Common Sense Bias Modeling for Classification Tasks

Miao Zhang[1]*, Zee Fryer[2], Ben Colman[2], Ali Shahriyari[2], Gaurav Bharaj[2]

[1]New York University
[2]Reality Defender Inc.
miaozhng@nyu.edu, {zee, ben, ali, gaurav}@realitydefender.ai

Abstract

Machine learning model bias can arise from dataset composition: correlated sensitive features can distort the downstream classification model's decision boundary and lead to performance differences along these features. Existing debiasing works tackle the most prominent bias features, such as colors of digits or background of animals. However, real-world datasets often include a large number of feature correlations that intrinsically manifest in the data as common sense information. Such spurious visual cues can further reduce model robustness. Thus, domain practitioners desire a comprehensive understanding of correlations and the flexibility to address relevant biases. To this end, we propose a novel framework to extract comprehensive biases in image datasets based on textual descriptions, a common sense-rich modality. Specifically, features are constructed by clustering noun phrase embeddings with similar semantics. The presence of each feature across the dataset is inferred, and their co-occurrence statistics are measured, with spurious correlations optionally examined by a human-in-the-loop module. Downstream experiments show that our method uncovers novel model biases in multiple image benchmark datasets. Furthermore, the discovered bias can be mitigated by simple data re-weighting to de-correlate the features, outperforming state-of-the-art unsupervised bias mitigation methods.

Introduction

Computer vision has been deployed with dramatically more diverse data (both realistic and generative) in recent years, which has drawn attention to data cleaning and bias reduction (Torralba and Efros 2011). One common bias is due to the frequent co-occurrence of target features with context features (illustrated in Figure 1), that downstream task models may rely on to make predictions even though the relationship is not robust and generalizable (Singh et al. 2020; Basu, Babu, and Pruthi 2023; Wang et al. 2024).

Current unsupervised algorithms detect bias informed by model performance. For example, to separate samples aligning with and conflict to bias in model latent space (Sohoni et al. 2020; Krishnakumar et al. 2021; Ahn, Kim, and Yun 2023; Li, Hoogs, and Xu 2022; Seo, Lee, and Han 2022), or by model gradient (Ahn, Kim, and Yun 2023). We observe that most approaches have only tackled limited bias types with prominent and delicate features, like the color of digits, the background of animals (Nam et al. 2020; Sagawa et al. 2019), or the gender of the person doing certain activities (Zhao et al. 2017). It remains to be an open question whether more general image features, including those coarse and subtle ones in a dataset will correlate with the target feature in an *unwanted* way and affect model predictions. For example, small objects like a keyboard nearby a cat, a person nearby a kite which we discover to compose short-cut learning (Figure 1), have been overlooked for bias modeling.

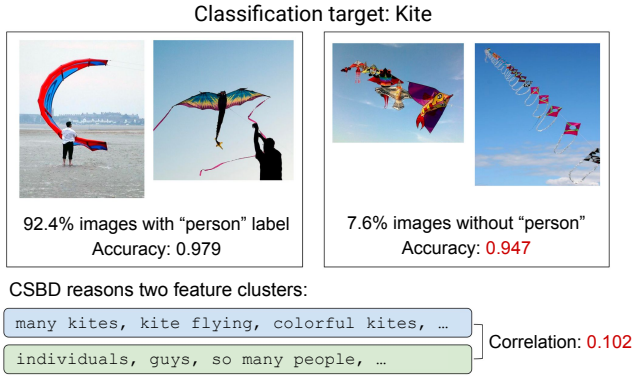

Figure 1: Spurious features are everywhere: Objects which co-occur frequently with the target can affect model prediction, for example, kites and people in MS-COCO. Spurious features like this are of multiple types and may cause different downstream biases. Our method aims to discover a comprehensive list of them based on common sense descriptions, and treat biases which have not been explored in literature.

These observations suggest to enrich bias signals, and we propose to model with additional data modalities, specifically text descriptions of images. Differentiating to labels, descriptions can cover a wider range of features that humans perceive (with respect to whole dataset instead of individual data samples as descriptions can be noisy). We show by experiments that debiasing based on supervision from description-derived bias obtains state-of-the-art results.

"Common sense reasoning" by natural language (Lake

*This work was completed during an internship at Reality Defender Inc.

et al. 2017; Antol et al. 2015; Radford et al. 2021; Diomataris et al. 2021) is a popular approach for high-level abstract understanding of images and reasoning for subtle image features. There have been some early explorations in using this approach for bias discovery. For example, studies encode verbalized spurious features into text embeddings, then aligning to model latent space to measure feature influence (Wu et al. 2023; Zhang et al. 2023). However, the use cases of the existing common sense bias modelings are restricted by the reliance on multi-modal models like CLIP (Radford et al. 2021) or generative models (Rombach et al. 2022) to align embedding space of image and text. This additional step of alignment is challenging: it has a high requirements for representation quality and may not capture features from fine-grained image regions. There are also studies that analyze bias in dataset captions (Van Miltenburg 2016; Wang et al. 2022), but their focus is on linguistic stereotypes within captions themselves, rather than using the captions to debias visual recognition.

To explore this direction, we design a novel description-based reasoning method for feature correlation and downstream model bias, called Common Sense Bias Discovery (CSBD). Given a description corpus of images (for example, captions), we unpack it into "features" through noun chunk clustering. These features are extracted from the semantic components of descriptions, which represent various image content, not limited by prior knowledge or cross-modal alignments. Then, correlations between features are quantified based on their co-occurrence across all data samples. It should note that discovered correlated features are not necessarily bias: for example, a correlation between "`teeth`" and "`smile`" in a dataset of face images is both expected and entirely benign. Therefore, we maintain a human-in-the-loop component to identify spurious correlations that should be addressed for specific tasks. However, following the identification step, our method does not require human intervention to mitigate the downstream model bias caused by these correlations. Our contributions are as follows:

1. A novel framework that uses textual descriptions obtained via humans or vision-language models that: (i) comprehensively analyzes spurious correlations contained in image datasets, and (ii) provides guidance on debiased downstream model training. Specifically:
2. A common sense reasoning approach that generates human interpretable "feature clusters". Based on the discovered clusters, the formulations to derive pairwise correlations and re-sampling weights for model debiasing.
3. Experimental results show that our approach discovers biases with respect to static spurious features on multiple benchmarks, which to the best our knowledge have not been previously addressed. The resulting mitigation achieves state-of-the-art performances.

Related Work

Unsupervised bias discovery. Many recent studies have improved model robustness without requiring sensitive feature annotation. One approach is to assume that easy-to-learn data samples lead to shortcut feature learning, thus resulting in biased classifiers. These bias-aligning samples have higher prediction correctness and confidence (Kim, Lee, and Choo 2021; Liu et al. 2021; Zhang et al. 2022b; Li, Vo, and Nakayama 2023), larger gradient (Ahn, Kim, and Yun 2023), or being fitted early in training (Nam et al. 2020; Lee et al. 2023). Feature clustering has also been used, which leverages the observation that samples with same the feature are located closely in model latent space. Bias is then identified based on the trained model's unequal performance across clusters (Seo, Lee, and Han 2022; Krishnakumar et al. 2021). Another approach finds bias related subgroups based on a certain amount of latent space directions that are highly correlated to model performance (Zhang et al. 2024b). Most methods rely on additional biased learning to infer discounted samples, which might not be robust to shifts in training algorithm or schedules. Also, the implicit bias discovery used in these methods lacks transparency and interpretability. We propose a novel direction to discover dataset bias by description reasoning, which is agnostic and generalizable to different downstream learning schemes.

Bias mitigation. Methods that identify bias-conflicting samples or subgroups use them as supervision to upweight low performers, thus training a debiased model (Nam et al. 2020; Krishnakumar et al. 2021; Liu et al. 2021; Ahn, Kim, and Yun 2023; Seo, Lee, and Han 2022; Lee et al. 2023; Zhang et al. 2024b). Alternatively, methods encourage models to learn similar representations for samples with the same target but different sensitive features via contrastive learning (Zhang et al. 2022b; Park et al. 2022; Zhang et al. 2022a), e.g., augmenting bias-aligned images to alter sensitive features while maintaining the target content (Kim, Lee, and Choo 2021; Ramaswamy, Kim, and Russakovsky 2021; Lim et al. 2023). Bias mitigation method using regularization loss penalizes models for violations of the Equal Opportunity fairness criterion (Li, Hoogs, and Xu 2022). These frameworks address bias identified during model learning process and perform interventions accordingly. We approach the problem from a new prospective, recognizing that bias can take various forms captured by the common sense information of a dataset. Analyzing feature connections from this information can lead to simple and effective bias mitigation.

Bias discovery with human knowledge. Building machine learning systems with human judgement is crucial for user-centric and responsible goals (Lake et al. 2017; Wang et al. 2019). It enables domain expertise to be involved for improving real-world performances with minimal cost (Wu et al. 2022). For bias discovery purposes, researchers have analyzed text-based item sets in visual question answering dataset and use association rules to interpret model behaviors (Manjunatha, Saini, and Davis 2019). Human is included to examine biases, including interpreting the semantic meaning of specific sensitive features (Li and Xu 2021) and analyzing bias information represented by the decomposition of sample subgroups in model latent space (Zhang et al. 2024b). The studies motivate us for interpretable bias distillation. Image descriptions or captions are data formats rich of cognitive and common sense knowledge, thus are of great interests to be utilized for debiasing supervision.

Method

As in Figure 2, we present our framework of discovering and treating bias supervised by image descriptions, **C**ommon-**S**ense **B**ias **D**iscovery method, referred to as CSBD. In brief, given a corpus of (image, text description) pairs, our method generates clusters of noun chunks to indicate presence of features across samples. Then, pairwise feature[1] correlations are analyzed. This is followed by an optional human-in-the-loop module to allow a domain expert to review highly correlated features and select those that may negatively impact the downstream task. Finally, for any selected correlation to treat, we mitigate model bias via data re-weighting, without a requirement for sensitive group labels. Detailed algorithm steps are in the appendix (Zhang et al. 2024a). The three major components of the method are described in the following.

Language-inferred feature distribution. Natural language image descriptions must first be split into information units to align with image objects. We use the spaCy "en_core_web_sm" model (Honnibal and Montani 2017) to extract noun chunks (noun and adjective grouped together), e.g. obtaining "`The girl`" and "`a big smile`" from "`The girl has a big smile`", from all descriptions. We then use a general purpose sentence embedding model, the Universal Sentence Encoder (Cer et al. 2018), to map the noun chunks into high-dimensional vectors that represent their semantics, denoted as $\boldsymbol{A} \in \mathbb{R}^d$. The selection of this text embedding model is based on accuracy and computing speed. Next, we implement agglomerative clustering on $\boldsymbol{A}$. It starts from single vector as a cluster and hierarchically merges pairs of clusters until a certain criteria is reached, given no prior knowledge of the cluster number. We set the criteria to be the maximum Euclidean distance between any vectors of two clusters, denoted as z. It is a tunable hyper-parameter that controls the granularity of the clusters, e.g. "`a boy`" and "`a young man`" may be clustered together with a larger z and be in separate clusters with a smaller z. Each obtained cluster is viewed as a feature that is consistent with common sense descriptions.

Discovering feature correlations. Having extracted a set of feature clusters in the previous step, the next step is understanding feature co-occurence within the dataset. This allows us to identify spurious correlations relevant to the target task, which may cause model bias. To quantify such correlations: first, we generate a binary indicator for whether each feature is present: $\boldsymbol{t}_f = [t_1, t_2, ..., t_N]$, where N is the size of the dataset, and $t_i = 1$ if feature f occurs in the i^{th} image's description, otherwise $t_i = 0$. Second, the phi coefficient ϕ (Cramér 1946, p. 282)[2] is used to measure the association between every two feature indicators. The ϕ between two indicators $\boldsymbol{t}_f$ and $\boldsymbol{t}_{f'}$ is defined as follows:

$$\phi_{\boldsymbol{t}_f \boldsymbol{t}_{f'}} = \frac{x_{11}x_{00} - x_{10}x_{01}}{\sqrt{x_{1*}x_{0*}x_{*0}x_{*1}}}, \tag{1}$$

where $c \in [0, C)$, $f, f' \in [0, F_c]$, and

$$\begin{aligned} x_{11} = \|\boldsymbol{t}_f \cdot \boldsymbol{t}_{f'}\|_1, \quad x_{01} = \|(\neg \boldsymbol{t}_f) \cdot \boldsymbol{t}_{f'}\|_1, \\ x_{10} = \|\boldsymbol{t}_f \cdot (\neg \boldsymbol{t}_{f'})\|_1, \quad x_{00} = \|(\neg \boldsymbol{t}_f) \cdot (\neg \boldsymbol{t}_{f'})\|_1, \\ x_{1*} = \|\boldsymbol{t}_f\|_1, \quad x_{0*} = \|\neg \boldsymbol{t}_f\|_1, \\ x_{*0} = \|\neg \boldsymbol{t}_{f'}\|_1, \quad x_{*1} = \|\boldsymbol{t}_{f'}\|_1. \end{aligned}$$

$\neg$ denotes element-wise negation, $\|\cdot\|_1$ denotes the L^1 norm.

Two features are positively correlated (likely to co-occur in an image) if ϕ is a positive value, and are negatively correlated (rarely co-occur) if ϕ is a negative value. A ϕ near zero indicates two features which co-occur randomly. Intuitively, features that have high correlation with any target features can become shortcuts for model learning and cause biased decision-making toward certain subgroups (Tian et al. 2022; Brown et al. 2023). Thus, feature pairs sorted by ϕ are returned for examination by humans, who have access to the common sense text description of each feature.

Strength of human evaluated bias. In many cases, the extracted feature pairs are naturally connected (e.g. "`teeth visible`" with "`a smile`", and most facial features with "`the face`"). These correlations are usually robust and generalizable, and thus viewed as benign. The benign correlations usually exist, besides, the function of correlations vary across different downstream learning scenarios and not all correlations need to be treated. Therefore, a human-in-the-loop step is necessary to ensure that only spurious correlations that may impact the target task are selected. The human-in-the-loop component of our pipeline, like previous efforts (Manjunatha, Saini, and Davis 2019; Li and Xu 2021; Wu et al. 2023; Zhang et al. 2023), allows flexibility and transparency for bias mitigation. The human involvement is *optional*, that if given any bias features blindly, our method will generate the mitigation approach as described below.

Bias mitigation via re-weighting. We use the feature correlation signals inferred from textual descriptions to mitigate vision model bias. Since adjusting sampling weights to intervene model learning for discounted samples is a frequently used approach (Wang et al. 2020; Ahn, Kim, and Yun 2023; Seo, Lee, and Han 2022), and our goal is *not* to design novel sampling strategies, we use a state-of-the-art sampling method (Qraitem, Saenko, and Plummer 2023) based on our analyzed feature presence ($\boldsymbol{t}$). We compute data sampling weights (or probabilities) with the goal to decorrelate the presence of the spurious feature[3] and the target feature. Specifically, We denote the dataset as D, the target feature indicator as $\boldsymbol{t}_y$, the spurious feature indicator as $\boldsymbol{t}_s$, and the and the conditional probability of $\boldsymbol{t}_s$ given $\boldsymbol{t}_y$ as $P_D(\boldsymbol{t}_s|\boldsymbol{t}_y)$. The new sampling probability P'_D for each element in D ensures that feature s is present in the same probability regardless of whether feature y is present or not:

$$P'_D(\boldsymbol{t}_s = 1|\boldsymbol{t}_y = 0) = P'_D(\boldsymbol{t}_s = 1|\boldsymbol{t}_y = 1) \tag{2}$$

[1] "Feature" in this work refers to a cluster of semantically similar phrases, e.g. "`beaming smile`" and "`wide smile`" are considered interchangeable and belong to one cluster/feature.

[2] Also known as the Matthews correlation coefficient (Matthews 1975).

[3] We use the term "spurious features" to reflect the fact that the primary biases of interest in literature often relate to features such as race/gender/disability/sexuality/etc; however, as we see in Table 2, our method is flexible enough to capture more unexpected types of unwarranted correlation, e.g. "`cat`" and "`couch`".

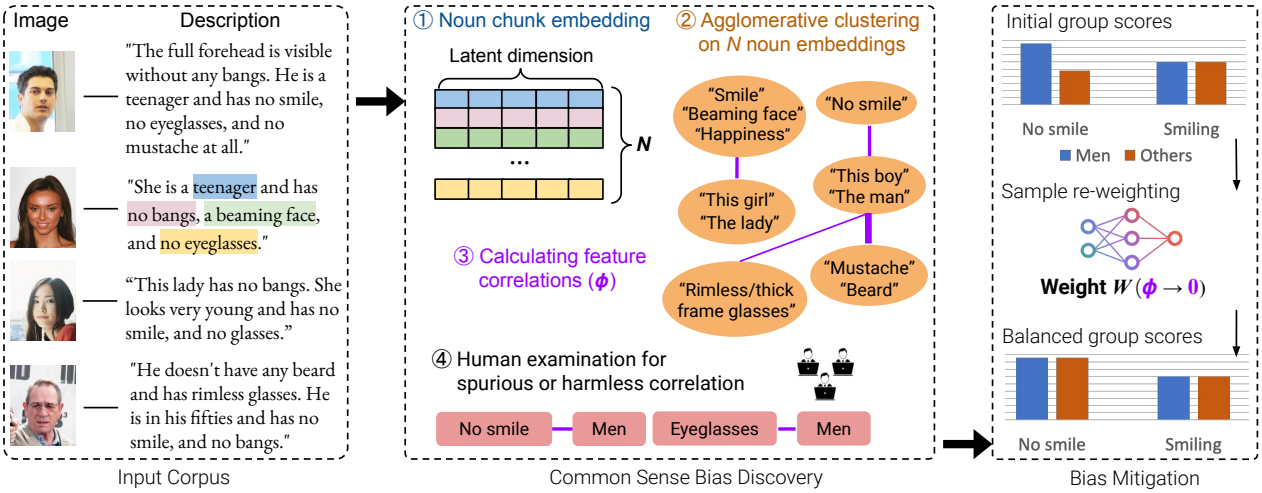

Figure 2: *Common Sense Bias Discovery (CSBD) system and the mitigation strategy overview*: (Left) Input corpus of image and corresponding description pairs are given. (Middle) (1) Descriptions are split into semantically meaningful noun chunks and encoded into representation vectors. (2) A hierarchical clustering on the vector set generates a list of common sense feature clusters. (3) Correlations between the presence of every two features are computed. (4) Highly-correlated features are examined by a human for mitigation. (Right) Finally, correlations are mitigated by adjusting image sampling weights during model training, automatically calculated from the derived feature presence statistics.

This equation ensures independence between s and y across the dataset (Qraitem, Saenko, and Plummer 2023). Additionally, we implement randomized augmentations alongside the sampling weights to enhance data diversity.

Experiments

Datasets

CelebA-Dialog (Jiang et al. 2021) is a visual-language dataset including captions for images in CelebA (Liu et al. 2015), a popular face recognition benchmark dataset containing 40 facial attributes. Each caption describes 5 labels: "Bangs", "Eyeglasses", "Beard", "Smiling", and "Age". The benefit of captions in natural language is that they also include other common sense information like pronouns and people titles (phrases like "the man" or "the lady"). We identify feature correlations for this dataset accordingly (see Figure 3). For classification tasks, we select different target labels to recognize and spurious feature labels to evaluate bias results (here "label" refers to the ground truth binary labels used only for evaluation). The labels are selected because their corresponding caption-inferred features are discovered by our method to be highly correlated and thus may cause downstream model bias.

MS-COCO 2014 (Lin et al. 2014) is a large-scale scene image datasets including annotations for 80 objects, and each image has 5 descriptive captions generated independently (Chen et al. 2015). We randomly select 1 caption for each image. Based on our feature correlation analysis, we select 9 target and spurious feature pairs for training downstream models and evaluating bias, as listed in Table 2. For example, we train a model for recognizing "Dog" and evaluate if the model shows performance disparity for images with/without "Frisbee". We perform binary classification on each target label, following (Wang and Russakovsky 2023). This is because although MS-COCO is more used for multi-label classification, certain labels will need to be treated as auxiliary features in order to reveal their influence on the target thus study bias caused by feature correlations.

LVLM-generated caption. We also apply our proposed method to image captions generated using large vision language model (LVLM). This data source is considered since human annotated caption is not always available and is especially expensive on large scale datasets. We use the open-source LLaVA-Next 7B parameter model (Liu et al. 2024) to caption images of CelebA dataset with prompt "`Please describe face attributes in this image`", and MS-COCO dataset with prompt "`Please describe the scene in this image`". We evaluate captioning quality of the model on MS-COCO against five ground truth captions. The SPICE (Anderson et al. 2016) metric on the validation set has an average score of 0.1583, showing a satisfying ability of the LVLM model to capture image features and their relations. Besides, we evaluate the common sense biases modeled from LVLM-generated caption when applying our proposed method, which is in general consistent with the results obtained with ground truth captions (Figure 3).

Implementation

For correlation reasoning, the noun chuncks from the description corpus are encoded as vectors in 512 dimensions using the Universal Sentence Encoder. Dimension reduction is performed using PCA before clustering, ensuring that the sum of the PCA components accounts for a high amount of variance (>90%). The distance criterion for generating

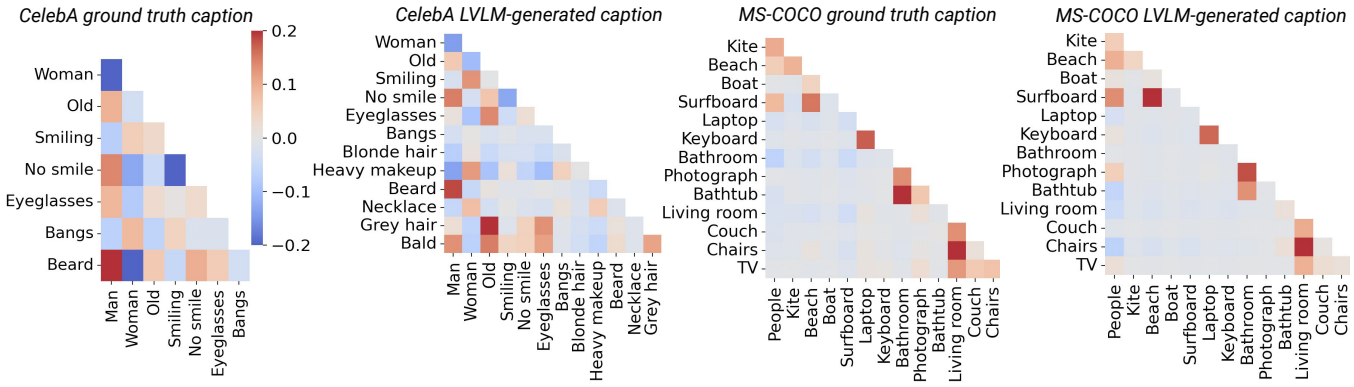

Figure 3: Part of correlated features and their correlation coefficients analyzed by CSBD on CelebA and MS-COCO datasets. Two types of common sense descriptions are used: human-generated (ground truth) and LVLM-generated caption.

feature clusters is set to $z = 1.0$; see ablation sections for sensitivity analysis. We use Chi-square test (Pearson 1900) to verify significance of derived correlation coefficients.

For the downstream training, we use the same training, validation, and testing split as in (Zhang et al. 2022b) for CelebA dataset, and as in (Misra et al. 2016) for MS-COCO dataset. However, because of label scarcity of individual objects relative to the entire dataset (for example, only 3% of images in MS-COCO contain "Cat" and 2% of images contain "Kite"), to avoid class imbalance problem introducing confounding bias to the classifier, we randomly take subsets of the dataset splits which include 50% of each target label. Experiments are based on three independent runs which provide the mean and standard deviation results.

Results

Bias discovery and mitigation are both needed for reliable model training. While the emphasis of the proposed method is to analyze bias from datasets, not primarily on mitigation algorithm, it is important to verify on downstream tasks that the method "discovers" correlations in the dataset that the model actually takes a shortcut for learning. So in this section, we first present bias discovery results on the two benchmark datasets. Then, we analyze robustness of the correlation derivation, with respect to potential errors and incompleteness of descriptions. Finally, we perform thorough downstream evaluations to show that reducing feature dependence analyzed by the proposed method serves as a practical and effective debiasing objective.

Correlation discovery results Figure 3 presents selected correlated feature pairs for CelebA and MS-COCO dataset. It shows that both human-annotated and LVLM-generated captions contain rich common sense information that can be distilled into distinctive features. The correlation between the obtained features are consistent across different image descriptions, highlighting the description-agnostic and generalization strength of the proposed method. We note that Figure 3 shows limited CelebA feature correlations with the ground truth caption because the caption is designed to annotate only five features.

The description-derived features include those not represented in the dataset labels and reveal highly correlated ones, such as "`city street`" and "`bus`". The spuriously correlated features are frequently seen but not generalizable to all samples, leading models to rely on context rather than appearance (e.g., recognizing a bus in an image based on a city street background). On CelebA dataset, facial features "`smiling`", "`eyeglasses`", "`bangs`", etc, are spuriously correlated to gender features "`man`", "`woman`", and age features "`young`", etc. Other strong correlations include "`Receding hairline`" and "`grey hair`" which aligns with common sense but lacks robustness. We designate naturally related features like "`mouth`" and "`smiling`" as benign. On MS-COCO dataset, benign correlations include "`frisbee`" and "`a game`", "`waves`" and "`ocean`", etc. The spurious ones include "`cat`" and "`couch`", "`kite`" and "`people`", "`bathroom`" and "`photograph`", etc. The full list of correlated feature pairs and their coefficient values are in the appendix (Zhang et al. 2024a).

Correlation robustness evaluation First, we test the robustness of analyzed feature presence, considering that unlike dataset label, description annotators may only select features relevant or noticeable to them, causing incompleteness or faults. Indeed, prior studies (Yu et al. 2022; Wang et al. 2022) have highlighted the incorrect and low-quality captions in MS-COCO. We randomly select 20 common sense features derived from MS-COCO captions and compare them with their corresponding labels. We compute the Pearson correlation coefficient between the presence of each feature and label across the dataset, obtaining a coefficient of 0.545 (95% confidence interval 0.153 to 0.937). This indicates a moderate correlation between label and feature distribution, which varies across different features. The inconsistency between common sense feature and label distribution is expected, influenced by both caption quality (captions may incorrectly describe ground truth label) and text embedding clustering (text with similar semantics may not be encoded closely in the representation space).

Next, we test the robustness of discovered feature correlations (or bias, to avoid confusion with the correlation test

Target	Spurious	No mitigation		LfF		DebiAN		PGD		BPA		CSBD (ours)	
		Wst. (↑)	Avg. (↑)	Wst.	Avg.	Wst.	Avg.	Wst.	Avg.	Wst.	Avg.	Wst.	Avg.
Smiling	Male	0.894	**0.919**	0.730	0.841	0.895	0.916	0.888	0.917	0.899	0.912	**0.908**	0.918
Eyeglasses	Male	0.962	0.986	0.912	0.963	0.967	0.979	0.966	0.984	0.971	0.983	**0.973**	**0.987**
Young *	Eyeglasses *	0.652	0.818	0.640	0.805	0.609	0.796	0.646	0.817	0.638	0.806	**0.655**	**0.820**
Receding hairline *	Grey hair *	0.186	0.650	0.174	0.648	0.198	**0.680**	0.225	0.672	**0.351**	0.521	0.337	**0.680**

Table 1: Treating common sense bias towards the target label (Target) among the spurious feature label (Spurious) on CelebA. The starred labels (*) represent that CSBD uses LVLM-generated descriptions to discover and mitigate bias. The metric for debiasing effectiveness: the worst performing group (Wst.), and average accuracy of all target-spurious feature groups (Avg.), are reported. The best results are in bold and the second best results are underlined.

Target	Spurious	No mitigation		LfF		DebiAN		PGD		BPA		CSBD (ours)	
		Wst.(↑)	Avg. (↑)	Wst.	Avg.	Wst.	Avg.	Wst.	Avg.	Wst.	Avg.	Wst.	Avg.
Cat	Couch	0.849	0.935	0.800	0.916	0.821	0.902	0.884	0.940	0.853	0.925	**0.906**	**0.941**
Cat	Keyboard	0.948	0.972	0.870	0.938	0.917	0.949	0.947	**0.977**	0.890	0.952	**0.953**	0.976
Dog	Frisbee	0.883	0.927	0.821	0.893	0.845	0.895	0.872	**0.940**	0.854	0.905	**0.903**	0.934
TV	Chair	0.726	0.893	0.690	0.867	0.718	0.876	0.821	0.907	0.758	0.887	**0.842**	**0.911**
Umbrella	Person	0.761	0.880	0.771	0.854	0.716	0.857	0.763	0.887	0.777	0.878	**0.814**	**0.892**
Kite	Person	0.937	**0.963**	0.854	0.923	0.941	0.960	0.917	0.955	0.921	0.957	**0.945**	0.960
Wine glasses *	Person *	0.848	0.888	0.834	0.869	0.838	0.873	0.828	0.869	0.844	0.890	**0.860**	**0.894**
Pizza *	Oven *	0.828	0.916	0.820	0.861	0.827	0.910	0.828	0.903	0.854	0.925	**0.862**	**0.932**

Table 2: Treating part of the discovered biases for binary object classification task on MS-COCO dataset.

here, we use the term "bias") under this circumstance. Bias existence between each feature and each of the remaining features is computed. The same bias measurement is then performed using labels. The correlation coefficient between the two measurements is 0.781 (95% confidence interval 0.417 to 1.0), which is much higher and shows lower variance than the result for feature presence. The analysis indicates that the proposed method operates not at the level of individual images but rather at the dataset level to identify trends in correlations, thus showing certain robustness to noise and incompleteness present in image descriptions.

Bias mitigation results Spurious feature correlations discovered by CSBD can indicate and help mitigate downstream model bias, as shown by image classification results on CelebA and MS-COCO dataset in Table 1 and Table 2. We use the worst group performance (Wst.) to evaluate model bias, and average group performance (Avg.) to evaluate model quality, following (Zhang et al. 2022b; Ahn, Kim, and Yun 2023; Seo, Lee, and Han 2022). The groups here are all combinations of target label and spurious feature label pairs as in (Zhang et al. 2022b). The comparison baselines are unsupervised bias mitigation methods. They include LfF (Nam et al. 2020) which upweights failure samples, DebiAN (Li, Hoogs, and Xu 2022) which uses additional bias predictor, and re-sampling methods PGD (Ahn, Kim, and Yun 2023) based on sample gradients, and BPA based on latent space clusters (Seo, Lee, and Han 2022).

Baseline model without bias mitigation presents a gap between Wst. and Avg. for all tasks, indicating model biases that align with our reasoning on the datasets. For example, images containing chair but no TV obtain an accuracy of 0.726 for TV classification, 19% lower than the average accuracy. Face images to recognize receding hairline with dark hair show an accuracy of 0.186, while the recognition with grey hair has an accuracy over 0.99 because of strong correlation between these two features. CSBD successfully mitigates model bias (most improved Wst.) on all targets, outperforming the baseline with no mitigation and other debiasing methods. Importantly, model classification quality is not degraded by using our method: CSBD always obtains the highest or the second Avg. Bias mitigation based on LVLM generated descriptions are effective on both datasets, which shows our method's generalizability to different description corpus and wide applicability to bias mitigation problems.

Discussion

We investigate the advantage of our method for bias mitigation by checking how well they separate target objects not affected by spurious features in latent space. We visualize image embeddings extracted from the last fully connected layer of the "Cat" classifier trained with MS-COCO. Model bias can be observed from the baseline (no mitigation) (Figure 4 (A)): while the "Cat" targets are separable for samples without the spurious feature "Couch", for images *with* the spurious feature, the model prediction boundary for the targets is blurred (blue and red dots are partially mixed).

The same embedding analysis is performed for bias mitigation methods[4]. As shown in Figure 4 (B-D), DebiAN, PGD, and BPA do not fix model's biased decision rule, that the model cannot distinguish two targets ("Cat" and 'No

[4]Here only showing methods outperforming no mitigation.

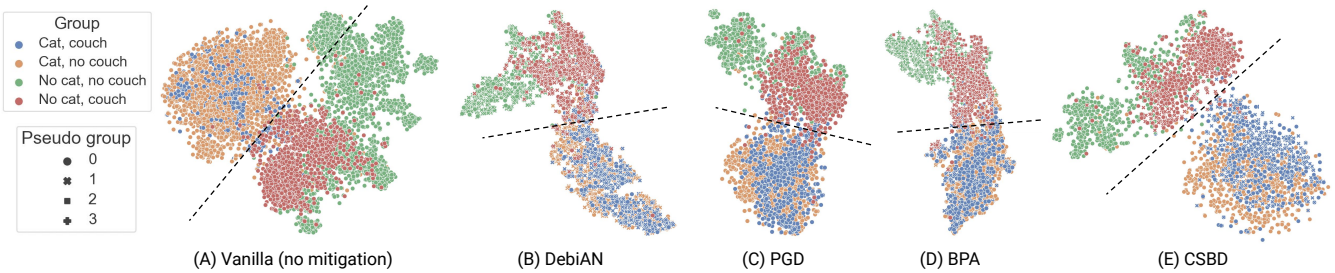

Figure 4: T-SNE results for sample embeddings of cat classifiers ("Couch" as sensitive label) trained with MS-COCO. Target label: "Cat" (blue & orange) and "No cat" (green & red) are most linearly separable (model decision boundary is indicated by the dotted straight line) with CSBD compared to other baselines. For bias discovery, methods are marked with their assigned bias-aligning and bias-conflict samples (Pseudo group). Only CSBD discovers thus treats the sensitive feature couch (E).

cat" samples), indicated by the dotted straight line. Instead, the methods only mix samples within a target class more closely. The reason is that the methods identify discrepancy between sample representations or gradients as bias and try to mitigate the discrepancy. However, the detected discrepancy fails to accurately capture the unwanted feature responsible for mis-classifying bias-conflicting samples (those near the decision boundary), so the black-box way of mitigation proves less effective for de-biasing.

z	0.5	0.8	0.9	1	1.2	1.5
ϕ	0.0608	0.0608	0.0908	0.0908	0.107	0.175
Wst. (↑)	0.971	0.972	0.973	0.973	0.973	0.972
Avg. (↑)	0.986	0.986	0.986	0.987	0.987	0.986

Table 3: Sensitivity analysis on the distance threshold z used for splitting feature clusters. z within a reasonable range can produce consistent correlated features ($\phi > 0$) and similar bias mitigation results (Wst.), between the example "`eyeglasses`" and "`man`" features on CelebA.

Ablation on the clustering threshold Feature clusters are obtained by agglomerative clustering on noun chunk embeddings so that the complete linkage between two clusters is lower than a hyper-parameter z. This tune-able parameter affects the granularity of discovered feature clusters and whether reasonable bias can be derived. As shown in Table 3, the bias mitigation (Wst.) and classification (Avg.) results are robust to different z, which indicates that feature clusters and their distributions are captured precisely in all cases. The granularity and content of feature clusters change with different z, which result in different correlation coefficient ϕ. Here, "`eyeglasses`" and "`sunglasses`" are one feature with $z = \{1.2, 1.5\}$ and shows a stronger correlation to "`man`" feature, but they are separate features with smaller z and "`eyeglasses`" alone shows weaker correlation to "`man`". The selection of z should be tuned based on the granularity of features that practitioners hope to derive and find biases with. For example, if z is too large (we test with $z = 5$), gender clusters will be replaced by a "`people`" cluster, thus the bias: man wearing glasses more will not be discovered on the CelebA dataset.

Ablation on re-sampling weights. We ablate the bias mitigation step by skewing the sampling weights W analyzed by CSBD. Specifically, for any images that need to be over-sampled ($W[i] = 1 + p$ for the i^{th} sample), we first double their increased weight by $W[i] = 1 + 2 * p$. The same downstream task to classify "Cat" with spurious attribute "Couch" on MS-COCO is performed. The model obtains Wst.: 0.904, Avg.: 0.940. Then, we only use half of the increased weight by $W[i] = 1 + 0.5 * p$, and the model obtains Wst.: 0.881, Avg.: 0.936. Compared to the results in Table 2, though training with skewed weights still outperforms the vanilla setting with no intervention, the bias mitigation and accuracy results are slightly discounted. It shows that the data re-sampling weights computed from CSBD is the key for effective bias mitigation.

Conclusion and Limitation

This work proposes to use textural descriptions of a dataset for systematic bias discovery and mitigation. The approach leads to a new perspective on bias problems regarding how we can use knowledge in natural language to precisely reduce a concerned feature's influence to image classification tasks. Our algorithm applies common sense reasoning to descriptions, which cover a wide range of image features not annotated by the label set, and indicates how feature correlations to the target will cause downstream bias. Experiment results show novel spurious correlations and model biases discovered for two vision benchmark datasets, and the state-of-the-art bias mitigation based on the discovery.

Limitations and future work. Though our method can model correlations between features beyond the labeled cohort, the discovered biases are limited to features included in textual descriptions. Besides, there are multiple topics to continue in this direction: First, we only study pairwise feature correlations, but higher degrees of correlations and how they affect model performance could be explored. Second, our method has not yet been tested in other vision recognition tasks and models, such as object detection and segmentation. However, since the method for discovering spurious feature correlations is agnostic to downstream tasks, this can be a direct next step.

References

Ahn, S.; Kim, S.; and Yun, S.-y. 2023. Mitigating dataset bias by using per-sample gradient. In *Eleventh International Conference on Learning Representations*. ICLR.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Basu, A.; Babu, R. V.; and Pruthi, D. 2023. Inspecting the Geographical Representativeness of Images from Text-to-Image Models. *arXiv preprint arXiv:2305.11080*.

Brown, A.; Tomasev, N.; Freyberg, J.; Liu, Y.; Karthikesalingam, A.; and Schrouff, J. 2023. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications*, 14(1): 4314.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Cramér, H. 1946. Mathematical methods of statistics.

Diomataris, M.; Gkanatsios, N.; Pitsikalis, V.; and Maragos, P. 2021. Grounding consistency: Distilling spatial common sense for precise visual relationship detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15911–15920.

Honnibal, M.; and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1): 411–420.

Jiang, Y.; Huang, Z.; Pan, X.; Loy, C. C.; and Liu, Z. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13799–13808.

Kim, E.; Lee, J.; and Choo, J. 2021. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14992–15001.

Krishnakumar, A.; Prabhu, V.; Sudhakar, S.; and Hoffman, J. 2021. Udis: Unsupervised discovery of bias in deep visual recognition models. In *British Machine Vision Conference (BMVC)*, volume 1, 3.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40: e253.

Lee, J.; Park, J.; Kim, D.; Lee, J.; Choi, E.; and Choo, J. 2023. Revisiting the importance of amplifying bias for debiasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14974–14981.

Li, J.; Vo, D. M.; and Nakayama, H. 2023. Partition-and-debias: Agnostic biases mitigation via a mixture of biases-specific experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4924–4934.

Li, Z.; Hoogs, A.; and Xu, C. 2022. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, 270–288. Springer.

Li, Z.; and Xu, C. 2021. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14970–14979.

Lim, J.; Kim, Y.; Kim, B.; Ahn, C.; Shin, J.; Yang, E.; and Han, S. 2023. Biasadv: Bias-adversarial augmentation for model debiasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3832–3841.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, E. Z.; Haghgoo, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 6781–6792. PMLR.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Manjunatha, V.; Saini, N.; and Davis, L. S. 2019. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9562–9571.

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta*, 405 2: 442–51.

Misra, I.; Lawrence Zitnick, C.; Mitchell, M.; and Girshick, R. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2930–2939.

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684.

Park, S.; Lee, J.; Lee, P.; Hwang, S.; Kim, D.; and Byun, H. 2022. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10389–10398.

Pearson, K. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed

to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302): 157–175.

Qraitem, M.; Saenko, K.; and Plummer, B. A. 2023. Bias Mimicking: A Simple Sampling Approach for Bias Mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20311–20320.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramaswamy, V. V.; Kim, S. S.; and Russakovsky, O. 2021. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9301–9310.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Seo, S.; Lee, J.-Y.; and Han, B. 2022. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16742–16751.

Singh, K. K.; Mahajan, D.; Grauman, K.; Lee, Y. J.; Feiszli, M.; and Ghadiyaram, D. 2020. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11070–11078.

Sohoni, N.; Dunnmon, J.; Angus, G.; Gu, A.; and Ré, C. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33: 19339–19352.

Tian, H.; Zhu, T.; Liu, W.; and Zhou, W. 2022. Image fairness in deep learning: problems, models, and challenges. *Neural Computing and Applications*, 34(15): 12875–12893.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528. IEEE.

Van Miltenburg, E. 2016. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*.

Wang, A.; Barocas, S.; Laird, K.; and Wallach, H. 2022. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 324–335.

Wang, A.; and Russakovsky, O. 2023. Overwriting Pretrained Bias with Finetuning Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3957–3968.

Wang, D.; Yang, Q.; Abdul, A.; and Lim, B. Y. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–15.

Wang, Y.; Sun, J.; Wang, C.; Zhang, M.; and Yang, M. 2024. Navigate Beyond Shortcuts: Debiased Learning through the Lens of Neural Collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12322–12331.

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.

Wu, S.; Yuksekgonul, M.; Zhang, L.; and Zou, J. 2023. Discover and Cure: Concept-aware Mitigation of Spurious Correlation. *arXiv preprint arXiv:2305.00650*.

Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; and He, L. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135: 364–381.

Yu, B.; Zhong, Z.; Qin, X.; Yao, J.; Wang, Y.; and He, P. 2022. Automated testing of image captioning systems. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, 467–479.

Zhang, F.; Kuang, K.; Chen, L.; Liu, Y.; Wu, C.; and Xiao, J. 2022a. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*.

Zhang, M.; Colman, B.; Shahriyari, A.; Bharaj, G.; et al. 2024a. Common-Sense Bias Discovery and Mitigation for Classification Tasks. *arXiv preprint arXiv:2401.13213*.

Zhang, M.; Sohoni, N. S.; Zhang, H. R.; Finn, C.; and Ré, C. 2022b. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*.

Zhang, Y.; HaoChen, J. Z.; Huang, S.-C.; Wang, K.-C.; Zou, J.; and Yeung, S. 2023. Diagnosing and rectifying vision models using language. *arXiv preprint arXiv:2302.04269*.

Zhang, Z.; Feng, M.; Li, Z.; and Xu, C. 2024b. Discover and Mitigate Multiple Biased Subgroups in Image Classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10906–10915.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.