Supplementary Materials for Understanding and Evaluating Racial Biases in Image Captioning

Dora Zhao Angelina Wang Olga Russakovsky
Princeton University
{dorothyz,angelina.wang,olgarus}@cs.princeton.edu

A. Comparing collected gender annotations with automatically derived ones

We explore extending the schema introduced in Sec. 3.2 for deriving gender labels from captions in three ways: 1) only labeling images where there is a `person` who has a bounding box greater than 5,500 pixels, 2) expanding the list of gendered words beyond "man" and "woman", and 3) having different cutoffs for how many captions (of the 5 per image) need to mention a gender for the image to be labeled. We call the use of the gendered set {man, woman} "few," and that of our expanded set "many."

Our expanded set "many" consists of the following words: ["male", "boy", "man", "gentleman", "boys", "men", "males", "gentlemen"] and ["female", "girl", "woman", "lady", "girls", "women", "females", "ladies"].

Our results in Fig. 1 show that while these extensions significantly increase both the number of images correctly labeled and the accuracy of labeled images, all methods are inaccurate and/or incomplete. The gender labels derived from captions remain highly imperfect, as expected, cautioning against automated means of gender derivation [40].

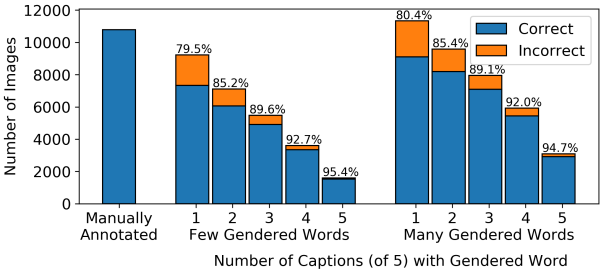

Figure 1: Comparison of various ways of automatically deriving image-level gender annotations from existing image captions. "Few" and "Many" gendered words refers to the size of the set of gendered words considered, and numbers refer to required captions that mention a gendered word, higher numbers limiting the images that can be labeled. Percentage over each bar indicates accuracy. All methods are imperfect and noisy, cautioning against the use of automatically-derived gender annotations.

B. Racial descriptors

When searching for descriptors of race and ethnicity in Sec. 4.1, we first convert the captions to lowercase. We then use the following keywords ["white", "Caucasian", "Black", "African", "Asian", "Latino", "Latina", "Latinx", "Hispanic", "Native", and "Indigenous"] — also in lowercase — to query the captions.

C. Caption performance

In Sec. 4.2 we assess the differences in caption performance for BLEU [55], CIDER [68], and SPICE [2] when evaluated on the COCO 2014 validation set. We extend this analysis by providing the overall scores across the four image captioning models and looking at two additional automated image captioning metrics.

To start, we look at the performance for our four models. As seen in Tbl. 1, **Oscar** has the best performance across all metrics. Further, we see that the newer transformer-based models outperform older models (e.g. **FC**, **Att2in**, and **DiscCap**) across all metrics as well.

We also report the differences in performance between `lighter` and `darker` images for two commonly used image captioning metrics — METEOR [5] and ROUGE [46] (Tbl. 2). Similar to the results for BLEU and CIDER, the differences for METEOR and ROUGE are greater for **Att2in**, **Transformer**, and **Oscar**. **AoANet** also shows some slight differences in performance for METEOR and ROUGE. This supports our observation that the better performing captioning models also tend to show greater discrepancies in performance between `lighter` and `darker` images.

Table 1: The caption performance as measured by BLEU [55], CIDEr [68], and SPICE [2] multiplied by 100 on the COCO 2014 validation dataset. Error bars represents 95% confidence intervals across random seeds used to train 5 models per architecture.

	BLEU	CIDEr	SPICE
FC [59]	28.2 ± 0.4	87.2 ± 0.4	16.9 ± 0.2
Att2in [59]	31.5 ± 0.2	94.2 ± 0.3	18.4 ± 0.1
DiscCap [50]	23.7 ± 0.7	71.1 ± 2.3	18.8 ± 0.2
Transformer [67]	33.9 ± 0.6	97.2 ± 2.2	20.2 ± 0.2
AoANet [33]	38.7 ± 1.6	116.2 ± 4.6	22.2 ± 0.6
Oscar [45]	40.0 ± 0.4	120.0 ± 0.5	23.0 ± 0.3

D. Vocabulary differences coefficients

In Sec. 4.4 we explore the different word choices in the captions describing `lighter` and `darker` images. We provide the most predictive words for `lighter` and `darker` across the manual captions and the automatically generated captions in Tbl. 3.

Table 2: The differences in captioning performance (score on `lighter`- score on `darker`) as measured by BLEU [55], METEOR [5], ROUGE [46], CIDEr [68], and SPICE [2] multiplied by 100 on the COCO 2014 validation dataset. Error bars represent 95% confidence intervals across random seeds used to train 5 models per architecture

	BLEU Δ	METEOR Δ	ROUGE Δ	CIDEr Δ	SPICE Δ
FC [59]	0.5 ± 0.5	-0.1 ± 0.2	0.0 ± 0.5	-0.8 ± 1.8	0.2 ± 0.3
Att2in [59]	2.4 ± 0.4	0.5 ± 0.1	0.8 ± 0.2	2.7 ± 0.7	0.0 ± 0.1
DiscCap [50]	0.3 ± 0.5	0.0 ± 0.3	0.0 ± 0.3	0.6 ± 0.7	0.0 ± 0.2
Transformer [67]	2.5 ± 0.9	0.6 ± 0.3	1.1 ± 0.4	3.2 ± 1.2	-0.1 ± 0.3
AoANet [33]	1.8 ± 0.8	0.3 ± 0.3	0.4 ± 0.3	1.9 ± 1.6	0.0 ± 0.2
Oscar [45]	3.0 ± 1.1	0.5 ± 0.4	0.8 ± 0.3	3.0 ± 1.1	0.1 ± 0.3

Table 3: The most predictive words of `lighter` and `darker` images for a logistic regression trained on manual and automated captions of images for which visual content has been controlled. Lower coefficients are more predictive of `lighter`. For the automated captions, the coefficients are averaged across the 5 models per architecture.

Human		FC [59]		Att2in [59]		DiscCap [50]		Transformer [67]		AoANet [33]		Oscar [43]	
Train	-0.13	Parked	-0.20	Table	-0.28	Tree	-0.26	Tree	-0.24	Bathroom	-0.22	Brown	-0.21
Cake	-0.12	Young	-0.15	Tree	-0.22	People	-0.17	Snow	-0.17	Tree	-0.19	Baseball	-0.18
Man	-0.11	Holding	-0.15	City	-0.22	Brown	-0.17	Two	-0.15	Baseball	-0.18	Bathroom	-0.16
Playing	-0.10	Wearing	-0.14	Field	-0.19	Two	-0.15	Bear	-0.14	Brown	-0.17	Person	-0.15
Covered	-0.07	Playing	-0.13	Girl	-0.17	Table	-0.13	Baseball	-0.12	Wearing	-0.16	City	-0.15
Pizza	0.12	Sign	0.17	Walking	0.16	Beach	0.15	Desk	0.12	Umbrella	0.15	Standing	0.14
Umbrella	0.13	Desk	0.18	Clock	0.17	Group	0.20	Cat	0.13	Player	0.16	Their	0.15
Tennis	0.14	Boy	0.18	Train	0.17	Yellow	0.20	Food	0.14	Sink	0.17	Computer	0.16
Clock	0.18	Umbrella	0.20	Sink	0.19	Umbrella	0.21	Men	0.14	Road	0.22	Sign	0.18
Cat	0.20	People	0.20	Food	0.19	Sitting	0.25	Sign	0.21	Green	0.23	Player	0.19