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

EMNLP Findings 2023

Background

- Image captioning models achieved notable benchmark performance in utilizing the correlation between **visual** and **co-occurring labels** to generate an accurate image description.
- However, this often results in a gender bias that relates to a specific gender, such as confidently identifying a woman when there is a kitchen in the image.

(a) Caption: a man eating a slice of pizza

(b) Caption: a woman and a baby at table with a cake

Background

- Image captioning models achieved notable benchmark performance in utilizing the correlation between **visual** and **co-occurring labels** to generate an accurate image description.
- However, this often results in a gender bias that relates to a specific gender, such as confidently identifying a woman when there is a kitchen in the image.

(a) Caption: a man eating a slice of pizza

(b) Caption: a woman and a baby at table with a cake

Contributions

- We investigate the gender bias object relation in image captioning systems. Our results show that only gender-specific objects have a strong gender bias.
- We propose a Gender Score that (1) discovers gender-to-object bias relation and (2) predicts the biased gender **without training or unbalancing** the dataset.

Visual information: pizza
Gender Score: **man 0.39** Woman 0.36

Visual information: dining table
Gender Score: man 0.14 **Woman 0.16**

Approach

Gender Score: Visual Bias Revision

- 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 this work, we revise the gender bias using object context from the image.

$$P(Q_c|Q_a) = P(Q_c)^{\alpha}$$

- **Hypothesis**: $\mathsf{P}(Q_c)$ Original belief - initial bias
- **Informativeness**: $1 - \mathsf{P}(Q_a)$ New information - object context
- **Similarities**: $\alpha = \left[\frac{1 - \text{sim}(a,c)}{1 + \text{sim}(a,c)}\right]^{1-\mathsf{P}(Q_a)}$ Degree of bias revision

Approach

Gender Score: Visual Bias Revision

- The Gender Score can be computed as the conditional probability of the caption g_y with the associated gender $a \in \{\text{man, woman}\}$ given the object information. $\mathcal{D}$ is the generated caption with the gender.
- α is a factor that calculates the degree of bias based on the similarity or relatedness between the object o and the caption with associated gender $sim(y, o)$. $\mathsf{P}(c_o)$ is confident of the bias object in the image.

$$\begin{aligned} \mathsf{GS}_a(y) &= \frac{1}{|\mathcal{D}|} \sum_{(y,o)\in\mathcal{D}} \mathrm{P}\left(g_y \mid c_o\right) \\ &= \mathrm{P}(g_y)^{\alpha}\ ,\ \alpha = \left(\frac{1-\mathrm{sim}(y,o)}{1+\mathrm{sim}(y,o)}\right)^{1-\mathsf{P}(c_o)} \end{aligned}$$

Approach

Gender Score: Visual Bias Revision

- The Gender Score can be computed as the conditional probability of the caption g_y with the associated gender $a \in \{$man, woman$\}$ given the object information. $\mathcal{D}$ is the generated caption with the gender.
- α is a factor that calculates the degree of bias based on the similarity or relatedness between the object o and the caption with associated gender $sim(y, o)$. $\mathrm{P}(c_o)$ is confident of the bias object in the image.

$$\mathsf{GS}_a(y) = \frac{1}{|\mathcal{D}|} \sum_{(y,o)\in\mathcal{D}} \mathrm{P}\left(g_y \mid c_o\right)$$

$$= \underbrace{\mathrm{P}(g_y)^{\alpha}}_{\text{initial bias}}, \quad \underbrace{\alpha = \left(\frac{1 - \mathrm{sim}(y, o)}{1 + \mathrm{sim}(y, o)}\right)^{1-\mathrm{P}(c_o)}}_{\text{degree of object bias}}$$

Approach

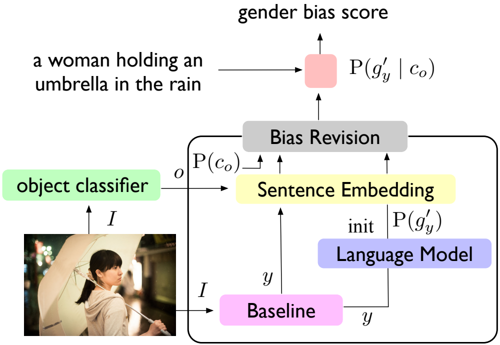
Gender Score: Visual Bias Revision

- The main components of the Gender Score are:

  Language Model: initial bias without visual information.

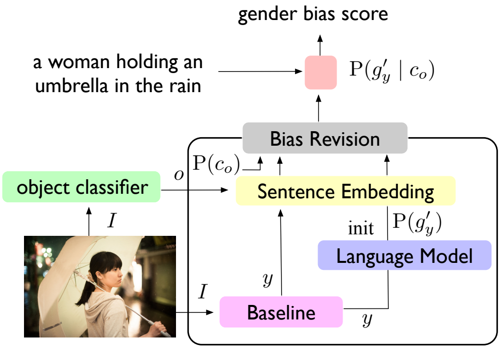
  Visual Concept: the bias object from the image.

  Similarity: measuring the degree of the object-to-gender bias.



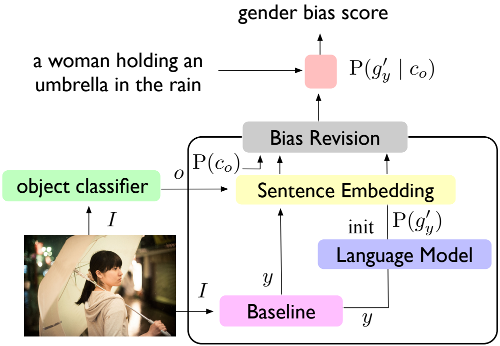
Approach

- **Language Model Block:** the hypothesis *i.e.* caption $\mathrm{P}(g_y)$ needs to be initialized by a common observation from general text $\mathrm{P}(g'_y)$.
- We employ GPT-2 (Radford et al., 2019) to initialize the hypothesis. This is an **initial bias** without any visual information.



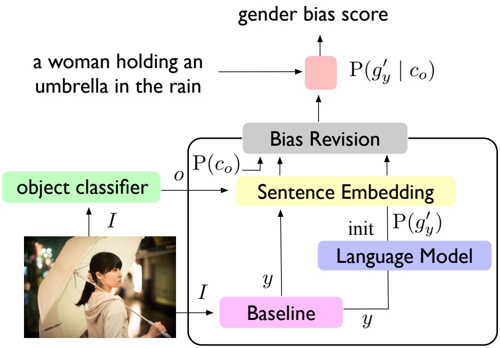
Approach

- **Visual Block:** the object informativeness o is the new information set with confident $\mathrm{P}(c_o)$ that causes the $\mathrm{P}(g'_y)$ caption bias revision.
- We leverage ResNet, CLIP, and Inception-ResNet v2 based faster R-CNN to extract the visual information from the image.



Approach

- **Similarity Block:** initial $P(g'_y)$ bias revision is more likely if there is a close relation between the caption y with the gender a and object o.
- We employ S-BERT to compute the Cosine Distance *i.e.* semantic similarity between the hypothesis y_a caption and its related o object.



Dataset

- **Dataset.** We investigate the relation between gender bias and the objects that are mainly used in image captioning systems, and more precisely, the widely used manually annotated image caption datasets: Flickr30K and COCO Captions datasets.
- **Visual Context.** We employ object classifiers to extract the top-k visual context: Resent152 1000 classes, CLIP, and Inception-ResNet FasterR-CNN 80 object categories (excluding *person* category).

Visual: tennis ball
Caption: a woman hitting a tennis ball on a tennis court

Visual: hotdog
Caption: a man eating a hot dog

Visual: umbrella
Caption: a woman holding an umbrella in the rain

Visual: motor scooter
Caption: a man riding a motorcycle on a road

Experimental Results

- We investigate the gender-to-object semantic relation for image captioning at the word level *i.e.* **gender-object** and the sentence level with captions *i.e.* **gender-caption** in the training dataset.
- The results show there is a slight bias toward men.
- GloVe and GN-GloVe (balanced) show identical results, *e.g.* bicycle-gender (GloVe: 0.31 and 0.27, ratio=0.53) and (GN-GloVe:0.15 and 0.13, ratio=0.53).

	COCO Captions					Flickr30K				
	Avg: Gender Object Distance			Ratio		Avg: Gender Object Distance			Ratio	
Model	+ person	+ man	+ woman	to-m	to-w	+ person	+ man	+ woman	to-m	to-w
Word2Vec (Mikolov et al., 2013)	0.101	0.116	0.124	0.48	0.51	0.116	0.142	0.154	0.47	0.52
GloVe (Pennington et al., 2014)	0.146	0.175	0.169	0.50	0.49	0.131	0.170	0.168	0.50	0.49
Fasttext (Bojanowski et al., 2017)	0.180	0.200	0.191	0.51	0.48	0.146	0.196	0.191	0.50	0.49
GN-GloVe (Zhao et al., 2018)	0.032	0.055	0.054	0.50	0.49	0.024	0.085	0.088	0.49	0.50
SBERT-NLI (Reimers et al., 2019)	0.124	0.155	0.128	0.54	0.45	0.121	0.167	0.129	0.56	0.43
SimCSE-RoBERTa (Gao et al., 2021)	0.194	0.137	0.093	0.59	0.40	0.189	0.140	0.107	0.56	0.43
InfoCSE-RoBERTa (Wu et al., 2022)	0.199	0.222	0.211	0.51	0.48	0.228	0.265	0.241	0.52	0.47

Experimental Results

- To evaluate the Gender Score, we compare our score against the existing approach Object Gender Co-Occ. Our Gender Score uses the object with context to predict the $<$ MASK $>$ biased 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

Experimental Results

- Example of the most common gender bias objects in COCO Captions Karpathy test split.
- The result shows that our score **bias ratio** aligns closely with the existing Object Gender Co-Occ approach when applied to the most gender-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

Qualitative Results

- Examples of Gender Score Estimation and Gender Object Distance *i.e.* Cosine Distance in predicting the biased gender.
- The proposed score predicts the strong gender object bias relation *e.g.* *paddle*, *surfboard*, and balances the object bias *e.g.* *tennis* and *laptop*.

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: 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: 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: 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

Contributions

- We investigate the relation between objects and gender bias in image captioning. Our results show that not all objects exhibit gender bias, and only in special cases does an object have a strong gender bias.
- We also propose a Gender Score that can be used as an additional metric to the existing Object-Gender Co-Occ method.

Thank You