Feature and Label Embedding Spaces Matter in Addressing Image Classifier Bias

Thong, W.; Snoek, C.G.M.

Publication date
2021

Document Version
Final published version

Published in
32nd British Machine Vision Conference 2021

Feature and Label Embedding Spaces Matter in Addressing Image Classifier Bias

William Thong *
w.e.thong@uva.nl
Cees G. M. Snoek
cgmsnoek@uva.nl

University of Amsterdam
Science Park 904
Amsterdam
The Netherlands

Abstract

This paper strives to address image classifier bias, with a focus on both feature and label embedding spaces. Previous works have shown that spurious correlations from protected attributes, such as age, gender, or skin tone, can cause adverse decisions. To balance potential harms, there is a growing need to identify and mitigate image classifier bias. First, we identify in the feature space a bias direction. We compute class prototypes of each protected attribute value for every class, and reveal an existing subspace that captures the maximum variance of the bias. Second, we mitigate biases by mapping image inputs to label embedding spaces. Each value of the protected attribute has its projection head where classes are embedded through a latent vector representation rather than a common one-hot encoding. Once trained, we further reduce in the feature space the bias effect by removing its direction. Evaluation on biased image datasets, for multi-class, multi-label and binary classifications, shows the effectiveness of tackling both feature and label embedding spaces in improving the fairness of the classifier predictions, while preserving classification performance.

1 Introduction

This paper strives to identify and mitigate biases present in image classifiers, with a focus on their feature and label embedding space. Adverse decisions from image classifiers can create discrimination against members of certain class of protected attribute, such as age, gender, or skin tone. Buolamwini and Gebru [7] importantly show that face recognition systems misclassify subgroups with darker skin tones. This also applies to object recognition, where performance is higher for high-income communities [10] mainly located in Western countries [41]. Similarly problematic, current classifiers perpetuate and amplify current discrimination present in society [8, 14]. For example, Kay *et al*. [27] highlight the exaggeration of gender bias in occupations by image search systems. These adverse decisions notably arise because image classifiers are prone to biases present in the dataset [16]. It is therefore essential to identify harmful biases in image representations and assess their effects on the classification predictions, as we do in this paper.

Addressing dataset biases is not enough, and classifier biases should also be addressed. Zhao *et al*. [54] importantly show that biases can actually be amplified during the image

* Currently affiliated with Sony AI, Switzerland.

classifier training. Even when balancing a dataset for the protected attribute gender, image classifiers can still surprisingly amplify biases when making a prediction [49]. This outcome emphasizes the importance of considering protected attributes during the training to avoid biased and adverse decisions. A first approach is to perform *fairness through blindness*, where the objective is to make the feature space blind to the protected attribute [1, 23, 53]. An alternative is to perform *fairness through awareness*, where the classifier label space is explicitly aware of the protected attribute label [12]. To better understand the effectiveness of these methods, Wang *et al*. [50] propose crucial benchmarks in biased image classification. They notably expose the shortcomings of these methods and show that a simple method with separate classifiers is more effective at mitigating biases. Building on this line of work, this paper first identifies a bias direction in the feature space, and secondly address bias mitigation in both label and feature spaces. Another important aspect concerns how to measure the fairness of image classifiers. We borrow from the general fairness literature [4, 12, 21] to ensure that predictions are similar for all members of a protected attribute, which complements the benchmarks introduced by Wang *et al*. [50] on image classification bias.

Contributions. Our main contribution is to demonstrate the importance of feature and label spaces for addressing image classifier bias. First, we identify a bias direction in the feature space of common classifiers. We aggregate class prototypes to represent every class of each protected attribute value, and show a main direction to explain the maximum variance of the bias. Second, we mitigate biases at both classification and feature levels. We introduce protected classification heads, where each head projects the features to a label embedding space specific to each protected attribute value. This differs from common classification, which usually considers a one-hot encoding for the label space [33, 40, 50]. For training, we derive a cosine softmax cross-entropy loss for multi-class, multi-label and binary classifications. Once trained, we apply in the feature space a bias removal operation to further reduce the bias effect. Experiments show the benefits on addressing classifier bias in both feature and label embedding spaces to improve fairness scores, while preserving the classification performance. The source code is available at: https://github.com/twuilliam/bias-classifiers.

2 Related Work

Biases in word embeddings. Assessing the presence of biases in word embeddings, especially the gender bias, has received large attention given their wide range of applications within and beyond natural language processing. The seminal and important work of Bolukbasi *et al*. [6] reveals that the difference between female and male entities in word2vec [35] contains a gender bias direction. This shows that word2vec implicitly captures gender biases, which in return creates sexism in professional activities. Caliskan *et al*. [8] further reveal that multiple human-like biases are actually present in word embeddings. Even contextualized word embeddings [37] are affected by a gender bias direction [56], which creates harmful risks [3]. To mitigate such gender bias, Bolukbasi *et al*. [6] propose a post-processing removal operation while Zhao *et al*. [55] derive regularizers to control the distance between relevant words during training. It is important to note that biases cannot be removed entirely as they can still be recovered to some extent [17]. As such, methods mainly mitigate biases in models rather than producing debiased models. Inspired by the literature on gender bias identification and mitigation in word embeddings, we pursue an analogous reasoning to show that biases are implicitly encoded in image classification models as well.

Biases in image datasets. As computer vision research relies heavily on datasets, they constitute a main source of biases. Torralba and Efros [45] identify that datasets have a strong built-in bias as they only represent a narrow view of the visual world, leading models to rely on spurious correlations and produce detrimental predictions. For fairness and transparency purposes, it becomes necessary to document the dataset creation [15, 24], as well as detecting the presence of potential biases and harms due to an unfair and unequal label sampling [5, 11, 41, 51]. Towards this end, Bellamy *et al*. [2] and Wang *et al*. [47] propose metrics to measure biases, and actionable insights to mitigate them in a dataset. Even though addressing biases when collecting a dataset is highly recommended, models can still produce unfair decisions [49]. In this paper, we focus on addressing image classifier bias.

Biases in image classifiers. Searching for a representative subset of image examples provides visual explanations of biases [28, 43]. In this paper, we rather identify that such bias exists in the feature space in image classifiers. To mitigate image classification bias, training with adversarial learning [19] makes the classifier blind to the protected attribute. Reducing the gender bias can be achieved by forcing a model to avoid looking at people to produce a prediction [23, 49]. Blindness can also be achieved in the feature space by removing the variation of the protected attribute [1, 53]. Though, Wang *et al*. [50] illustrate that adversarial approaches tend to be detrimental as they decrease the performance by making image classifiers less discriminative. At the same time, non-adversarial approaches tend to amplify biases less, while performing well on image classification. Wang *et al*. [50] notably show that encoding the protected attribute into separate heads better mitigates biases. We build on this literature and propose to mitigate biases at both classification and feature levels.

Biases benchmarking. No consensus exists (yet) in mitigating image classifier bias, which makes apple-to-apple comparisons complicated: (a) benchmarks become no longer valid because datasets are taken down for ethical reasons [36] (e.g., Racial faces in-the-wild [48] derives from the problematic MS-Celeb-1M [20], and Diversity in Faces [34] has received complaints); (b) datasets are introduced without benchmarks of debiasing methods (e.g., FairFace [26] mainly evaluates commercial facial classification systems); (c) related works come with differing evaluation settings (e.g., Wang *et al*. [49] train MLP probes to measure model leakage). While addressing algorithm bias in face verification [18, 42, 52] is crucial, we focus on image classification [25, 29, 49, 50]. Therefore, we adopt in this paper the benchmarks introduced by Wang *et al*. [50] and Kim *et al*. [29] in multi-class, multi-label and binary classifications for their comprehensiveness and reproducibility.

3 Identifying a Bias Direction

Problem formulation. We consider the task of image classification where every image $\boldsymbol{x}$ is assigned a label $y \in \mathcal{Y}$. For every image, there also exists a protected attribute value $v \in \mathcal{V}$, on which the classifier should not base its decision. In other words, classifiers should not discriminate against specific members of a protected attribute. In this paper, we consider discrete variables for protected attribute values, and limit the problem to binary values with $\mathcal{V}=\{0,1\}$. For example, we only consider the values "female" and "male" to describe the protected attribute *gender*. It is important to note that this formulation is a simplification of the real world where protected attributes go beyond binary values, and are non-discrete.

Image classifiers are typically composed of a base encoder and a projection head. First, a base encoder $f(\cdot)$ extracts the feature representations of images $\boldsymbol{x}$. In our case, this corresponds to a convolutional network and results in $\boldsymbol{h}=f(\boldsymbol{x})$. Second, a projection head $g(\cdot)$

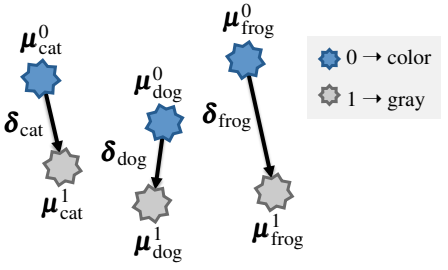
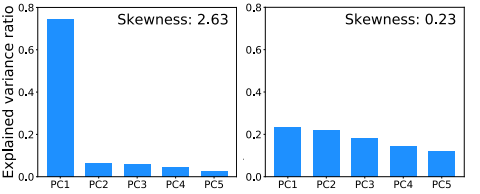


Figure 1: **2D toy visualization** of the feature space, where class prototypes $\boldsymbol{\mu}$ represent three categories with a color bias (✸ *vs.* ✸). A bias vector $\boldsymbol{\delta}$ is computed for every class.



(a) $\boldsymbol{\Delta}$. (b) Random $\boldsymbol{\Delta}$.

Figure 2: **Bias direction** in the feature space. (a) The PCA of $\boldsymbol{\Delta}$ shows the maximum variance as the bias direction. (b) On a random $\boldsymbol{\Delta}$, the direction disappears and the explained variance is no longer skewed.

maps the features $\boldsymbol{h}$ to a discriminative space where a class is assigned. In our case, this corresponds to a linear projection, or a multilayer perceptron, and results in $\boldsymbol{z}=g(\boldsymbol{h})$ with $\boldsymbol{z} \in \mathbb{R}^M$. For example, in a one-hot encoding, M equals the number of classes.

During training, we are given access to the protected attribute labels and can incorporate it in model formulations. We denote the triplet $(\boldsymbol{x}_i, y_i, v_i)$ as the i-th sample in the training set. During the evaluation, models only have access to the images. In this section, we show that common image classifiers – that do not leverage protected attribute labels during training – still implicitly encode their information in the feature space.

Protected class prototypes. Once a model has been trained, we extract the features $\boldsymbol{h}$ from the training set. We then aggregate prototypes $\boldsymbol{\mu}^v_y$ for every class y and specific to each protected attribute value v, coined as protected class prototypes. For example in Figure 1, the class y=cat has two prototypes in the feature space, one for v=*color* images and one for v=*gray*. For any class y with any protected attribute value v, we compute the protected class prototypes as their average representation in the feature space from the training set:

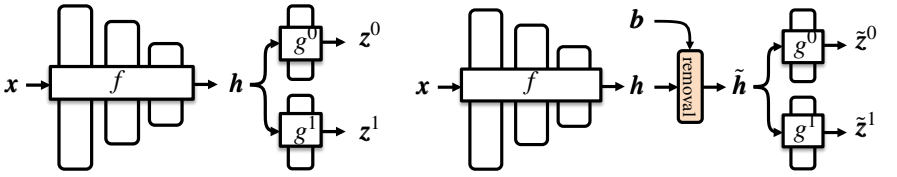
$$\boldsymbol{\mu}^v_y = \frac{1}{N^v_y} \sum_i \mathbb{I}[y_i = y \cap v_i = v] f(\boldsymbol{x}_i), \tag{1}$$

where N^v_y is the number of training images of class y with protected attribute v, and $\mathbb{I}[\cdot]$ is the indicator function. Once all protected class prototypes are computed, we extract a subspace that captures the variance of the bias related to the protected attribute.

Bias direction. To identify a bias direction, we experiment with a standard convolutional network trained with a softmax cross-entropy loss on CIFAR-10S [50]. This dataset provides a simple testbed to measure biases in images, as certain classes are skewed towards *gray* images, while others are skewed towards *color* images. Once trained, we aggregate the difference between class prototypes of each protected attribute value for every class:

$$\boldsymbol{\Delta} = \{\boldsymbol{\delta}_y | y \in \mathcal{Y}\} = \{\boldsymbol{\mu}^1_y - \boldsymbol{\mu}^0_y | y \in \mathcal{Y}\}. \tag{2}$$

Note that for multi-label classification, we consider all binary labels to define $\mathcal{Y}$. Figure 2(a) shows the principal component analysis (PCA) of $\boldsymbol{\Delta}$. When computing the ratio of explained variance of every principal component (PC), a main direction of variance appears. The first PC is more important than the others, which yields a high skewness. Figure 2(b) depicts the same analysis on a random $\boldsymbol{\Delta}$, where no main direction appears. Hence, there exists a subspace in the feature space where the bias information is maximized.



(a) Protected embeddings for classification. (b) Bias removal in the feature space.

Figure 3: **Mitigating biases in classification predictions.** (a) For classification, we mitigate biases with protected label embeddings where each protected attribute value has its own space. (b) In the feature space, we include a removal operation of the bias direction $\boldsymbol{b}$ once the model has been trained, where $\boldsymbol{b}$ is computed from the training set.

4 Mitigating Biases

Figure 3 illustrates our approach to mitigate biases in class predictions at both classification and feature levels. For the classification level, we create two protected label embedding spaces, one for each value of the binary protected attribute. For the feature level, we propose a bias removal operation once the model has been trained. The proposed method works for multi-class, multi-label and binary settings.

Protected label embeddings. We project features $\boldsymbol{h}$ into embedding spaces, one for each protected attribute value. This results in the embedding representation $\boldsymbol{z}^v = g^v(\boldsymbol{h}) \in \mathbb{R}^M$, where classification occurs. During training, each projection head $g^v(\cdot)$ only sees samples from its assigned attribute value, which creates a protected embedding. By only seeing samples of one protected value, class boundaries are better separated [40].

We further push these properties by relying on a cosine softmax cross-entropy loss for classification. $\boldsymbol{z}$ constitutes a discriminative embedding representation with semantic information about classes. This differs from related approaches in domain adaptation [33, 40] or bias mitigation [50], which also show the benefits of separate projection heads with a standard softmax but with a one-hot encoding label space. Below we derive a cosine softmax with protected embeddings for both multi-class, multi-label and binary classifications.

Multi-class classification assigns a label $y \in \mathcal{Y}$ to an image $\boldsymbol{x}$. We introduce a protected weight matrix $\boldsymbol{W}^v \in \mathbb{R}^{|\mathcal{Y}| \times M}$, where M is the size of the embedding space and $v \in \mathcal{V}$ is the protected attribute value. Every row $\boldsymbol{W}^v_{y,:}$ acts as a latent real-valued semantic representation for every class y of each protected attribute v. The objective is then to maximize the cosine similarity, denoted as "sim", between an embedding representation $\boldsymbol{z}^v$ and its corresponding weight representation. This results in the probabilistic model:

$$p(y|\boldsymbol{z}^v, v) = \frac{\exp\left(\text{sim}(\boldsymbol{W}^v_{y,:}, \boldsymbol{z}^v)/\tau\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\text{sim}(\boldsymbol{W}^v_{y',:}, \boldsymbol{z}^v)/\tau\right)}, \tag{3}$$

where τ is a temperature scaling hyper-parameter. For training, we minimize the cross-entropy loss over the training set of size N: $\mathcal{L} = -\frac{1}{N}\sum_i^N \sum_{v' \in \mathcal{V}} \mathbb{I}[v_i = v'] \log p(y_i|\boldsymbol{x}_i, v_i)$. During inference, the attribute value label is not present. Thus, we perform an ensemble prediction over both heads to predict $\hat{y} = \arg\max_y \sum_{v' \in \mathcal{V}} p(y|\boldsymbol{x}, v')$.

Multi-label classification assigns multiple binary labels $\boldsymbol{y}$ to an image $\boldsymbol{x}$. This typically occurs when we want to predict the presence of multiple binary attributes in an image. We denote as $y^{(c)} \in \{0, 1\}$ the label of attribute c. Similar to multi-class classification, we introduce a protected weight matrix $\boldsymbol{W}^{v,c} \in \mathbb{R}^{2 \times M}$ where the two rows correspond to the absence

and presence of attribute c for protected attribute v. The resulting probabilistic model is:

$$p(y^{(c)}|\boldsymbol{z}^v, v) = \frac{\exp\left(\text{sim}\left(\boldsymbol{W}^{v,c}_{y,:}, \boldsymbol{z}^v\right)/\tau\right)}{\sum_{y'\in\{0,1\}}\exp\left(\text{sim}(\boldsymbol{W}^{v,c}_{y',:}, \boldsymbol{z}^v)/\tau\right)}, \tag{4}$$

which corresponds to a classifier for two classes. Compared with a binary classifier with a sigmoid function, the softmax function offers more flexibility for the model to represent the negatives. We minimize the cross-entropy loss over all C attributes of the training set of size N: $\mathcal{L} = -\frac{1}{N\cdot C}\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{v'\in\mathcal{V}}\mathbb{I}[v_i = v']\log p(y_i^c|\boldsymbol{x}_i, v_i)$. During inference, we also perform an ensemble prediction to compute the probability score for the presence of every attribute $\hat{y}^c = \sum_{v'\in\mathcal{V}} p(y^{(c)} = 1|\boldsymbol{x}, v')$. Binary classification is a special case where C=1.

Bias removal in the feature space. Once trained, we perform the same analysis as in Section 3 where we collect protected class prototypes in the feature space from the training set and also apply a principal component analysis on their differences $\boldsymbol{\Delta}$. We refer to the direction of the first principal component of $\boldsymbol{\Delta}$ as $\boldsymbol{b}$. Following Bolukbasi *et al*. [6], we first project features $\boldsymbol{h}$ on the bias direction $\boldsymbol{b}$ to obtain $\boldsymbol{h_b}$. Then, we neutralize the bias effect by removing $\boldsymbol{h_b}$ from the features $\boldsymbol{h}$, resulting in the mitigated features $\tilde{\boldsymbol{h}}$. Mathematically, this bias removal operation corresponds to: $\tilde{\boldsymbol{h}} = \boldsymbol{h} - \boldsymbol{h_b} = \boldsymbol{h} - \frac{\boldsymbol{h}\cdot\boldsymbol{b}}{\|\boldsymbol{b}\|}\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}$. Once $\tilde{\boldsymbol{h}}$ is computed, we can further feed it to each head to get the mitigated protected embeddings $\tilde{\boldsymbol{z}}^v = g^v(\tilde{\boldsymbol{h}})$.

Relation with *Domain Independent* [50]. Our proposed method builds on the observation from Wang *et al*. [50] that separate classification heads improve the fairness of the predictions. We differ by demonstrating how feature and label spaces also matter for addressing biases. We find the feature space implicitly encodes a bias direction (Section 3) and we derive a bias removal operation to reduce its influence. As distances matter in the feature space, this motivates us to switch from a one-hot encoding to a real-valued vector representation for the label space, where classification now occurs through a cosine embedding softmax.

5 Experiments

5.1 Fairness Metrics

Bias amplification measures whether spurious correlations present in the dataset have been amplified by the model during training [54]. Following Zhao *et al*. [54], the bias amplification score corresponds to: $\frac{1}{|\mathcal{Y}|}\sum_{v\in\mathcal{V}}\sum_{y\in\mathcal{Y}}\mathbb{I}_{s(y,v)>\frac{1}{|\mathcal{Y}|}}\frac{P_y^v}{P_y^0+P_y^1} - s(y,v)$, where P_y^v is the number of images positive for class y with a protected attribute v predicted by the model, and $s(y,v)$=$N_y^v/(N_y^0+N_y^1)$ is the ratio of training images N_y^v of class y with a protected attribute v. Intuitively, the score should be as low as possible: a positive value indicates a bias amplification while a negative value indicates a bias reduction. When training and testing sets are not *i.i.d.*, we follow Wang *et al*. [50] and compute: $\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\frac{\max(P_y^0,P_y^1)}{P_y^0+P_y^1} - 0.5$.

Demographic parity assesses the independence between a prediction $\hat{y}$ and a protected attribute v such that $p(\hat{y}$=$y'|v$=$0)$=$p(\hat{y}$=$y'|v$=$1)$ [12, 21]. Following Beutel *et al*. [4], a statistical parity difference score is derived: $\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\left|\frac{\text{TP}_y^1+\text{FP}_y^1}{N^1} - \frac{\text{TP}_y^0+\text{FP}_y^0}{N^0}\right|$, where TP_y^v and FP_y^v are the number of true positives and false positives of class y with protected attribute v, and N^v is the number of images with protected attribute v in the evaluation set. When the score tends to zero, the model makes the same rate of predictions for class y' regardless of the protected attribute value.

Model	Loss	Acc. (%,↑)	Bias (↓)	Parity (%,↓)	Opp. (%,↓)	Odds (%,↓)
BASELINE	N-way softmax	88.5±0.3	0.074±0.003	2.90±0.11	13.07±0.37	7.19±0.21
OVERSAMPLING	N-way softmax	89.1±0.4	0.066±0.002	2.77±0.67	12.58±0.19	6.91±0.11
ADVERSARIAL	w/ confusion [1, 46]	83.8±1.1	0.101±0.007	4.14±0.28	16.71±1.37	9.28±0.73
	w/ ∇ rev. proj. [13]	84.1±1.0	0.094±0.011	3.60±0.46	14.13±1.43	7.89±0.81
DOMAIN DISCRIMINATIVE	joint ND-way softmax	90.3±0.5	0.040±0.002	1.65±0.06	7.27±0.32	4.02±0.17
DOMAIN INDEPENDENT	N-way softmax×D	**92.0**±0.1	**0.004**±0.001	0.20±0.04	1.07±0.22	0.59±0.12
This paper	N-way cos softmax×D	91.5±0.2	**0.004**±0.000	**0.15**±0.01	**0.83**±0.12	**0.46**±0.07

Table 1: **Multi-class classification comparison** on N=10 classes of CIFAR-10S. Despite a small loss in the accuracy score, our proposed approach with a cosine softmax, rather than a common softmax as in DOMAIN INDEPENDENT, improves the fairness of the model in multi-class classification.

Equality of opportunity assesses the conditional independence on a particular class y' between a prediction $\hat{y}$ and a protected attribute v such that $p(\hat{y}=y'|y=y',v=0)=p(\hat{y}=y'|y=y',v=1)$ [21]. Following Beutel *et al.* [4], a difference of equality of opportunity score is derived: $\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\left|\frac{\text{TP}_y^1}{\text{TP}_y^1+\text{FN}_y^1}-\frac{\text{TP}_y^0}{\text{TP}_y^0+\text{FN}_y^0}\right|$, where FN_y^v is the number of false negatives of class y with protected attribute v. When the score tends to zero, the model classifies images as class y' correctly regardless of the protected attribute value.

Equalized odds assesses the conditional independence on any class y' between a prediction $\hat{y}$ and a protected attribute v such that $p(\hat{y}=y'|y=y,v=0)=p(\hat{y}=y'|y=y,v=1)$ [21]. Following Bellamy [2], a difference of equalized odds score is derived: $0.5\cdot(|FPR_y^1-FPR_y^0|+|TPR_y^1-TPR_y^0|)$, where FPR_y^v is the false positive rate of class y with protected attribute v and TPR_y^v is the true positive rate. When the score tends to zero, the model exhibits similar true positive and false positive rates for both protected attribute values.

5.2 Multi-class Classification

Setup. We evaluate multi-class classification on the CIFAR-10S dataset [50], which is a biased version of the original CIFAR-10 dataset [31]. A color bias is introduced in the training set, where 5 classes contain 95% gray images and 5% color images, and conversely for the 5 other classes. Visual examples for every class in their dominant color bias are present in the appendix. This creates simple spurious correlations that still affect common classifiers. Two versions of the testing set are considered: one with only gray images and another one with only color images. Although this breaks the *i.i.d.* assumption between training and testing sets, it allows the assessment of the color bias in a controlled manner. We report the per-class accuracy over 5 runs. We rely on ResNet18 [22] as the encoding function f and set each projection function g^v as a fully-connected layer of size M=128 followed by a linear activation. Training is done from scratch with stochastic gradient descent with momentum [44] for 200 epochs, and the following hyper-parameters: learning rate of 0.1 with a momentum of 0.9, batch size of 128, weight decay of 5e-4, and temperature of 0.1. The learning rate is reduced by a factor 10 every 50 epochs. Note that this setup is identical for all models we compare with, as benchmarked by Wang *et al.* [50].

Results. Table 1 compares our method with four other approaches. BASELINE is a standard model trained with an N-way softmax while OVERSAMPLING balances out the training by sampling more often underrepresented values of the protected attribute. ADVERSARIAL

blinds the feature space to the protected attribute. This is achieved either with a uniform confusion loss [1, 46] or a gradient reversal layer [13]. DOMAIN DISCRIMINATIVE makes the classification aware of the protected attribute label by assigning a class for every category and protected attribute pair [12]. DOMAIN INDEPENDENT creates two classification heads, one head for each value of the protected attribute [50]. Reported accuracy and bias amplification scores correspond to Wang *et al*. [50], while we reproduce their experiments from the source code for the demographic parity, equality of opportunity, and equalized odds scores.

Our proposed approach improves upon the other alternatives in the fairness scores. Only in the accuracy metric our model yields slightly lower results compared with DOMAIN INDEPENDENT. This shows that there might exist a trade-off between the downstream task and the fairness of the classifier, as improving both remains challenging. It is interesting that ADVERSARIAL produces worse results than simple methods such as BASELINE or OVERSAMPLING. As ADVERSARIAL blurs the distinction between both protected attribute values, it also alters the class boundaries, which makes the model less discriminative. DOMAIN DISCRIMINATIVE achieves a lower performance than our model and DOMAIN INDEPENDENT. This highlights the importance of separating the classification heads for each protected attribute value. Overall, our proposed approach with a cosine softmax, rather than a common softmax as in DOMAIN INDEPENDENT, reduces the bias direction in the feature space (see appendix) and improves the fairness in multi-class classification.

5.3 Multi-label Classification

Setup. We evaluate multi-label classification on the "Align and Cropped" split of the CelebA dataset [32], which contains 202,599 face images labeled with 40 binary attributes. Following Wang *et al*. [50], we consider the gender as the protected attribute and train models to predict the other 39 attributes. Visual examples of attributes with a high gender skewness are presented in the appendix. During the testing phase, only 34 attributes are considered as the other 5 do not contain both genders. We report the weighted mean average (mAP) precision across the selected attributes. Every positive man image is weighted by $(N_m + N_w)/(2N_m)$ while every positive woman image by $(N_m + N_w)/(2N_w)$, where N_m and N_w are the man and woman image counts in the test set. This weighting ensures a balanced representation of both genders in the evaluation of every attribute.

We rely on ResNet50 [22] pre-trained on ImageNet [39] as the encoding function f. We remove the final classification layer and replace it with two fully-connected layers (one for each protected attribute v) of size M=128 followed by a linear activation as the projection function g^v. Training is done with stochastic gradient descent with momentum [44], and the following hyper-parameters: learning rate of 0.1 with a momentum of 0.9, batch size of 32, and temperature of 0.05. The best model is selected according to the weighted mAP score on the validation set. Compared with the benchmarks introduced by Wang *et al*. [50], our model training only differs by the optimizer, as we notice some overfitting issues when using Adam [30]. The backbone and the rest of the hyper-parameters are similar.

Label space. Table 2 compares the different formulations of the label embedding space. Relying on a real-valued embedding space learned with a cosine similarity function improves the fairness of the predictions compared with the common one-hot representation. Labels now correspond to a real-valued vector instead of a binary value, which enables a distributed class representation. Switching to a softmax function instead of a sigmoid provides a weight representation for negatives, which in return helps the classification performance. The benefit of negative representations is further highlighted when applying the bias removal opera-

Loss	mAP	Bias	Parity	Opp.	Odds
N sigmoids × D	75.4	-0.039	17.74	14.87	9.19
N cos sigmoids × D	75.5	0.001	11.63	10.29	5.79
+ bias removal	74.7	-0.020	7.43	7.00	**4.00**
N cos softmax × D	**76.3**	-0.006	11.97	10.18	6.06
+ bias removal	75.3	**-0.041**	**6.71**	**6.73**	4.10

Table 2: **Label space** comparison on CelebA. An embedding learned with a cosine similarity improves the fairness upon common sigmoids. A softmax with bias removal in the feature space further improves fairness.

Embedding	Cos softmax	mAP	Bias	Parity	Opp.	Odds
Single	N	74.5	-0.039	10.65	14.02	7.77
Single	N × D	67.7	**-0.070**	19.26	21.02	13.54
Protected	N × D	**75.3**	-0.041	**6.71**	**6.73**	**4.10**

Table 3: **Single *vs*. protected embedding** comparison on CelebA. Separating the gender information into protected heads results in an increased classification and fairness performance over a single head.

Model	Loss	mAP (%,↑)	Bias (↓)	Parity (%,↓)	Opp. (%,↓)	Odds (%,↓)
BASELINE	N sigmoids	74.7	0.010	23.32	24.34	14.28
ADVERSARIAL	w/ confusion [1, 46]	71.9	0.019	23.73	28.66	16.69
DOMAIN DISCRIMINATIVE	ND sigmoids	73.8	0.007	22.34	25.35	14.69
DOMAIN INDEPENDENT	N sigmoids × D	**75.4**	-0.039	17.74	14.87	9.19
This paper	N cos softmax × D	75.3	**-0.041**	**6.71**	**6.73**	**4.10**

Table 4: **Multi-label classification comparison** of N=34 attributes in CelebA. Despite a small loss in the mAP score, our proposed embedding – learned with a cosine softmax rather than a common softmax with one-hot encoding as in DOMAIN INDEPENDENT – improves the fairness of the model in multi-label classification.

tion in the feature space, even though a small drop in the classification score occurs. Overall, learning an embedding with a softmax cross-entropy, plus the bias removal, preserves the performance of the downstream task while improving the fairness of the predictions.

Single *vs*. protected embeddings. Table 3 assesses the importance of having protected embeddings, with one projection function g^v for each value v of the protected attribute gender. We evaluate the single head setting with and without the protected attribute label in the loss function. When the protected attribute information is available, we basically have two cosine softmax losses, one for each value. Mixing the two losses in one single head is detrimental to the performance as the model gets confused on where to project the inputs in the embedding space. Protected embeddings better separate the gender information for the classification of every attribute as illustrated by the improved performance, and fairness scores overall.

Results. Table 4 compares our model with four other approaches, similarly to the comparison in Table 1. Reported mAP and bias amplification scores correspond to Wang *et al*. [50], while we reproduce their experiments to measure demographic parity, equality of opportunity, and equalized odds scores. Our proposed approach yields the fairer scores across all evaluated models. And similar to multi-class classification, we also notice a small drop in the downstream task when measuring the mAP. The ADVERSARIAL produces again the worst results across all metrics. This indicates that current methods applying an adversarial training remove more information than the bias, which is detrimental for both the downstream task and the fairness of the model. DOMAIN DISCRIMINATIVE and BASELINE result in a similar performance. Interestingly, a trade-off between the mAP and fairness scores is also present in DOMAIN INDEPENDENT. Our proposed approach improves over DOMAIN INDE-

(a) Gender prediction (age protected)

Method	Trained on *EB1*		Trained on *EB2*	
	EB2	*Test*	*EB1*	*Test*
BASELINE	59.86	84.42	57.84	69.75
Alvi *et al*. [1]	63.74	85.56	57.33	69.90
Kim *et al*. [29]	68.00	86.66	64.18	74.50
This paper	**70.85**	**88.73**	**80.59**	**83.65**

(b) Age prediction (gender protected)

Trained on *EB1*		Trained on *EB2*	
EB2	*Test*	*EB1*	*Test*
54.30	77.17	48.91	61.97
66.80	75.13	64.16	62.40
54.27	77.43	62.18	63.04
35.93	**77.67**	**65.90**	**73.08**

Table 5: **Binary classification comparison** on IMDB face dataset. Our formulation of the label embedding space improves the binary classification accuracy (%) with an extreme bias over methods that impose an invariance to the protected attribute in the feature space.

PENDENT in the fairness scores by a large margin. Mitigating the bias in both feature and label embedding spaces is then preferred over methods that only address one of the two.

Binary classification. We evaluate binary classification on the "cropped" split of the IMDB face dataset [38]. Following Kim *et al*. [29], we create three sets with an extreme bias: *EB1* comprises women $\leq$ 29 years old (yo) and men $\geq$ 40 yo; *EB2* has women $\geq$ 40 yo and men $\leq$ 29 yo; and *Test* has women and men $\leq$ 29 yo and $\geq$ 40 yo. They contain 36,004, 16,800 and 13,129 face images of celebrities. Similar to Kim *et al*. [29], we learn to predict the gender with age as a protected attribute (and conversely), and rely on ResNet18 [22] pre-trained on ImageNet [39] as the encoding function f. We add a fully-connected layer of size M=128 with linear activation for each projection function g^v. Training is done with stochastic gradient descent with momentum [44], and a learning rate of 0.1 with momentum of 0.9 and an exponential decay of 0.999, batch size of 128, and temperature of 0.1. Given the extreme bias, we update both protected heads instead of only one as done previously.

Table 5 compares our model with three other approaches. BASELINE is also a standard model trained with binary cross-entropy. Both Alvi *et al*. [1] and Kim *et al*. [29] mitigate the extreme bias by making the feature space invariant to the protected attribute. Kim *et al*. [29] rely on an adversarial formulation [9, 13], improving over Alvi *et al*. [1]. Given the binary classification setting, we did not apply a bias removal operation, as a PCA on two samples is not pertinent. Still, our formulation of the label space improves the performance in both the gender and age settings. Only when predicting age and training on *EB1*, our model struggles a bit as it tends to overfit quickly. This binary classification comparison further confirms that simpler alternatives to adversarial losses can better mitigate biases in image classifiers.

6 Conclusion

Reducing the effect of adverse decisions involves the identification and mitigation of biases within model representations. In this paper, we focus on biases coming from binary protected attributes. First, we identify a direction in the feature space of common image classifiers, where the first principal component of the difference of protected class prototypes captures bias variation. Second, building on this observation, we mitigate bias with protected projection heads that learn a label embedding space for each protected attribute value. This formulation trained with a cosine softmax cross-entropy loss improves the fairness in multi-class, multi-label and binary classifications compared with a common one-hot encoding. Removing the bias direction in the feature space reduces even further the bias effect on the classifier predictions. Overall, addressing image classifier bias on both feature and label spaces improves the fairness of predictions, while preserving the classification performance.

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCVw*, 2018.

[2] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. In *arXiv:1810.01943*, 2018.

[3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *FAccT*, 2021.

[4] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *FAT/ML*, 2017.

[5] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *WACV*, 2021.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NeurIPS*, 2016.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, 2018.

[8] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 2017.

[9] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016.

[10] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *CVPRw*, 2019.

[11] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *AIES*, 2018.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCSC*, 2012.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1), 2016.

[14] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, 115(16), 2018.

[15] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. In *FAT/ML*, 2018.

[16] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 2020.

[17] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL*, 2019.

[18] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*, 2020.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[20] Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[21] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, 2016.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[23] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[24] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *FAccT*, 2021.

[25] Sunhee Hwang, Sungho Park, Pilhyeon Lee, Seogkyu Jeon, Dohyung Kim, and Hyeran Byun. Exploiting transferable knowledge for fairness-aware image classification. In *ACCV*, 2020.

[26] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021.

[27] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *CHI*, 2015.

[28] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. Examples are not enough, learn to criticize! Criticism for interpretability. In *NeurIPS*, 2016.

[29] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, 2019.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[31] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[33] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019.

[34] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv:1901.10436*, 2019.

[35] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[36] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv:2108.02922*, 2021.

[37] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.

[38] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 126(2-4):144–157, 2018.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115 (3), 2015.

[40] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018.

[41] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NeurIPSw*, 2017.

[42] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *AAAI*, 2020.

[43] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *ECCV*, 2018.

[44] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

[45] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[46] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015.

[47] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. In *ECCV*, 2020.

[48] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, 2019.

[49] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019.

[50] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020.

[51] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *FAccT*, 2020.

[52] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, 2019.

[53] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AIES*, 2018.

[54] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[55] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *EMNLP*, 2018.

[56] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *NAACL*, 2019.