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

Paula Antequera, Egoitz Gonzalez,
Marta Grasa, Martijn van Raaphorst

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

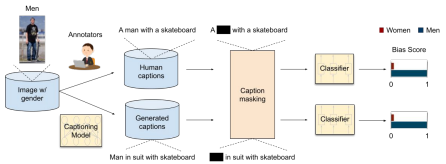
Bias is often present in large-scale captioning models, however there is no standard metric to measure it. In (Hirota et al., 2022), the *Leakage for Image Captioning* metric (LIC) is proposed for this objective. We aim to reproduce their results, leading to a confirmation or denial of the following claims stated in the paper:

- LIC is robust against encoders
- All models amplify both gender and race bias
- Racial bias is not as apparent as gender bias
- NIC+Equalizer (Burns et al., 2018) increases gender bias, but not racial bias

We extend these results by also performing experiments on age bias

Procedure

We use the following procedure to compute the LIC metric of a captioning model:



Pipeline for computing the bias score of a single captioned image, which is then used to compute the LIC metric of the model from which the caption was generated.

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

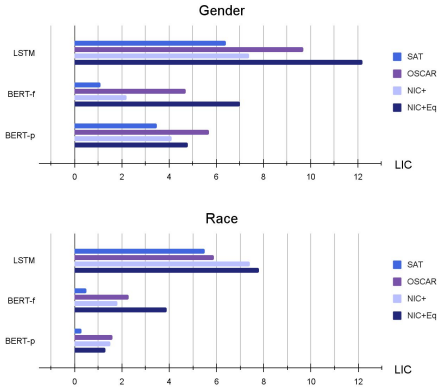
$$\mathrm{LIC}_D = \frac{1}{|\mathcal{D}|} \sum_{(y^*,a)\in\mathcal{D}} s^*_a(y^*) \mathbf{1}[f^*(y^*) = a]$$

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

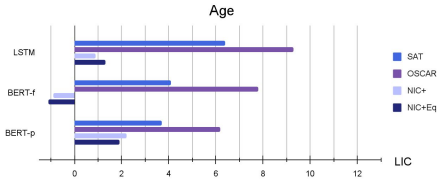
$\mathcal{D}$: test data
a : protected attribute
f : classifier
y : predicted attribute
s : **bias score**

Results

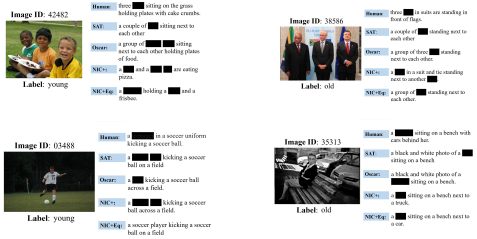
We replicate the experiments on gender and racial bias for the same captioning models and classifiers. We perform additional tests on the most relevant models to also measure age bias.



For gender and racial bias, we replicate the experiments presented in the original paper. For each model, we mask the words which indicate either gender or race in both the model generated and human produced captions. We then use one of three encoders to see if they can correctly predict the label. The most representative models are shown above



We also use a manually labeled dataset on age (examples below) to see if these results extend to other protected attributes. The results for this additional experiments are presented in the last bar plot.



Some examples of images and (masked) captions in our manually labeled dataset

Conclusions

Our experiments support all claims made in the original paper (Hirota et al., 2022), laid out in the abstract.

Regarding the additional experiments, the first and last claims still hold true when considering age bias, as LIC scores are consistent across classifiers for all models and NIC+Equalizer does not significantly increase bias compared to the baseline. However, the second claim is not supported, as not all models increase age bias. As for the third claim, age bias is not consistently more or less apparent than either gender or race bias.

References

Y. Hirota, Y. Nakashima, and N. Garcia (2022). "Quantifying Societal Bias Amplification in Image Captioning." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13450– 13459.

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

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan (2015). "Show and tell: A neural image caption generator." In: CVPR.

W. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al (2020). "Oscar: Object-semantics aligned pre-training for vision-language tasks." In: ECCV.