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

Machine Learning Reproducibility Challenge (MLRC), NeurIPS 2023

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

Paper to reproduce

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

Claims

1. LIC is robust against encoders. Its overall tendency is maintained across all language models (*LSTM, BERT-ft, BERT-pre*).
2. All models amplify both gender and race bias.
3. Racial bias is not as apparent as gender bias
4. *NIC+Equalizer* increases gender bias, but not racial bias, with respect to the baseline (*NIC+*).

LIC metric

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

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

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

$\mathcal{D}$: test data

a : protected attribute

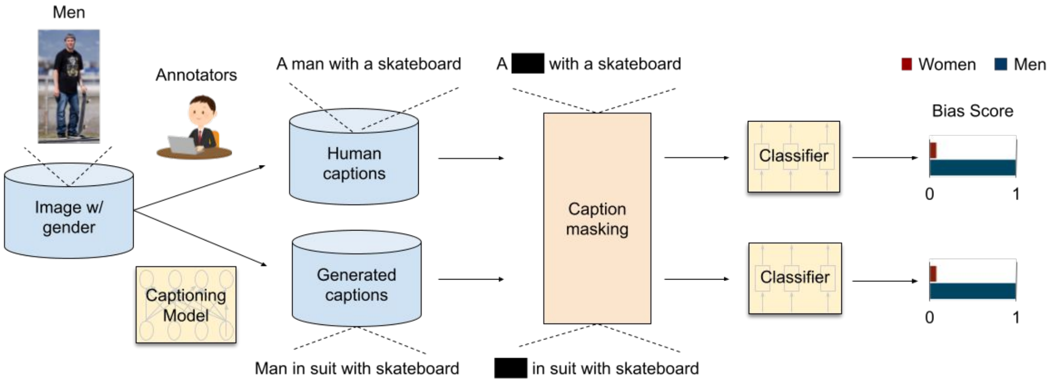
f : classifier

y : caption

s : **bias score**

Setup

- COCO dataset
- Gender + race bias



Hirota et al. 2022

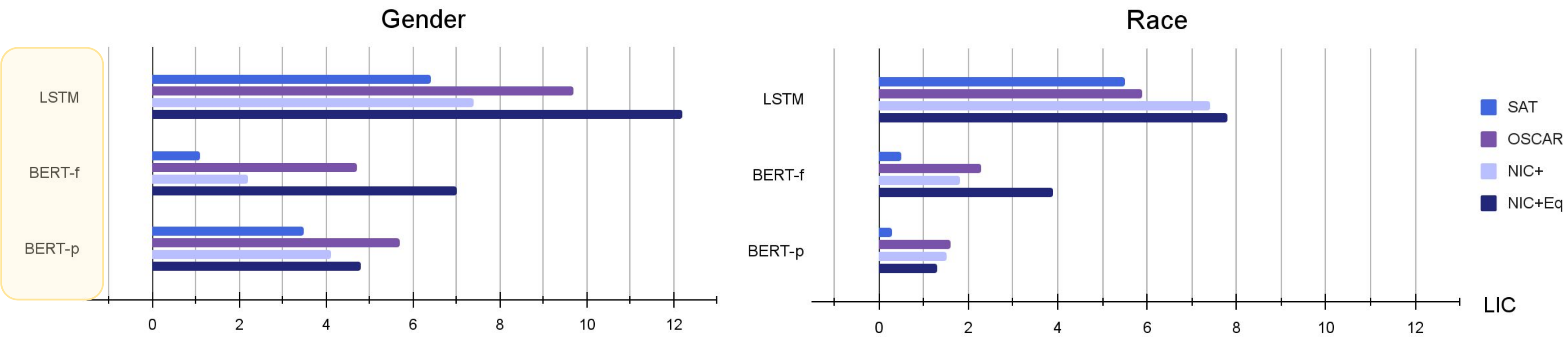
Models:

- CNN enc + LSTM dec: ***NIC, SAT, FC, Att2in, updn***
- Transformers: ***transformer, OSCAR***
- ***NIC+***: *NIC* + trained on gender bias
- ***NIC+Equalizer***: *NIC+* + gender bias mitigation

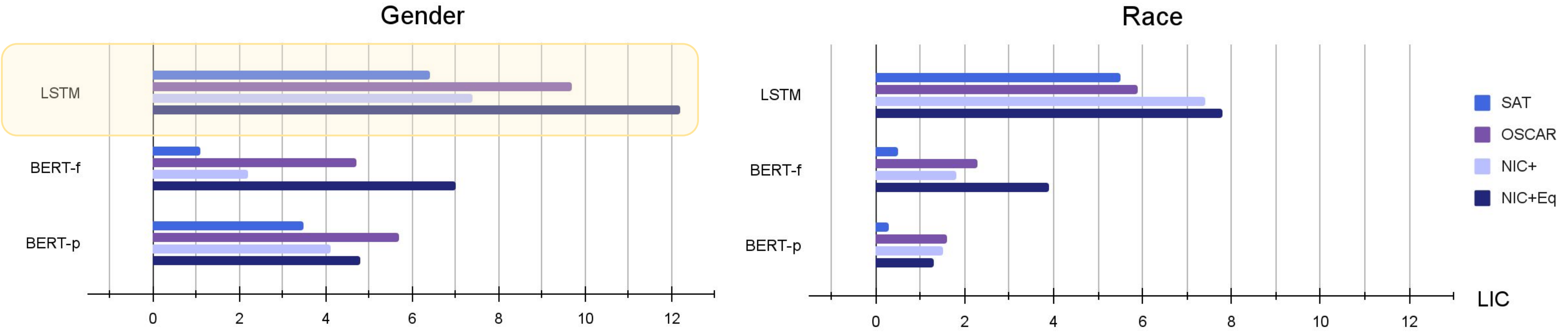
Classifiers:

- ***LSTM***
- ***BERT fine-tuned***
- ***BERT pre-trained***

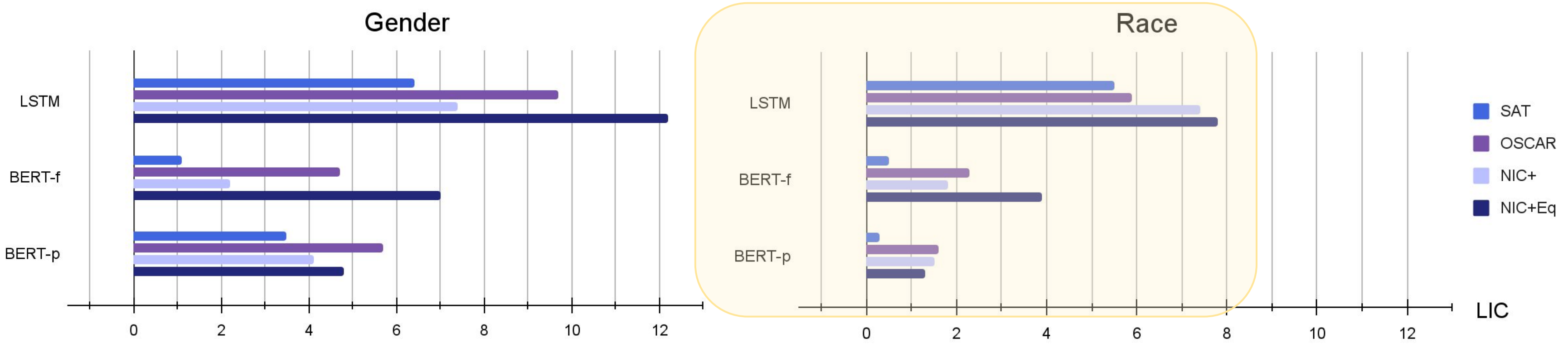
Claim 1: LIC is robust against encoders



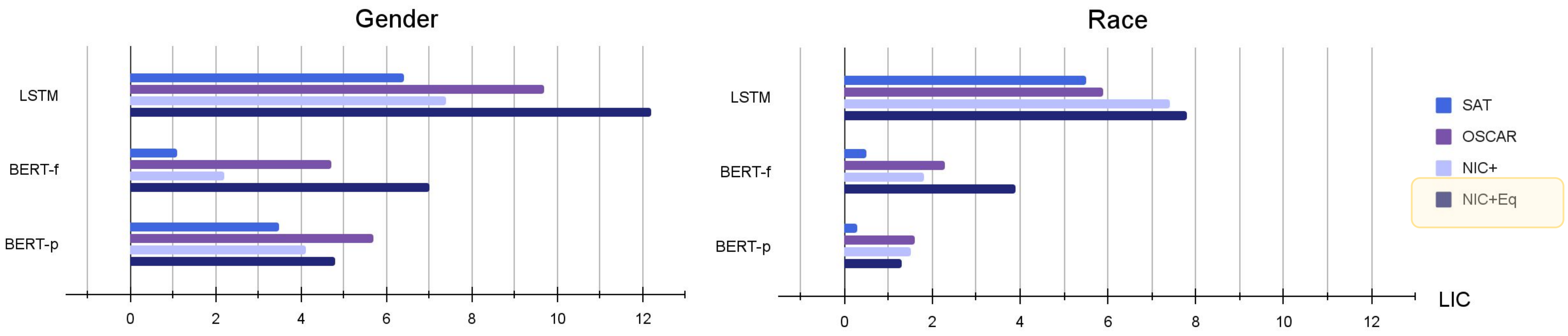
Claim 2: All models amplify gender and race bias



Claim 3: Racial bias is not as apparent as gender bias



Claim 4: *NIC+Equalizer* increases gender bias, but not racial bias, with respect to the baseline (*NIC+*)



Age: Hand-annotated data

- COCO dataset hand-annotated using our own tool

Image ID: 03488

Label: young

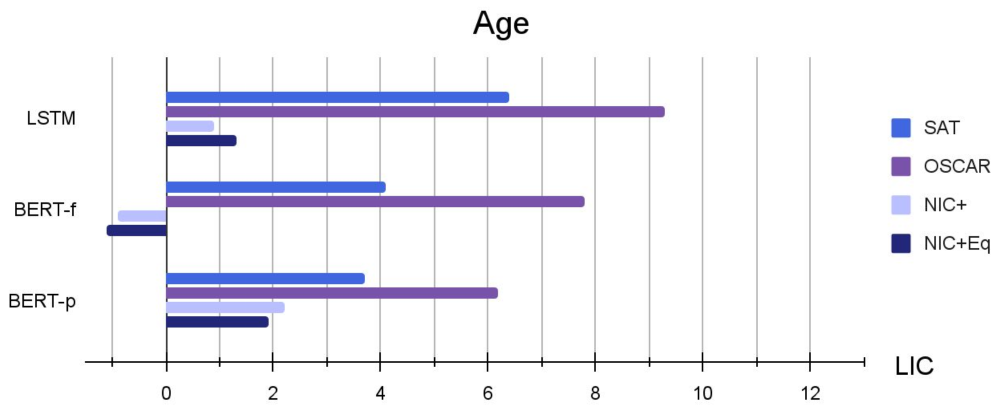
Human:	a [redacted] in a soccer uniform kicking a soccer ball.
SAT:	a [redacted] [redacted] kicking a soccer ball on a field
Oscar:	a [redacted] kicking a soccer ball across a field.
NIC+:	a [redacted] [redacted] kicking a soccer ball across a field.
NIC+Eq:	a soccer player kicking a soccer ball on a field

Image ID: 42752

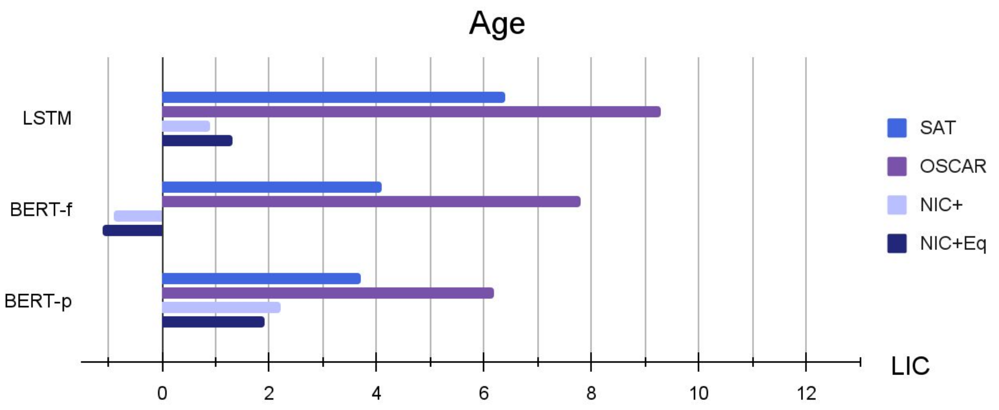
Label: old

Human:	a [redacted] using a laptop at a desk, with a briefcase next to it.
SAT:	a [redacted] sitting at a desk in a room
Oscar:	a [redacted] sitting at a table with a tray of food in front of him.
NIC+:	a group of [redacted] in a kitchen preparing food.
NIC+Eq:	a [redacted] sitting in front of a laptop computer.

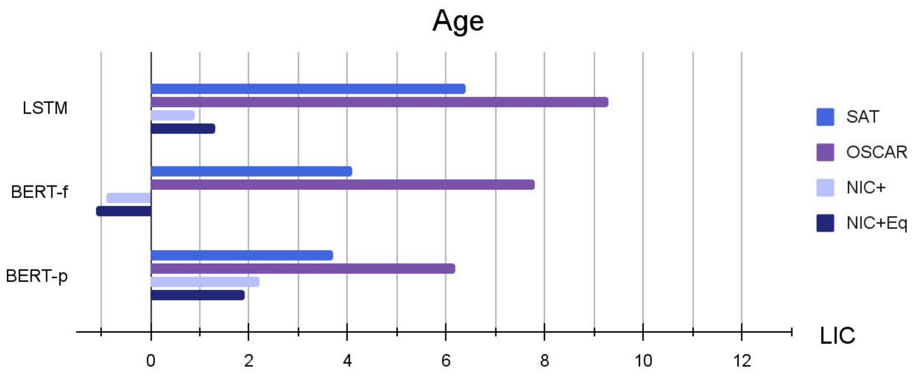
Claim 1: LIC is robust against encoders

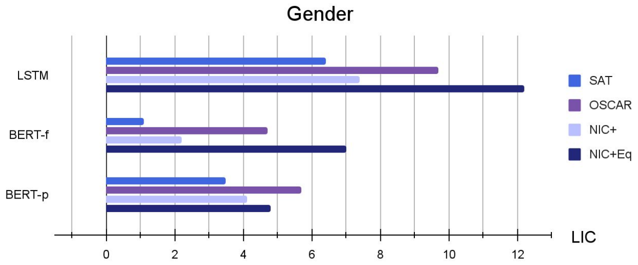


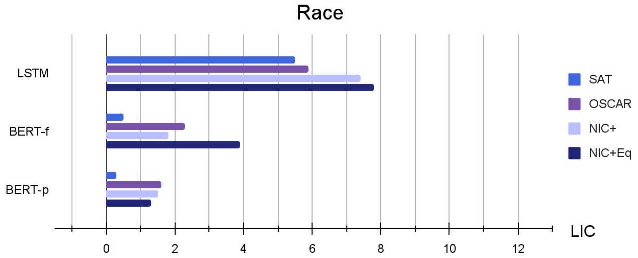
Claim 2: All models amplify age bias



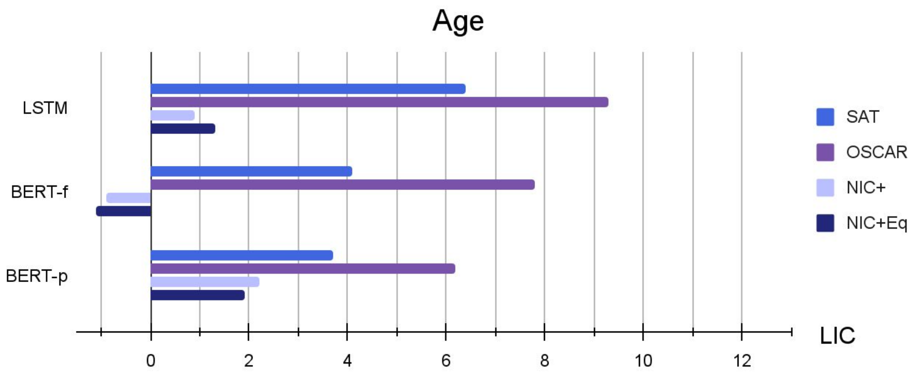
Is age bias as apparent as gender and race bias?

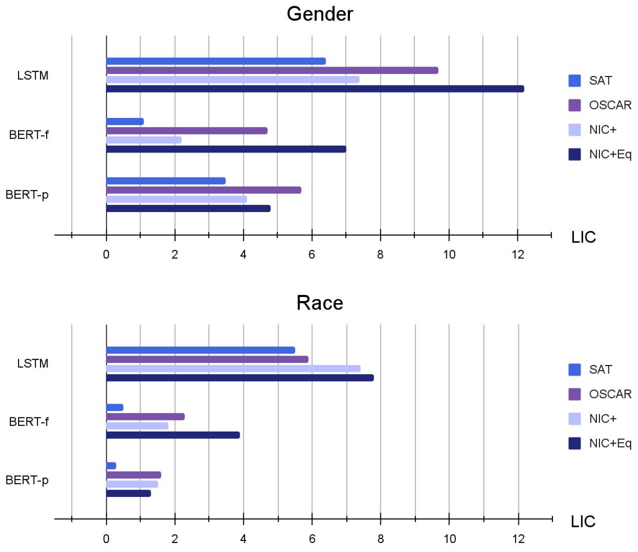






Does *NIC+Equalizer* increase age bias with respect to the baseline (*NIC+*)?





Conclusions

- Not difficult to reproduce
- Results align with original paper
- Extension also shows same trend

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

Machine Learning Reproducibility Challenge (MLRC), NeurIPS 2023

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