Contrastive Learning for Fair Representations

Anonymous ACL submission

Abstract

Trained classification models can unintentionally lead to biased representations and predictions, which can reinforce societal preconceptions and stereotypes. Existing debiasing methods for classification models, such as adversarial training, are often expensive to train and fragile to optimise. Here, we propose a method for mitigating bias in classifier training by incorporating contrastive learning, in which instances sharing the same class label are encouraged to have similar representations, while instances sharing a protected attribute are forced further apart. In such a way our method learns representations which capture the task label in focused regions, while ensuring the protected attribute has diverse spread, and thus has limited impact on prediction and thereby results in fairer models. Extensive experimental results on three tasks show that: our method achieves fairer representations larger bias reduction than competitive baselines; it does so without sacrificing main task performance; and it generalizes across modalities and binary- and multi-class classification tasks, being conceptually simple and agnostic to network architecture, and incurring minimal additional compute cost.

1 Introduction

Neural methods have achieved great success for classification tasks in NLP and computer vision. However, datasets which neural models are trained on embody cultural and societal stereotypes from the real world. Models trained on such datasets often capture spurious correlations between target labels and protected attributes, leading to biased predictions (i.e., models perform unequally for different sub-groups) and leakage of authorship-related sensitive information from learned representations (i.e., attackers can recover the demographic information from learned representations). This kind of unfairness has been identified in various tasks, such as twitter sentiment analysis (Blodgett et al., 2016; Han et al., 2021b), part-of-speech tagging (Hovy and Søgaard, 2015; Li et al., 2018; Han et al., 2021a), and image activity recognition (Wang et al., 2019; Zhao et al., 2017).

To mitigate bias associated with protected attributes, various kinds of methods have been proposed (Zhao et al., 2018, 2017; Li et al., 2018). Data manipulation, such as balancing the dataset with respect to the protected attribute (Wang et al., 2019) and augmenting a gender-biased dataset with gender-swapped sentences (Zhao et al., 2018), can reduce bias at the input level. However it can be costly in terms of time and compute resources, and has been shown to have a limited debiasing effect. Adversarial training is a popular method for mitigating bias by preventing a discriminator from reverse engineering protected attribute information from learned representations (Elazar and Goldberg, 2018; Resheff et al., 2019; Han et al., 2021b,a; Li et al., 2018). However, it is often difficult to optimise and increases model complexity and, consequently, computational cost.

We propose a novel debiasing method based on contrastive learning (Oord et al., 2018; Li et al., 2021a; Tian et al., 2020; Henaff, 2020; Bui et al., 2021; Li et al., 2021b; Chen et al., 2020b), which is both effective and efficient. Driven by the intuition that good and fair representations for classification should pull instances together *only* if they belong to the same class but not based on shared protected attributes (such as gender or race), we present an effective debiasing method based on contrastive learning. Specifically, our proposed method combines two contrastive loss components with a cross-entropy loss, thereby maximising the similarities of instance pairs which share a main task label and minimising the similarities of pairs with a shared protected attribute. To the best of our knowledge, our work is the first to integrate contrastive loss components to obtain fairer representations. We demonstrate the effectiveness of our method across three tasks, spanning NLP and computer vision. To ensure reproducibility, our code with this research will be released on publication. Our contributions in this work are:

1. We present a debiasing method based on contrastive learning, combining cross-entropy loss with two contrastive loss components;
2. Experimental results over two NLP and one computer vision task show that our proposed method achieves the best accuracy–fairness tradeoff in each case;
3. Our method is simple to implement and agnostic to model architectures, and incurs minimal additional computing cost.

2 Related Work

We briefly review research in the two most related areas: debiasing methods and contrastive learning.

2.1 Debiasing Methods

Prior debiasing methods fall into three categories. First, data manipulation aims to balance the input, followed by re-training the model on a fairer dataset (Wang et al., 2019; Badjatiya et al., 2019; De-Arteaga et al., 2019; Elazar and Goldberg, 2018). However, it has been shown to be both computationally prohibitive for large datasets and models, and ineffective in ensuring fair models (De-Arteaga et al., 2019; Wang et al., 2019). Second, post-processing methods "bleach" sensitive information from learnt representations after main task training. In the third category, approaches augment the original training objective, to encourage the model to learn representations that are oblivious to protected attributes. Adversarial models are the prime example (Li et al., 2018; Zhang et al., 2018; Resheff et al., 2019; Wang et al., 2019; Barrett et al., 2019; Han et al., 2021b), in leveraging one (Li et al., 2018; Elazar and Goldberg, 2018) or more (Han et al., 2021b) discriminators to encourage the main model to learn representations that do not reveal protected information. Our method also introduces an augmented objective, however, unlike adversarial methods, it does not add additional model parameters, and hence is computationally much lighter weight.

There are studies directly optimising fairness measures during training (Madras et al., 2018a; Zhao et al., 2020; Cho et al., 2020a), such as demographic parity (Feldman et al., 2015; Zafar et al., 2017; Cho et al., 2020b), equalized odds (Cho et al., 2020b; Hardt et al., 2016; Madras et al., 2018b), and equal opportunity (Hardt et al., 2016; Madras et al., 2018b). For example, Cho et al. (2020b) use kernel density estimation to approximate equalized odds during training, where this method is tailored to binary classification and can lead to a poor performance–fairness tradeoff in high-dimensional settings. Our proposed contrastive loss method can be shown to optimise for equal opportunity, encouraging the model to achieve the same true positive rate across two subgroups for instances with the same main task label (§3.4).

2.2 Contrastive Learning

The basic idea behind contrastive learning is to pull similar instances together and push dissimilar instances apart by maximising the similarities of similar instances and minimising those of dissimilar pairs within the unit feature space (Oord et al., 2018; Tian et al., 2020; Li et al., 2021a; Grill et al., 2020; Chen et al., 2020a; Henaff, 2020). It has been particularly successful in computer vision, where positive (similar) instance pairs can be generated via data augmentation (i.e., systematic, meaning-invariant manipulation of an input image such as cropping or blurring (Chen et al., 2020a; Fang et al., 2020; Cubuk et al., 2019)), and negative (dissimilar) instance pairs correspond to different items in the original data. More recently, supervised contrastive learning (SCL) was proposed in the context of classification, where positive instances belong to the same class, and negative instances belong to different classes (Khosla et al., 2020). When combined with a cross entropy loss, it has been shown to improve model robustness to noise and data sparsity (Gunel et al., 2021), as well as adversarial attacks (Bui et al., 2021). We adapt SCL to *fair* supervised learning, and present evidence of its effectiveness in learning debiased representations and fair classifiers.

3 Fair & Supervised Contrastive Learning

Our proposed method equips supervised contrastive learning with an improved loss function which simultaneously encourages data separation in terms of the main class labels, and discourages the differentiation of data points on the basis of their protected attributes. Fair contrastive learning is illustrated in Figure 1, and is compatible with different classifier architectures and data modali-

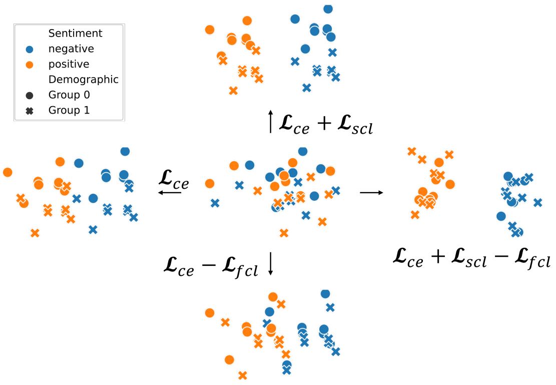

Figure 1: Illustration of our proposed method in the context of sentiment classification, where $\mathcal{L}_{\text{ce}}$ is cross-entropy loss, $\mathcal{L}_{\text{scl}}$ is contrastive loss based on main task, and $\mathcal{L}_{\text{fcl}}$ is contrastive loss based on the protected attribute.

ties, such as language and vision. Our architecture consists of three components:

1. An *embedding module*, $\boldsymbol{e} = \texttt{Embed}(\boldsymbol{x})$, which maps an input instance $\boldsymbol{x}$ (e.g., a document or an image) to a vector representation $\boldsymbol{e}$, which is in turn used as input to the encoder network;
2. An *encoder network*, $\boldsymbol{h} = \texttt{Enc}(\boldsymbol{e})$, which maps the input representation to the final hidden representation;
3. An *aggregated objective* ($\mathcal{L}_*$), which is a weighted combination of a cross-entropy loss, contrastive loss based on main task labels, and contrastive loss based on protected attribute labels, as described next.

3.1 Cross-entropy Loss

The cross-entropy loss is defined as

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{Y} y_{i,c}\log\hat{y}_{i,c},$$

where Y is the number of main task classes; $y_{i,c}$ denotes that the ith instance belongs to the main task class c; $\hat{y}_{i,c}$ denotes the predicted probability of the ith instance belonging to the main task class c; and $\hat{y}_{i,c}$ is obtained after softmax normalization of the classifier output, whose input is $\boldsymbol{h}$. However, cross-entropy loss focuses on maximising the predicted probability of the ith instance belonging to the gold-standard class, but not on ensuring larger distances in representation space between dissimilar instances than between similar ones (Figure 1, left). In this work, we explicitly model the similarity of instances in the representation space via supervised contrastive learning.

3.2 Contrastive Losses

Given a mini-batch with a set of N randomly sampled instances, positive instance pairs (those which represent the same concept) and negative instance pairs (those representing distinct concepts) are formed. We use two different criteria for creating these pairs: their main task label, and their protected attribute, as described below. Assuming a batch of positive and negative pairs, the contrastive loss is computed as,

$$\mathcal{L}_{\text{scl}} = \sum_{i=1}^{N}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\tilde{\boldsymbol{h}}_i\cdot\tilde{\boldsymbol{h}}_p/\tau)}{\sum_{q\in Q(i)}\exp(\tilde{\boldsymbol{h}}_i\cdot\tilde{\boldsymbol{h}}_q/\tau)},$$

where $i{=}1\ldots N$ is the index of an instance in the mini-batch, and $Q(i) \equiv \{1\ldots N\}\setminus\{i\}$; $\tilde{\boldsymbol{h}}_i = l_2(\texttt{Enc}(\texttt{Embed}(\boldsymbol{x}_\texttt{i})))$ is the normalised representation; and $\tau > 0$ is a scalar temperature parameter controlling smoothness. $P(i) \equiv \{p \in Q(i) : y^p = y^i\}$ is the set of instances that result in positive pairs with the ith instance, and $|P(i)|$ is its cardinality. We next describe how positive/negative pairs are created.

Supervised Contrastive Loss: $\mathcal{L}_{\text{scl}}$ is computed on positive and negative samples constructed based on main task labels (e.g., POS vs NEG sentiment), where instances in the mini-batch belonging to the same main task class are used to construct positive samples; otherwise, they are used to form negative samples. The intuition behind this loss component is that representations that are well-separated for the main task are more desirable, as illustrated in the top quadrant of Figure 1, where the main task labels are indicated in blue and orange, and are separated into distinct clusters.

Fair Contrastive Loss: $\mathcal{L}_{\text{fcl}}$ is based on positive and negative samples with respect to protected attribute labels (e.g., MALE vs FEMALE), where instances belonging to the same protected attribute class form positive samples; otherwise, they are used to construct negative samples. Our goal is to infer latent representations which are oblivious to the protected attribute of an instance. We enforce representations of instances with different protected attribute values to mix together by discouraging the model from effectively contrasting those instances, with the goal of reducing the correlation between the main task and protected attribute. We do not condition the loss on the class label, as this leads to an increase in leakage in preliminary experiments. The intuition behind this loss is illustrated in Figure 1 (bottom).

3.3 Objective Function

Our final objective incorporates both contrastive learning methods, to produce task-indicative and protected-attribute-agnostic representations, as illustrated in the right quadrant of Figure 1, formulated as a weighted average of $\mathcal{L}_{\text{ce}}$, $\mathcal{L}_{\text{scl}}$, and $\mathcal{L}_{\text{fcl}}$,

$$\mathcal{L}_* = \alpha\mathcal{L}_{\text{ce}} + \beta\{\mathcal{L}_{\text{scl}} - \mathcal{L}_{\text{fcl}}\}.$$

The second term, $\mathcal{L}_{\text{scl}}$, pulls instances from the same main task label closer together, and pushes instances from different classes further apart, while the third term, $\mathcal{L}_{\text{fcl}}$, encourages instances with the same protected attribute to disperse and instances from different groups to mix together. Our $\mathcal{L}_{\text{fcl}}$ can directly extend to non-binary protected attributes which cover more than two groups, in which case negative instances would be sampled at random from any alternative subgroup. α and β are hyperparameters that control the relative importance of the cross entropy and contrastive learning terms. Here, we adopt the same β for both $\mathcal{L}_{\text{scl}}$ and $\mathcal{L}_{\text{fcl}}$ as they are similar conceptually as well as in magnitude, and weighing them equally balances performance with bias reduction, as confirmed in extensive preliminary experiments.

We experiment with two versions of the presented model. Con_* learns all components in an end-to-end fashion. In addition, we present a pipelined setup, where we first train the $\mathtt{Enc}(\cdot)$ module using the two contrastive loss components $\mathcal{L}_{\text{scl}} - \mathcal{L}_{\text{fcl}}$, and then use its output to train a logistic classifier for the main classification task. This method is denoted as $\mathsf{Con}_*^{\text{ft}}$, which separates the representation learning and classifier training and is more efficient.

Our method differs from existing debiasing methods in that fairer representations and predictions are: (1) achieved via contrastive learning rather than data manipulation; (2) jointly trained with the base classifier, rather than removing protected attribute information through post-processing, such as with INLP (Ravfogel et al., 2020); and (3) obtained without the need to train an additional network, as necessary for adversarial methods (Li et al., 2018). We show in extensive experiments that our model is superior to adversarial and post-processing methods in terms of the performance–fairness tradeoff, and faster to train than adversarial debiasing.

3.4 Theoretical Connection

$\mathcal{L}_{\text{ce}}$-$\mathcal{L}_{\text{fcl}}$ (Figure 1, lower quadrant) optimises for demographic parity (Zafar et al., 2017; Cho et al., 2020b), where the prediction of models is independent of the protected attribute value. Our full loss adds $\mathcal{L}_{\text{scl}}$ to $\mathcal{L}_{\text{ce}}$-$\mathcal{L}_{\text{fcl}}$ (Figure 1, right quadrant), thus encouraging instances from different groups *within the same class* to be treated equally. This corresponds to equal opportunity (Hardt et al., 2016; Madras et al., 2018b), conforming to theoretical motivation and well-connected with target fairness metric (GAP, see Section 4.2). The learnt representations (Figure 4) corroborate this argument empirically.

4 Experiments

We vary the architecture of $\mathtt{Embed}(\cdot)$ across different tasks, and do not finetune it during training.[1] The architecture of $\mathtt{Enc}(\cdot)$ consists of two fully-connected layers with a hidden size of 300. All models are trained and evaluated on the same dataset splits, and models are selected based on their performance on the development set. For fair comparisons, we finetune the learning rate, batch size, and extra hyperparameters introduced by the corresponding debiasing methods for each model on each dataset. Details of the hyperparameters for each model and dataset, such as the number of layers and activation functions, are included in Supplementary Material. For all experiments, we use the Adam optimiser (Kingma and Ba, 2015) and early stopping with a patience of 5. In the absence of a standardised method for performing model selection in fairness research (noting the complexity of model selection given the multi-objective accuracy–fairness tradeoff), we determine the best-achievable accuracy for a given model, and select the hyperparameter settings that minimise GAP while maintaining accuracy as close as possible to the best-achievable value (all based on the dev set). The development of a robust, reproducible, standardised model selection method is desperately needed in fairness research, and something that we plan to investigate in future work.

4.1 Baselines

We compare our method with various baselines:

[1] For image activity recognition, $\mathtt{Embed}(\cdot)$ is first finetuned to obtain task-specific representations, and then fixed in later stages of training.

1. CE: train $\texttt{Enc}(\cdot)$ with cross-entropy loss and no explicit bias mitigation.
2. INLP: train $\texttt{Enc}(\cdot)$ with cross-entropy loss, and apply iterative null-space projection ("INLP": Ravfogel et al. (2020)) to the learned representations. Specifically, a linear discriminator is iteratively trained over the protected attribute to project the representation onto the discriminator's null-space, thereby reducing protected attribute information from the representation.
3. Adv: jointly train $\texttt{Enc}(\cdot)$ with cross-entropy loss and an ensemble of 3 adversarial discriminators over the protected attribute, with an orthogonality constraint applied to each pair of sub-discriminators to encourage them to learn different aspects of the representations (Han et al., 2021b).

4.2 Evaluation Metrics

To evaluate the performance of models on the main task, we adopt **Accuracy** for all three datasets. We measure model bias in a number of different ways, via bias in the model predictions or linear leakage over hidden or logit representations.

True positive rate (TPR) GAP measures the difference in TPR between binary protected attribute a and $\neg a$ (such as FEMALE vs. MALE, or AAE vs. SAE) for each main task class. It is defined as $\text{GAP}^{\text{TPR}}_{a,y} = |\text{TPR}_{a,y} - \text{TPR}_{\neg a,y}|, y \in Y$, where $\text{TPR}_{a,y} = \mathbb{P}\{\hat{y} = y | y, a\}$. Here $\hat{y}$ and y are the predicted and gold-standard main task labels; Y is the set of main task labels. $\text{TPR}_{a,y}$ measures the percentage of correct predictions among instances with main task label y and protected attribute a. $\text{GAP}^{\text{TPR}}_{a,y}$ measures the absolute difference between the two different groups represented by the protected attribute, with a larger absolute value indicating larger bias. A difference of 0 indicates a fair model, as the prediction $\hat{y}$ is conditionally independent of protected attribute a. Note that this formulation of the metric does not generalise to multiclass protected attributes, but in all three datasets used in this paper, all protected attributes are binary. To be able to evaluate fairness where the main task label is multiclass, we follow De-Arteaga et al. (2019) and Ravfogel et al. (2020) in calculating the root mean square of $\text{GAP}^{\text{TPR}}_{a,y}$ over all classes $y \in Y$, to get a single score:

$$\text{GAP} = \sqrt{\frac{1}{|Y|}\sum_{y \in Y}(\text{GAP}^{\text{TPR}}_{a,y})^2}$$

Linear leakage measures the ability of a linear classifier to recover the protected attribute from a model's output hidden representations or logits.

1. Leakage@**h**: based on the final hidden representation before the classifier layer.
2. Leakage@$\hat{\mathbf{y}}$: based on the main task output $\hat{y}$ (logits).

In each case, we train a linear-kernel SVM on outputs generated for the training instances, and measure leakage over the test instances. Lower values indicate a fairer model.

Tradeoff is a single aggregate measure comprising model performance as well as the three fairness metrics (GAP and leakage at $\boldsymbol{h}$ and $\hat{\boldsymbol{y}}$). Before aggregation, we scale each metric to the unit interval by dividing the model-specific values by their respective maximum ($N(\cdot)$), so that normalized values reflect the performance of each model relative to the best result. Next we assign predictive performance and overall fairness equal weights. Between fairness measures, we weigh prediction bias equal to overall leakage, leading to: Tradeoff$=\frac{1}{2}N(\text{Accuracy}) + \frac{1}{4}N(1-\text{GAP}) + \frac{1}{8}N(1-\text{Leakage@}\mathbf{h}) + \frac{1}{8}N(1-\text{Leakage@}\hat{\mathbf{y}})$. The best achievable Tradeoff is 1, which indicates that a model outperformed all other models with respect to all metrics.

Efficiency measures the GPU time required to train a model to achieve the reported results, averaged over 10 runs.

We apply our models across 3 datasets, covering NLP and vision tasks, in the form of both binary and multi-class main task classification tasks. We report results in terms of accuracy, fairness (GAP and leakage), and efficiency across all tasks. We additionally explore the accuracy–fairness tradeoff in detail for one binary NLP task (**Moji**) and one multi-class computer vision task (**imSitu**).

4.3 Experiment 1: Sentiment Analysis

4.3.1 Task and Dataset

The task is to predict the binary sentiment for a given English tweet, based on the dataset of Blodgett et al. (2016) (**Moji** hereafter), where each tweet is also annotated with a binary private attribute indirectly capturing the ethnicity of the tweet author as either African American English (AAE)

Dataset	Model	Accuracy↑	GAP↓	Leakage@**h**↓	Leakage@$\hat{\mathbf{y}}$↓	Tradeoff↑	Time↓
Moji	CE	72.09±0.65	40.21±1.23	85.75±0.46	70.96±2.11	0.77	1.0×
	INLP	72.81±0.01	36.81±3.49	68.15±1.98	67.80±1.80	0.84	–
	Adv	74.47±0.68	30.59±2.94	81.98±2.90	65.04±1.49	0.84	6.5×
	Con$_*^{\text{ft}}$	**75.99**±0.20	14.40±1.83	57.01±2.41	55.42±1.14	0.99	0.2×
	Con$_*$	75.84±0.16	**13.92**±0.44	**55.75**±0.21	**55.32**±0.25	**1.00**	1.5×
Bios	CE	**82.19**±0.04	16.68±0.46	99.24±0.05	92.72±0.85	0.76	1.0×
	INLP	79.42±0.28	15.45±1.05	92.77±6.22	67.01±0.77	0.85	–
	Adv	79.72±1.02	16.78±0.87	71.41±7.44	69.54±6.62	0.92	2.8×
	Con$_*^{\text{ft}}$	56.57±0.97	**7.35**±1.18	**66.66**±2.29	**61.06**±1.34	0.84	0.2×
	Con$_*$	81.69±0.07	16.83±0.36	75.20±1.10	66.38±1.12	**0.93**	0.9×
imSitu	CE	**58.97**±0.66	11.77±0.73	72.78±0.70	64.96±0.30	0.94	1.0×
	INLP	57.36±0.47	10.53±0.87	**60.10**±2.04	59.06±0.38	0.97	–
	Adv	58.38±0.50	10.58±0.60	67.31±0.94	62.37±0.73	0.97	4.5×
	Con$_*^{\text{ft}}$	57.67±0.30	**9.41**±1.12	71.04±0.83	**58.34**±0.47	0.97	0.1×
	Con$_*$	57.14±0.83	10.41±0.77	64.44±1.37	59.51±1.47	**0.98**	0.9×

Table 1: Experimental results on the three datasets (averaged over 10 runs). The best result for each dataset is indicated in bold. Here, ↑ and ↓ indicate that higher and lower performance, resp., is better for the given metric.

or Standard American English (SAE). Following previous studies (Ravfogel et al., 2020; Han et al., 2021b), the training dataset is balanced with respect to both sentiment and ethnicity but skewed in terms of sentiment–ethnicity combinations (40% HAPPY-AAE, 10% HAPPY-SAE, 10% SAD-AAE, and 40% SAD-SAE, respectively).[2] The dataset contains 100K/8K/8K train/dev/test instances.

4.3.2 Implementation Details

Following previous work (Elazar and Goldberg, 2018; Ravfogel et al., 2020; Han et al., 2021b), we use DeepMoji (Felbo et al., 2017), a model pretrained over 1.2 billion English tweets, as Embed(·) to obtain text representations. The parameters of DeepMoji are fixed in our experiments.

4.3.3 Results

Table 1 (**Moji**) presents the results. Compared to the CE model, INLP moderately reduces model bias across all metrics while retaining comparable accuracy, and Adv improves main task accuracy compared to CE while simultaneously reducing model bias. Both versions of our model, Con$_*$ and Con$_*^{\text{ft}}$, lead to the largest gain in accuracy and also the largest bias reduction across all metrics, requiring less GPU time compared to Adv. With leakage scores around 55, our model approaches the lower-bound value of 50 (indicating that an attacker would guess the binary protected attribute at exactly chance level). Overall, our methods achieve the best accuracy–fairness tradeoff. It is encouraging to see that incorporating debiasing techniques can contribute to improvement on the main task. We hypothesise that incorporating debiasing techniques (either in the form of adversarial training or contrastive loss) acts as a form of regularisation, leading to greater robustness over the training dataset skew relative to the unbiased test set.

Accuracy–Fairness tradeoff. We plot the tradeoff between Accuracy and Leakage@**h** for INLP, Adv, and Con$_*$ on the test set in Figure 2 (left), where points in red circles are Pareto frontiers for each model.[3] The results are obtained by varying the most-sensitive hyperparameter for each model: the number of iterations for INLP, the weight for adversarial loss for Adv, and β for our method Con$_*$. We can see that our proposed method achieves the best performance in terms of both Accuracy and Leakage@**h**.

4.4 Experiment 2: Profession Classification

4.4.1 Task and Dataset

The task is to predict a person's profession given their biography, based on the dataset of De-Arteaga et al. (2019), consisting of short online biographies which have been labelled with one of 28 professions (main task label) and binary gender (protected

[2] Note that the dev and test set are balanced in terms of sentiment–ethnicity combinations.

[3] Given two predictions whose Accuracy and Leakage@**h** are (a_1, b_1) and (a_2, b_2), if $a_1 > a_2$ and $b_1 < b_2$, we say the prediction (a_2, b_2) is dominated by the prediction (a_1, b_1); otherwise, they are non-dominated predictions, and form part of the Pareto frontier.

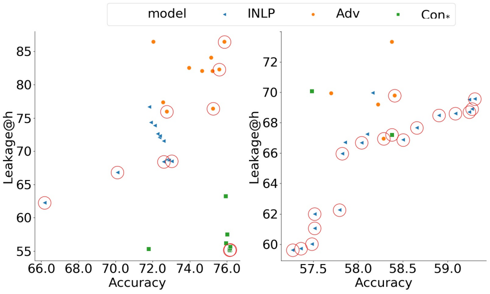

Figure 2: Accuracy vs. Leakage@**h** of different models on the **Moji** (left) and **imSitu** (right) test set, as we vary the most sensitive hyperparameter for each model. Note that points in red circles are pareto-optimal for each model.

attribute). We use the dataset split of (De-Arteaga et al., 2019; Ravfogel et al., 2020), consisting of 257K/40K/99K train/dev/test instances.[4]

4.4.2 Implementation Details

Following the work of Ravfogel et al. (2020), we use the "CLS" token representation of the pretrained uncased BERT-base (Devlin et al., 2019) as Embed(·), without any further finetuning.

4.4.3 Results

Table 1 (**Bios**) shows the results on the test set. We can see that INLP achieves the best performance in terms of GAP, but the absolute bias reduction is small compared to CE. Worryingly, both Adv and Con_* marginally increase GAP. We hypothesise that this is because of the multi-class setting (28 classes), where the large number of main task classes inhibits the ability of adversarial training and contrastive learning to mitigate bias in the model under joint training. Con_* achieves the best performance in terms of Leakage@$\hat{\mathbf{y}}$ at similar accuracy to CE, while Adv achieves the best performance in terms of Leakage@**h**. While Con_*^{ft} substantially reduces bias across the three fairness metrics, it comes at the cost of a large drop in accuracy, indicating the necessity of explicitly incorporating class information during training for this task. Overall, Con_* once again achieves the best Tradeoff of all the models with less GPU time.

4.5 Experiment 3: Activity Recognition

4.5.1 Task and Dataset

We include action recognition, a computer vision task, to demonstrate the generality of our method. Given an image, the model predicts the activity depicted in the image. We use the imSitu dataset (Wang et al., 2019; Zhao et al., 2017; Yatskar et al., 2016), which contains 211 activity classes and binary gender labels. The dataset contains only about 110 instances per activity, making it difficult to obtain decent performance without finetuning the backbone model. Therefore, we group these fine-grained labels according to their corresponding coarse-grained labels, where similar verbs are grouped into one class according to the FrameNet label hierarchy (Baker et al., 1998). The resulting dataset contains 12 target labels, and 12K/3K/2K train/dev/test instances.

4.5.2 Implementation Details

Following Wang et al. (2019) and Zhao et al. (2017), we use a standard ResNet-50 encoder (He et al., 2016) pretrained on ImageNet to extract activity-capturing representations. The classifier layer is first trained with a learning rate of 0.0001 and a batch size of 128. Then ResNet-50 is finetuned with a learning rate of 1e-5 and a batch size of 64 for at most 60 epochs. The best-performing snapshot evaluated on the dev set is used as Embed(·) to obtain image representations.

4.5.3 Results

Table 1 (**imSitu**) shows the results on the test set. INLP and Adv decrease GAP and leakage to varying degrees, with INLP achieving better performance in terms of Leakage@$\hat{\mathbf{y}}$, and Adv achieving better performance in terms of Leakage@**h**. On the other hand, Con_* achieves the best performance in terms of Leakage@$\hat{\mathbf{y}}$ and Leakage@**h** with less GPU training time. Surprisingly, Con_*^{ft} achieves the best performance in terms of GAP and Leakage@$\hat{\mathbf{y}}$, which we attribute to the fact that the classifier for the main task is disconnected from the encoder training, thereby leading to better bias reduction. Once again, Con_* is best overall in terms of Tradeoff.

Accuracy–Fairness tradeoff. Figure 2 shows the tradeoff plot between Accuracy and Leakage@**h** on the test set. The models exhibit distinct tradeoff curves, with INLP achieving the lowest leakage at high levels of accuracy.

4.6 Analysis

Effect of Loss Components To explore the impact of $\mathcal{L}_{\text{scl}}$ and $\mathcal{L}_{\text{fcl}}$, we conduct ablation studies on the **Moji** and **Bios** datasets by ablating one of

[4] There are slight differences between our dataset and that used by De-Arteaga et al. (2019) and Ravfogel et al. (2020) as a small number of biographies were no longer available on the web when we scraped them.

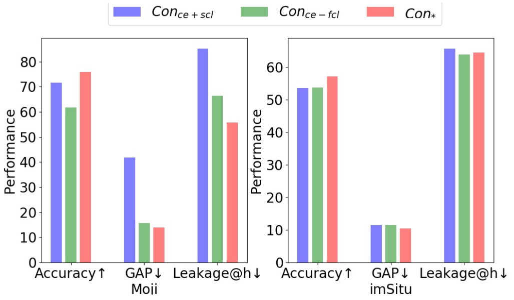

Figure 3: Effects of contrastive loss components.

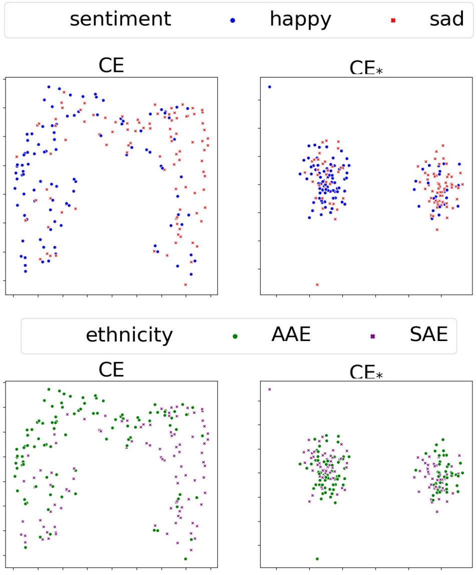

Figure 4: t-SNE scatter plots of learned representations of CE and Con_* over the **Moji** dataset (based on 100 random samples from each main task class; best viewed in colour). Red and blue colours indicate that they have different sentiment (main task) labels: red → SAD and blue → HAPPY. Green and purple colours indicate that they have different ethnic groups (protected attribute): green → SAE and purple → AAE.

the two contrastive loss components. We denote the model trained with $\alpha\mathcal{L}_{\text{ce}} + \beta\mathcal{L}_{\text{scl}}$ as $\text{Con}_{\text{ce+scl}}$, and the model trained with $\alpha\mathcal{L}_{\text{ce}} - \beta\mathcal{L}_{\text{fcl}}$ as $\text{Con}_{\text{ce}-\text{fcl}}$.

The results are shown in Figure 3. We can see that Con_* achieves the best performance across all evaluation metrics on the **Moji** dataset. On the **imSitu** dataset, Con_* also achieves the best accuracy, while roughly equalling the best bias results. This illustrates the advantage of incorporating both contrastive loss components.

Visualising Representations In Figure 4, we show t-SNE plots of the learned representations of CE and Con_* on the **Moji** training set from the perspectives of the main task labels and protected attribute values. We can clearly see that for CE, the positive (HAPPY) instances are mostly on the left of the figure and negative (SAD) instances are mostly on the right of Figure 4 (upper left). From the ethnicity perspective, AAE instances are more likely towards the left and instances with SAE are most likely to be towards the right of Figure 4 (bottom left). For Con_*, the resulting representations show that instances belonging to the same class cluster together in terms of sentiment (top right), and instances belonging to the different classes mix together in terms of ethnicity (bottom right), affirming our motivation.

4.7 Limitations

A limitation of our proposed approach is that the method is designed to remove information related to protected attributes based on the assumption that the attacker model will be a linear classifier. We leave the investigation of protecting against attacks by non-linear classifiers to future work. In our work, $\texttt{Embed}(\cdot)$ is not learned or fine-tuned together with $\texttt{Enc}(\cdot)$ and the classification layer in an end-to-end fashion. However, finetuning the $\texttt{Embed}(\cdot)$ has the potential for better task-specific or semantic-preserving representations of text and images, which may further remove biases encoded in the the pretrained models. We presented diverse experiments including existing data sets across language and vision, and balanced and imbalanced data, but acknowledge several simplifying assumptions: we restrict to binary protected attributes, implying the adoption of an oversimplified binary notion of gender. Exploring attributes of higher arity, and more complex and realistic bias dimensions is an important direction for future work.

5 Conclusion

Biased representations and predictions can reinforce existing societal biases and stereotypes. Based on the intuition that similar instances belonging to the same main task class should be pulled together and similar instances belonging to the same protected attribute class should be pushed apart in the representation space, we proposed to combine cross-entropy loss with two contrastive loss components in optimising neural networks. Experimental results over NLP and vision datasets demonstrate the effectiveness of our proposed method. Further analysis and ablation studies indicate the necessity of incorporating both contrastive loss components in bias reduction, to maintain main task accuracy.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6330–6335.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Anh Bui, Trung Le, He Zhao, Paul Montague, Seyit Camtepe, and Dinh Phung. 2021. Understanding and achieving efficient robustness with adversarial supervised contrastive learning. *arXiv preprint arXiv:2101.10027*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. In *Proceedings of Advances in Neural Information Processing Systems 33*.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. 2020a. A fair classifier using kernel density estimation. In *NeurIPS*.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. 2020b. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems*.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In *Proceedings of Advances in Neural Information Processing Systems 33*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021a. Decoupling adversarial training for fair NLP. In *Findings of the Association for Computational Linguistics*, pages 471–477.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021b. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 2: Short papers)*, pages 483–488.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of Advances in Neural Information Processing Systems 33*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021a. Prototypical contrastive learning of unsupervised representations. In *Proceedings of the 9th International Conference on Learning Representations*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 25–30.

Yunfan Li, Peng Hu, Jerry Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021b. Contrastive clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 8547–8555.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018a. Learning adversarially fair and transferable representations. In *ICML*, pages 3381–3390.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. 2018b. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning,*, pages 3381–3390.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Yehezkel S. Resheff, Yanai Elazar, Moni Shahar, and Oren Sar Shalom. 2019. Privacy and fairness in recommender systems via adversarial training of user representations. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, pages 476–482.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Proceeding of the 16th European Conference on Computer Vision*, pages 776–794.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 962–970.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. 2020. Conditional learning of fair representations. In *ICLR*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 15–20.

A Adv Settings

Each sub-discriminator consists of two MLP layers with a hidden size of 256, where the first layer is accompanied with LeakyReLU activation function. A subsequent classifier layer is used to predict the protected attribute. Sub-discriminators are optimised for at most 100 epochs after each epoch of $\texttt{Enc}(\cdot)$ training, leading to extra training time.

B Hyperparameter Settings

B.1 Twitter Sentiment Analysis

For CE, the learning rate is 3e-3, and the batch size is 2,048. For Adv, the learning rate is 1e-3, the batch size of 2,048, the number of sub-discriminators is 3, λ_{adv} is 0.5, and λ_{diff} is 1e-3. For INLP, following Ravfogel et al. (2020), we use 300 linear SVM classifiers. For Con_* and $\mathsf{Con}_*^{\text{ft}}$, the learning rate is 7e-5, the batch size is 1,024, $\tau = 0.01$, and $\alpha = \beta = 0.5$.

B.2 Occupation Classification

For CE, the learning rate is 3e-3, and the batch size is 2,048. For Adv, the learning rate is 0.01, the batch size is 1,024, the number of sub-discriminators is 3, λ_{adv} is 0.01, and λ_{diff} is 1e4. For INLP, we use 300 linear SVM classifiers. For Con_* and $\mathsf{Con}_*^{\text{ft}}$, the learning rate is 3e-3, the batch size is 512, $\tau = 0.01$, $\alpha = 0.91$, and $\beta = 0.09$.

B.3 imSitu Activity Recognition

For CE, the learning rate is 5e-4, and the batch size is 512. For Adv, the learning rate is 1e-3, the batch size is 256, the number of sub-discriminators is 3, λ_{adv} is 0.01, and λ_{diff} is 1.0. For INLP, we use 300 linear SVM classifiers. For Con_* and $\mathsf{Con}_*^{\text{ft}}$, the learning rate is 5e-3, the batch size is 512, $\tau = 0.01$, $\alpha = 0.95$, and $\beta = 0.05$.