the BERT model trained on the age dataset. Darker hue indicates larger attention weight.

5 Discussion

5.1 Main claims

The reproduced results reproducing are slightly different from the original paper's results. One reason for this is only training the classifiers with 3 seeds per model instead of 10 due to limited computational resources. Our results are therefore slightly less reliable. Even though the numbers are different, our results support all four main claims from the original paper.

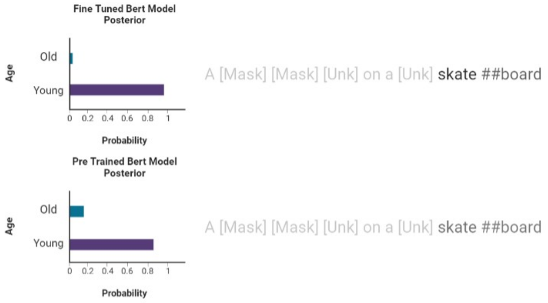

Figure 2. Attention maps and confidence levels for BERT-ft and BERT-pre trained on the age dataset. Darker hue indicates larger attention weight.

When inspecting the results, it is remarkable that for all protected attributes, BERT-ft generally has the lowest LIC score in comparison to the other classifiers. Specifically, the LIC_D scores for BERT-ft are higher, indicating that this classifier learns to predict the protected attribute better from a masked human caption. While the LIC_M scores for BERT-ft are also higher, showing that BERT-ft is the best performing classifier in general, the increase in LIC_D is larger than for LIC_M, resulting in a lower LIC score. BERT-ft's proficiency in learning the dataset bias is attributed to its attention mechanism having been adjusted to the dataset during the finetuning stage.

5.2 Encoder analysis using attention visualizer

Visualizing the attention heads of each encoder layer allows us to observe which tokens of a caption are strongly correlated with the class token. We performed a qualitative analysis on the attention heads of the pretrained and finetuned BERT models using our implemented attention visualizer. Thus we make sure that the two BERT encoders act as expected in two ways: they should not look at the masked words to make predictions, and the correlation between the detected class and the words associated with a class should be strong. The results of our qualitative analysis show that the models indeed pay attention to the words we expect them to look at. This is illustrated in Figure 1.
Figure 2 illustrates how the attention weights of BERT are adapted to the task, showing that the finetuned encoder manages to turn its attention to class-hinting tokens much better than the pretrained one. This analysis allows us to understand why the finetuned BERT performs better than the other encoders. The advantage of finetuned attention is also visible in the confidence scores of the prediction, with the pretrained BERT being less confident (albeit still correct) with its predicted class.

5.3 Reviewing LIC and the original claims about it in the age context

The results obtained for the age experiments can be attributed to a number of key factors. Firstly, there appears to be bias mitigation in multiple models and under the use of multiple encoders, which cannot be said about the reproduced experiments. The high values of the LIC_D scores are mostly responsible for this, seeing as the other component of the LIC score, the model bias, does not appear to have changed from the other experiments. As such, we attribute the bias mitigation effect to the encoders being more capable of learning the bias in the dataset. This effect can also be attributed to the age attribute being mostly expressed as an adjective. Due to this syntactic trait, the bias can be learnt with more ease.
To re-evaluate the authors' claims using the information gathered from the extension experiment, we first look at claims 1 and 2, which state that all models amplify racial

and gender bias. As has been discussed, age bias does not appear to be amplified by all models. Instead, only UpDn, Transformer and Oscar appear to amplify age bias.
LIC's robustness against encoders was also further tested in our extension. This can be evaluated by observing whether it fluctuates for one model depending on the employed encoder. In the age experiment's case, we notice some minor fluctuations in some models, but generally the LIC scores for the three encoders have the same polarity. Keeping in mind we already discussed how BERT-ft learns the bias in the dataset better, so slightly different results are to be expected, we maintain the claim that LIC is robust against encoders.

5.4 Limitations of age experiment

While the age attribute has significantly contributed to our understanding and confidence in the claims from the main paper, it must be noted that this experiment also has some weaknesses. Besides the smaller dataset size, the used captions dataset was not even specialized for the age experiment. Instead, the gender dataset was used. A dataset built specifically for age bias experiments would be more suitable for the task, and should result in more valuable findings. Moreover, the attribute masking was done automatically for this experiment, and it did not perform without issues. When consecutive words contained information about the protected attribute, multiple masks were applied. Therefore, some sentences contained two concatenated mask tokens, instead of simply using one mask token. Due to the stochasticity of the models used in the experiments, it is highly probable that using one mask token for multiple words would lead to different results.

5.5 What was easy

Operational code — The base code ran right out of the box, and required no modification on our part to make it operational.
Hyperparameters — The hyperparameters used to train models were clearly provided in the supplementary materials section for all model configurations.
Data availability — The data is linked by the author, and the instructions for where to place it were clear and allowed for a quick setup.
Run arguments — The provided scripts have an extensive argument parser, making it easy to modify relevant fields such as the captions dataset, hyperparameter settings, and random seeds. This allowed us to quickly set up our experiments.

5.6 What was difficult

Lack of comments — The provided code's documentation contains room for improvement. As a result, we had difficulties interpreting it.
Repetitive code — The scripts for the gender and race attributes and different encoders are very similar. Extending the code mainly involved duplicating the BERT and LSTM gender scripts and modifying them by changing relevant keywords/variable names, which is not an optimal way of adding new attributes.
Unavailable model weights — The authors retrained all image captioning models on only the MSCOCO training dataset, as they intended to use the validation set for evaluation. However, the retrained model weights were not available. As a result, we had to use the generated captions provided by the authors instead of generating them ourselves.

5.7 Communication with original authors

We did not communicate with the original authors, neither about our reproduction of their paper nor about the extensions that we implement and test.

References

1. Y. Hirota, Y. Nakashima, and N. Garcia. "Quantifying Societal Bias Amplification in Image Captioning." In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. 2022, pp. 13450–13459.
2. J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." In: Jan. 2017, pp. 2979–2989. DOI: 10.18653/v1/D17-1323.
3. T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. **Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations**. 2018. DOI: 10.48550/ARXIV.1811.08489. URL: https://arxiv.org/abs/1811.08489.
4. D. Zhao, A. Wang, and O. Russakovsky. "Understanding and Evaluating Racial Biases in Image Captioning." In: Oct. 2021, pp. 14810–14820. DOI: 10.1109/ICCV48922.2021.01456.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: **ArXiv** abs/1810.04805 (2019).
6. S. Hochreiter and J. Schmidhuber. "Long Short-term Memory." In: **Neural computation** 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.
7. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and tell: A neural image caption generator." In: June 2015, pp. 3156–3164. DOI: 10.1109/CVPR.2015.7298935.
8. K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." In: **International Conference on Machine Learning**. 2015.
9. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. "Self-Critical Sequence Training for Image Captioning." In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)** (2016), pp. 1179–1195.
10. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. "Bottom-Up and Top-Down Attention for Image Captioning and VQA." In: (July 2017).
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need." In: **Advances in Neural Information Processing Systems**. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
12. X. Li et al. "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks." In: **European Conference on Computer Vision**. 2020.
13. L. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. "Women Also Snowboard: Overcoming Bias in Captioning Models: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III." In: Sept. 2018, pp. 793–811. DOI: 10.1007/978-3-030-01219-9_47.
14. D. P. Kingma and J. Ba. **Adam: A Method for Stochastic Optimization**. 2014. DOI: 10.48550/ARXIV.1412.6980. URL: https://arxiv.org/abs/1412.6980.
15. I. Loshchilov and F. Hutter. **Decoupled Weight Decay Regularization**. 2017. DOI: 10.48550/ARXIV.1711.05101. URL: https://arxiv.org/abs/1711.05101.
16. X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. **Microsoft COCO Captions: Data Collection and Evaluation Server**. 2015. DOI: 10.48550/ARXIV.1504.00325. URL: https://arxiv.org/abs/1504.00325.
17. L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. "Women also snowboard: Overcoming bias in captioning models." In: **Proceedings of the European conference on computer vision (ECCV)**. 2018, pp. 771–787.

A Supplementary materials

A.1 Masked words

The following words are masked for age:

- *Young*: *kid child, young, boy, little, baby, babies, childhood, babyhood, toddler, adolescence, adolescent, teenager, schoolboy, schoolgirl, youngster, infant, preschooler, toddler, student, girl,* and their plurals.
- *Old*: *elder, man, woman, old, elderly, grandma, grandpa, mom, dad, father, ancient, aged, senior, grandparent, senior,* and their plurals.

See Hirota et al. [1] for a list of masked words for gender. No words are masked for race, as it is generally not explicitly mentioned in captions.