Counterfactually Measuring and Eliminating Social Bias in Vision-Language Pre-training Models

Yi Zhang∗
School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
yi.zhang@bjtu.edu.cn

Junyang Wang
School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
21120406@bjtu.edu.cn

Jitao Sang†
∗School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
†Peng Cheng Lab, Shenzhen 518066, China
jtsang@bjtu.edu.cn

ABSTRACT
Vision-Language Pre-training (VLP) models have achieved state-of-the-art performance in numerous cross-modal tasks. Since they are optimized to capture the statistical properties of intra- and inter-modality, there remains risk to learn social biases presented in the data as well. In this work, we (1) introduce a counterfactual-based bias measurement CounterBias to quantify the social bias in VLP models by comparing the [MASK]ed prediction probabilities of factual and counterfactual samples; (2) construct a novel VL-Bias dataset including 24K image-text pairs for measuring gender bias in VLP models, from which we observed that significant gender bias is prevalent in VLP models; and (3) propose a VLP debiasing method FairVLP to minimize the difference in the [MASK]ed prediction probabilities between factual and counterfactual image-text pairs for VLP debiasing. Although CounterBias and FairVLP focus on social bias, they are generalizable to serve as tools and provide new insights to probe and regularize more knowledge in VLP models.

CCS CONCEPTS
• Social and professional topics → Computing / technology policy; • Applied computing → Law, social and behavioral sciences; • Computing methodologies → Machine learning.

KEYWORDS
Fairness in Machine Learning, Responsible Artificial Intelligence

ACM Reference Format:
Yi Zhang, Junyang Wang, and Jitao Sang. 2022. Counterfactually Measuring and Eliminating Social Bias in Vision-Language Pre-training Models. In Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548396

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10...
https://doi.org/10.1145/3503161.3548396

1 INTRODUCTION

Figure 1: We propose to measure social bias in VLP model by comparing the [MASK]ed prediction probabilities of factual and counterfactual samples.

The past few years have witnessed the rapid development of Vision-Language Pre-training (VLP) models [2, 4, 17, 38], and task-specific finetune on VLP models has become a new and state-of-the-art paradigm in many multimedia tasks [20, 21, 32]. Beyond accuracy, fairness which concerns about the discrimination towards socially protected or sensitive groups plays a critical role in trustworthy deployment of VLP models in downstream tasks. This is highly endorsed by many VLP model designers [16]. Yet, very limited studies have begun to examine the fairness problem in VLP models, which typically apply the bias measurement designed for the unimodal pre-training (e.g., text-based language models [1]). The characteristics of VLP models regarding the interaction between visual modality and textual modality are largely ignored. This work is thus devoted to measuring and eliminating the social bias in VLP models by emphasizing/addressing its multimodal characteristics, with the goal of helping understand the existed social bias and inspiring the development of fair VLP models.

Measuring the social bias in pre-training model essentially involves the correlation between bias and target concepts. The cross-modal interaction of VLP models poses two main challenges for bias measurement. The first challenge is on modeling the bias and target concepts. The characteristics of VLP models regarding the interaction between visual modality and textual modality are largely ignored. This work is thus devoted to measuring and eliminating the social bias in VLP models by emphasizing/addressing its multimodal characteristics, with the goal of helping understand the existed social bias and inspiring the development of fair VLP models.
calculating the distance in the embedding space as in [26] not applicable. Inspired by the Masked Language Modeling (MLM) pre-training task commonly used in VLP models, we employ prompt-based query to probe the modeling of bias and target concepts in VLP models. As illustrated in Figure 1, the concepts to be modeled are masked before issuing into VLP models. The resultant MLM prediction probability of [MASK]ed token $P([\text{MASK}]=\text{“shopping”})$ manifests VLP model’s modeling of the concept given the input image-text pair.

The second challenge is on quantifying the bias-target association. With MLM prediction probability for concept probing, a natural idea is to compare the prediction probabilities of the target concept in the observed image-text pairs of different bias concepts (e.g., the difference of $P([\text{MASK}]=\text{“shopping”})$) between male and female inputs. However, relying on the observed data suffers from two problems: (1) There exists no adequate number of image pairs with only different bias concepts. For example, when measuring the social bias of “shopping” over gender, it is difficult to collect pairs of identical shopping images except for different genders. Other information beyond gender like the different background concepts can largely affect the MLM prediction probability, as detailed discussed in Section 3 for the example in Figure 2. (2) The bias measurement result is highly sensitive to the observed dataset. Examining image-text datasets with different distributions will lead to inconsistent observations. To address this, we propose a counterfactual-based bias measure method (CounterBias) to quantify the social bias in VLP models exempted from the limitation in observed data pairs. Specifically, we design a model-specific adversarial attack scheme to generate counterfactual images by altering the bias concept from factual images (e.g., attacking woman to man as illustrated in Figure 1). This successfully alters the bias concept without changing other information. The resultant MLM prediction difference between factual and counterfactual image-text pairs thus manifests the social bias of the examined VLP models.

Since there exists no off-the-shelf dataset specially designed to analyze the social bias in VLP models, we propose VL-Bias, a dataset to facilitate the study on understanding model bias in VLP models. VL-Bias contains 52 activities and 13 occupations with a total of 24k image-text pairs. Using CounterBias, we examined two typical VLP architectures, dual-stream and single-stream, on VL-Bias dataset. Several key observations include: (1) Social bias is prevalent in the examined three VLP models, and exists in both visual and language modalities; (2) The gender bias contained in VLP models is basically consistent with human gender stereotypes; (3) The examined single-stream ViLT and dual-stream ALBEF/TCL show different cross-modal conformities regarding the observed unimodal social bias; (4) The examined VLP models contain stronger gender bias than unimodal pre-training model BERT.

Inheriting the idea of measuring social bias, we propose a simple counterfactual-based method to eliminate the social bias in VLP models. Specifically, we first generate counterfactual samples of bias concepts in visual and language modalities, respectively. Then, the difference in the predicted [MASK]ed probability between factual and counterfactual image-text pairs is minimized, so that preventing the model from learning the association between bias and target concepts. Experimental results demonstrate the effectiveness in benchmark datasets.

Our contributions can be summarized as follows:

- We propose to employ model-specific adversarial attacks on bias concepts to generate counterfactual samples, and counterfactualy measure social bias at instance-level by comparing the [MASK]ed prediction probabilities of factual and counterfactual samples.
- We introduce VL-Bias, a dataset for studying gender bias in VLP models. Observations are derived regarding typical VLP architectures, the consistency with human stereotypes, and the comparison with unimodal pre-training models.
- We propose a counterfactual bias elimination method, which eliminates bias by minimizing [MASK]ed probability discrepancy between factual and counterfactual samples.

2 BACKGROUND AND RELATED WORKS

2.1 Vision-Language Pre-training(VLP)

Benefited from the cross-modal interactions performed by the transformer [33], Vision-Language Pre-training (VLP) has improved performance on various joint vision-and-language(V+L) downstream tasks [8, 12, 20, 25, 29]. There are two mainstream architectures for bridging the cross-modal semantic gap: dual-stream architecture and single-stream architecture. The former first encodes visual and language modalities separately and then fuses the two representations through a transformer network; the latter collectively operates on a concatenation of image and text inputs. In the work, we consider ALBEF [16] and TCL [36] as example to examine the former, and ViLT [13] as example to examine the latter.

2.2 Bias Measurement and Elimination in VLP

Much research work [1, 41] has been done on evaluating bias in unimodal language models(LM). [1] proposed cosine-based word embedding bias measure, and showed that word embeddings like GloVe [24] encode gender biases. Furthermore, [27] investigates two unimodal biases in VL-BERT [28]. However, there exists very limited work for such study in multimodal bias of VLP due to the cross-modal interaction. As far as we know, only [26] applied the cosine-based word embedding bias measure to multimodal bias of VLP models. However, [15] proved this method is not suitable for the contextual setting, as direct word will be affected by the context of image and text, resulting in inconsistent measure result.
Factual target probability

\[ P(T | V-L_f) \]

CounterFactual target probability

\[ P(T | V-cf) \]

Bias Measure

\[ \frac{P(T | V-L_f) - P(T | V-cf)}{P(T | V-L_f) - 1} \]

| Bias Source | Factual target probability | CounterFactual target probability | Bias Measure |
|-------------|---------------------------|----------------------------------|--------------|
| Multimodal  | \[ P(T | V-L_f) \]          | \[ P(T | V-cf) \]                |              |
| Vision      | \[ \text{The man is [mask]} \] | \[ P(T | V-cf) \]                |              |
| Language    | \[ \text{The woman is [mask]} \] | \[ P(T | V-cf) \]                |              |

Figure 3: The proposed CounterBias to counterfactually measure social bias between target concept \( T \) and bias concept \( B \).

This motivates us to design bias measure scheme to consider the multimodal nature of the input in VLP models. For bias elimination, there has been works [30, 35] focused on eliminating bias in unimodal language models. However, eliminating the social bias in VLP models remain unexplored till now.

## 3 COUNTERFACTUALLY MEASURING SOCIAL BIAS IN VLP

The pre-training task of Masked Language Modeling (MLM) has been recognized as confidentially manifesting VLP model behavior by maximizing the mutual information between [MASK]ed token and the image-text input [16], and [15, 27] utilized [MASK]ed token to probe the model’s response of its context to measure unimodal contextual bias. These provide theoretical basis and inspire us to use the prediction of [MASK]ed token to probe VLP model’s understanding of the bias and target concepts. Specifically, when we construct the input to VLP model with image and text template masking off certain target concept \( T \) (e.g., template of "the woman is [MASK]") masking off the activity concept "shopping" in Figure 1), the output prediction probability can indicate how VLP model probes the concept from the given factual image-text pair.

We denote the calculation of factual target probability as follows:

\[
P(T | V-L_f) = P([MASK] = T | V_f, L_f) \tag{1}
\]

Social bias essentially refers to how the change of bias concept in the input influences the model’s inference on the target concept. Note that ideally only the involved bias concept is changed but fixing all the other information, so as to accurately examine the inherent influence from changed bias concept to target concept inference. However, as discussed in Section 1, it is almost impossible to perfectly control and keep the other variables unchanged. Figure 2 shows some "shopping" images with their predicted factual target probability. We can see among images with the same bias concept (in the same row), the diversity of other variables like environment and background objects imposes obvious influence on the predicted target probability, even more significant than the bias concepts (between the two rows). This indicates relying on the observed data for bias measurement is unreliable. To avoid the negative effects of image diversity as a confounder, we propose CounterBias to counterfactually change the bias concepts from observed image-text pairs without changing other information, and quantify the social bias by examining the difference between the predicted target probability of observed factual data \( P(T | V-L_f) \) and generated counterfactual data \( P(T | V-cf) \).

Regarding how to counterfactually change the bias concepts in the visual modality, a direct idea is to use GANs to edit images. However, It is generally accepted that GANs-based editing can unintentionally change other information (e.g., image background) [3, 7, 10, 11, 23], which defeats the purpose of controlling other information unchanged.

**Model-specific adversarial attacks.** To avoid changing information other than bias concept, we propose model-specific adversarial attacks to counterfactually change the bias concepts in image. In addition to affecting other information as little as possible, model-specific adversarial attacks ensure that all bias features that the VLP model relies on are counterfactually altered. Specifically, we create a bias prompt template with masking the factual bias concept, "a [MASK] is in the picture", as text input to guide the VLP model to classify bias concept \( B \) (e.g., woman) in the image input. Then we use the prediction probability \( P(B | V_f) \) of the [MASK]ed bias concept word to represent the bias concept that the model extracts from the factual image \( V_f \):

\[
P(B | V_f) = P([MASK]=B|V_f, bias prompt) \tag{2}
\]

where \( B \) is the factual bias concept of input image. Similar to traditional adversarial attacks [6, 14, 22], using cross-entropy loss between \( P(B | V_f) \) and counterfactual goal \( B' \), and search the adversarial perturbation \( \delta \) based on gradient of VLP model so that bias concept of \( V_f + \delta \) is close to \( B' \):

\[
\delta = \arg \min_{||\delta|| \leq e} \mathcal{L} \left( P([MASK]|V_f + \delta, bias prompt), B' \right) \tag{3}
\]

where \( B' \) represents counterfactual goals, if \( B \) is woman, \( B' \) is man. And \( V_f + \delta \) is generated counterfactual image \( V_{cf} \).

Regarding how to counterfactually change bias concepts in language modality, we manually replace the bias concept word in factual text \( L_f \), e.g. gendered word woman is replaced by man. The generated counterfactual text is denoted as \( L_{cf} \).

Note that the bias concept is continuous in images rather than discrete, e.g., the intensity of gender features in images of woman is continuous, so the change of \( P(B|V-L) \) also needs to be considered in the measure of social bias. The bias concepts in factual and counterfactual images, \( P(B | V_f) \) and \( P(B | V_{cf}) \), are formulated as the predicted probability of the [MASK]ed biased concept word, and 4998
similar to Equation 2. Considering that the bias concept in text is binarized, $P(B | L_f)$ can thus be directly defined as 1, and $P(B | L_{cf})$ is 0. The bias concept $P(B | V, L)$ in the image-text pairs can be formulated as the average of the $P(B | V)$ and $P(B | L)$.

Once factual probability and counterfactual probability in target concept and bias concept are calculated, we can measure how the counterfactual change of bias concept influences the model’s inference on the target concept, as summarized in Figure 3. Specifically, the involved social bias between bias concept $B$ and target concept $T$ can be measured as:

$$bias(T, B) = \frac{P(T | V_{cf}, L_f) - P(T | V, L_f)}{P(B | V_{cf}, L_f) - P(B | V, L_f)}$$

(4)

Considering the multi-modal characteristic of VLP, we are interested in measuring social bias both in multimodality, and also in visual and language unimodality. The difference between measuring multimodal and unimodal biases lies in the counterfactual calculations. The following two subsections described measuring multimodal bias and unimodal bias in detail.

3.1 Measuring Multimodal Bias

Multimodal Bias refers to how the change of bias concept in both the visual and language modalities influences the model’s inference on the target concept. Given counterfactual image-text pair $(V_{cf}, L_{cf})$ as input, we calculate the definition of counterfactual target probability as follows:

$$P(T | V_{cf}, L_{cf}) = P([MASK] = T | V_{cf}, L_{cf})$$

(5)

The multimodal bias $bias_{VL}(T, B)$ can be formulated as:

$$bias_{VL}(T, B) = \frac{P(T | V_{cf}, L_{cf}) - P(T | V, L_f)}{P(B | V_{cf}, L_{cf}) - P(B | V_f)}$$

(6)

$bias_{VL}(T, B)$ is an instance-level bias measure for model bias. Positive values indicate target concept $T$ biased towards $B$, and negative values indicate that the bias direction of $T$ is opposite to $B$. Further, we also measure dataset-level social bias, by applying indicator function to align the bias directions represented by positive values, and calculate the average bias over a set of instances $S_T$ which contain the specific target concept $T$:

$$bias_{VL}(T) = \mathbb{E}_{S_T} \{ I_{B_0} \cdot bias_{VL}(T, B) + I_{B_1} \cdot bias_{VL}(T, B) \}$$

(7)

where $B_0$ and $B_1$ are pre-defined, and the sign of $bias_{VL}(T)$ indicates the bias direction - positive for $B_0$, and negative for $B_1$. If $B_0$ is male, positive values indicate $T$ biased towards male.

3.2 Measuring Unimodal Bias

Unimodal bias refers to how the change of bias concept in unimodality influences the model’s inference on the target concept. We use counterfactual images and counterfactual texts to discover visual unimodal bias and language unimodal bias, respectively.

For visual unimodal bias, we calculate the counterfactual target probability from given counterfactual visual $V_{cf}$ input and factual language input $L_f$ as follows:

$$P(T | V_f, L_{cf}) = P([MASK] = T | V_{cf}, L_f)$$

(8)

Then, visual bias $bias_V(T, B)$ can be calculated as follows:

$$bias_V(T, B) = \frac{P(T | V_{cf}, L_f) - P(T | V_f, L_f)}{P(B | V_{cf}, L_f) - P(B | V_f)}$$

(9)

For the language unimodal bias, we calculate the counterfactual target probability from given counterfactual visual $V_f$ input and factual language input $L_{cf}$ as follows:

$$P(T | V_f, L_{cf}) = P([MASK] = T | V_f, L_{cf})$$

(10)

Then, language bias $bias_L(T, B)$ can be calculated as follows:

$$bias_L(T, B) = \frac{P(T | V_f, L_{cf}) - P(T | V_f, L_f)}{P(B | L_{cf}) - P(B | V_f)}$$

(11)

4 MEASURING RESULTS

4.1 VL-Bias Dataset

Fairness research in CV mainly focuses on human face analysis [9, 34, 39], since existing face datasets such as CelebA [19] and UTKface [40] have detailed annotation of both bias and target concepts of faces. For fairness in NLP models, researchers can measure the social bias of NLP models by means of manually constructing language templates [31]. However, in vision-language research bridging CV and NLP, few studies focus on fairness problems due to the lack of both bias and target concepts annotation.

To address this issue, we construct VL-Bias to facilitate the study on analyzing social bias in VLP models. Using gender as the bias concept example, we selected 52 activities and 13 occupations related to humans. For image collection, in addition to downloading human-related images from the Internet, we also collected human-related images from the existing image datasets, such as MS-COCO [18], Flickr [37].

Figure 4: Example images in the collected VL-Bias dataset.

Table 1: Templates for caption generation.

| Type                      | Template                                                                 |
|---------------------------|--------------------------------------------------------------------------|
| Sentence w/o specification| The [gender] is [MASK]                                                   |
| Sentence w/ specification | The [gender] in the photo is [MASK]                                     |
| Phrase w/o specification  | A [gender] who is [MASK]                                                 |
| Phrase w/ specification   | A photo of a [gender] who is [MASK]                                     |
Then, we annotated gender concepts, activity concepts, and occupation concepts of the images and removed the images that did not explicitly contain these concepts. We illustrate some annotated examples in Figure 4.

We also generate corresponding captions for each image according to the annotation. Consider different languages and grammars have different effects on the prompt of [MASK], we use four templates described in Table 1 to generate captions. Finally, for each template, we have collected 24k image-text pairs, including 13K for the 52 activities and 11K for the 13 occupations.

### 4.2 The Observed Social Bias in VLP

Current VLP models can be divided into dual- and single-stream models according to architectures [13]. In dual-stream VLP models, the visual and textual inputs are first processed by two uni-modal encoders respectively, then the resulting representations are fed into cross-modal Transformer layers. For single-stream VLP models use the concatenation of the image-text pair directly as the input of cross-modal Transformer layers. To deeply analyze the social bias in VLP models, we selected the representative VLP models of these two architectures respectively. For dual-stream, we consider ALBEF and TCL, and for single-stream, we consider ViLT. Taking gender bias as an example, we evaluate each VLP model on our VL-Bias dataset.

![Figure 5: The human gender stereotypes (Left) and the observed multimodal bias (B@V-L) of three examined VLP models on different activities.](image)

![Figure 6: Example of the calculated social bias in VLP model ALBEF.](image)

| Activity 13K | Occupation 11K |
|-------------|---------------|
| B@V-L | B@V | B@L | B@V | B@L |
| ALBEF |        |        |
| S/W/O | 10.29 | 9.44 | 3.98 | 13.31 | 9.62 | 2.38 |
| S/W | 11.33 | 6.54 | 7.15 | 16.00 | 10.46 | 6.46 |
| P/W/O | 10.75 | 8.21 | 4.27 | 11.85 | 7.85 | 1.69 |
| P/W | 9.85 | 7.54 | 3.31 | 11.31 | 8.23 | 2.08 |
| Avg. | **10.56** | **7.93** | **4.68** | **13.12** | **9.04** | **3.15** |
| TCL |        |        |
| S/W/O | 7.62 | 6.96 | 3.67 | 12.00 | 10.85 | 3.92 |
| S/W | 8.94 | 6.15 | 5.50 | 16.46 | 12.01 | 6.00 |
| P/W/O | 7.52 | 5.90 | 2.65 | 10.15 | 6.85 | 4.77 |
| P/W | 6.88 | 5.38 | 2.37 | 8.38 | 5.46 | 2.85 |
| Avg. | **7.74** | **6.10** | **3.55** | **11.75** | **8.79** | **4.84** |
| ViLT |        |        |
| S/W/O | 8.62 | 6.52 | 2.37 | 16.15 | 17.23 | 3.62 |
| S/W | 9.48 | 8.71 | 2.35 | 14.92 | 17.35 | 4.62 |
| P/W/O | 14.65 | 14.38 | 2.27 | 12.92 | 17.61 | 3.38 |
| P/W | 11.19 | 11.85 | 1.69 | 13.23 | 15.15 | 4.46 |
| Avg. | **10.99** | **10.37** | **2.17** | **14.31** | **16.84** | **4.02** |

Using the multimodal bias measurement as example, given target concept T and bias concept B, instance-level social bias can be calculated according to Equation 6. One example is illustrated in Figure 6. We can see the instance-level measurement provides intuitive understanding of the involved social bias. By aggregating instances with the same target concept, we can derive the final dataset-level social bias according to Equation 7. The following elaborates some of the key observations.

**Social bias is prevalent in the examined VLP models.** We report the gender bias of the 52 activities and 13 occupations in Table 2, where each element represents the average absolute value of the measured bias results. Main observations include: (1) all three VLP models involve with significant gender bias. Taking multimodal bias (B@V-L) of ALBEF in activities as an example, 10.56% means that by altering the gender information of the input, the model’s predicted probability of [MASK]ed activity will change by an average of 10.56%. (2) Gender bias exists in both modalities of VLP model and is more serious in visual modality (B@V) than in language modality (B@L). We conjecture that this is because images have richer information than text, which leads the model to capture
more statistical properties within the visual modality. This further leads to stronger social bias in visual modality than in language modality. (3) Bias measurement is sensitive to the choice of text input template, and using the average(Avg. in Table 2) of multiple templates as the measurement result is more generalized.

The gender bias contained in VLP models is basically consistent to human gender stereotypes. To investigate whether the social bias involved in VLP models matches that from human stereotypes, we recruited 100 Amazon Mechanical Turk workers for labeling the stereotypes for the 52 activities and retained 19 activities with agreed gendered stereotyped label \(^7\) (male or female-biased, illustrated in Figure 5 (Left)).

In Figure 5 (Right), we plot the multimodal bias (B@V-L) of examined VLP models with each activity, marked with human stereotypes using colored points(red for male-biased, green for female-biased, gray for neutral). The sign of B@V-L indicates the direction of gender bias, positive for “male,” and negative for “female”. We observed that the social bias B@V-L manifests is basically consistent with human gender stereotypes, which indicates that VLP models inherit human stereotypes from the training data. In the three examined VLP models, compared to ALBEF and TCL, ViLT slightly deviates from human stereotypes in some of the activities, such as ‘skating’. This may be related to the inconsistency in the bias direction between visual and language modalities in the following analysis.

The examined dual-stream ALBEF/TCL and single-stream ViLT show different cross-modal conformities regarding the observed unimodal social bias. Unlike unimodal pre-training models, there are complex vision-language cross-modal interactions in VLP models. The relationship between the social bias in two modalities may be of interest. Taking dual-stream ALBEF and single-stream ViLT as examples, in Figure 7, we use scatter plots to compare the difference between visual modal bias (x-axis) and language modal bias (y-axis). Each point corresponds to an activity, with three colors representing different human stereotypes. For ALBEF (Figure 7(a)), it is easy to find the strong conformity

\(^7\) For each activity, we asked 100 distinct AMT workers to label it as male-biased, female-biased or neutral. An activity obtains gender-biased stereotypes label only when at least 90% works offer the same label. The rest 33 activities are associated with neutral label in the resultant human stereotype.

### Table 3: The social bias in language modality of three VLP models and language model BERT.

| Target concept | ALBEF | TCL | ViLT | BERT |
|----------------|------|-----|------|------|
| 52 Activities  | 0.53 | 0.59| 0.39 | 0.15 |
| 13 Occupations | 0.69 | 0.71| 0.48 | 0.20 |

between visual modal bias and language modal bias: all activities are basically located in the first and third quadrants of the coordinate system, which means that the discovered gender biases in two modalities follow the same direction. Moreover, the bias of two modalities in 84% of activities is consistent with human stereotypes. However, the bias in ViLT is different from ALBEF, as shown in Figure 7(b), visual modal bias and language modal bias have large differences in direction, whit activities widely spreading across the four quadrants. Bias in unimodality is also inconsistent with human stereotypes in some activities, such as visual bias in skating is also inconsistent with humans.

The conformity between visual modal bias and language modal bias in dual-stream ALBEF is possibly attributed to the explicit align constraints between the two unimodal representations (i.e., Image-Text Contrastive Learning pre-training task). By aligning the two modalities in dual-stream, ALBEF forces the two modalities to learn only shared information between visual and textual inputs, resulting in more consistent social biases contained in the two modalities, and biases in both two modalities are also consistent with human stereotypes. Moreover, the inconsistency between the learned information of the two modalities in ViLT again explains the reason for the slight deviation of multimodal bias(B@V-L) and human stereotypes.

The examined VLP models contain stronger gender bias than unimodal pre-training model BERT. It is common practice for VLP models to use BERT \([5]\) as backbone. We investigate the difference in language modal bias between BERT-based VLP models and original BERT.

For a fair comparison, we remove visual input when predicting [MASK]ed target concepts, which ensures that both VLP models and BERT rely only on language input:

\[
P(T | L_f) = P(\text{[MASK]} = T | V_{\text{none}}, L_f) \\
P(T | L_{cf}) = P(\text{[MASK]} = T | V_{\text{none}}, L_{cf})
\]  

(12)
Then, considering that the absence of visual input may result in formulation of MLM, we can utilize both the image $V_f$ between factual and counterfactual samples. Minimize the difference in the prediction of [MASK]ed words $L$.

For both VLP models and BERT, we use Equation 13 to calculate language modal bias. We report the language modal bias of the VLP models with different debiasing pre-training tasks.

| Method       | Activity 13K B@V-L | Activity 13K B@V | Activity 13K B@L | Occupation 11K B@V-L | Occupation 11K B@V | Occupation 11K B@L |
|--------------|---------------------|------------------|------------------|-----------------------|-------------------|-------------------|
| vanilla      | 13.02               | 11.10            | 4.89             | 14.84                 | 14.23             | 3.15              |
| GS           | 11.21               | 9.25             | 4.03             | 12.47                 | 12.35             | 2.98              |
| DR           | 11.17               | 9.54             | 4.39             | 13.52                 | 12.19             | 3.08              |
| CycleGAN     | 10.92               | 9.23             | 4.32             | 12.07                 | 11.35             | 3.01              |
| FairVLP-V    | 8.41                | 7.23             | 3.91             | 9.11                  | 7.98              | 2.86              |
| FairVLP-L    | 8.53                | 7.46             | 3.82             | 11.23                 | 9.61              | 1.77              |
| FairVLP      | 6.97                | 6.76             | 3.29             | 7.74                  | 7.15              | 1.72              |

5 COUNTERFACTUALLY ELIMINATING SOCIAL BIAS IN VLP

5.1 Method

Section 4 observes that social bias is prevalent in the examined VLP models, which can lead to unintended social consequences in downstream tasks. Inferring the idea of counterfactually measuring social bias, we propose a fair Vision-Language Pre-training task FairVLP that eliminates social bias by minimizing the difference in the predicted [MASK]ed probability between factual and counterfactual image-text pairs.

Following the setup of masked language modeling (MLM) pre-training tasks commonly used in VLP models, our FairVLP is applicable to both single- and dual-stream structures without being restricted to a specific VLP architecture. For the image-text pairs input at each training step $i$, we randomly mask out the text input tokens with probability $q$ and replace them with the special token [MASK]. The image and [MASK]ed text are denoted as $V_f^i$ and $L_f^i$.

**Counterfactual samples generation.** Similar to Section 3, for factual image $V_f^i$ we use model-specific adversarial attack to generate counterfactual image $V_f^{cf}$ by altering the bias concept of $V_f$. And we directly replace the words about bias concept (e.g., man is replaced to woman) in factual text $L_f^i$ to generate counterfactual text $L_f^{cf}$.

Minimize the difference in the prediction of [MASK]ed words between factual and counterfactual samples. Following the formulation of MLM, we can utilize both the image $V$ and the text $L$ to predict [MASK]ed words $P([\text{MASK}] | V, L)$. As shown in Figure 8, we use factual and counterfactual samples to predict $P([\text{MASK}] | V_f^i, L_f^i)$ and $P([\text{MASK}] | V_f^{cf}, L_f^{cf})$.

At current training step, our FairVLP method eliminates social bias in VLP models by minimizing the Kullback-Leibler(KL) divergence in [MASK]ed word prediction distributions between the factual and counterfactual image-text pairs:

$$L_{KL} = D_{KL}(P([\text{MASK}] | V_f^i, L_f^i) || P([\text{MASK}] | V_f^{cf}, L_f^{cf}))$$

(14)

With the basic cross-entropy loss $\mathcal{H}$ for the factual and counterfactual samples in MLM task:

$$L_{mlm} = \mathcal{H}(\mathbf{M}^i, P([\text{MASK}] | V_f^i, L_f^i)) + \mathcal{H}(\mathbf{M}^i, P([\text{MASK}] | V_f^{cf}, L_f^{cf}))$$

(15)

where $\mathbf{M}^i$ is the [MASK]ed word, and the final training objective is to minimize $L_{FairVLP}$:

$$L_{FairVLP} = (1 - \alpha) \cdot L_{mlm} + \alpha \cdot L_{KL}$$

(16)

where $\alpha$ is the coefficient weight to control $L_{KL}$.

By iteratively generating counterfactual samples and minimizing $L_{FairVLP}$. The generated counterfactual samples $V_f^{cf}$ and $L_f^{cf}$ are guaranteed to have counterfactual properties for the VLP model at training step $i$. Minimizing the difference in prediction between factual and counterfactual samples, as a result, the VLP model’s inference on target concept is independent of the bias concept at training step $i$. Throughout pre-training process, FairVLP consistently prevents the VLP from learning social bias.

5.2 Experimental Results

**Experiment Setup.** We follow dual-stream ALBEF architecture and common MLM and ITM (image-text matching) pre-training tasks to pretrain debiasing VLP model on the MS-COCO dataset. To evaluate the debiasing performance of FairVLP, we consider two common and effective debiasing methods in language models as baselines: Gender Swapping(GS) [42] and Dropout Regularization(DR) [35]. GS swaps gendered words in text input for mitigating the correlation between gender and other concepts. DR eliminates model over-fitting to gender concepts by increasing dropout rate. We migrated the idea of GS to a visual modality and edited the concept of face gender in the images using CycleGAN [43]. We
also consider two variants of our FairVLP for ablation study, eliminating bias in visual modality by using only visual counterfactual samples (FairVLP-V) and eliminating bias in language modality by only using language counterfactual samples (FairVLP-L). For all methods, the learning rate is 1e-4 with 0.02 weight decay, and the batch size is 256.

**Debiasing performance.** Table 4 summarizes the debiasing performance of different methods measured on the VL-Bias dataset. Main observations include: (1) All the proposed three settings of FairVLP obtain remarkable debiasing performance than that of baselines. Among three settings, by simultaneously performing counterfactual debiasing on both two modalities, FairVLP demonstrates the best performance in eliminating social bias. (2) SW and CycleGAN show very limited debiasing results due to the lack of minimizing the difference in prediction of [MASK]ed word between the original sample and gender-swapping sample. (3) DR underperforms SW, we owe this result to the random dropout does not change the statistical relationship in the training data.

Moreover, We observe that FairVLP-V, which is designed to eliminate the bias of visual modality, not only achieves a significant debiasing effect in the visual modality, but also demonstrates a remarkable debiasing effect in language modality. And FairVLP-L also performs debiasing for visual modality. We conjecture that this is because gender conflict in image-text pairs is an implicit form of data augmentation, which effectively makes the modeling of other concepts independent of gender concept.

**Downstream task performance.** For the downstream tasks of the pre-training model, we follow the most commonly used Image-Text Retrieval among V+L tasks for model performance evaluation. We evaluate pre-trained models on the Flickr30K and COCO benchmarks, and fine-tune the pre-trained model using the training samples from each dataset.

Table 5 report results on fine-tuned image-text retrieval, respectively. Our FairVLP pre-training task achieves comparable if not better performance with Vanilla. This suggests that FairVLP can leverage reasonably constructed counterfactual samples to effectively eliminate social bias without affecting the model’s learning of reliable knowledge. This phenomenon further indicates that the performance of the downstream task can be compatible with social bias elimination.

| Method     | Flickr30K (1K test set) | MSCOCO (5K test set) |
|------------|-------------------------|----------------------|
|            | TR | IR | TR | IR | TR | IR | TR | IR | TR | IR | TR | IR |
| Vanilla    |    |    |    |    |    |    |    |    |    |    |    |    |
| FairVLP    |    |    |    |    |    |    |    |    |    |    |    |    |

7 CONCLUSION

In this work, we propose to measure social bias in VLP models by comparing how the model of target concept changes factual and counterfactual samples with different bias concepts. A novel VL-Bias dataset is introduced for studying social bias in multimodal models. Taking gender as an example, our bias measure results on VLP models indicate that gender bias is prevalent in the examined VLP models. We propose to counterfactually eliminate social bias via minimizing the modeling differences between factual and counterfactual samples with different bias concepts. In the future, we are working towards counterfactually probing more knowledge as well as regulating the learning of VLP models.

**ACKNOWLEDGMENTS**

This work is supported by the National Key R&D Program of China (Grant No. 2018AAA0100604).
