Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search

Yunhe Feng, Chirag Shah
Information School, University of Washington
yunhe@uw.edu, chirags@uw.edu

Abstract

Gender bias is one of the most common and well-studied demographic biases in information retrieval, and in general in AI systems. After discovering and reporting that gender bias for certain professions could change searchers' worldviews, mainstreaming image search engines, such as Google, quickly took action to correct and fix such a bias. However, given the nature of these systems, viz., being opaque, it is unclear if they addressed unequal gender representation and gender stereotypes in image search results systematically and in a sustainable way. In this paper, we propose adversarial attack queries composed of professions and countries (e.g., 'CEO United States') to investigate whether gender bias is thoroughly mitigated by image search engines. Our experiments on Google, Baidu, Naver, and Yandex Image Search show that the proposed attack can trigger high levels of gender bias in image search results very effectively. To defend against such attacks and mitigate gender bias, we design and implement three novel re-ranking algorithms – epsilon-greedy algorithm, relevance-aware swapping algorithm, and fairness-greedy algorithm, to re-rank returned images for given image queries. Experiments on both simulated (three typical gender distributions) and real-world datasets demonstrate the proposed algorithms can mitigate gender bias effectively.

Introduction

The web's biggest image search engines, such as Google, and Bing, provide an important information-seeking interface for people to explore the world. According to Internet Live Stats[1], Google processes more than 3.5 billion queries per day and 1.2 trillion searches per year. Google image searches account for 22.6% of all searches[2]. Given the volume and importance in our daily lives, image search results can significantly influence how people perceive and view the world. Images are often more than useful objects of information; they provide a visual representation of a phenomenon, a concept, and a perceived reality of the world around us. Given this, it is not sufficient to assess the quality of image search results using relevance metrics; we also need to consider how this visual information carries various perceptions and prejudice. For example, Lam et al. (2018) showed that searching for 'CEO' in Google Image Search resulted in predominantly white males. While the same query in a web search provides a diverse set of information objects (definition, Wikipedia article, questions, and answers), image search results are mono-media and appeal to one's visual perceptions, which can more quickly affect their worldview (Hibbing and Rankin-Erickson 2003). If these results carry biases such as those shown by Lam et al. (2018) and Kay, Matuszek, and Munson (2015), they are much easier to perpetuate than regular web search results. Therefore, while evaluating image search results, we need to look beyond their relevance. We must also look at inherent disparity and biases carried out in them.

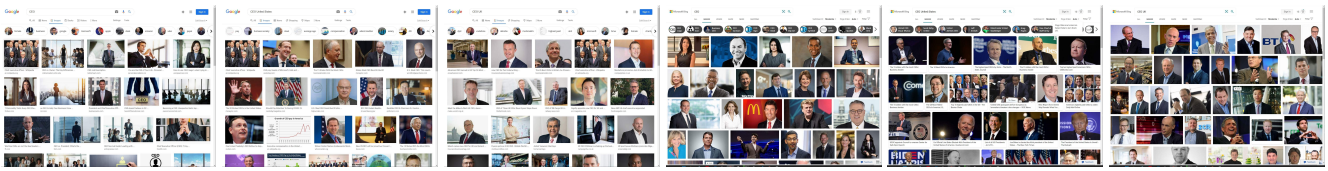
Among different types of image biases, gender bias is one of the most common and well-studied demographic biases. Not surprisingly, this also gets scrutiny and attention from scholars and media. That often makes the service providers take immediate actions to fix such biases in an ad hoc manner. For example, after the work described before (Kay, Matuszek, and Munson 2015) received a lot of attention, Google shifted the gender distributions in image search results for CEO and some other occupations. For instance, the famous query of CEO in image search has been fixed for a long time (see Figure 1(a) and Figure 1(d)). However, Mozilla's recent Internet Health Report (Mozilla 2021) points out that the default internet user is still viewed as white, male, and cisgender, and big tech has not done enough to fix it (Griffith 2021). But that is only one way the gender bias problem is not fixed.

In this paper, we revisit gender bias in image search results for professional occupations. For some search terms, such as 'CEO,' gender fairness is observed in image search results. But have image search engines mitigated gender biases in search results systematically? To further illustrate this research question, we present the relevant adversarial attack queries of the CEO in Google and Bing, as shown in Figure 1. Both Google and Bing image search engines have already fixed the gender stereotypes for the occupation of CEO. However, such gender bias resurfaces when appending the country names, such as United States and UK, to the original keyword of CEO. This finding inspires us to dive into the exploration of gender fairness in image search engines to reveal superficial bias mitigation.

Ten occupations, also investigated by Kay, Matuszek, and Munson (2015) seven years ago, are chosen as image search terms in our study. We design two search keywords for each occupation, i.e., the original occupation name and the adversarial attack keyword that consists of occupation name and the country name of United States. The latter aims to trigger gender bias in image search results. For each keyword, we retrieve the top 200 images (if available) from four widely

[1]https://www.internetlivestats.com/google-search-statistics/

[2]https://firstsiteguide.com/google-search-stats/



(a) CEO - Google (b) CEO U.S. - Google (c) CEO UK - Google (d) CEO - Bing (e) CEO U.S. - Bing (f) CEO UK - Bing

Figure 1: Image search results of CEO, CEO United States (almost all males), and CEO UK (all males) by Google and Bing.

used image search engines, namely, Google from USA, Baidu from China, Naver from South Korea, and Yandex from Russia. In total, we collected more than 18,000 images.

When image retrieval is complete, we attempt to leverage image gender detection APIs to recognize the genders of people in these images. To be specific, five popular gender detection APIs, including Amazon Rekognition, Luxand, Face++, Microsoft Azure, and Facebook DeepFace, are selected to calculate gender distributions of returned occupation images. We compared the gender labels detected by APIs with human annotations, and found that only Amazon Rekognition APIs were acceptable but still failed to handle the images with a low ratio of detected faces. Therefore, we propose a hybrid approach to detect face genders in search images by combining Amazon Rekognition results and crowdsourced human annotations through Amazon Mechanical Turk.

To mitigate gender bias, we present three generalized re-ranking algorithms, including the epsilon-greedy algorithm, relevance-aware swapping algorithm, and fairness-greedy algorithm, to balance the trade-off between gender fairness and image relevance. Evaluations of the proposed gender bias mitigation algorithms on both simulated and real-world datasets demonstrated that it is feasible (and advisable) to address bias in image search (and perhaps in other types of search as well) in a systematic and more meaningful way than doing individual query fixes in an ad hoc manner.

Our contributions are summarized as follows.

- We design adversarial attacks with regard to gender bias in image search and determine that gender bias is not fixed systematically by search engines.
- CIRF, an open-sourced Cross-search-engine Image Retrieval Framework to collect images from multiple search engines for given search terms, is developed.
- We find image gender detection APIs cannot always perform well on search images in the wild, so a hybrid approach combining automatic gender detection and manual annotation is presented.
- We propose and validate three re-ranking algorithms to mitigate gender bias in image search results.

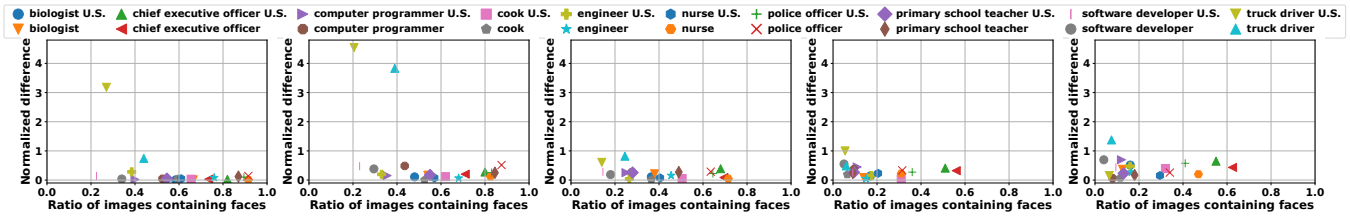
Related Work

This section presents the importance of gender fairness in image search results, summarizes the gender bias-related research findings from multiple perspectives, discusses the existing approaches to mitigate image gender biases, and highlights the difference between existing works and ours.

The gender fairness or biases demonstrated in image search results affect people's perceptions and views significantly (Ellemers 2018; Metaxa et al. 2021). Kay, Matuszek, and Munson (2015) are one of the first to investigate gender biases in professional occupation image search results. They reported that such image search results for occupations slightly exaggerate gender stereotypes, and people thought image search results were better if they agreed with the stereotype. More importantly, this research work pointed out that the biased representation of gender in image search results could shift people's perceptions about real-world distributions. Otterbacher, Bates, and Clough (2017) proposed a trait adjective checklist inspired method further to identify the existence of gender biases in image search. They found that images of men were more often retrieved for agentic traits whereas warm traits were demonstrated in photos of women. In addition, photos of stereotype-incongruent individuals exhibited a backlash effect, e.g., 'competent women' were less likely to be portrayed positively. Otterbacher et al. (2018) measured the user perception of gender bias in image search from the perspective of sexism, and found search engine users, who were more sexist, were less likely to perceive gender biases.

There also exist many research studies exploring gender bias in different types of images. By detecting gender labels of the photographs of U.S. members of Congress and their tweeted images, Schwemmer et al. (2020) concluded Google Cloud Vision (GCV) could produce correct and biased labels at the same time because a subset of many possible true labels was selectively reported. Wijnhoven (2021) found a gender bias toward stereotypically female jobs for women but also for men when searching jobs via Google search engine. By examining four professions across digital platforms, Singh et al. (2020) concluded: 1) gender stereotypes were most likely to be challenged when users acted directly to create and curate content, and 2) algorithmic approaches for content curation showed little inclination towards breaking stereotypes. Makhortykh, Urman, and Ulloa (2021) conducted a cross-engine comparison of racial and gender bias in the visual representation of the search term 'artificial intelligence' and gender representation of AI is more diverse than its racial representation. Hashemi and Hall (2020) reported no gender bias was identified when detecting criminal tendency based on mugshot images of arrested individuals.

In the last couple of years, many approaches have been proposed to detect and mitigate gender bias in images for training deep learning models (Wang et al. 2020; Xu et al. 2020; Hwang et al. 2020; Adeli et al. 2021). For example, Serna et al. (2021) showed how bias in face images impacted the activations of gender detection models and developed InsideBias to detect biased models. To reduce gender bias in deep image representations, an adversarial method for the removal of features associated with a protected variable (gender) from



(a) Amazon Rekognition (b) Luxand (c) Face++ (d) Microsoft Azure (e) Facebook DeepFace

Figure 2: Normalized female ratio difference (compared with MTurk results) vs. the ratio of detected faces in images.

the intermediate convolutional neural network based representations was presented (Wang et al. 2019). Many other image gender bias mitigation approaches, such as a posterior regularization based gender de-biasing framework (Jia et al. 2020), a fairness-aware disentangling variational autoencoder (FD-VAE) (Park et al. 2021), and an adversarial gender de-biasing algorithm (AGENDA) (Dhar et al. 2020), are also proposed. Besides, some post-processing bias mitigation methods, such as FA*IR (Zehlike et al. 2017) and multi-task learning for fair regression (Zhao and Chen 2019), have been proposed.

Our study adds to the literature on exploring gender bias in image search results in the following ways. First, similar to Kay, Matuszek, and Munson (2015), we investigate the gender distribution in professional occupation image results. Still, we also design an adversarial search attack by adding the country information into the occupation search terms. We find evidence that image search engines do not fix the reported gender bias in search results systematically. Second, we not only examine the performance of five popular image gender detection APIs, but also propose a hybrid approach that combines automatic detection (Amazon Rekognition services) and manual annotations (Amazon Mechanical Turk) to improve gender distribution estimation. Finally, we develop three re-ranking algorithms, i.e., epsilon-greedy, relevance-aware swapping, and fairness-greedy methods, to mitigate gender biases in image search results.

Image Retrieval and Gender Detection

In this section, we describe how to build the image search datasets, examine the performance of image based gender detection APIs, and propose a hybrid approach to strike a balance between detection accuracy and efficiency.

Search Image Retrieval

We propose and develop an open-sourced Cross-search-engine Image Retrieval Framework (CIRF [3]) to automatically collect images from multiple search engines for given search terms. CIRF mainly consists of three components: URL Builder, Data Downloader, and Image Parser.

URL Builder To enable automatic data download, we first construct image search URLs based on search-engine-specific URL templates for given search terms. For example, we use *https://www.google.com/search?q=**keyword**&source=lnms&tbm=isch* as a URL template for Google Image Search, where ***keyword*** is the placeholder of search terms. As CIRF is able to handle multiple search engines, we also design similar URL templates for popular search engines, including Baidu from China, Naver from South Korea, and Yandex from Russia. The ***keyword*** can be written in any language because browsers will encode it in the UTF-8 format.

Data Downloader We collect two types of search image data based on the built URLs: (i) the web page HTML file that captures the layout and names of images returned by search engines; (ii) individual image files embedded in the image gallery. CIRF adopts the web framework Selenium WebDriver [4] to open URLs in a Chrome browser with incognito mode. To display and cache more search images, CIRF scrolls up and down the web page by sending PAGE_UP and PAGE_DOWN commands to the HTML entities. CIRF leverages PyAutoGUI [5], a cross-platform GUI automation module, to save the web page HTML file and all supplementary materials including images.

Image Parser This component is responsible for extracting the images and their orders from the downloaded HTML file. In general, three types of images are collected: standard images, Base64 encoded images, and image URLs. For the latter two types of images, CIRF decodes them into standard images and retrieves images via URLs respectively. When all images are ready, CIRF renames them according to their orders in the HTML file for further analysis.

Gender Detection

Image-based gender detection has widely been adopted in diverse domains, so many commercial and open-sourced gender detection APIs have been developed and released. Considering the scalability and efficiency, we intended to rely on these available tools to label search images automatically. To evaluate their performance, we randomly selected and searched ten occupations and their corresponding adversarial attack search terms (i.e., appending 'United States') in Google Image Search. Then we conducted an IRB-approved user study to recruit participants from Amazon Mechanical Turk (MTurk) to build the ground truth of genders. We paid each participant $0.5 for annotating 50 images, and each image was assigned to three workers. Five popular gender detection APIs, including Amazon Rekognition, Luxand, Face++, Microsoft Azure, and Facebook DeepFace, were chosen to

[3] https://github.com/YunheFeng/CIRF

[4] https://www.selenium.dev/documentation/webdriver/

[5] https://pyautogui.readthedocs.io/

calculate gender distributions of the top 200 Google search images for each given search term.

Figure 2 demonstrates the normalized female ratio difference between MTurk results and face gender detection APIs. In general, Amazon Rekognition outperforms the rest of APIs in terms of face detection ratios (see X-axis) and the female ratio errors (see Y-axis). When the face detection ratio is above 0.5, the normalized difference of Amazon Rekognition is below 15%. Therefore, Amazon Rekognition was chosen to identify the genders of people in images. Thus, we propose a two-step hybrid method to annotate image gender labels: 1) use Amazon Rekognition to detect image genders; 2) for search terms that suffer from a low face detection ratio (below 0.5), we still rely on MTurk to manually label them.

Exploring Unsystematic Gender Bias Fixing

We investigate whether gender bias in image search results is systematically fixed by designing adversarial search attacks and measuring the degree of gender fairness.

Adversarial Search Attack Design

As mentioned before, we are motivated to investigate whether image search engines fix gender bias in different occupation queries systematically. Therefore, we follow the occupation list on U.S. Bureau of Labor Statistics [6] and choose occupation names as the baseline search keywords. When constructing adversarial searches, we append the country name of 'United States' to each occupation name to build the attacking search term. If both baseline and attacking searches demonstrate no difference with the gender distribution ground truth for one occupation, we think search engines mitigate gender bias systematically for that occupation. Otherwise, we argue that such fixes and mitigation of gender bias are just hit-or-miss.

As the previously existing gender biases in image search queries, such as 'CEO,' drew huge attention of the public and academia, mainstream search engines had already mitigated such biases accordingly. However, our analytics show that gender biases crossing over all occupations are not fixed in a systematic way.

Gender Bias Measurement

It is very intuitive and straightforward to compare the normalized difference between gender probability distribution P in image search results and the ground truth gender probability T for each occupation. For top k images returned by search engines, we calculate the Kullback-Leibler divergence $D_{KL}(T \parallel P^k)$ between these k images and the ground truth. The average Kullback-Leibler divergence is used to represent the existing bias.

$$d = \frac{\sum_{k=1}^{N} D_{KL}(T \parallel P^k)}{N} \tag{1}$$

Algorithms to Mitigate Gender Bias

We propose three interpretable and lightweight re-ranking algorithms to mitigate gender biases in image search results.

[6] https://www.bls.gov/oes/current/oes_nat.htm#00-0000

Epsilon-greedy Algorithm

Inspired by the exploitation and exploration trade-off idea in reinforcement learning (Berry and Fristedt 1985; Sutton and Barto 2018) and re-ranking (Gao and Shah 2020), we propose the epsilon-greedy re-ranking algorithm, which swaps items in the image rank list with a controllable degree of randomness. The randomized swapping breaks original gender distributions and might improve fairness especially when items with the same attribution values are gathered together densely (e.g., male CEO images fully occupy the top 20 CEO image search results). This algorithm has two main advantages – simplicity and generalizability. It is straightforward and simple to randomly shuffle items without considering other factors. In addition, no prior knowledge, such as the optimal gender distribution, is required to apply this algorithm.

In the proposed epsilon-greedy algorithm, the randomness is specified by the parameter $\epsilon \in (0, 1]$, representing the probability of swapping two items. Each item has a probability of ϵ to exchange positions with a random item that follows it. A larger ϵ introduces more randomness, leading to a re-ranked list that is more different from the original list.

Relevance-aware Swapping Algorithm

Normally, the items with large relevance weights are ranked at the top of search engines' returned image list. If unrelated or less related images are ranked high, it harms user experience and utility. The epsilon-greedy algorithm is very straightforward and simple, but it ignores the relevance of search items during re-ranking. Therefore, we propose the relevance-aware swapping algorithm to consider both the randomness and relevance weight of image items to re-rank the image list. To keep the utility of the re-ranked list, an image with a larger relevance weight is less likely to be swapped with an image item that follows it.

Relevance Weight Modeling We grade an image's relevance weight based on its index in the image list returned by search engines. Similar to the Mean Reciprocal Rank (MRR) (Voorhees et al. 1999), the relevance weight of an image with a rank index i can be modeled by its reciprocal rank $\frac{1}{i}$. However, such relevance weight decreases too fast with the growth of the rank index i. Instead, we model the relevance weight distribution in a linear manner. Suppose we have an image list $\mathbf{L}$ containing $|\mathbf{L}|$ images. The linear relevance weight of the i^{th} image is estimated as $1 - \frac{i}{|\mathbf{L}|}$. Inspired by the Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen 2002), we further introduce a discount factor of $\log_2(i+1)$ to smooth the decay of relevance weights of bottom images in $\mathbf{L}$. Finally, the relevance weight of image $\mathbf{L}_i$ is expressed as:

$$\mathbf{W}_i = \frac{1 - \frac{i}{|\mathbf{L}|}}{\log_2(i+1)} \tag{2}$$

Swapping Probability The swapping probability of image $\mathbf{L}_i$ is determined by its relevance weight $\mathbf{W}_i \in [0, 1]$. To ensure that the image with a high relevance weight is less likely to be swapped, we can use $1 - \mathbf{W}_i$ to represent the swapping

Algorithm 1: Relevance-aware Swapping Algorithm

Input: $\mathbf{L}$: the original image list; ρ: the sensitivity of swapping two items;
Output: $\mathbf{R}$: the re-ranked image list;
$\mathbf{R} \leftarrow \varnothing$; // initialize $\mathbf{R}$ as empty
for $i = 1 \rightarrow |\mathbf{L}|$ **do**
  $\mathbf{W}_i = \frac{1-\frac{i}{|\mathbf{L}|}}{\log_2(i+1)}$; // relevance weight
  $p \leftarrow$ a random number between 0 and 1;
  if $p <= \rho * (1 - \mathbf{W}_i)$ **then** // swap items
    $temp \leftarrow \mathbf{L}_i$;
    $j \leftarrow$ a random number between $i+1$ and $|\mathbf{L}|$;
    $\mathbf{L}_i \leftarrow \mathbf{L}_j$;
    $\mathbf{L}_j \leftarrow temp$;
    append $\mathbf{L}_i$ to $\mathbf{R}$; // add swapped item
  else // keep the original item
    append $\mathbf{L}_i$ to $\mathbf{R}$; // add unswapped item
  end
end
return $\mathbf{R}$

probability for image $\mathbf{L}_i$. In addition, we design a coefficient $\rho \in (0, 1]$ to further control the swapping sensitivity and the swapping probability of image $\mathbf{L}_i$ is expressed as $\rho(1 - \mathbf{W}_i)$. The detailed implementation is illustrated in Algorithm 1.

Fairness-greedy Algorithm

Considering more than 90% of users do not go past the first page of the Google search results (Sharma et al. 2019) and the first three items displayed in Amazon search results account for 64% of all clicks (Baker 2018), we think it is of great significance to ensure gender fairness in images ranked top in search results. Therefore, we propose the fairness-greedy algorithm to guarantee gender fairness in the first few pages with high priority. Accordingly, the gender distribution of images displayed on the last pages, to which users pay lesser attention, is given less consideration.

The main idea of the fairness-greedy algorithm is to narrow the difference in gender distributions between top-ranked images and the ground truth by moving images up and down. Unlike epsilon-greedy and relevance-aware swapping algorithms, the fairness-greedy algorithm needs to know the ground truth T (i.e., the gender distribution of search terms in real life) and a list of gender labels $\mathbf{G}$ for returned images $\mathbf{L}$ in search engines. The ground truth of a searched profession is usually available through open data, such as census data. The image gender labels, which can be estimated by available computer vision based gender APIs, are required to calculate the gender distribution of top-ranked images.

The detailed implementation of the fairness-greedy algorithm is shown in Algorithm 2. To make our algorithm more general, we use $\mathcal{X}$ to represent all involved features, such as gender features of female and male. Note that the fairness-greedy algorithm is capable of handling more than two different features. We keep the first item in the original rank list as it is at the beginning (see line 3). Starting from the second item, we calculate the gender distribution P over the latest re-ranked list $\mathbf{R}$. P_x represents the ratio of feature $x \in \mathcal{X}$,

Algorithm 2: Fairness-greedy Algorithm

1 **Input**: $\mathbf{L}$: the original image list; T: the ground truth of gender distribution; $\mathcal{X}$: the set of gender features;
2 **Output**: $\mathbf{R}$: the re-ranked image list;
3 $\mathbf{R} \leftarrow [\mathbf{L}_1]$; // initialize $\mathbf{R}$ as $[\mathbf{L}_1]$
4 **for** $i = 2 \rightarrow |\mathbf{L}|$ **do**
5 $P \leftarrow$ gender distribution on $[\mathbf{G}_{\mathbf{R}_1}, ..., \mathbf{G}_{\mathbf{R}_{i-1}}]$;
6 $flag \leftarrow False$;
7 $x_{min} \leftarrow None$; // most underrep. feat.
8 $\mathcal{C} \leftarrow \varnothing$; // set checked features as $\varnothing$
9 **while** $(flag = False)$ *and* $(\mathcal{C} \neq \mathcal{X})$ **do**
10 $d_{min} \leftarrow 0$;
11 add x_{min} to $\mathcal{C}$; // update $\mathcal{C}$
  /* select most underrep. feature */
12 **for** $x \in \mathcal{X} - \mathcal{C}$ **do**
13 $d = P_x - T_x$; // diff. in feat. x
14 **if** $d \leq d_{min}$ **then**
15 $x_{min} \leftarrow x$; // underrep. feat.
16 **end**
  /* find 1^{st} item w/ underrep. feat. */
17 **for** $j = i \rightarrow |\mathbf{L}|$ **do**
18 **if** $\mathbf{G}_{\mathbf{L}_j} = x_{min}$ **then** // find an item
19 $temp \leftarrow \mathbf{L}_j$; // save $\mathbf{L}_j$
20 **for** $k = i + 1 \rightarrow j$ **do**
21 $\mathbf{L}_k \leftarrow \mathbf{L}_{k-1}$ // move down
22 **end**
23 $\mathbf{L}_i \leftarrow temp$; // update $\mathbf{L}_i$
24 append $\mathbf{L}_i$ to $\mathbf{R}$; // update $\mathbf{R}$
25 $flag \leftarrow True$; // find the item
26 break;
27 **end**
28 **end**
29 **end**
30 **return** $\mathbf{R}$

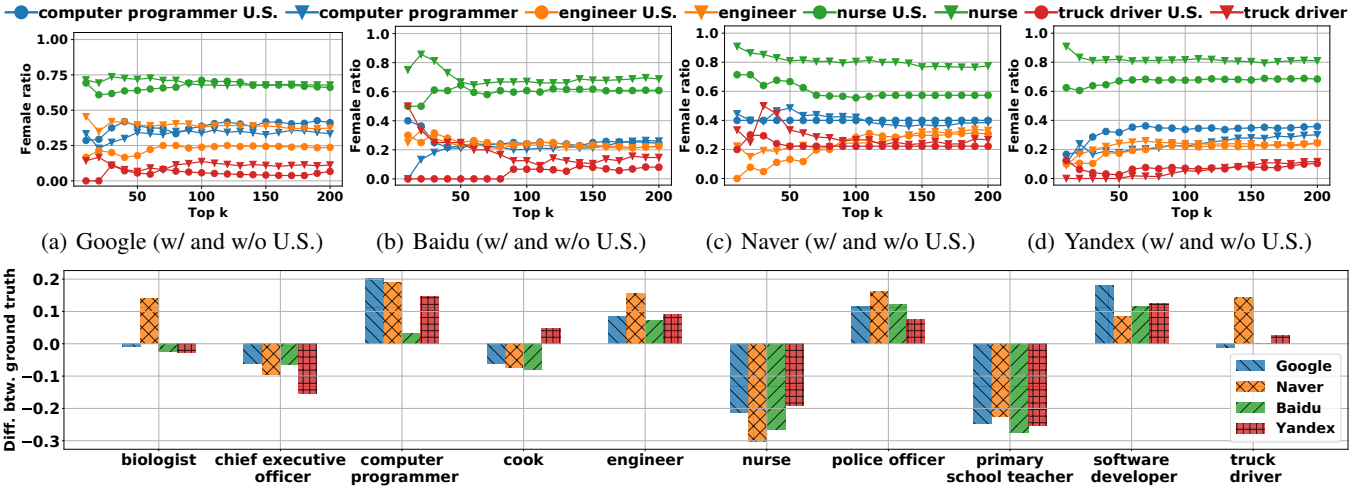
and T_x is the ground truth of feature x in the real world. Next, we take a two-step re-ranking method to mitigate feature biases. Step 1: identify the most underrepresented feature x_{min} by comparing the difference between P_x and T_x (see line 12-16). Step 2: find the first item $\mathbf{L}_j$ with a feature of x_{min} (i.e., $\mathbf{G}_{\mathbf{L_j}} = x_{min}$) in $\mathbf{L}_{\mathbf{i} \rightarrow |\mathbf{L}|}$ and move it forward as the new $\mathbf{L}_i$ (see line 17-27). If such an item $\mathbf{L}_j$ does not exist, we exclude the feature x_{min} by adding it into the checked feature set $\mathcal{C}$ and continue the re-ranking (see line 9-11).

Experiments and Evaluation

This section presents the evaluations of the three proposed bias mitigation approaches on synthetic and real datasets.

Evaluation on Synthetic Data

We generated three synthetic datasets with different gender distribution patterns: 1) Uniform Dataset: female and male items are distributed evenly across the whole list; 2) Heavy-headed Dataset: female items are aggregated at the top of the list; 3) Heavy-tailed Dataset: female items are aggregated at the bottom of the list. For these experiments, we created a list $\mathbf{L}$ with a length of 200 and set the female ratio as 0.5, i.e., 100 items are labeled as female. On the heavy-headed and heavy-tailed datasets, the 100 female items are distributed at



(a) Google (w/ and w/o U.S.) (b) Baidu (w/ and w/o U.S.) (c) Naver (w/ and w/o U.S.) (d) Yandex (w/ and w/o U.S.)

(e) Difference in the female ratio between image search results (w/ United States in search terms) and the ground truth

Figure 3: Gender distributions of ten occupations in Google, Baidu, Naver, and Yandex image search engines.

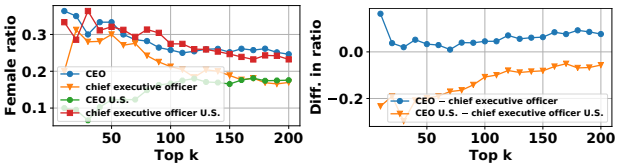
the top 50% and the bottom 50% on the list respectively. We set the ground truth of gender distribution T as $\{female : 0.5, male : 0.5\}$.

The bias mitigation performance of the three proposed algorithms (with 1000 runs) and a widely used fair top-k ranking algorithm named FA*IR (Zehlike et al. 2017) is shown in Table 1. Recall that we used Equation 1 to measure the bias. As expected, neither epsilon-greedy nor relevance-aware swapping algorithms can mitigate bias by introducing randomness on the uniform dataset, because the original list has already been randomized entirely. For the same reason, FA*IR also fails to improve the fairness of the uniform dataset. On heavy-headed and heavy-tailed datasets, if more randomness is introduced (a larger ϵ in epsilon-greedy algorithm and a larger ρ in relevance-aware swapping algorithm), the bias is more mitigated. FA*IR also reduces gender bias significantly. The fairness-greedy algorithm performs best on all three datasets.

Evaluation on Real-world Data

We conducted adversarial attacks on gender fairness in four major image search engines, where various gender distributions are observed for the same search term. We also found that image search engines are sensitive to the search term variants that convey the same semantics. Finally, we evaluated the performances of the three proposed bias mitigated algorithms on the collected dataset. This subsection presents the details of these evaluations.

Gender Bias in Cross-culture Search Engines Besides the Google image search engine, we evaluated the same occupation terms in Baidu from China, Naver from South Korea, and Yandex from Russia. Using the hybrid image gender detection method (see the subsection of Gender Detection), similar to Google, all the above three image search engines are deemed to have a gender bias with search terms that include 'United States' in them (see Figure 3(e), where a positive value indicates over-representing females and a negative value indicates under-representing females). The effectiveness of the proposed adversarial attack approach in cross-culture search engines is demonstrated in Figure 3(a) to Figure 3(d), where the difference in female ratios between search terms with and without 'United States' is evident, especially among the top 50 items. We can also observe that distinct occupations demonstrate different gender distribution patterns in the same search engine, and the same occupation may demonstrate different patterns across search engines. These findings led us to consider that such gender bias exists across cultures and needs attention globally.



(a) CEO vs Chief Exec. Officer (b) Difference in female ratio

Figure 4: Sensitive to variant search terms.

Sensitive to Variant Search Terms Another evidence of the unsystematic mitigation of gender bias is that image search engines are sensitive to variant search terms. As shown in Figure 4, the female ratios of image search results between CEO and chief executive officer are significantly different, especially when search terms include 'United States.' However, with the increase of top k, the difference in the female ratio demonstrates a trend to become stable and small, especially for search terms containing 'United States.' (see Figure 4(b)).

Gender Bias Mitigation We deployed the three proposed algorithms on the image search datasets collected from Google, Baidu, Naver, and Yandex. To illustrate how the proposed algorithms work, we take the epsilon-greedy algorithm as an example to show the dynamic fairness achievements on 'biologist' datasets, as shown in Figure 5. As ϵ increases (i.e.,

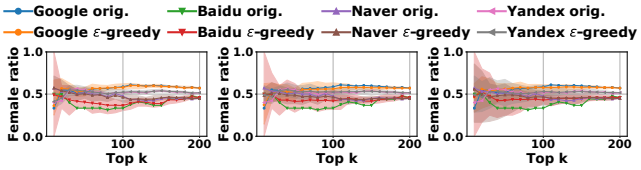
	Original	Epsilon-greedy			Relevance-aware Swapping			Fair-greedy	FA*IR
		ϵ=0.2	ϵ=0.4	ϵ=0.6	ρ=0.2	ρ=0.4	ρ=0.6		p=0.5 α=0.1
Uniform	0.066	0.059±0.019	0.055±0.025	0.052±0.028	0.065±0.013	0.064±0.018	0.063±0.022	**0.020**	0.066
Heavy-headed	2.046	0.426±0.189	0.203±0.107	0.105±0.063	0.553±0.222	0.316±0.143	0.198±0.095	**0.020**	0.142
Heavy-tailed	2.046	0.423±0.199	0.194±0.096	0.102±0.061	0.548±0.219	0.312±0.136	0.198±0.098	**0.020**	0.142

Table 1: Bias mitigation performance on synthetic datasets. The bias value in the table is measured by Equation 1.

	Original	Epsilon-greedy			Relevance-aware Swapping			Fair-greedy	FA*IR
		ϵ=0.2	ϵ=0.4	ϵ=0.6	ρ=0.2	ρ=0.4	ρ=0.6		p=0.5 α=0.1
biologist U.S.	0.138	0.102±0.044	0.087±0.046	0.071±0.049	0.128±0.032	0.108±0.046	0.114±0.046	**0.018**	0.072
ceo U.S.	0.172	0.175±0.055	0.160±0.082	0.144±0.087	0.169±0.048	0.167±0.052	0.160±0.054	**0.021**	0.084
comp. programmer U.S.	0.114	0.119±0.027	0.120±0.030	0.135±0.062	0.113±0.021	0.114±0.030	0.120±0.035	**0.034**	0.071
cook U.S.	0.149	0.131±0.051	0.109±0.064	0.101±0.070	0.148±0.049	0.133±0.052	0.128±0.064	**0.017**	0.102
engineer U.S.	0.04	0.044±0.011	0.053±0.019	0.063±0.036	0.045±0.022	0.048±0.016	0.052±0.022	**0.02**	0.027
nurse U.S.	0.115	0.119±0.011	0.119±0.015	0.128±0.023	0.118±0.009	0.121±0.015	0.124±0.017	**0.066**	0.076
police officer U.S.	0.049	0.053±0.015	0.054±0.016	0.055±0.018	0.048±0.008	0.047±0.011	0.046±0.013	**0.015**	0.088
prim. school teacher U.S.	0.135	0.136±0.007	0.136±0.010	0.137±0.011	0.137±0.006	0.136±0.008	0.137±0.009	0.1	**0.085**
software developer U.S.	0.189	0.193±0.066	0.171±0.078	0.156±0.082	0.193±0.035	0.180±0.061	0.184±0.067	**0.055**	0.094
truck driver U.S.	0.056	0.067±0.044	0.088±0.062	0.088±0.067	0.070±0.044	0.074±0.048	0.087±0.064	**0.007**	0.02

Table 2: Bias mitigation performance on Google occupation image datasets. The bias value is measured by Equation 1.

more randomness is introduced), the gender distribution of the re-ranked list becomes more likely to be different from the original one (see the shaded range), implying more fairness will be achieved if the raw image search list suffers from severe gender bias. With the increase of top k, the female ratio becomes more stable and finally converges when top k reaches 200.



(a) epsilon = 0.2 (b) epsilon = 0.4 (c) epsilon = 0.6

Figure 5: Performance of the epsilon-greedy algorithm on Google, Baidu, Naver, and Yandex 'biologist' datasets.

Similar to the evaluations on synthetic datasets, we explored the performance of our algorithms and FA*IR (Zehlike et al. 2017) on real-world datasets. Table 2 illustrates the gender mitigation performance of each algorithm on 10 Google image datasets, which were collected with the search keywords of 10 occupations plus 'United States.' When the original bias is larger than 0.1 (e.g., biologist United States), gender bias normally decreases along with the increase of ϵ in the epsilon-greedy algorithm and ρ in the relevance-aware swapping algorithm. However, if the original bias is small (e.g., engineer United States), epsilon-greedy algorithm and relevance-aware swapping algorithm cannot mitigate gender bias. We can observe that the fairness-greedy algorithm consistently achieves a low bias because it gives the highest priority to fairness during re-ranking. FA*IR also demonstrates a stable and good performance regardless of the original bias. In addition, comparing the result columns of Original and Fairness-greedy in Table 2 can tell the degree of gender bias hidden in the original image list.

Conclusion and Limitation

Bias in AI systems has become an increasingly prevalent and complex issue to address. Often the system developers fix a problem by creating a superfluous solution without addressing the underlying issue. In this paper, we used an adversarial query attack method by appending additional information like country names to trigger potential gender bias in image search. An open-sourced Cross-search-engine Image Retrieval Framework (CIRF) was developed to retrieve data from Google, Baidu, Naver, and Yandex. To recognize the gender of people in photos, five popular image gender detection APIs, namely Amazon Rekognition, Luxand, Face++, Microsoft Azure, and Facebook DeepFace, were evaluated. Although these APIs are endorsed by AI giants, they could not always handle images in the wild with high accuracy. Therefore, a hybrid method combining automatic gender detection APIs and crowdsourced human workforce was designed to label image genders. To mitigate gender bias, we proposed three lightweight and interpretable re-ranking algorithms and evaluated their performance on both synthetic and real-world datasets. Our results demonstrated that it is possible (and advisable) to address bias in image search (and perhaps in other types of search as well) in a systematic, sustainable, and more meaningful way than doing individual query fixes in an ad hoc fashion.

In this paper, we treated gender as a binary attribute inferred by either gender APIs or humans. However, we acknowledge that gender is different from biological sex, and is non-binary. It is also something that a third-party — be it a human or a program — is not always in a position to detect genders correctly. Our reliance on the binary gender norm and third-party annotation is a limitation of this research.

References

Adeli, E.; Zhao, Q.; Pfefferbaum, A.; Sullivan, E. V.; Fei-Fei, L.; Niebles, J. C.; and Pohl, K. M. 2021. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2513–2523.

Baker, L. 2018. Amazon's Search Engine Ranking Algorithm: What Marketers Need to Know. https://www.searchenginejournal.com/amazon-search-engine-ranking-algorithm-explained/265173/. Accessed: 2022-02-10.

Berry, D. A.; and Fristedt, B. 1985. Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability). *London: Chapman and Hall*, 5(71-87).

Dhar, P.; Gleason, J.; Souri, H.; Castillo, C. D.; and Chellappa, R. 2020. Towards Gender-Neutral Face Descriptors for Mitigating Bias in Face Recognition. *arXiv preprint arXiv:2006.07845*.

Ellemers, N. 2018. Gender stereotypes. *Annual review of psychology*, 69: 275–298.

Gao, R.; and Shah, C. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management*, 57(1): 102138.

Griffith, E. 2021. Algorithms Still Have a Bias Problem, and Big Tech Isn't Doing Enough to Fix It. https://www.pcmag.com/news/algorithms-still-have-a-bias-problem-and-big-tech-isnt-doing-enough-to. Accessed: 2022-02-10.

Hashemi, M.; and Hall, M. 2020. RETRACTED ARTICLE: Criminal tendency detection from facial images and the gender bias effect. *Journal of Big Data*, 7(1): 1–16.

Hibbing, A. N.; and Rankin-Erickson, J. L. 2003. A picture is worth a thousand words: Using visual images to improve comprehension for middle school struggling readers. *The reading teacher*, 56(8): 758–770.

Hwang, S.; Park, S.; Kim, D.; Do, M.; and Byun, H. 2020. FairfaceGAN: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*.

Järvelin, K.; and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4): 422–446.

Jia, S.; Meng, T.; Zhao, J.; and Chang, K.-W. 2020. Mitigating gender bias amplification in distribution by posterior regularization. *arXiv preprint arXiv:2005.06251*.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

Lam, O.; Broderick, B.; Wojcik, S.; and Hughes, A. 2018. Gender and Jobs in Online Image Searches. *Pew Social Trends. Retrieved March*, 14: 2020.

Makhortykh, M.; Urman, A.; and Ulloa, R. 2021. Detecting race and gender bias in visual representation of AI on web search engines. In *International Workshop on Algorithmic Bias in Search and Recommendation*, 36–50. Springer.

Metaxa, D.; Gan, M. A.; Goh, S.; Hancock, J.; and Landay, J. A. 2021. An image of society: Gender and racial representation and impact in image search results for occupations. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23.

Mozilla. 2021. Internet Health Report 2020. https://creativecommons.org/licenses/by/4.0. Accessed: 2021-09-15.

Otterbacher, J.; Bates, J.; and Clough, P. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems*.

Otterbacher, J.; Checco, A.; Demartini, G.; and Clough, P. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR conference on research & development in information retrieval*, 933–936.

Park, S.; Hwang, S.; Kim, D.; and Byun, H. 2021. Learning Disentangled Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2403–2411.

Schwemmer, C.; Knight, C.; Bello-Pardo, E. D.; Oklobdzija, S.; Schoonvelde, M.; and Lockhart, J. W. 2020. Diagnosing gender bias in image recognition systems. *Socius*, 6: 2378023120967171.

Serna, I.; Peña, A.; Morales, A.; and Fierrez, J. 2021. InsideBias: Measuring bias in deep networks and application to face gender biometrics. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 3720–3727. IEEE.

Sharma, D.; Shukla, R.; Giri, A. K.; and Kumar, S. 2019. A brief review on search engine optimization. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 687–692. IEEE.

Singh, V. K.; Chayko, M.; Inamdar, R.; and Floegel, D. 2020. Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology*, 71(11): 1281–1294.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Voorhees, E. M.; et al. 1999. The TREC-8 question answering track report. In *Trec*, volume 99, 77–82. Citeseer.

Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.-W.; and Ordonez, V. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5310–5319.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8919–8928.

Wijnhoven, F. 2021. Search engine gender bias. *Frontiers in big Data*, 4: 29.

Xu, T.; White, J.; Kalkan, S.; and Gunes, H. 2020. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, 506–523. Springer.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1569–1578.

Zhao, C.; and Chen, F. 2019. Rank-based multi-task learning for fair regression. In *2019 IEEE International Conference on Data Mining (ICDM)*, 916–925. IEEE.