RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] RELIC: Reproducibility and Extension on LIC metric in quantifying bias in captioning models

Paula Antequera[1], Egoitz Gonzalez[1], Marta Grasa[1], and Martijn van Raaphorst[1]

[1]Universiteit van Amsterdam, Amsterdam, Netherlands

Edited by
Koustuv Sinha,
Maurits Bleeker,
Samarth Bhargav

Received
04 February 2023

Published
20 July 2023

DOI
10.5281/zenodo.8173741

Reproducibility Summary

Scope of Reproducibility — In this work we reproduce and extend the results presented in "Quantifying Societal Bias Amplification in Image Captioning" by Hirota et al. [1]. This paper introduces LIC, a metric to quantify bias amplification by image captioning models, which is tested for gender and racial bias amplification.
The original paper claims that this metric is robust, and that all models amplify both gender and racial bias. It also claims that gender bias is more apparent than racial bias, and the Equalizer variation of the NIC+ model [2] increases gender but not racial bias. We repeat the measurements to confirm these claims. We extend the analysis to whether the method can be generalized to other attributes such as bias in age.

Methodology — The authors of the paper provided a repository containing the necessary code. We had to modify it and add several scripts to be able to run all the experiments. The results were reproduced using the same subset of COCO [3] as in the original paper. Additionally, we manually labeled images according to age for our specific experiments. All experiments were ran on GPUs for a total of approximately 100 hours.

Results — All claims made by the paper [1] seem to hold, as the results we obtained follow the same trends as those presented in the original paper even if they do not match exactly. However, the same cannot always be said of the additional experiments.

What was easy — The paper was clear and matched the implementation. The code was well organized and was easy to run using the command interface provided by the authors. This also made it easy to replicate and expand upon it by adding our own new features. The data was also readily available and could be easily downloaded with no need for preprocessing.

What was difficult — We had to run several iterations of the same code, using different seeds and models, to get the results with the same conditions as in the original paper, which made use of time and resources. Our own experiments required additional time to hand-annotate data due to lack of data for new features.

Correspondence should be addressed to Marta Grasa (marta.grasa@student.uva.nl)
The authors have declared that no competing interests exist.
Code is available at https://github.com/eggonz/relic-caption-bias – DOI https://doi.org/10.5281/zenodo.7930598. – SWH swh:1:dir:01b16d04ee6ff9480c0fab3f65ea8957997e05c5.
Open peer review is available at https://openreview.net/forum?id=9_hCoP3LXwy.

Communication with original authors — There was no contact with the authors, since the code and the experiments were clear and did not need any additional explanation.

1 Introduction

It has been shown that large-scale computer vision models are often biased [4], and detecting and addressing these biases is nowadays an active research field. It has also been shown [5] that the bias in the results originates not only from biased data, but also from the models themselves.
Even though these biases are well known, there is no standard metric to measure them. The paper we aim to reproduce [1] introduces the *Leakage for Image Captioning* metric (LIC), which measures how much bias is introduced by the model considering the bias present in the dataset as a baseline.

2 Scope of reproducibility

The original paper [1] introduces the LIC score as a measure of how much bias is introduced by the model. It makes some claims on the performance of this metric.

- **Claim 1:** LIC is robust against encoders. Its overall tendency is maintained across all language models (LSTM [6], BERT-ft, BERT-pre [7]). This claim is supported by the data presented on Tables 1 and 2.
- **Claim 2:** All models amplify both gender and race bias. This is supported by the results presented in Section 4.1.
- **Claim 3:** Racial bias is not as apparent as gender bias. This can be concluded by comparing the data presented in Table 1 and Table 2.
- **Claim 4:** NIC + Equalizer [2] increases gender bias, but not racial bias, with respect to the baseline (NIC+). This effect is shown by the data presented in the last row of Tables 1 and 2.

3 Methodology

To replicate the results of the original paper [1], we used the source code provided by the authors, only making minimal changes. We also used the same datasets (Section 3.2). We ran the experiments on a GPU.
For the experiments beyond the original paper, we re-used the code provided by the authors of the paper and prepared our own dataset. The data was processed to match the input data format so that the modifications in the code were minimal.

3.1 Model descriptions

The aim of this study is to test the use of LIC introduced in the original paper [1] as a metric for bias in captioning models. This score is defined as

$$\mathrm{LIC} = \mathrm{LIC}_M - \mathrm{LIC}_D \tag{1}$$

where

$$\mathrm{LIC}_M = \frac{1}{|\hat{\mathcal{D}}|} \sum_{(\hat{y},a)\in\hat{\mathcal{D}}} \hat{s}_a(\hat{y}) \mathbf{1}[\hat{f}(\hat{y}) = a], \tag{2}$$

$$\mathrm{LIC}_D = \frac{1}{|\mathcal{D}|} \sum_{(y^*,a)\in\mathcal{D}} s_a^*(y^*)\mathbf{1}[f^*(y^*) = a]. \tag{3}$$

$\mathcal{D}$ represents the set of samples, y represents their provided caption annotations, f is a classifier and s is a confidence score [1]. All variables with ˆ apply to the captions predicted by the captioning models, while those with * apply to the human-captioned database. The ground truth of the protected attribute of the image is represented by a.
We computed the LIC score of nine image captioning models: NIC [8], SAT [9], FC [10], Att2in [10], UpDn [11], Transformer [12], OSCAR [13] (which is transformer-based), NIC+ [2], and NIC+Equalizer [2], which is a variation of NIC+ which aims to reduce gender bias. We didn't use these models directly, but the captions produced by them available in the source code of the original paper.
We limited our experiments on age to the models SAT, OSCAR, NIC+ and NIC+Equalizer. These models give a good overview of the trends when applied to other attributes.
To obtain the predicted protected attribute, we use three different classifiers. We first use an LSTM [6] without pre-computed word embeddings. We then use BERT [7], both fine-tuned for this task and its pre-trained version.

3.2 Datasets

We used the same dataset as the original paper, that is, a subset of the MSCOCO dataset [14] with binary gender and race annotations provided in [3]. For further experiments we used a subset of the original data which had to be hand-annotated for age. Further description of the dataset can be found in the experiments section (Section 4.2) and Appendix A. Links to download these datasets are available in the link to the code provided in Section 3.4.
The data we used consists of 10,780 images for gender; 10,969 for race and 7,036 for age. Instead of the whole available dataset, we used a balanced split for each attribute, which consisted of 6,628 total images for gender; 2,192 for race and 4,870 for age. These subsets where then randomly divided into train and test splits with 90% being used for training and 10% for testing in all cases.

3.3 Hyperparameters

We used the same hyperparameters described in the original paper [1], which are learning rate of $1 \cdot 10^{-5}$ and batch size of 64.
For the LSTM model, we used embeddings of dimension 100 and 2 layers with 256 hidden dimensions, with a dropout rate of 0.5, and ran it for 20 epochs.
For both variants of the BERT model [7], we used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\varepsilon = 1 \cdot 10^{-6}$. We limited the sequence length to 64 and ran the model for 5 epochs.

3.4 Experimental setup and code

The code used to replicate and extend the experiments can be found in our GitHub repository[1]. It contains all the scripts, pipelines and instructions necessary to replicate the results.

3.5 Computational requirements

Due to our limited resources, we did not train the captioning models in order to replicate the experiments presented in the original paper. We instead used the data provided in its source code. We did, however, fine-tune a version of BERT [7] and compared its

[1] https://github.com/eggonz/relic-caption-bias

performance to the pre-trained version. We ran every experiment with 10 seeds and present the results as the mean and standard deviation of all runs.
Half of the experiments on both BERT versions were performed using the GPUs available on *Google Colab*, while the other half was ran using an NVIDIA Titan RTX GPU. This took approximately 14 hours for gender bias and 5 hours for racial bias for the pre-trained version and 34 hours for gender bias and 13 hours for racial bias for fine-tuning and experimenting on the fine-tuned version.
The experiments on an LSTM were run locally on an NVIDIA GeForce MX150 GPU. This took approximately 22 hours for gender bias and 6 hours for racial bias.
The additional experiments ran using age as a protected attribute were all ran on an NVIDIA Titan RTX GPU, which took approximately 1 hour for LSTM, 5.5 hours for the fine-tuned version of BERT and 2 hours for the pre-trained version.

4 Results

4.1 Results reproducing original paper

All claims made by the authors are also supported by our experiments. We notice that even though the results obtained do not match those in the original paper [1], the overall trends present do support the claims stated in Section 2.

Gender bias scores according to LIC — The results we obtained by replicating the experiments on gender bias are shown in Table 1.

Model	LSTM			BERT-ft			BERT-pre		
	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC
NIC [8]	42.4 ± 0.8	39.6 ± 0.9	2.8	49.0 ± 1.5	48.0 ± 1.1	1.0	38.0 ± 0.7	35.1 ± 0.4	2.9
SAT [9]	45.7 ± 1.6	39.3 ± 1.0	6.4	48.6 ± 1.0	47.5 ± 1.3	1.1	39.4 ± 1.0	35.9 ± 0.7	3.5
FC [10]	46.3 ± 1.1	37.9 ± 0.9	8.4	48.5 ± 1.9	45.9 ± 1.3	2.6	41.1 ± 1.4	35.1 ± 0.6	6.0
Att2in [10]	45.7 ± 0.8	38.4 ± 1.0	7.2	47.7 ± 1.6	46.8 ± 1.2	0.9	40.4 ± 1.2	35.3 ± 0.6	5.1
UpDn [11]	48.3 ± 1.5	39.0 ± 0.8	9.3	52.6 ± 1.2	47.5 ± 1.2	5.1	41.8 ± 1.5	35.8 ± 0.7	6.0
Transformer [12]	48.7 ± 1.5	39.8 ± 0.9	9.0	53.7 ± 1.7	48.4 ± 1.2	5.2	40.0 ± 1.9	36.1 ± 0.6	3.9
OSCAR [13]	49.0 ± 1.4	39.3 ± 0.9	9.7	52.4 ± 1.6	47.7 ± 1.3	4.7	40.8 ± 1.1	35.1 ± 0.6	5.7
NIC+ [2]	46.8 ± 1.6	39.4 ± 0.8	7.4	49.9 ± 2.2	47.7 ± 1.2	2.2	39.1 ± 1.5	35.0 ± 0.6	4.1
NIC+Equalizer [2]	51.5 ± 1.5	39.4 ± 0.9	12.2	54.6 ± 1.5	47.6 ± 1.4	7.0	39.9 ± 1.5	35.1 ± 0.6	4.8

Table 1. Gender bias scores according to LIC, LIC_M, and LIC_D for several image captioning models. Captions are encoded with LSTM [6], BERT-ft (fine-tuned) or BERT-pre (pre-trained) [7].

We can see that these results support the claims presented in Section 2. We see that, as the first claim states, the overall tendency of LIC values for gender bias is consistent across language models.
We also notice that all models increase gender bias, as claim 2 maintains. In particular, NIC+Equalizer [2] has a higher LIC score than NIC+ [2], which supports claim 4.

Racial bias scores according to LIC — The results we obtained by replicating the experiments on racial bias are shown in Table 2.
These results also show that the LIC metric is consistent across models, and that all models (with small exceptions only when using BERT as a language model) do increase racial bias, as stated in claims 1 and 2.
Regarding claim 3, it is supported by comparing Table 1 and Table 2. As can be seen, the bias scores are consistently lower for all models for racial bias then for gender bias.
We notice that claim 4 also holds. Unlike in the case of gender bias, we can see that NIC+Equailizer [2] does not increase bias with respect to NIC+ [2].

Model	LSTM			BERT-ft			BERT-pre		
	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC
NIC [8]	32.4 ± 1.5	27.8 ± 1.1	4.6	36.4 ± 1.7	36.8 ± 1.1	-0.5	28.3 ± 2.0	29.0 ± 0.9	-0.7
SAT [9]	32.4 ± 1.9	26.8 ± 0.8	5.5	37.0 ± 2.0	36.5 ± 1.1	0.5	28.9 ± 1.8	28.6 ± 0.7	0.3
FC [10]	33.5 ± 2.4	26.2 ± 0.7	7.3	39.2 ± 2.7	36.3 ± 1.2	2.9	30.4 ± 1.9	28.4 ± 0.8	1.9
Att2in [10]	35.3 ± 2.5	26.8 ± 0.8	8.5	40.6 ± 2.7	36.3 ± 1.0	4.3	31.0 ± 2.1	28.3 ± 0.9	2.7
UpDn [11]	34.4 ± 2.3	26.8 ± 1.0	7.6	41.0 ± 2.7	36.7 ± 0.9	4.3	31.0 ± 1.3	28.7 ± 1.1	2.3
Transformer [12]	33.6 ± 1.9	27.3 ± 0.6	6.4	40.3 ± 2.1	37.4 ± 1.5	2.8	30.4 ± 1.6	28.9 ± 0.8	1.5
OSCAR [13]	33.1 ± 1.8	27.2 ± 1.2	5.9	39.0 ± 2.1	36.7 ± 1.2	2.3	30.2 ± 2.8	28.6 ± 1.0	1.6
NIC+ [2]	34.8 ± 2.1	27.4 ± 1.3	7.4	41.7 ± 5.4	39.9 ± 5.0	1.8	30.2 ± 2.2	28.7 ± 1.1	1.5
NIC+Equalizer [2]	35.1 ± 2.3	27.3 ± 0.8	7.8	43.4 ± 8.0	39.5 ± 5.0	3.9	30.2 ± 2.0	28.9 ± 0.7	1.3

Table 2. Racial bias scores according to LIC, LIC_M, and LIC_D for several image captioning models. Captions are encoded with LSTM [6], BERT-ft (fine-tuned) or BERT-pre (pre-trained) [7].

4.2 Results beyond original paper

After running the experiments for gender and racial bias, we decided to study whether the same results are also obtained for a new protected attribute, age. We want to determine if the trends that we observed in the previous experiments and the main claims of the original paper hold true when considering age bias. To do this, we computed the LIC score of SAT [9] and OSCAR [13] as representative models of classical and state-of-the-art image captioning, respectively; and NIC+ [2] and NIC+Equalizer [2], to be able to compare the results of the fourth claim applied to age bias. We used the same three classifiers LSTM [6], BERT fine-tuned and BERT pre-trained [7], all with the same hyperparameters as described above.

Additional Result 1 — We observe that some of the claims of the original paper [1] exposed in section 2 hold when using age as a protected attribute, while others do not.

Model	LSTM			BERT-ft			BERT-pre		
	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC
SAT [9]	48.5 ± 1.4	42.2 ± 1.3	6.4	52.1 ± 1.9	48.0 ± 1.4	4.1	37.0 ± 1.4	33.3 ± 0.5	3.7
OSCAR [13]	52.9 ± 1.9	43.6 ± 4.1	9.3	55.9 ± 2.0	48.1 ± 0.8	7.8	39.6 ± 1.8	33.4 ± 0.9	6.2
NIC+ [2]	43.0 ± 1.6	42.1 ± 1.0	0.9	46.6 ± 1.7	47.5 ± 1.1	-0.9	35.1 ± 1.1	32.9 ± 1.6	2.2
NIC+Equalizer [2]	43.3 ± 1.5	42.0 ± 1.1	1.3	46.7 ± 1.6	47.9 ± 1.1	-1.1	34.8 ± 1.8	33.0 ± 0.9	1.9

Table 3. Age bias scores according to LIC, LIC_M, and LIC_D for several image captioning models. Captions are encoded with LSTM [6], BERT-ft (fine-tuned) or BERT-pre (pre-trained) [7].

We first notice that claim 1 holds, as the LIC scores are consistent across classifiers for all studied models. The data also implies that claim 4 holds beyond the use of only race as an alternative protected attribute to gender, as NIC+Equalizer [2] does not significantly increase age bias with respect to the baseline. This supports the claim that it only increases gender bias.
On the other hand, we can observe that the data does not support claim 2, as not all models increase age bias. This opens the possibility of further study on whether the different models increase bias for different protected attributes.
Regarding claim 3, we see that age bias is not consistently more or less apparent than either gender or race bias.

5 Discussion

Although the results from the experiments do not replicate the exact results which the authors of the original paper [1] obtained, they do follow the same trend that is observed

in the original results of the paper. The results obtained in our experiments therefore support the claims made by the authors. The source code was also hardly modified to run the experiments and the same seeds were used to reproduce the results.
For the experiments beyond the original paper, we could not find a dataset which contained images with captions which were annotated for age. This led us to hand-annotating the dataset for age. Due to a lack of time most images that were annotated were not checked and only annotated by a single person. The results of the experiments were also run on different GPU's which might cause inconsistency. The LIC scores obtained by running the experiments on this dataset also support most of the claims stated in the original paper, with the exception of claim 2.

5.1 What was easy

The original paper [1] was clear in stating the claims supported by the results they obtained, which made it easy for us to identify what we should be looking for when replicating them.
It also provided us with the source code needed to reproduce the experiments and extensive instructions on how to run it, which made the work easier. We only needed to make minor changes to be able to reproduce the given results. The source code was also clear enough that we were able to easily identify the parts that needed to be modified to run the experiments on a new protected attribute (age) and update them.

5.2 What was difficult

The main difficulty of this study was the amount of hours needed to run all of the experiments. As we did not always have reliable access to a GPU, we had to run them in small batches over the course of weeks, which was time-consuming. In addition, even running the same seeds, we still got noticeably different results than the original paper [1], even if our results still support their claims on the LIC metric.
In the beginning we also had a bit of trouble with installing the environment needed to run the code. Although worked for some of us locally, we had some trouble getting it to work on the more powerful GPUs we had available, which made running the code take even longer since we had to run it on *Google Colab*.
We also notice that the results of the experiments vary greatly when ran using different seeds, which led to us running them many times and increasing the number of hours of GPU access needed.
Another difficulty we encountered was the lack of age annotations in the provided dataset, which we needed in order to perform the additional experiments. We solved this by annotating ourselves the subset of the MSCOCO dataset [14] for which we had the captions produced by the models we wanted to study.

5.3 Communication with original authors

There was no contact with the authors, since the code and the experiments were clear and did not need any additional explanation.

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. K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, and A. Rohrbach. "Women also snowboard: Overcoming bias in captioning models." In: **ECCV** (2018).

3. D. Zhao, A. Wang, and O. Russakovsky. "Understanding and Evaluating Racial Biases in Image Captioning." In: **International Conference on Computer Vision (ICCV)**. 2021.
4. J. Buolamwini and T. Gebry. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In: **ACM FAccT** (2018).
5. C. D'ignacio and L. F. Klein. **Data feminism**. MIT press, 2020.
6. S. Hochreiter and J. Schmidhuber. "Long short-term memory." In: **Neural computation** 9.8 (1997).
7. J. Delvin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: pre-training of deep bidirectional transformers for language understanding." In: **NAACL-HLT** 1 (2019).
8. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and tell: A neural image caption generator." In: **CVPR** (2015).
9. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In: **ICML** (2015).
10. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. "Self-critical sequence training for image captioning." In: **CVPR** (2017).
11. 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 visual question answering." In: **CVPR** (2018).
12. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need." In: **NeurIPS** (2017).
13. W. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." In: **ECCV** (2020).
14. X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. "Microsoft COCO captions: Data collection and evaluation server." In: (2015). arXiv:1504.00325.

Appendices

A The age-annotated dataset

To run the additional experiments of the LIC metric for bias in captioning models introduced in [1], we needed a captioned images dataset annotated for the protected attribute we wanted to study: age.
We decided to use the same images and captions used in the original paper, that is, a subset of the MSCOCO dataset [14]. As we were unable to find age annotations for this dataset, we manually classified 7,036 images as either "young" or "old" to use in our experiments. These images correspond to a subset of the intersection of the images used in the original paper by the models we wanted to study (SAT [9], OSCAR [13], NIC+ [2] and NIC+Equalizer [2]).

Figure 1. Some examples of elements of the age-annotated caption dataset. Each elements contains the id of the corresponding image, a label (either "young" or "old", five captions written by human annotators (one of which is shown) and captions predicted by the studied models: SAT, OSCAR, NIC+ and NIC+Equalizer. These examples have already had their agewords masked.

B List of age-related words

To better asses the usefulness of the LIC metric for measuring age bias, we masked some words related to age in the captions, and substituted them with a special token. The list of words we masked is:

child, children, young, baby, babies, kid, kids, little, boy, boys, girl, girls, old, man, men, woman, women, lady, ladies, gentleman, gentlemen, person, people, guy, guys, teenager, teenagers, teen, teens, adult, adults, elderly, elder.

C Using different vocabularies

One of the problems that arises when comparing human-generated and AI-generated captions is that humans tend to use a much richer vocabulary. The experiments conducted in the original paper [1] and the age extension presented in this study solve this by masking all words which do not appear in the model vocabulary with a special token when presenting them to the classifier.
However, we also tried a different approach. Instead of using the model vocabulary, we tried using the most common words in the human captions, that is, masking every word not among the n most common where n is the size of the vocabulary used by the studied model. Table 4 shows the LIC scores obtained when using this vocabulary.

Model	LSTM			BERT-ft			BERT-pre		
	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC
SAT [9]	48.7 ± 1.8	43.2 ± 1.1	5.5	51.7 ± 1.5	48.9 ± 1.3	2.8	37.0 ± 1.4	33.2 ± 0.8	3.8
OSCAR [13]	53.3 ± 1.7	43.2 ± 1.0	10.1	55.9 ± 2.0	50.3 ± 2.0	5.6	39.6 ± 1.8	33.5 ± 0.8	6.1
NIC+ [2]	43.1 ± 1.5	43.3 ± 1.1	-0.2	46.6 ± 1.7	48.9 ± 1.2	-2.4	35.5 ± 1.1	33.3 ± 0.7	2.1
NIC+Equalizer [2]	43.2 ± 1.6	43.0 ± 1.1	0.2	46.7 ± 1.6	48.8 ± 1.3	-2.1	34.8 ± 1.8	33.3 ± 0.8	1.5

Table 4. Age bias scores according to LIC, LIC_M, and LIC_D for several image captioning models, when using vocabularies consisting on the most common words in the human-produced captions. Captions are encoded with LSTM [6], BERT-ft (fine-tuned) or BERT-pre (pre-trained) [7].

We notice that these results also support the claims stated in the original paper, even if the LIC scores are noticeably different. Further research may be conducted on how the use of different vocabularies affect the LIC score.