Mitigating Gender Bias in Captioning Systems

Ruixiang Tang  Mengnan Du  Yuening Li  Zirui Liu  Xia Hu
Department of Computer Science and Engineering
Texas A&M University
{rxtang, dumengnan, liyuening, tradigrada, xiahu}@tamu.edu

Abstract

Recent studies have shown that captioning datasets, such as the COCO dataset, may contain severe social bias which could potentially lead to unintentional discrimination in learning models. In this work, we specifically focus on the gender bias problem. The existing dataset fails to quantify bias because models that intrinsically memorize gender bias from training data could still achieve a competitive performance on the biased test dataset. To bridge the gap, we create two new splits: COCO-GB v1 and v2 to quantify the inherent gender bias which could be learned by models. Several widely used baselines are evaluated on our new settings, and experimental results indicate that most models learn gender bias from the training data, leading to an undesirable gender prediction error towards women. To overcome the unwanted bias, we propose a novel Guided Attention Image Captioning model (GAIC) which provides self-guidance on visual attention to encourage the model to explore correct gender visual evidence. Experimental results validate that GAIC can significantly reduce gender prediction error, with a competitive caption quality. Our codes and the designed benchmark datasets are available at https://github.com/CaptionGenderBias2020.

1 Introduction

Automatically understanding and describing visual contents is an important and challenging interdisciplinary research topic [1, 2, 3, 4, 5, 6]. Over the past years, a lot of efforts have been dedicated to improving the overall caption quality, without considering unwanted bias learned by models. In this work, we focus on gender bias and investigate the gender words (e.g., woman and man) generated by the learning models. A biased model may rely on incorrect visual features to provide gender descriptions, e.g., describing a person as a woman because of the kitchen background, which leads to incorrect gender prediction when evaluating on samples without a learned prior.

Gender bias widely exists in captioning models mainly because of two reasons. First, many caption datasets originally contain severe gender bias. For example, COCO dataset [7] has an imbalanced 3:1 men to women ratio and the object-gender joint distribution further exacerbates the imbalance, e.g., 95% skateboard images co-occur with men. The biased dataset not only results in a biased model, but also makes it hard to detect the bias learned by models. For instance, in COCO dataset, images with women comprise a much smaller portion that gender prediction error could be ignored. Second, existing training paradigm makes models prone to replicate bias in datasets. Our preliminary experimental results show that most models inevitably learn bias existing in the training dataset.

A straightforward solution to mitigate the gender bias is to train the model based on a gender-balanced dataset. Unfortunately, our experimental results indicate that simply balancing image numbers for each gender has limited improvement in bias mitigation. From the training perspective, an alternative approach is to increase the loss weight of gender words, which also doesn’t achieve a satisfactory result. In addition to designing mitigation algorithms, another challenge is that gender bias may be underestimated when models are evaluated on the test set with similar bias to training samples.

Preprint. Under review.
To bridge the gap, we propose new benchmark datasets to quantify gender bias. In this work, we focus on the COCO captioning dataset and create two new splits. COCO-GB v1 is designed to systematically measure gender bias in existing models. COCO-GB v2 is designed to further assess capabilities of models in gender bias mitigation. We also propose a novel Guided Attention Image Captioning model (GAIC) to mitigate gender bias. GAIC has two complementary streams to encourage the model to explore correct gender features. The training pipeline can seamlessly add extra supervision to accelerate the self-exploration process. Besides, GAIC is model-agnostic and can be easily applied to various captioning models.

We report the performance of several baseline models on our new benchmark datasets. The key observation is that most models learn or even exaggerate gender bias contained in the training data, which causes a significantly higher gender prediction error for women. Another important finding is that common evaluation metrics, such as BLEU [8] and CIDEr [9], mainly focus on the overall caption quality and are not sensitive to gender error. Our experimental results show that a model achieves competitive caption quality scores even when it has misclassified 27% women into men, which further demonstrates that the high performance achieved by current models should be revisited. Experimental results validate that our proposed GAIC method can significantly reduce gender prediction error, with a competitive caption quality. Visualization evaluations further indicate that GAIC learns to utilize correct gender evidence for gender word generation.

2 COCO-GB: Dataset Creation and Analysis

The COCO-GB dataset are created for quantifying gender bias in models. We construct COCO-GB v1 based on a widely used split and create a gender-balanced secret test dataset. COCO-GB v2 is created by reorganizing the train/test split so that the gender-object joint distribution in training set is very different from testing set. Procedures of dataset creation are shown as follows.

Gender Labeling: An important reason why previous work ignores the gender bias problem is that COCO dataset doesn’t have explicit gender annotations. Hence our first step is to annotate the gender of people in the images. Because many images do not have a clear human face, normal face recognition models cannot provide a reliable gender prediction. Alternatively, inspired by [10], we utilize the 5 training captions available for each image to annotate the gender. Images are labeled as “men” if at least one of the five descriptions contain the male gender words and do not include female gender words. Similarly, images are labeled as “women” if at least one of the five descriptions contain the female gender words and do not include male gender words. Images that mentioned both “man” and “woman” are discarded. To further improve the labeling accuracy, we only consider images that contain one major person and also ignore the picture in which person in the image is too small. (Gender words list and examples of gender labeling are shown in Sec. A.1).

COCO-GB v1: COCO-GB v1 is created for evaluating gender bias in existing models. Hence we follow a widely used Karpathy split [3], and create a gender-balanced secret test dataset. Captioning models trained on karpathy split can directly evaluate their gender bias on this secret test set without retraining. Comparing to previous work [10] which only balance the image number of each gender, we further balance the gender-object joint distribution in the secret dataset. In this way, we can comprehensively show the bias learned by the model. We compute the gender bias towards men for each object category by metrics proposed in [11]:

$$\frac{\text{count}(\text{object, men})}{\text{count}(\text{object, men}) + \text{count}(\text{object, women})},$$

where men and women refers to images labeled as “men” and “women”. Ideally, an object with a bias ratio of 0.5 indicates that women and men has the same probability to co-occur with it. We show gender bias of several objects in Fig. [4]. Our statistical analysis show that COCO training dataset has a severe gender bias that training data has an average bias ratio of 0.65, and 90% of objects are biased towards men. Similar bias is also found in the original test split, which demonstrates that directly evaluating models on original test dataset can underestimate the bias learned by models. We then utilize a greedy algorithm to select a secret test dataset from the original test split so that each category has a nearly balanced gender ratio. Finally, our selected secret test dataset has 500 images for each gender and obtains an average bias ratio of 0.52. We then utilize this COCO-GB v1 dataset to evaluate existing models and quantify the unwanted gender bias captured by models.
Figure 1: We select several objects and show their gender bias in (a) COCO training set, (b) COCO original test set and (c) COCO-GB v1 secret test set. There is a significant bias in training set, more than 90% objects have higher probability to co-occur with men. Similar bias also exists in the original test set, while the secret test dataset in COCO-GB v1 has balanced the gender ratio.

COCO-GB v2: This dataset is designed to further assess the robustness of captioning models when exposed to novel gender-object pairs at test time. To create the new split, we first sort 80 object categories in COCO according to their gender bias. Unlike creating a balanced test dataset in COCO-GB v1, we start from the most biased object and greedily add selected data into the test set. As a result, the distribution difference has been dramatically enlarged between the training and testing dataset. We guarantee that there are sufficient images from each category during training, but at test time model will face novel compositions of gender-object pairs, e.g., women with the skateboard. The final split has 118,062/5,000/10,000 images in train/val/test respectively. We utilize this dataset to further evaluate our proposed bias mitigation approach. (Gender-object joint distribution of COCO, COCO-GB v1 and COCO-GB v2 are shown in Sec. A.2)

3 Benchmarking Captioning Models on COCO-GB v1

To reveal the gender bias in existing models, we utilize the gender prediction performance to quantify bias learned by models [10, 11]. Models are trained on Karpathy split, obtain caption quality from original test split, and evaluate gender predicition performance on the COCO-GB v1 secret test dataset.

Baselines: We consider models both with and without attention mechanism in the experiment where “attention” refers to models have ability to focus on specific image regions [12]. For non-attention models, we consider FC [13] which initialize LSTM with features extracted directly by CNN, and LRCN [14] which leverages visual information at each time step. For attention models, we select Att [6], which firstly applies visual attention mechanism in caption generation. AdapAtt [15], which automatically determines when and where to utilize visual attention. Att2in [13], which modifies the architecture of Att, and inputs the attention features only to the cell node of the LSTM. Besides, we also choose models that utilize extra visual grounding information. TopDown [16] proposes a specific ?top-down attention? mechanism based on Faster R-CNN model [17]. NBT [18] generates the sentence ?template? with slot locations, and fill slots by an object detection model.

Learning Objective and Implementation: Baselines are trained with cross-entropy loss. Besides, FC, LRCN, Att2in and TopDown model are also trained by self-critical loss [13], which uses reinforcement learning to directly optimize the non-differentiable CIDEr metric. For a fair comparison, all baselines utilize visual features extracted by the ResNet-101 network, NBT and TopDown model obtain extra grounded information from the Faster R-CNN model.

Results analysis: In Tab. 1, we report caption quality as well as gender prediction performance where gender is predicted correctly/incorrectly or “neutrally” when no gender-specific words are generated. An unbiased model should have similar outcome and low error rate for each gender. From the results, we reach the following conclusions.

- Across all models, the error rate of women is substantially higher than men. The average error rate of all models is 16.7% and 7.7% for women and men respectively. An interesting finding is that models with attention mechanism have much higher women errors (average of 22.6%) compared to non-attention models (average of 15.8%). One possible explanation is that attention mechanism
Table 1: Gender bias analysis on COCO-GB V1 split. We utilize BLEU-4 (B-4), CIDEr(C) and METEOR(M) to evaluate captions qualities, all results are generated with beam size 5. Caption quality is obtained from test dataset, and gender bias is evaluated on COCO-GB V1 secret test dataset. † denotes the models that utilize extra grounded information from the Faster R-CNN network. (*) denotes the models that are trained with self-critical loss.

| Model         | B-4 | C    | M    | Women       |        | Men        |        |
|---------------|-----|------|------|-------------|--------|------------|--------|
|               |     |      |      | correct     | wrong  | neutral    | correct| wrong  | neutral|
| FC            | 31.4| 95.8 | 24.9 | 60.5        | 14.8   | 24.7       | 64.3   | 10.3   | 25.4   |
| LRCN          | 30.0| 90.8 | 23.9 | 61.7        | 16.8   | 21.6       | 64.7   | 11.2   | 24.0   |
| Att           | 31.0| 95.1 | 24.8 | 42.8        | 24.7   | 32.5       | 61.0   | 4.3    | 34.7   |
| AdapAtt       | 31.1| 98.2 | 25.2 | 47.9        | 26.8   | 27.4       | **75.7** | **3.8** | 20.5   |
| Att2in        | 32.8| 102.0| 23.3 | 61.5        | 16.3   | 22.1       | 70.2   | 6.5    | 23.3   |
| TopDown†      | 34.6| 107.6| 26.7 | 65.5        | **9.0**| 25.5       | 63.5   | 7.4    | 29.0   |
| NBT†          | 34.1| 105.1| 26.2 | **72.3**    | 9.3    | 18.3       | 77.6   | 4.3    | 18.0   |
| FC*           | 33.4| 103.9| 25.0 | 61.6        | 20.7   | 17.7       | 71.9   | 6.8    | 21.3   |
| LRCN*         | 29.4| 93.0 | 23.5 | 68.1        | 11.3   | 20.6       | 60.6   | 15.5   | 23.9   |
| Att2in*       | 33.6| 106.7| 25.7 | 61.8        | 19.7   | 18.5       | 73.8   | 6.2    | 20.0   |
| TopDown†*     | **34.9**| **117.2**| **27.0**| 69.0        | 15.1   | 15.8       | 73.3   | 7.5    | 19.1   |

enhances the model’s ability to capture visual features which also makes models easier to learn vision bias. Models utilizing extra visual grounding information have a much lower women error rate (average of 9.1%), which indicates that the extra visual features provided by Faster R-CNN are unbiased. The shortcoming is that NBT and TopDown models require training multiple sub-models, and the gender accuracy highly depends on the extra object detection model.

- Models with a high gender error rate can still receive competitive caption quality scores, e.g., AdapAtt model obtains a decent caption quality performance in three quality metrics, but has the highest women error rate (26.8%) in all models, which means model misclassifies 26.8% of images with women into men. This experimental results demonstrate that current caption quality metrics mainly focus on the overall quality and are not sensitive to the gender prediction error.

- Self-critical loss can improve models’ overall caption quality but also amplify the gender error rate. After training with self-critical loss, the metric CIDEr obtains an average improvement of 6.2%. However the error rate of women in FC, Att2in, and TopDown model increases by an average of 5.1% and the error rate of men in LRCN model increases by 4.2%.

4 Image Captioning Model with Guided Attention

We propose a novel Guided Attention Image Captioning model (GAIC) to mitigate gender bias by self-supervising on model’s visual attention. Attention mechanism has been widely used in image captioning task, which significantly improves the quality of generated descriptions [16, 15, 6]. However, most existing work uses the visual attention mainly for improving the overall caption quality. Few studies have considered the potential bias learned by the attention mechanism. GAIC considers a situation with no grounded supervision in which the model explores the correct gender evidence by model itself. Model GAIC_{es} considers a semi-supervised scenario that a small amount of labeled data is available, and fine-tunes the model attention with the extra supervision.

4.1 Self-Guidance on the Caption Attention

To achieve the goal of self-supervision, we design a two-stream training pipeline. As shown in Fig. 2, GAIC contains caption generation stream $S_{cg}$ and gender evidence mining stream $S_{gm}$. The two streams share the same parameters. The purpose of stream $S_{cg}$ is to generate high-quality descriptions as well as synthesize attention maps of gender words. The stream $S_{gm}$ then forces the attention to be focused on the correct gender evidence. The two complementary streams encourage the attention regions of gender words to gradually move to correct gender features, such as the human appearance, and keep away from biased features, e.g., laptops and umbrellas.

In stream $S_{cg}$, after inputting an image $I$, the captioning model will generate the corresponding description which is supervised by the Language Quality Loss $L_{lg}$, such as cross-entropy loss. At
the same time $S_{cg}$ will generate the visual attention maps $\alpha = \{\alpha_i | i \in [1, \text{caption length}]\}$ for each inferred word. We can directly get attention maps from models with attention mechanism. For non-attention models, attention maps can be obtained by post-hoc interpretation method, such as Grad-CAM [19] and saliency maps [20]. Here we focus on the attention maps of gender words which is denoted by $\alpha_g$. We utilize the $\alpha_g$ to generate a soft mask and apply the mask on the original input image. In this way, features which are related to gender inference for $S_{cg}$ stream are removed from images. $I^{*g}$ denotes the image without gender features captured by $S_{cg}$ which is defined as follows:

$$I^{*g} = I - (T(\alpha^g) \odot I),$$

where $\odot$ denotes the element-wise multiplication. $T$ is a masking function based on a thresholding operation. To normalize it, we use Sigmoid function as an approximation defined in Eq. (3).

$$T(\alpha^g) = \frac{1}{1 + \exp(-w(\alpha^g - E))},$$

where $E$ is a matrix where all elements equal to a threshold $\sigma$. Parameter $w$ is a scale ensuring $T(\alpha^g)$ approximately equals to 1 when $T(\alpha^g)$ is larger than $\sigma$, or to 0 otherwise. We feed $I^{*g}$ into stream $S_{gm}$ and generate corresponding captions. The loss on the $S_{gm}$ is denoted as Gender Evidence Loss $\mathcal{L}_{ge}$, which is defined as follows:

$$\mathcal{L}_{ge} = \arg\max_\theta \sum_{t=1}^T \log p(y_t|y_1, \ldots, y_{t-1}, I^{*g}), \ y_t \leftarrow g_n \text{ if } y_t \in g_w \cup g_m,$$

where $y_t$ denotes the $t^{th}$ caption word, $g_n$, $g_w$ and $g_m$ refer to neutral, female and male gender words respectively. We replace $y_t$ into $g_n$ if $y_t$ belongs to male or female gender words. With the guidance of $\mathcal{L}_{ge}$, the GAIC model learns to focus on correct gender evidence. Suppose the stream $S_{cg}$ utilizes biased gender features for gender prediction, then $I^{*g}$ generated by $S_{cg}$ will remove the biased visual features, e.g., a laptop. Thus the stream $S_{gm}$ will generate a incomplete caption because of missing the important features, and receives the penalty from $\mathcal{L}_{ge}$. Loss $\mathcal{L}_{ge}$ can be minimized only when model focus on correct gender features such as person appearance (An example is shown in Fig. 2).

Besides capturing correct gender features, we also expect model to predict cautiously and use gender neutral words such as people, person when gender evidence is vague. Since $I^{*g}$ has removed most gender features, we correspondingly change the gender words in $S_{gm}$ by following replacements:

- woman/man/etc $\rightarrow$ person, women/ment/etc $\rightarrow$ people
- boy/girl/etc $\rightarrow$ child, boys/girls/etc $\rightarrow$ children

Finally, we combine $\mathcal{L}_{tg}$ and $\mathcal{L}_{ge}$ as the self-guidance gender discrimination loss:

$$\mathcal{L}_{self} = \mathcal{L}_{tg} + \alpha \mathcal{L}_{ge},$$

where $\alpha$ is the weighting parameter, and we use $\alpha = 0.1$ in all our experiments. With the joint optimization of $\mathcal{L}_{self}$, the model learns to generate high quality captions as well as focuses on real visual features that contribute to gender recognition.

Figure 2: GAIC has two streams of networks that share parameters. Stream $S_{cg}$ finds out regions that help model to classify the gender, and Stream $S_{gm}$ tries to make sure all selected regions are correct gender evidence features. The attention map is online generated and two streams are trained by the Language Quality Loss and Gender Evidence Loss jointly. GAIC$_{ex}$ model seamlessly adds a small amount of extra supervision to further refine model attention which denotes as $S_{ex}$. 
4.2 Integrating with Extra Supervision

In addition to self-exploration training, we also consider the semi-supervised scenarios where a small amount of extra labeled data is added to accelerate the self-exploration process. More specifically, we utilize the pixel-level person segmentation masks to regularize attention, and denotes the model as GAIC_{es}. GAIC_{es} utilizes these masks as a boundary regularization and force the attention of gender words to be limited in the mask. In addition to $\mathcal{L}_{\text{self}}$, GAIC_{es} has another loss, Gender Attention Loss $\mathcal{L}_{\text{ga}}$, for the attention supervision, which is defined as follows:

$$
\mathcal{L}_{\text{ga}} = 1 - \sum_i (\alpha_i^g \odot (1 - M_i)),
$$

where $M_i$ is a binary mask where 1 indicates pixels belonging to a person and 0 represents the background. Loss $\mathcal{L}_{\text{ga}}$ forces the attention maps of gender words to be limited in the mask. The final loss for GAIC_{es} is defined as follows:

$$
\mathcal{L}_{\text{es}} = \mathcal{L}_{\text{self}} + \beta \mathcal{L}_{\text{ga}},
$$

where $\beta$ is the weighting parameter that adjusts the strength of $\mathcal{L}_{\text{ga}}$. Since labeling pixel-level segmentation maps is extremely time-consuming, we consider to use a small amount of data with external supervision, such as 10% in our experiment. In Fig. 2, we utilize $S_{\text{es}}$ to denote this stream, and all three streams share same parameters and optimize in an end-to-end manner.

5 Experiments

5.1 Settings and Baselines

Att model [6] is a classic captioning model and many captioning models are developed from its architecture. Our benchmarking experiments indicate that Att model has a severe gender bias. Thus we adopt Att model as the baseline as well as the base model to build GAIC. For GAIC_{es} model, we add 10% images with person segmentation masks as extra supervision and set the $\beta = 0.1$. We compare our proposed model with the following common debiasing approaches.

Balanced: A subset is selected from the training data which has 1:1 men to women ratio. Then we train a new model on this instance-level balanced dataset. In order to get a subset with a sufficient number of training data, we can only balance the image number of each gender. Note that unless we rebuilt the whole dataset and add more images with women, the gender bias existing in gender-object joint distribution cannot be effectively removed.

UpWeight: We also conduct an experiment that we amplify the weight of gender words’ cross-entropy loss during training. Specifically, we label the gender word position for each sentence in training dataset and train the model by multiplying a constant value 10 on loss for gender words. Intuitively, it will encourage models to accurately predict gender words.

Pixel-level Supervision (PixelSup): As a variant of GAIC_{es} model, we remove the self-exploration streams and directly use Eq. 6 to fine-tune model attention maps with 10% extra data.

5.2 Evaluation Metrics

Gender Accuracy: Unlike the traditional binary gender classification tasks, prediction results of captioning models can be correct, wrong and neutral. Models should firstly have low error rate. On this basis, to make decision more discriminative, captioning models should obtain low neutral rate and model generates gender neutral words only when the visual evidence are too vague to distinguish.

Gender Divergence: Besides high accuracy, we also expect our model to treat each gender fairly so that the outcome distribution for each gender should be similar. Here, we utilize Cosine Similarity between correct/wrong/neutral rate for men and women to measure the fairness of the model, which resembles to the fairness definition proposed in Equality of Odds [21]. Attention Correctness: To measure whether attention focuses on the correct gender features, we compare the attention maps of gender words with the person segmentation mask. We adopt two evaluation metrics: Pointing Game [22], which counts the highest attention point contained in the person masks, and Attention Sum [23], which calculates the attention weights in the masks.
### Table 2: Gender bias analysis on COCO-GB V1 split.

| Model  | C   | M   | Woman correct | wrong | neutral | Men correct | wrong | neutral | D     |
|--------|-----|-----|----------------|-------|---------|-------------|-------|---------|-------|
| Baseline | 95.1 | 24.8 | 42.8          | 24.7  | 32.5    | 61.0        | 4.3   | 34.7    | 0.075 |
| Balanced | 93.9 | 24.9 | 54.3          | 20.3  | 25.3    | 69.7        | 7.4   | 22.8    | 0.032 |
| UpWeight | 95.8 | 24.9 | **70.6**      | 26.4  | 2.9     | **81.4**    | 10.5  | 8.9     | 0.028 |
| PixelSup | 92.0 | 24.5 | 46.8          | 20.7  | 32.5    | 58.0        | 7.3   | 34.7    | 0.031 |
| GAIC   | 93.7 | 24.6 | 62.0          | 16.9  | 21.1    | 77.3        | 7.0   | 15.7    | **0.021** |
| GAIC<sub>es</sub> | 94.6 | 24.7 | 64.1          | **13.1** | 22.8 | 75.3        | 5.2   | 19.5    | **0.011** |

### Table 3: Attention Correctness on COCO-GB V1 split. We adopt Point Game [22] and Attention Sum [23] as evaluation metrics. GAIC and GAIC<sub>es</sub> model can significantly improve attention correctness.

| Accuracy | Women | Men | Average |
|----------|-------|-----|---------|
| Baseline | 25.5  | 21.2| 23.4    |
| Balanced | 25.0  | 21.4| 23.2    |
| Upweight | 26.7  | 23.3| 25.0    |
| PixelSup | 30.0  | 28.1| 29.0    |
| GAIC     | 27.4  | 24.3| 25.6    |
| GAIC<sub>es</sub> | **32.5** | **28.5** | **30.1** |

(a) Attention Sum (b) Point Game

#### 5.3 Experimental Results

**Gender Accuracy:** We report the gender prediction performance of COCO-GB v1 in Tab. 2. Our key observation is that GAIC significantly improves the gender prediction performance compared to the baseline, where the gender accuracy of women increases from 42.8% to 62.0% and error rate of woman reduce from 24.7% to 16.9%. Although the UpWeight method obtains the highest accuracy of both women and men, it also causes a significant high error rate towards each gender. There is no substantial difference between the Balanced model and baseline model, and similar trend has been found by [24], which indicates that models learn gender bias mainly from feature-level (gender-object co-occurrence), and balancing number of each gender can not remove bias in dataset. PixelSup obtains sensible improvements, which indicates that supervising directly on attention maps is also helpful. GAIC<sub>es</sub> obtains consistently better performance than GAIC model and we empirically find that adding extra supervision can accelerate self-exploration process and makes training more stable. For fairness evaluation, we compare different model’s gender divergence. GAIC and GAIC<sub>es</sub> obtain the lowest divergence, which indicates that models treat each gender in a fair manner.

Experiments on COCO-GB v2 have almost similar trends in COCO-GB v1 (Detailed analysis of COCO-GB v2 results is in Sec C). All baselines in COCO-GB v2 receive a worse gender prediction results. This is mainly because the unseen gender-object pairs in the test dataset increase the prediction difficulty. In comparison, our proposed GAIC and GAIC<sub>es</sub> model obtain a comparable performance with COCO-GB v1, which further improves the robustness of the self-exploration training strategy.

**Caption Quality:** Besides high gender accuracy, we also expect our model to obtain decent caption quality. We use METEOR(M) and CIDEr(C) to evaluate caption qualities. As the result shows in Tab. 2, GAIC and GAIC<sub>es</sub> only cause a minor performance drop compared to the baseline (from 95.1 to 94.6 on METEOR and from 24.8 to 24.7 on CIDEr). In Fig. 3 examples show that the sentences generated by GAIC and GAIC<sub>es</sub> are linguistically fluent with more correct gender descriptions.

**Attention Correctness:** To quantitatively evaluate attention correctness, we extract attention maps of gender words and compare them with person segmentation masks. Quantitative results are shown in Tab. 3. We observe that GAIC and GAIC<sub>es</sub> receive consistent improvement over the baseline model and all model variants, which indicates that our proposed models can focus on the described person for the gender word prediction. Qualitative comparison is shown in Fig. 3 We observe that baseline model may utilize biased visual features for gender prediction and thus makes incorrect gender prediction, e.g., describe a woman as a boy because of the tennis ball. In comparison, GAIC and GAIC<sub>es</sub> correctly focus on the described person for gender prediction.
6 Related Work

Gender Bias in Dataset: Gender bias in dataset has been studied in a variety of domains [25, 24, 26, 27, 28, 29], especially in natural language processing [10, 50, 51, 52, 53, 54, 11]. For image captioning, several studies have considered gender bias problem in the dataset. [35] analyzes the stereotype in Flicker30k dataset and mentions the gender-related issues. [10] indicates that caption annotators in COCO dataset tend to infer a person’s gender from the image context when gender cannot be confirmed, e.g., a baseball player is labeled as “man” even if gender evidence is occluded.

Vision and Language Bias: Bias in captioning models can be divided into vision and language bias [12]. Vision bias refers to learn wrong visual evidence while language bias refers to capture unwanted language priors. For example, the word “on the beach” always follows the word “Surfboard.” Due to the RNN’s recurrent mechanism, models can learn this language prior and always generate the phase “on the beach” if the word “Surfboard” has been inferred. We notice that gender words are usually mentioned at the beginning of the sentence (on the average at position 2 with average sentence length 9), and words before gender words, e.g., “a” and “the,” do not have the gender preference. Hence gender bias in captioning systems should mainly come from the vision part.

Mitigating Gender Bias: Few initial attempts have been made to design captioning models to overcome gender biases in datasets. One solution is to break the task into two steps [36]. It firstly locates and recognizes the person in the image. Then a language model utilizes grounded information to generate captions. Hence the gender accuracy of this approach highly depends on the extra object detection model. In another work, two novel losses are designed to reduce unwanted bias towards specific words, including gender words. Their approach requires segmentation masks for each image, which is costly and unpractical for many dataset [10].

7 Conclusions

In this paper, two novel COCO splits are created for studying gender bias problem in image captioning task. We provide extensive baseline experiments for benchmarking different models, training strategies, as well as a comprehensive analysis of the dataset. Our experimental results indicate that many captioning models have a severe gender bias problem, leading to an undesirable gender prediction error towards women. We propose a novel training framework GAIC which can significantly reduce gender bias by self-guided supervision. Besides, $GAIC_e$ model can seamlessly add extra supervision and further improves the gender prediction accuracy. Quantitative and qualitative results further validate that our proposed model can focus on correct visual evidence for gender prediction.
8 Broader Impact

In this work, we reveal the severe gender bias problem widely existing in most captioning models, which leads to an undesirable gender prediction error towards women. The experimental results remind researchers to revisit the high performance achieved by current captioning models and encourages the community to put more effort into promoting the fairness of captioning systems. Two novel COCO splits proposed in this work enable future work to efficiently quantify gender bias in systems, which will have a strong impact on promoting a more fair platform for emerging and future computer vision as well as natural language processing systems. The self-exploration training strategy proposed in this work can significantly reduce gender bias in learning models and understanding how the interpretation methods such as attention mechanism could be utilized to further advance fairness of machine learning methods, thereby broadly impacting the Machine Learning field.

This work also plays an integral part in educating and training students. The research will also be tightly integrated with related courses on data science at the author’s university. The course will show how the bias in data could be learned by models, and potentially leads to unintentional discrimination in learning models. The course will have a section to teach students to design a fair machine learning algorithm. We will actively encourage undergraduate participation in this bias mitigation project.

References

[1] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1473–1482, 2015.

[2] MD Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR), 51(6):118, 2019.

[3] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137, 2015.

[4] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632, 2014.

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–3164, 2015.

[6] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057, 2015.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.

[9] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4566–4575, 2015.

[10] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In European Conference on Computer Vision, pages 793–811. Springer, 2018.

[11] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2979–2989, 2017.
[12] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4035–4045, 2018.

[13] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7008–7024, 2017.

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2625–2634, 2015.

[15] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 375–383, 2017.

[16] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6077–6086. IEEE, 2018.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015.

[18] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7219–7228, 2018.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

[20] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. 2014.

[21] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Advances in neural information processing systems, pages 3315–3323, 2016.

[22] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. International Journal of Computer Vision, 126(10):1084–1102, 2018.

[23] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[24] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems, pages 4349–4357, 2016.

[25] Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6:587–604, 2018.

[26] Shikha Bordia and Samuel Bowman. Identifying and reducing gender bias in word-level language models. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 7–15, 2019.

[27] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In International Conference on Machine Learning, pages 803–811, 2019.

[28] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91, 2018.

[29] Joel Escudé Font and Marta R Costa-jussà. Equalizing gender bias in neural machine translation with word embeddings techniques. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 147–154, 2019.
[30] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7):2966–2981, 2019.

[31] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, 2018.

[32] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018.

[33] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, 2018.

[34] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, 2019.

[35] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E Papalexakis, and Amit K Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1856–1864, 2018.

[36] Shruti Bhargava. *Exposing and correcting the gender bias in image captioning datasets and models*. PhD thesis, 2019.
A COCO-GB dataset

A.1 Gender Annotation

We show the gender word list in Tab. 4. Gender words are selected based on the word frequency in COCO dataset and we delete the less frequent gender words. Word “woman” and “man” are most frequent gender-specific word and account for more than 60% of the total gender-specific words.

| female words ($g_w$)          | woman, women, girl, sister, daughter, wife, girlfriend |
|-------------------------------|--------------------------------------------------------|
| male words ($g_m$)            | man, men, boy, brother, son, husband, boyfriend       |
| gender neutral words ($g_n$)  | people, person, human, baby                            |

Table 4: Gender words list

Some labeled examples are shown in Fig. 4. We label an image as "women" when at least one sentence mentioned female words and label an image as "men" when at least one sentence mentioned male words. Images that both mention male words and female words are discarded. In Fig. 4(c), we show that when gender evidence is occluded, annotators may provide the gender prediction based on context cues or social stereotype. Hence our gender annotations may contain this kind of social bias.

![Image](image_url)

Figure 4: (a): An image is labeled as women. (b): An image is labeled as men. (c) An image is labeled as men, however the gender evidence is actually occluded.

A.2 COCO Gender Distribution

We show the gender-object joint distribution of COCO training dataset, COCO testing dataset, COCO-GV v1 secret testing dataset, COCO-GB v2 testing dataset as follows. We sort the objects according to bias rate in training dataset. For COCO-GB dataset, we choose 63 object from 80 objects in COCO dataset according to their image numbers. We observe that COCO-GB v1 has a balanced gender-object joint distribution while COCO-GB v2 has a distribution opposite to the training set.

![Diagram](diagram_url)

Figure 5: Gender-object joint distribution of COCO training dataset
B Caption Generation with Visual Attention

Given an image $I$ and the corresponding caption $y = \{y_1, ..., y_t\}$, the objective of an encoder-decoder image captioning model is to maximize the following formulas:

$$\arg \max_{\theta} \sum_{(I,y)} \log p(y|I; \theta) = \sum_{t=1}^{T} \log p(y_t|y_1, ..., y_{t-1}, I),$$  \hspace{1cm} (8)

where $\theta$ are trainable parameters of captioning model. We utilize chain rule to decompose the joint probability distribution into ordered conditionals. A recurrent neural network (RNN) predicts each conditional probability as follows:

$$\log p(y_t|y_1, ..., y_{t-1}, I) = f(h_t, c_t),$$  \hspace{1cm} (9)
where \( f \) is a nonlinear function, and we adopt the classic LSTM as function \( f \) in this paper. \( h_t \) is the hidden state of RNN at \( t \) steps. \( c_t \) is the visual context vector extracted from image \( I \) for \( y_t \). Generally, \( c_t \) is an important extra information in image captioning models, which can provide visual evidence for caption generation. We follow the work [6] to compute \( c_t \). A CNN extracts a set of image features from last convolutional layer which we denote them as \( v \). \( v = \{v_1, \ldots, v_L\} \) corresponds to the features extracted at different image locations, where \( v_i \in \mathbb{R}^D \). We calculate attention values for each \( v_i \) as follows:

\[
e_{ti} = f_{\text{att}}(a_i, h_{t-1}), \quad \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})},
\]

(10)

where \( f_{\text{att}} \) is a multi-layer perceptron conditioned on the previous hidden state. For each location \( i \), \( \alpha_{ti} \) represents the importance of region \( i \) for generating \( t^{th} \) word. Once we obtain the attention weight, the context vector \( c_t \) is computed by

\[
c_t = \sum_{i=1}^{L} \alpha_{ti} v_i.
\]

(11)

### C Experimental Results on COCO-GB v2

| Model      | C   | M   | Woman          | Men          | D    |
|------------|-----|-----|----------------|--------------|------|
|            |     |     | correct | wrong | neutral | correct | wrong | neutral |     |
| Baseline   | 98.2| 27.2| 51.6    | 28.3  | 20.1    | 77.9   | 4.9   | 17.1    | 0.094|
| Balanced   | 97.5| 27.3| 57.9    | 25.5  | 26.6    | 71.1   | 11.5  | 17.4    | 0.034|
| UpWeight   | 95.8| 26.9| 72.2    | 26.1  | 1.7     | 86.1   | 11.7  | 2.1     | 0.023|
| PixelSup   | 96.8| 27.1| 54.2    | 25.1  | 20.5    | 76.4   | 6.2   | 17.2    | 0.062|
| GAIC       | 97.8| 26.9| 67.1    | 18.0  | 14.9    | 68.9   | 10.7  | 20.3    | 0.008|
| GAIC\textsubscript{es} | 98.1| 27.0| 69.1    | 15.2  | 15.7    | 71.4   | 8.1   | 20.5    | 0.007|

Table 5: Gender bias analysis on COCO-GB V2 split.

Compared to COCO-GB v1, all baseline and common debiasing approaches obtain a higher error rate of women (average increase of 3.25%) on COCO-GB v2. GAIC model improves the gender prediction accuracy of woman from 41.6% to 67.1% and reduce the error rate of women from 28.3% to 18.0%. Although the UpWeight method obtains the highest accuracy of both women and men, it has an undesirable high error rate towards each gender. There is no substantial difference between the Balanced model and baseline model, and similar trend has been found in COCO-GB v1. GAIC\textsubscript{es} obtains consistently better performance than GAIC model. For fairness evaluation, we compare different model’s gender divergence. GAIC and GAIC\textsubscript{es} obtain the lowest divergence, which indicates that models treat each gender in a fair manner. For caption quality, GAIC and GAIC\textsubscript{es} only cause a minor performance drop compared to the baseline.

### D More on Implementation Details

**Benchmarking Baselines:** All the baseline models obtain visual features extracted from the fourth layer of ResNet-101. All models except for NBT, TopDown, Att and AdaptAtt are implemented in the same open source framework from https://github.com/ruotianluo/self-critical.pytorch, and we directly use the test caption results from https://github.com/LisaAnne/Hallucination. For other models, we implement by our-self and make sure that the caption quality is close to the results reported in paper. Caption quality is evaluated by the official COCO evaluation tool https://github.com/tylin/coco-caption.

**GAIC model and debiasing approaches:** We select a subset from original training set which contains 4,000 images for each gender. Baseline Att model are trained on COCO for 5 epoches. For Balanced baseline, we directly fine-tune the original Att model on this subset. For PixelSup baseline, we fine-tune the model with the subset and extra person segmentation annotations. For GAIC model, we use the two streams pipeline to fine-tune the model on the subset, and set \( \alpha = 0.1 \). For GAIC\textsubscript{es} model, we fine-tune the dataset with extra person segmentation annotations, and set \( \alpha = 0.1 \) and \( \beta = 0.1 \). For above-mentioned models, we fine-tune the model for 1 extra epoch on the subset.
### More Qualitative Results

| Original Image | Baseline | GAIC | GAICes |
|---------------|----------|------|--------|
| ![Original Image](image1.png) | ![Baseline](image2.png) | ![GAIC](image3.png) | ![GAICes](image4.png) |
| a man in a suit and tie holding a cell phone | a woman is holding a cell phone in her hand | a woman is sitting in a chair talking on a cell phone |
| ![Original Image](image5.png) | ![Baseline](image6.png) | ![GAIC](image7.png) | ![GAICes](image8.png) |
| a young boy standing in front of a refrigerator | a woman is standing in front of a refrigerator | a woman is standing in front of a refrigerator |
| ![Original Image](image9.png) | ![Baseline](image10.png) | ![GAIC](image11.png) | ![GAICes](image12.png) |
| a woman holding a yellow umbrella in her hand | a person with a hat on top of her head | a young boy with a hat on his head holding a banana |
| ![Original Image](image13.png) | ![Baseline](image14.png) | ![GAIC](image15.png) | ![GAICes](image16.png) |
| a woman in a red dress holding a cell phone | a person sitting on a bed using a laptop computer | a person is sitting on a bed with a laptop |

Figure 9: Qualitative comparison of baselines and our proposed model. At the top, we show success cases that our proposed modes predict correct gender and utilize correct visual evidence. The bottom case shows that when gender evidence is vague, our model tend to use neutral gender words, such as "person" to describe the gender of the person.