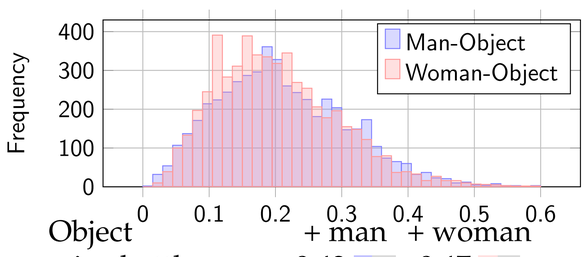
# Women Wearing Lipstick: Measuring the Bias Between an Object and Its Related Gender

Ahmed Sabir and Lluís Padró
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
asabir@cs.upc.edu, padro@cs.upc.edu

## Motivation

- Image captioning systems utilize the correlation between visual and co-occurring labels to predict an accurate static story in an image. However, this will result in a **gender bias** that relates to a specific gender.
- In this work, our primary focus is to examine whether there is a **stronger correlation** between the object and gender within image captioning systems, aiming to measure the degree of gender bias *e.g.* motorcycle bias toward men.
  - Therefore, in this work, unlike previous works, we examine the problem from a semantic perspective between the object and the gender.
- We thus propose a Gender Score via inspired-human judgment named Belief Revision, which can be used to (1) discover bias and (2) predict gender bias without *training* or *unbalancing* the dataset.



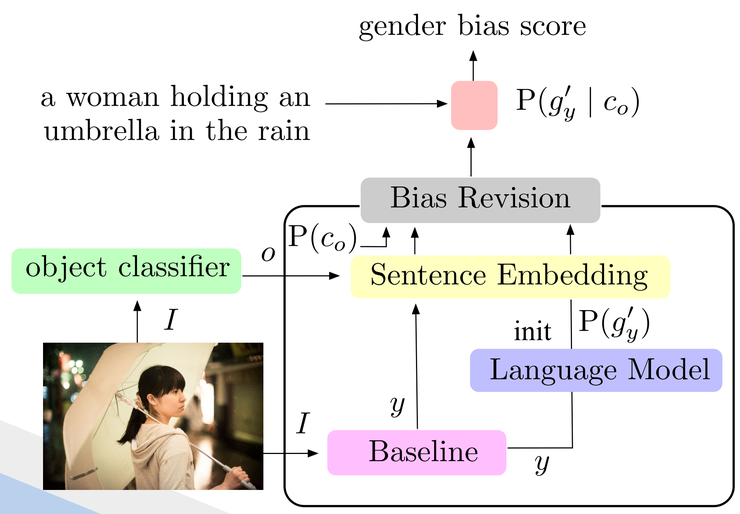

| Object | + man | + woman |
|---|---|---|
| wine bottle | 0.43 | 0.47 |
| umbrella | 0.28 | 0.30 |
| ski | 0.54 | 0.45 |
| cellular telephone | 0.55 | 0.54 |
| television | 0.35 | 0.30 |
| restaurant | 0.28 | 0.37 |

## Gender Score

The Gender Score is based on the visual likelihood revisions score (Sabir et al., 2022). This approach utilized Belief Revision to convert the similarity into a probability measure (Blok et al., 2003). Belief Revision is a process of formatting a belief by bringing into account a **new piece** of information. In our scenario, we revise gender bias using visual context and visual-to-gender similarity score.

**Model Architecture**: The main components of the object to gender Visual Bias Revisions *i.e.* Gender Score as follows:

- Language Model: initial bias without visual information.
- Visual Concept: the bias visual context object from the image.
- Similarity: measuring the degree of the object-to-gender bias.




The Visual likelihood Bias Revision block for Gender Score:

$$GS_a(y) = \frac{1}{|\mathcal{D}|} \sum_{(y,o)\in\mathcal{D}} P\left(g_y \mid c_o\right) = P(g_y)^{\alpha} \text{ where } \alpha = \left(\frac{1-sim(y,o)}{1+sim(y,o)}\right)^{1-P(c_o)}$$

- **Hypothesis** ($g$): The prior probabilities of original belief. As this approach is inspired by human-judgment *i.e.* condition is to start with an initial hypothesis. The hypothesis $P(g)$ is the baseline predicted output caption $y$ with the associated gender $a \in \{\text{man}, \text{woman}\}$ and needs to be initialized by a common observation general text such as a Language Model. We employ GPT-2 (Radford et al., 2019) to initialize the hypothesis $P(g')$, and we consider this as an **initial bias** without visual information.
- **Informativeness** ($c$): The information from the image causes the gender bias based likelihood revisions. We leverage visual classifiers to extract visual context *object* $o$ information from the image.
- **Similarities** $sim(y, o)$: The cause of the initial bias $P(g')$ revision is more likely if there is a close relation between $y$ with the gender $a$ and $o$ new visual information. We employ Sentence BERT (Reimers and Gurevych, 2019) to compute the Gender Object Distance *e.g.* similarity between the caption $y$ with the associated gender $a$ and its related visual context $o$ from the image.

## Experiments

### Dataset and Visual Context

- **Dataset:** We investigate the relation between gender and the objects that are mainly used in image captioning systems and datasets: Flickr30K and COCO datasets. For testing, we employ the standard Karpathy test split.
- **Visual Context:** We employ object classifiers to extract the visual context: (1) Resent152 with 1000 classes, (2) CLIP, and (3) Inception-ResNet FasterR-CNN (Huang et al., 2017) with 80 COCO categories (excluding *person* category).

### Experimental Results

Comparison results between Object Gender Co-Occ and our Gender Score Estimation on the Karpathy test split. Our score uses the object with context to predict the $<$ MASK $>$ gender. The proposed score measures gender bias more accurately, particularly when there is a strong object to gender bias relation.

| | Gender | | Bias Ratio | |
|---|---|---|---|---|
| Model | man | woman | to-m | to-w |
| Object Gender Co-Occ (Zhao et al., 2017) | | | | |
| Transformer (Vaswani et al., 2017) | 792 | 408 | 0.66 | 0.34 |
| AoANet (Huang et al., 2019) | 770 | 368 | 0.67 | 0.32 |
| Vilbert (Lu et al., 2020) | 702 | 311 | 0.69 | 0.30 |
| OSCAR (Li et al., 2020) | 845 | 409 | 0.67 | 0.32 |
| BLIP (Li et al., 2022) | 775 | 385 | 0.66 | 0.33 |
| TraCLIPS-Reward (Cho et al., 2022) | 769 | 381 | 0.66 | 0.33 |
| BLIP-2 (Li et al., 2023) | 695 | 356 | 0.66 | 0.33 |
| Gender Score (Gender Score Estimation) | | | | |
| Transformer | 616 | 217 | 0.73 | 0.26 |
| AoANet | 527 | 213 | 0.71 | 0.28 |
| Vilbert | 526 | 161 | 0.76 | 0.23 |
| OSCAR | 630 | 237 | 0.72 | 0.27 |
| BLIP | 554 | 240 | 0.69 | 0.30 |
| TraCLIPS-Reward | 537 | 251 | 0.68 | 0.31 |
| BLIP-2 | 498 | 239 | 0.67 | 0.32 |

### Qualitative Results

Example of the most common gender bias objects in Karpathy test split. The result shows that our score has similar results (bias ratio) to the existing Object Gender Co-Occ approach on the most biased objects toward men.

| | Bias Ratio Toward Men | | | | Bias Ratio Toward Women | | | |
|---|---|---|---|---|---|---|---|---|
| Model | skateboard | kitchen | motorcycle | baseball | skateboard | kitchen | motorcycle | baseball |
| Object Gender Co-Occ (Zhao et al., 2017) | | | | | | | | |
| Transformer | 0.96 | 0.50 | 0.83 | 0.75 | 0.05 | 0.50 | 0.16 | 0.25 |
| AoANet | 0.97 | 0.51 | 0.85 | 0.81 | 0.02 | 0.48 | 0.14 | 0.18 |
| Vilbert | 0.96 | 0.47 | 0.84 | 0.66 | 0.03 | 0.52 | 0.15 | 0.33 |
| OSCAR | 0.97 | 0.58 | 0.82 | 0.90 | 0.02 | 0.41 | 0.18 | 0.09 |
| BLIP | 0.96 | 0.52 | 0.88 | 0.97 | 0.03 | 0.47 | 0.11 | 0.02 |
| TraCLIPS-Reward | 0.89 | 0.48 | 0.93 | 0.50 | 0.10 | 0.51 | 0.06 | 0.50 |
| BLIP-2 | 0.94 | 0.57 | 0.88 | 0.90 | 0.05 | 0.42 | 0.11 | 0.10 |
| Gender Score | | | | | | | | |
| Transformer | 0.96 | 0.51 | 0.83 | 0.61 | 0.03 | 0.48 | 0.16 | 0.38 |
| AoANet | 0.97 | 0.46 | 0.84 | 0.82 | 0.02 | 0.53 | 0.15 | 0.17 |
| Vilbert | 0.96 | 0.53 | 0.84 | 0.65 | 0.03 | 0.46 | 0.15 | 0.34 |
| OSCAR | 0.98 | 0.42 | 0.78 | 0.83 | 0.01 | 0.57 | 0.21 | 0.16 |
| BLIP | 0.96 | 0.50 | 0.86 | 0.98 | 0.03 | 0.49 | 0.13 | 0.01 |
| TraCLIPS-Reward | 0.88 | 0.43 | 0.92 | 0.50 | 0.11 | 0.56 | 0.07 | 0.49 |
| BLIP-2 | 0.93 | 0.56 | 0.82 | 0.89 | 0.06 | 0.43 | 0.17 | 0.10 |

**Visual:** tennis ball
**Caption:** a $<$ MASK $>$ hitting a tennis ball on a tennis court
**Gender Object Distance:** man 0.44 woman **0.46**
**Gender Score:** man 0.45 woman 0.45

**Visual:** umbrella
**Caption:** a $<$ MASK $>$ holding an umbrella in the rain
**Gender Object Distance:** man 0.20 woman 0.20
**Gender Score:** man 0.12 woman **0.13**

**Visual:** laptop
**Caption:** a $<$ MASK $>$ sitting on a couch with two laptops
**Gender Object Distance:** man 0.43 woman 0.42
**Gender Score:** man 0.25 woman 0.25

**Visual:** paddle
**Caption:** a $<$ MASK $>$ riding a wave on top of a surfboard
**Gender Object Distance:** man 0.16 woman 0.11
**Gender Score:** man 0.33 woman 0.30

## Conclusion

- We investigate the bias between objects and gender in image captioning.
- Our results show that not all objects have a strong gender bias, and only in special cases does the object have a strong gender bias.
- We also propose a Gender Score as an additional metric to the Object-Gender Co-Occ method, which can used without training or unbalancing the dataset.

**Code:** https://github.com/ahmedssabir/GenderScore