Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations

Sheng Liang, Philipp Dufter, Hinrich Schütze
Center for Information and Language Processing (CIS)
LMU Munich, Germany
{shengliang, philipp}@cis.lmu.de

Abstract

Pretrained language models (PLMs) learn stereotypes held by humans and reflected in text from their training corpora, including gender bias. When PLMs are used for downstream tasks such as picking candidates for a job, people's lives can be negatively affected by these learned stereotypes. Prior work usually identifies a linear gender subspace and removes gender information by eliminating the subspace. Following this line of work, we propose to use DensRay, an analytical method for obtaining interpretable dense subspaces. We show that DensRay performs on-par with prior approaches, but provide arguments that it is more robust and show that it preserves language model performance better. By applying DensRay to attention heads and layers of BERT we show that gender information is spread across all attention heads and most of the layers. Also we show that DensRay can obtain gender bias scores on both token and sentence level. Finally, we demonstrate that we can remove bias multilingually, e.g., from Chinese, using only English training data.

1 Introduction

Word embeddings, which represent the semantic meaning of text data as vectors, are used as input in natural language processing tasks. It has been found that word embeddings exhibit biases such as gender bias, which are present in their training corpora (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018). Contextual word embedding models, such as BERT (Devlin et al., 2019), have become increasingly common and achieved new state-of-the-art results in many NLP tasks. Researchers have also found gender bias in contextualized embeddings (Zhao et al., 2019; May et al., 2019). It has been found that these biases have a big impact on downstream tasks (Vanmassenhove et al., 2018; Moryossef et al., 2019; Rudinger et al., 2018; Zhao et al., 2018).

A common approach for removing gender information in static embeddings is to identify a linear gender subspace (e.g., a gender direction) and subsequently setting all values on the gender direction to zero. Successful approaches rely on simple principal component analysis (Bolukbasi et al., 2016; Mu and Viswanath, 2018). Bolukbasi et al. (2016) require pairs of gendered words to compute a direction (e.g., "man"-"woman") and Mu and Viswanath (2018) rely on computing a PCA of a set of gender words hoping that the main variation direction is the gender direction. We propose to use DensRay (Dufter and Schütze, 2019): the main advantage is that DensRay only requires two or multiple groups of gendered words. In contrast to Bolukbasi et al. (2016), it does not require explicit pairs. Compared to Mu and Viswanath (2018), it has explicit supervision with gender labels. We show in §2.5 that this enables DensRay to identify better gender directions.

In summary, our contributions are: i) We adjust DensRay to be usable to debias contextualized embeddings. We evaluate the approach on two tasks – a set of templates we constructed and and association tests – and show that DensRay performs comparable to prior approaches. ii) We provide arguments why DensRay is more robust and find that applying DensRay preserves language model performance on downstream tasks better. iii) We analyze how gender information is processed in BERT by applying DensRay to attention heads and layers: we conclude that there is no single attention head responsible for processing gender information. In addition we show in a qualitative analysis that token-level gender scores can be obtained. iv) We apply our debiasing method to the multilingual-BERT

(mBERT) model: we show that English training data can be used to effectively debias Chinese. The source code of our experiments is available. [1]

2 Methodology

2.1 Debiasing Conceptor

Karve et al. (2019) introduced conceptor debiasing. Given a set of gendered words $V := \{v_1, v_2, \ldots, v_n\}$ and their embeddings E, gender bias can be mitigated by multiplying the debiasing conceptor matrix $\neg C = I - C$, where C is the conceptor matrix that minimizes the objective

$$||E - CE||_F^2 + \alpha^{-2}||E||_F^2, \tag{1}$$

where α is a parameter. C has an analytical solution given by

$$C = \frac{1}{d}EE^T(\frac{1}{d}EE^T + \alpha^{-2}I)^{-1}. \tag{2}$$

Intuitively, C is a soft projection matrix on the linear subspace where embeddings have the maximum bias.

2.2 Hard Debasing

We will also compare with the hard debiasing method proposed by Mu and Viswanath (2018), which is originally a postprocessing technique for improving word representations. Karve et al. (2019) adopted it as a method for debiasing. Hard debiasing relies upon the assumption that the first principal component of the embedding vectors associated to gendered words is a meaningful gender direction. The first principal component of E is the gender direction q.

2.3 DensRay

DensRay (Dufter and Schütze, 2019) is an analytical method for identifying the embedding subspace of certain linguistic features. It identifies the "gender subspace" using a set of gendered words $V := \{v_1, v_2, \ldots, v_n\}$ and their embeddings $E \in R^{n \times d}$. In contrast to the above approaches it uses a function l for the gender attribute: $l : V \to \{-1, 1\}$; e.g. $l(\text{father}) = 1$, $l(\text{sister}) = -1$. The objective of DensRay is to find an orthogonal matrix $Q \in R^{d \times d}$ such that the rotated space EQ the dimensions are ordered by there degree of correlation with gender, e.g., the first dimension can then be interpreted as a gender subspace.

Let $L_= := \{(v, w) \in V \times V | l(v) = l(w)\}$ and define $L_{\neq}$ analogously. The DensRay objective in Eq. 3 is to maximize the distance of the word pairs from the same gender group ($L_=$) and minimize the distance of the word pairs from the different gender group ($L_{\neq}$).

$$\max_q \sum_{(v,w) \in L_{\neq}} \alpha_{\neq} ||q^\intercal d_{vw}||_2^2 - \sum_{(v,w) \in L_=} \alpha_= ||q^\intercal d_{vw}||_2^2 \tag{3}$$

where $d_{vw} := e_v - e_w$. The objective can be simplified to

$$\max_q q^\intercal (\sum_{(v,w) \in L_{\neq}} \alpha_{\neq} ||d_{vw} d_{vw}^\intercal||_2^2 - \sum_{(v,w) \in L_=} \alpha_= ||d_{vw} d_{vw}^\intercal||_2^2) q =: \max_q q^\intercal A q \tag{4}$$

As stated in (Dufter and Schütze, 2019) q is simply the eigenvector of A corresponding to the largest eigenvalue. $\alpha_{\neq}, \alpha_= \in [0, 1]$ are hyperparameters to balance the different optimization terms.

2.4 Removing Gender Information

Hard debiasing and DensRay yield a gender dimension $q \in R^d$. In a contextualized language model like BERT each layer yields a contextualized embedding matrix $X \in R^{t \times d}$ where t is the length of the sentence. To debias representations we simply zero out the projected values on q for each position, that is we set $x_i^{\text{debiased}} = x_i - (x_i^\intercal q)q$ for each position i. To debias X with conceptor one simply performs one matrix multiplication $X^{\text{debiased}} = X(I - C)$. Then we use X^{debiased} as the input of the next layer.

[1] https://github.com/liangsheng02/densray-debiasing/

2.5 Comparison of the Approaches

Figure 1 shows artificially created two dimensional embeddings. The lines show the gender directions identified by hard debiasing and DensRay. In this example the first principal component does not correlate with gender, while DensRay is able to handle this as it explicitly uses the labels.

As a soft projection, Conceptor debiasing does not compute a single gender direction, but always debias on a linear combination of all directions, it is not obvious how to depict the direction in the figure. We argue that having a single gender direction as in DensRay is more interpretable and easier to use than the Conceptor, e.g., it allows to assign token level gender scores easily as shown in Table 5. In addition DensRay affects language model performance less than Conceptor, as we show in Table 3.

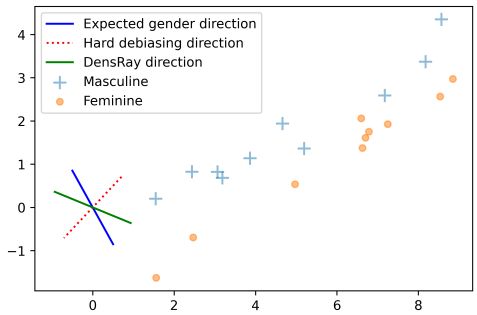


Figure 1: Gender direction for a set of gendered words computed for hard debiasing and DensRay.

2.6 Adapting DensRay to Contextualized Language Models

We now describe how we adapt DensRay to contextualized language models. Given a set of gendered words V, we run a contextualized language model with M layers on a set of sentences that each contains at least one of the gender words, i.e., $t_1, \ldots, t_j, \ldots, t_n$ (where $t_j \in V$) to get their embeddings. For each t_j in each sentence we compute the contextualized representations $e_j^m, 1 \leq m \leq M$, one for each layer. We then use the vectors e_j^m in Eq. 4. to compute a gender direction q_m for the m-th BERT layer.

3 Experiments

3.1 Setup and Data

In the experiments, we lowercase all text and use the BERT models "bert-base-uncased" and "bert-large-uncased". We implemented all experiments using the transformers library (Wolf et al., 2019).

To compute the rotation matrices by DensRay, we need a gendered word list and a corpus. For the word list, we get 23 masculine words and 23 feminine words from the "family" category,[2] of the Google analogy test set (Mikolov et al., 2013). As the input corpus, we collect text data from Wikipedia that contains 5,000 (resp. 10,000) occurrences of words in the gendered list for the BERT base (resp. large) model. We carefully balance the occurrences such that the number of male and female samples are equal, then we set $\alpha_{\neq} = \alpha_{=} = 0.5$. We compare with hard debiasing (Mu and Viswanath, 2018) and debiasing conceptor (Karve et al., 2019) to eliminate gender bias as adapted to contextualized embeddings. In the experiments, we found that DensRay and hard debiasing applied to all BERT layers yield better results than only applying to the last layer. In contrast, when applying debiasing conceptor to all layers, language modeling performance (as measured by perplexity) is extremely poor. We therefore applied the DensRay and hard debiasing to all BERT layers while debiasing conceptor only to the last layer.

3.2 Evaluation

We use two evaluation datasets to measure gender bias: Association tests (Section 3.2.2) and OCCTMP (occupation templates). OCCTMP is a method to use existing occupation datasets for evaluation, based on the templates that we created specifically to evaluate contextualized language models. It has the added advantage that results are easier to interpret than those for Association tests.

[2] `https:\\download.tensorflow.org/data/questions-words.txt`

3.2.1 OCCTMP

To construct OCCTMP, we start with 320 occupation names provided by Bolukbasi et al. (2016). Each occupation name is converted into the format "[MASK] is an *occupation*." We measure gender bias as the mean of the differences between the log probability of BERT predicting [MASK] as "he" vs. "she"

$$\textit{diff} = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}}(\log p_{max} - \log p_{min}),$$

$$\text{where } p_{max} = \max\{p(\text{he}|T), p(\text{she}|T)\}; p_{min} = \min\{p(\text{he}|T), p(\text{she}|T)\}$$

where $\mathcal{T}$ is the set of 320 templates. We can also use the standard deviation of the differences std to indicate the stability of debiasing, and the average sum probability $sum = prob(he|\mathcal{T}) + prob(she|\mathcal{T})$ as an indication of language model performance. We will also report the standard deviation of the differences. We find that for most experiments and templates, the probability of "he" is higher than "she", which qualitatively indicates that gender bias can be identified using the OCCTMP templates. We also find that in most cases the sum of the two probabilities is higher than 0.7; thus, this evaluation task is a good fit for BERT because it has learned that a pronoun is likely to occur in the masked position. Our templates can be easily extended to other languages such as Chinese.

3.2.2 Association Tests

Association tests originated in sociological research. Greenwald et al. (1998) proposed the Implicit Association Test (IAT) to quantify societal bias. In IAT, response times were recorded when subjects were asked to match two concepts. For example, subjects were asked to match black and white names with "pleasant" and "unpleasant" words. Subjects tended to have shorter response times for concepts they thought associated.

Based on IAT, Caliskan et al. (2017) proposed the Word Embedding Association Test (WEAT). It uses word similarities between targets and attributes instead of the response times to get rid of the requirement for human subjects. Consider two sets of target words X_1, X_2 with equal size $|X_1| = |X_2|$, and two sets of attribute words A_1, A_2. The null hypothesis in the WEAT statistical test is: there is no difference in the similarity between X_1, X_2 and A_1, A_2. In the prior literature, it has been argued that if the null hypothesis cannot be rejected, there is no significant gender bias. The WEAT test statistic is defined as

$$s(X_1, X_2, A_1, A_2) = \sum_{x\in X_1} s(x, A_1, A_2) - \sum_{x\in X_2} s(x, A_1, A_2),$$

$$\text{where } s(x, A_1, A_2) = \text{mean}_{a\in A_1} cos(\vec{x}, \vec{a}) - \text{mean}_{a\in A_2} cos(\vec{x}, \vec{a}),$$

in which $cos(\vec{x}, \vec{a})$ denotes the cosine similarity between embedding vector $\vec{x}$ and $\vec{a}$. Intuitively, $s(x, A_1, A_2)$ measures the association of a word with the attributes, so the test statistic measures the differential association of the two target sets with the attributes.

Let $\{(X_{1i}, X_{2i})\}_i$ denote all the partitions of $X_1 \cup X_2$. The one-sided p-value of the permutation test is defined as

$$p = Pr_i[s(X_{1i}, X_{2i}, A_1, A_2) > s(X_1, X_2, A_1, A_2)].$$

The effect size d-value is a normalized measure of how separated the two distributions of associations between the target and attribute are. It is defined as

$$d = \frac{s(X_1, X_2, A_1, A_2)}{\text{std}_{x\in X_1\cup X_2} s(x, A_1, A_2)}.$$

To extend WEAT to contextual embeddings, Karve et al. (2019) extracted contextual embeddings from the template "[MASK] is *word*". May et al. (2019) proposed the Sentence Embedding Association Test (SEAT), for which they designed more complex templates to extract word embeddings.

In these association tests, we measure the effect size d-value and the one-sided p-value of the permutation test. A d-value closer to zero indicates less gender bias. We also prefer a high p-value (at least 0.05) that aims to not reject the null hypothesis, i.e., we do not reject that there is no gender bias. We use the three categories C6: career/family, C7: math/arts, C8: science/arts, following Karve et al. (2019)'s WEAT setup and May et al. (2019)'s SEAT setup.

3.2.3 Model Performance

It is crucial that debiasing methods do not harm the downstream performance of BERT models. Thus we test the perplexity of language modeling on Wikitext-2 (Merity et al., 2016), a subset of Wikipedia with 2 million words. We also test on GLUE (Wang et al., 2018), using the same setup as (Wolf et al., 2019).[3] Since the deviation already exists in the data of these tasks, the overall performance will decrease after debiasing. Our expectation here is to cause less damage to the model performance after debiasing.

4 Results

4.1 Debiasing Results

Table 1 gives results on OCCTMP. Two OCCTMP examples are given in Table 4. We see that DensRay can mitigate the gender bias in BERT as measured by diff: bias between predicting he/she drops by large margin (e.g., for bert-base from 1.98 to 0.36). The table indicates that DensRay outperforms the other two methods on OCCTMP. Note that the prediction probabilities of debiasing conceptor are quite low for both "he" and "she" indicating that conceptor debiasing might affect language model performance. However, in relative terms, "he" is still strongly favored compared to "she". Table 2 shows the results on association tests. We observe that association tests have a high variance depending on which category is used. Overall the debiasing performance of the three methods are comparable with DensRay and Conceptor having the best performance three times and hard-debiasing having the best performance 5 times. Overall the debiasing performance of the three methods are comparable, with DensRay and Conceptor both having the best performance three times and hard-debiasing having the best performance five times.

model	debiasing	$p(\text{he}\|\mathcal{T})$	$p(\text{she}\|\mathcal{T})$	sum	diff	std
bert-base	-	0.66	0.19	0.85	1.98	1.39
	hard	0.35	0.42	0.77	0.42	0.09
	conceptor	0.18	0.11	0.28	0.68	0.26
	densray	0.48	0.37	0.86	**0.36**	0.07
bert-large	-	0.63	0.19	0.82	1.82	1.30
	hard	0.40	0.23	0.63	0.69	0.30
	conceptor	0.43	0.18	0.61	1.03	0.53
	densray	0.47	0.31	0.77	**0.49**	0.13

Table 1: BERT debiasing results on OCCTMP, measure by *diff*, *std*, and *sum* described in 3.2.1

		bert-base				bert-large			
		WEAT		SEAT		WEAT		SEAT	
category	debiasing	d	p	d	p	d	p	d	p
C6	-	0.66	0.08	1.04	$<10^{-2*}$	1.57	$<10^{-2*}$	0.50	$<10^{-2*}$
	hard	0.15	0.38	**-0.08**	0.67	0.80	0.06	0.07	0.35
	conceptor	**0.07**	0.46	0.77	$<10^{-2*}$	1.33	$<10^{-2*}$	**0.06**	0.37
	densray	0.62	0.12	0.36	0.02^{*}	**0.76**	0.07	0.13	0.22
C7	-	0.60	0.11	0.17	0.15	-0.40	0.75	0.38	0.01^{*}
	hard	**-0.07**	0.56	**-0.06**	0.64	-0.51	0.83	0.38	0.01^{*}
	conceptor	0.54	0.14	-0.25	0.93	-0.32	0.73	**-0.60**	0.99
	densray	0.09	0.45	-0.47	0.99	**0.06**	0.05	-0.73	0.99
C8	-	0.78	0.08	0.81	$<10^{-2*}$	-0.60	0.87	-0.30	0.95
	hard	-0.29	0.68	**-0.10**	0.71	0.78	0.06	**-0.03**	0.56
	conceptor	0.62	0.14	0.50	$<10^{-2*}$	**0.12**	0.39	0.30	0.94
	densray	**0.03**	0.47	0.41	0.01^{*}	0.20	0.33	-0.66	0.99

Table 2: BERT debiasing results on association tests. * shows significant gender bias. Only models without significant gender bias ($p > 0.05$) can be accepted.

[3] https://huggingface.co/transformers/

4.2 Model Performance

Table 3 shows that DensRay debiasing gets comparable results with the original models on Wikitext-2 and GLUE tasks. In most tasks on bert-base and all tasks on bert-large, DensRay performs better than hard debiasing, so DensRay affects model performance less. Similarly, in most tasks on bert-base and all tasks but one on bert-large, DensRay performs better than debiasing conceptor. Overall we find that DensRay affects model performance the least among the considered methods.

model	debiasing	Wikitext-2	CoLA	SST-2	MRPC	STS-B	RTE	WNLI	GLUE avg
bert-base	-	3.77	49.15	92.09	85.86	82.66	62.82	52.11	70.78
	hard	3.95	45.53	**91.74**	82.48	**82.60**	63.54	**56.34**	70.37
	conceptor	4.46	**48.31**	91.43	84.08	81.37	59.57	**56.34**	70.18
	densray	**3.81**	48.04	**91.74**	**84.89**	82.43	**63.90**	53.52	**70.75**
bert-large	-	3.29	47.93	94.90	89.30	87.60	70.10	65.10	75.82
	hard	3.85	47.45	93.95	85.01	82.33	67.12	63.02	73.15
	conceptor	4.13	**49.44**	93.87	87.67	83.44	62.45	56.34	72.20
	densray	**3.35**	48.91	**94.02**	**88.84**	**85.63**	**67.78**	**64.48**	**74.94**

Table 3: Language modeling perplexity and GLUE tasks performance.

4.3 Examples

In Table 4 we show two OCCTMP examples. Again, DensRay works best as measures by diff. The sum probabilities of "he" and "she" on the debiasing conceptor are around 0.5, indicating that the language model has lost part of its ability to predict that a pronoun is likely to occur in the masked position.

template T	debiasing	$p(he\|T)$	$p(she\|T)$	sum	diff
[MASK] is a professor.	-	0.84	0.13	0.97	1.86
	hard	0.37	0.55	0.92	0.40
	conceptor	0.28	0.23	0.51	0.20
	densray	0.53	0.37	0.90	0.36
[MASK] is a dancer.	-	0.22	0.72	0.94	1.19
	hard	0.27	0.64	0.91	0.86
	conceptor	0.20	0.33	0.53	0.50
	densray	0.42	0.52	0.94	0.21

Table 4: OCCTMP examples with prediction probabilities.

4.4 Analysis

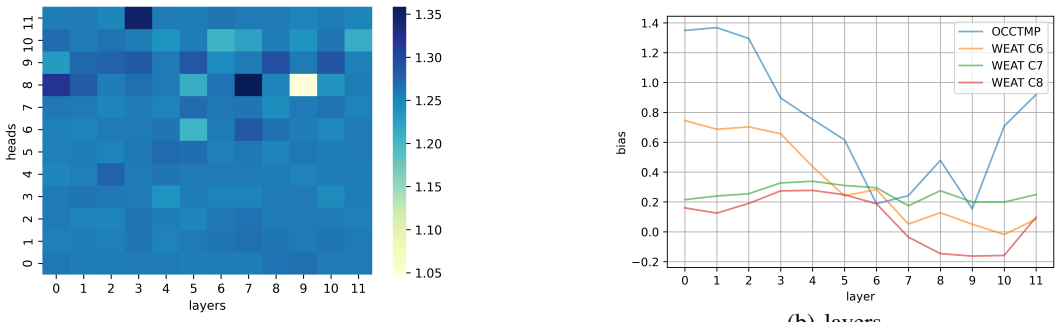
Debiasing on Attention Heads and Layers

We now apply DensRay to single attention heads in BERT and investigate the debiasing effect on OCCTMP. The heatmap Figure 2(a) shows that the debiasing effect of one single attention head is not apparent, with diff scores all in (1.0,1.4). We conjecture that there is no single attention head that is responsible for processing gender information. This conjecture is similar to the experiment from Bau et al. (2019), which showed modifying some neurons activations in NMT did not help in controlling gender, and speculated that gender property is very distributed on neurons.

So far we have always applied DensRay to all BERT layers simultaneously. Figure 2(b) illustrates the effect of debiasing a single layer on OCCTMP and the three WEAT categories. In contrast to the attention heads we observe a different debiasing effect across different layers. We see that the debiasing effect is stronger in layers 7–10 than in the other layers in bert-base. This indicates that gender information is processed on BERT layers, especially the upper layers.

Quantifying Gender Bias with DensRay

DensRay can be used to quantify gender bias for sentences and tokens. We use the distance to the origin in gender subspace as the measure. In BERT, we use the average bias score of tokens to quantify the whole sentence. Table 5 compares DensRay with the log probability score (Kurita et al., 2019), which quantifies gender bias based on templates of form "[TARGET] is a [ATTRIBUTE]". We regard zero as a



(a) attention heads (b) layers

Figure 2: (a): DensRay debiasing on each single attention head in bert-base, measured by diff on OCCTMP. (b): DensRay debiasing on each single layer, measured by diff on OCCTMP and d-value on WEAT.

balance point without bias. Contrary to the log probability score, a positive DensRay score represents the level of female bias. These examples show that DensRay is more versatile, it can quantify bias both on the token and on the sentence level in contrast to the log-probability score. In the sentence "The professor asked the nurse ." one can immediately see that the model has a male bias on "professor" and a female bias on "nurse", although the sentence itself is completely gender-neutral.

DensRay	Avg.	log probability score
[MASK] cooked dinner .	0.14	-1.04
[MASK] is a professor .	-0.45	0.63
[MASK] is a nurse .	1.58	-5.44
The professor asked me .	-0.19	not applicable
The professor asked the nurse .	0.43	not applicable
The child played with the car .	0.03	not applicable
The child played with the doll .	0.49	not applicable

Table 5: Examples for quantifying bias (model: bert-base), where red(blue) denotes female(male) bias.

Alternatives to Removing the Gender Dimension

In our experiments, we zero out the dimensions of the gender subspace to remove gender bias. We also explored three alternatives to zeroing out. 1) Replace the first dimension of the gender subspace with the mean value of the first dimension of the training samples. 2) Standardize the first dimension. 3) Replace the first dimension with a small random variable sampled from a Gaussian distribution. All of them did not perform well. Using higher dimensional gender subspaces also did not improve the debiasing results while harming the model performance.

4.5 Multilingual Debiasing

We now show that, in a multilingual contextualized language model like mBERT, we can use DensRay for zero-shot debiasing. Specifically, we train a DensRay model on English and use it to debias Chinese.

In our opinion, multilingual debiasing involves two distinct problems. First, how can bias be removed from the underlying model? Second, how does bias manifest in a particular language? The removal of bias from the underlying model can be argued to be largely independent of the language whereas the way bias manifests is highly language-dependent. For example, Chinese does not mark gender and most German nouns describing people are gender-specific. So on the surface, Chinese are gender-neutral and German cannot be gender-neutral (e.g., *Studentinnen* and *Studenten*). However, these particularities of surface form of individual languages do not change the underlying problem of biased language models: Chinese and German language models are still biased and should be debiased to avoid unfair and biased impact caused by deployed NLP systems.

We use bert-base-multilingual-uncased from (Wolf et al., 2019) with the same setup as for bert-base-uncased in our previous experiments. As before, we compute the rotation matrices using the English gendered words from the "family" category of the Google analogy test set (Mikolov et al., 2013). Since

Chinese is a language that does not mark gender, we construct OCCTMP by directly translating from the English templates. We use the following form: "[MASK]是一个*occupation*。" We translate the occupations using Tencent Translation[4] and make some manual adjustments to the translation. After removing duplicates (e.g. firefighter and fireman have the same translation is Chinese), 302 Chinese templates remain.

Table 6 gives results for the Chinese templates. Two examples are given in Table 7. We see that DensRay trained with English can mitigate gender bias in mBERT: the diff score drops from 1.39 to 1.22 on Chinese templates.

debiasing	$p(he\|\mathcal{T})$	$p(she\|\mathcal{T})$	sum	diff	std
-en	0.51	0.14	0.65	1.66	0.90
densray-en	0.33	0.12	0.46	1.33	0.81
-zh	0.24	0.07	0.31	1.39	0.77
densray-zh	0.12	0.04	0.17	1.22	0.69

Table 6: Results of OCCTMP on mBERT after applied DensRay. Models with *-en* are tested on English templates, and those with *-zh* are tested on Chinese templates.

template T	debiasing	p(他$\|T$)	p(她$\|T$)	sum	diff
[MASK]是一个客座教授。 [MASK] is an adjunct professor.	-en	0.68	0.16	0.84	1.45
	densray-en	0.51	0.18	0.70	1.04
	-zh	0.52	0.11	0.63	1.55
	densray-zh	0.30	0.08	0.38	1.31
[MASK]是一个管理员。 [MASK]is an administrator.	-en	0.53	0.17	0.70	1.14
	densray-en	0.35	0.13	0.48	0.99
	-zh	0.68	0.16	0.84	1.45
	densray-zh	0.51	0.18	0.69	1.04

Table 7: Sanity check on the Chinese templates, where 他 means *he* and 她 means *she*. Corresponding English translations are shown below the Chinese.

5 Related Work

5.1 Quantifying Gender Bias

Association tests are commonly used to measure gender bias. The original WEAT is used on static embeddings. To extend WEAT to contextual embeddings, some template-based processes (Karve et al., 2019; Kurita et al., 2019; Tan and Celis, 2019) were constructed to obtain word embeddings from the context. SEAT further improved the templates and computed the similarities between sentences instead of words. Guo and Caliskan (2020) proposed CEAT (Contextual Embedding Association Test), which samples the occurrences from corpus to get the embeddings, and measures bias by a random-effects model. However, we find that it is difficult to obtain a stable result on CEAT, since there are too many hyper-parameters to be controlled, thus in this paper we use WEAT and SEAT for our experiments. Kurita et al. (2019) proposed a template-based log probability bias score to measure the association between targets and attributes in BERT. Since it can only be applied on specific templates, we compare this method with DensRay as a measure of gender bias in §4.4.

An alternative way to measure gender bias is to evaluate on **downstream tasks**. For coreference resolution, Zhao et al. (2018) designed Winobias and Rudinger et al. (2018) designed Winogender schemas. Webster et al. (2018) released GAP, a balanced corpus of Gendered Ambiguous Pronouns, which measures gender bias as the ratio of F1 score on masculine to F1 score on feminine. However the ratio close to 1.0 (Chada, 2019; Attree, 2019) making it hard to compare debiasing systems. For sentiment analysis, Equity Evaluation Corpus (EEC) (Kiritchenko and Mohammad, 2018) was designed to measure gender bias by the difference in emotional intensity predictions between gender-swapped sentences. Since the measures of gender bias in these datasets are not intuitive, we use association tests in this work.

[4] `fanyi.qq.com/`

5.2 Debiasing Methods

Many methods to remove gender bias have been proposed. The most common way is to define a gender direction (or, more generally, a subspace) by a set of gendered words and debias the word embeddings in a post-processing projection. Bolukbasi et al. (2016) propose (i) *hard debiasing*: use the gendered words to compute the difference embedding vector as the gender direction, and remove the gender subspace component of the neutral words while preserve it for the gendered words; and (ii) *soft debiasing*, a machine learning based method that combines the inner-products objective of word embedding and an objective to project the word embedding into an orthogonal gender subspace. It has been found to work better than soft debiasing. Prost et al. (2019) proposed a variant of the hard debiasing algorithm by simply removing gender subspace component of all words in the vocabulary, we also applied this strategy in our approach. Dev and Phillips (2019) explored partial projection and some simple tricks to improve the hard debiasing method. Zhao et al. (2019) applied the data augmentation and debiasing method of Bolukbasi et al. (2016) to mitigate gender bias on ELMo (Peters et al., 2018). Karve et al. (2019) proposed the debiasing conceptor, which shrinks each principal component of the covariance matrix of the embeddings to achieve a soft debiasing. They also introduced a simple and intuitive hard debiasing method proposed by (Mu and Viswanath, 2018), which identified the gender subspace by PCA and projected the first principal component off.

The debiasing conceptor and the Mu and Viswanath (2018) hard debiasing produce gender direction by gendered word list mixed with male and female words. In contrast, the Bolukbasi et al. (2016) hard debiasing used two groups of gendered words for definition and another two groups for alignment, to identify the gender direction by male-female pairs. The method we use, DensRay, is similar to the Bolukbasi et al. (2016) hard debiasing in this aspect. However, DensRay uses only one male and one female word list, and it can be solved efficiently in a closed form. So it would be more stable to be applied to contextual models than the Bolukbasi et al. (2016) hard debiasing.

Gonen and Goldberg (2019) showed that except the gender direction, biases on static embeddings also come from the association to other implicitly gendered terms. We assume that this phenomenon will be weakened in contextual embedding, so we only focus on the gender direction, this may be a limitation.They also proposed some experiments to evaluate the remaining bias after debiasing, among which the gender classifier shared the same idea with OCCTMP.

6 Conclusion

We introduced DensRay debiasing on BERT. Our experiments showed that this method can effectively mitigate gender bias on OCCTMP and the Association Tests, while maintained the performance of BERT on language modeling and GLUE tasks. We applied DensRay to BERT attention heads, showed that gender information is processed in all attention heads, there is no single attention head responsible for processing gender information. We also used DensRay to obtain interpretable gender bias scores, to quantify bias on token and sentence level for all representations. Finally, we demonstrated that we can remove bias multilingually, we used only English training data to effectively debias Chinese.

References

Sandeep Attree. 2019. Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 134–146.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, Apr.

Rakesh Chada. 2019. Gendered pronoun resolution using bert and an extractive question answering formulation. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 126–133.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887.

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.

Philipp Dufter and Hinrich Schütze. 2019. Analytical methods for interpretable ultradense word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1185–1191.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3635–E3644.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Wei Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *ArXiv*, abs/2006.03955.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ArXiv*, abs/1301.3781.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy, August. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, page 13209–13220.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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, Volume 2 (Short Papers)*, pages 15–20.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634.

Appendix A. Number of Training Samples:

1 <<< pd: not sure whether we need this appendix? >>>

DensRay is essentially a supervised learning method. It is difficult to extract useful features in the case of insufficient or unbalanced labels. As shown in Figure 3, the debiasing results improve with an increased number of training samples.

As a projection-based debiasing method, the premise of DensRay debiasing is that the gender direction should be correct. Unbalanced samples will lead to incorrect gender direction biased towards either male or female, resulting in reversing the gender bias during debiasing. The figure also shows that a balanced training set improves debiasing performance.

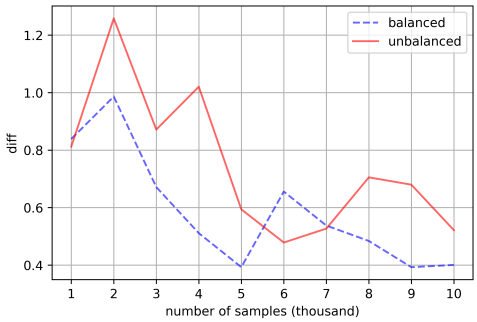


Figure 3: DensRay debiasing results on OCCTMP with different number of samples and unbalanced/balanced data.