Equalize Don't Minimize: Robust and Efficient Mitigation of Social Bias in Pre-trained Language Models

James Hou
The Bishop's School
houjames05@gmail.com

Marcus Jaiclin
The Bishop's School
marcus.jaiclin@bishops.com

Abstract

In recent years, paired with the rapid success of contextual word embeddings or pre-trained language models, have been growing concerns about the presence of bias within them. Past studies have found that language models like BERT not only inherit but also amplify social biases implicitly embedded in its training corpora. Research has moved to combat these biases, working towards fairer applications and a more equitable society. Past methods have taken various approaches towards a solution but have mostly encountered the same roadblock: aggressive debiasing leads to catastrophic forgetting, while model preserving methods fail to make significant bias removal. In this work, we propose a novel debiasing technique with 1) an *Equivalence Loss* that more naturally encapsulates fairness as equal affinity rather than a minimization of the protected-attribute space done by past methods, and 2) the application of the Adapter. Evaluated on the StereoSet bias metric, our method achieves state-of-the-art bias mitigation, removing 69.0% of the total bias while maintaining robust language modeling ability, breaking the aforementioned barrier. Moreover, our method proves highly computation- and data-efficient, making it a prime candidate to push widespread language model debiasing efforts.

1 Introduction

In the last decade, the field of Natural Language Processing (NLP) has witnessed significant progress thanks to the rapid developments of the word embedding. In general, word embeddings are learned representations of human text, which capture the semantic similarity between words in the medium of a vector space. Word embeddings such as Mikolov et al. (2013a,b) and Pennington et al. (2014) are trained on large corpora to obtain robust representations of the vocabulary. Words that frequently appear next to each other are represented with vectors in close proximity—a common measure is the cosine similarity between vectors. Moreover, analogous relations such as "man is to king as woman is to queen" are preserved in taking the differences between the vector representations. A bridge between human text and numerical vectors interpretable by computers, they prove incredibly effective serving as the backbone to many NLP tasks like text classification, and Q&A. As a result, they have been increasingly applied in real-world scenarios including resume screening, legal work, clinical notes processing and loan approval (Hansen et al., 2015; Dale, 2019; Alsentzer et al., 2019; Stevenson et al., 2021).

Word embeddings have opened the way for many more conveniences in using AI in text-relevant social applications. However, studies have shown that these word embeddings often inherit and even amplify social, gender and racial prejudices present within the data they learn from. Since word embeddings aim to capture the distribution of words present within a corpora, biased commentary and underrepresentation of minority groups within the source text can lead to skewed representations of marginalized groups (Caliskan et al., 2017; Zhao et al., 2017). Notably, Bolukbasi et al. (2016) demonstrated that the word *man* had the same relation to *computer programmer* as *woman* had to *homemaker*, despite the fact that neither occupation is gender specific. In response, studies have been conducted to debias these word embeddings such as that of Bolukbasi et al. (2016) which projects the protected attribute (race, gender, etc.) space away from the neutral attributes such as occupation, pleasantness, competency, etc.

In recent years, a more powerful class of word representations was introduced: contextual word embeddings or pre-trained language models. Unlike traditional word embeddings that always output the same vector for a word no matter the context, contextually-aware word embeddings such as, ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) adapt the representation of a word with respect to the surrounding words within the sentence. In addition to producing word representations, they also output sentence embeddings. These property has enabled them to outperform the state-of-the-art of many language tasks and become a staple in real-world applications. However, they are not exempt from the racial biases present in traditional word embeddings. Recent research has demonstrated that there similarly exists a significant amount of bias within these models (Kurita et al., 2019; May et al., 2019; Tan and Celis, 2019; Zhao et al., 2019).

A few studies have proposed techniques for debiasing language models like BERT but the task remains difficult. There have been works applying the projection techniques from traditional word embeddings to post-process the generated contextual word embeddings (Liang et al., 2020; Ravfogel et al., 2020). However, these techniques designed for static word embeddings aren't well-suited for the dynamic output of contextual word embeddings. Others have proposed Counterfactual Data Augmentation (Zmigrod et al., 2019; Lauscher et al., 2021), which extends the pre-training process for the model to learn from more diverse examples, but being without a debiasing objective, the method is neither comprehensive nor resource efficient. Loss functions penalizing bias have also been introduced but works such as Kaneko and Bollegala (2021) often result in catastrophic forgetting that damage the language model's performance in the process of debiasing.

In this work, our main focus is to address all three challenges of efficiency, effectiveness, and model preservation faced by past techniques aimed toward debiasing contextual word embeddings. Our main contributions are:

1. A novel *Equivalence Loss* function that applies a more natural interpretation and measurement of bias—measuring the difference in the association of a neutral attribute to two distinct races—to maximize the effectiveness of bias mitigation and model preservation during a debiasing fine-tuning process.

2. An innovative application of the Adapter (Pfeiffer et al., 2020) to our loss-based debiasing process that protects the language model from forgetting and reduces computational complexity.

3. We test and benchmark our debiasing technique against others with the StereoSet Benchmark (Nadeem et al., 2021) and obtain state-of-the-art racial bias removal while maintaining comparable language modeling capabilities.

2 Related Works

In this section, we will outline the various debiasing methods proposed by past studies and evaluate their strengths and shortcomings.

2.1 Post-hoc Techniques

In an early work on bias in traditional word embeddings, Bolukbasi et al. (2016) proposed to debias with a projection technique. The approach is a post-hoc method where the procured embeddings are put through post-processing to minimize bias. To do this, they iterate over protected-attribute sensitive (e.g. gender-, race-related words) embeddings and apply Principal Component Analysis (PCA) to identify the protected-attribute space. Following, they project the protected-attribute space away from embeddings as a post-processing process, which they show to remove bias.

As a new generation of language models—contextual word embeddings—were invented and found to hold similar biases, studies have sought to adapt the debiasing techniques of traditional word embeddings for newer models. This includes Liang et al. (2020), which directly applies Bolukbasi et al. (2016)'s approach and uses template sentences to estimate the protected-attribute space. Though this method demonstrates some debiasing ability there is more to be desired.

This post-hoc approach has been called into question by Gonen and Goldberg (2019)'s study, which demonstrates that it only "superficially" removes bias, while actually leaving most of the biased associations untouched in the embedding space. To combat this, Ravfogel et al. (2020) proposed the Iterative Nullspace Projection, attempting to predict a comprehensive protected-attribute space for better debiasing projection.

However, tested on metrics (Meade et al., 2022), the technique still does not produce a desired debiasing. We believe this to be the nature

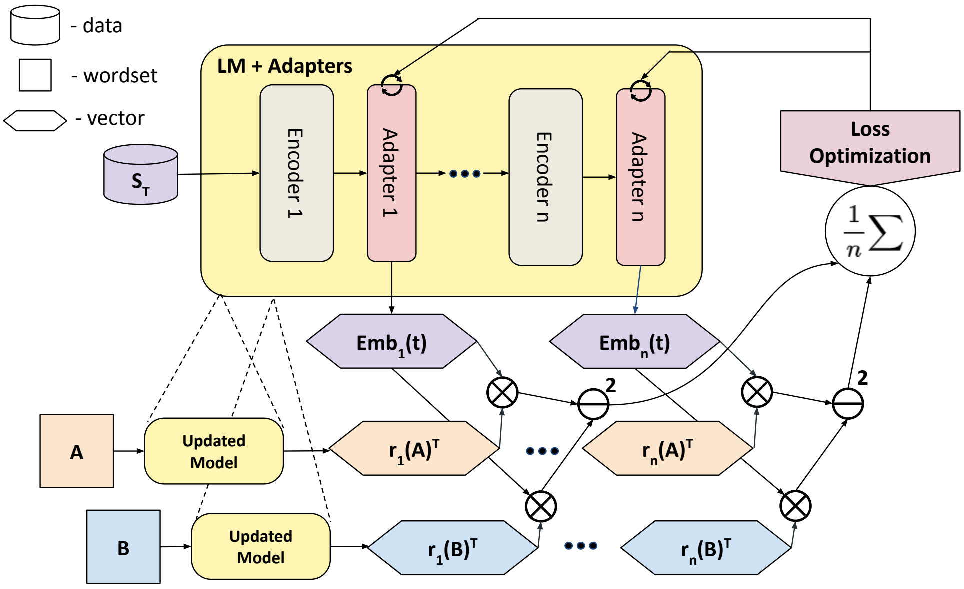

Figure 1: This is an overall schematic of our proposed debiasing process. The original language model (LM) is modified with injected Adapter layers. During training, model outputs are used to compute the *Equivalence Loss*, which motivates debiasing optimization. As defined in Section 3.1, A and B are protected attribute wordsets and T is a collection of neutral attributes. Emb_i, r_i represents the embedding obtained at layer i.

of post-hoc methods, which are not well suited for the dynamic and context-dependent word embeddings. Therefore, for our method, we looked to create a debiasing framework that naturally integrates into current language models.

2.2 Counterfactual Data Augmentation

Data augmentation has long been a practiced technique in machine learning to source additional training samples without collecting more data. As the name suggests, data samples are altered in a substantial but character-preserving way. Recently, Zmigrod et al. (2019) introduced the Counterfactual Data Augmentation (CDA) method that builds upon this principle to debias language models.

As many have identified, the bias problem is largely caused by an inequality of representation of protected groups in the training corpus, with some appearing frequently next to positive words and vice versa for others. Thus, CDA attempts to balance the representation of each protected group in the dataset, creating duplicate training samples but with race- or gender-sensitive words swapped (e.g. *he is strong* ⇒ *she is strong*).

Following the data augmentation process, the dataset is fed through the same training process as the original word embedding. This process is shown to maintain strong language modeling ability but still lacks effective debiasing without an explicit debiasing motivator (Meade et al., 2022). Moreover, it is costly in data and computation.

2.3 Loss-Based Debiasing

In contrast with CDA, there have also been studies proposing to alter the training process for bias removal. This notably includes methods using a loss function to discourage bias.

Kaneko and Bollegala (2021) takes a similar approach to the post-hoc techniques, opting to completely disassociate the vocabulary relating to protected groups (gender, race, religion, etc.) from vocabulary vulnerable to biases (competency, likeability, occupation). They accomplish this by estimating the general embeddings of both vocabularies and taking their dot product as the motivator to the loss function. When the loss is minimized, the two vocabularies are orthogonal, or in other words unrelated.

Though this method exhibits strong debiasing ability, it proves highly destructive to language modeling and unstable, which made the proposed method unfeasible. The method proposed to add an L2 regularization term to the loss to prevent drastic shifts in model weights, but it does not reflect well in the results discussed in Section 4.

3 Method

In this section, we detail our proposed approach to removing bias, specifically racial bias, from contextual language models that 1) conducts comprehensive and effective debiasing as a natural extension of the original model, 2) maintains comparable levels of language modeling ability as before debiasing, 3) efficiently uses data and computational resources. These objectives are achieved by a novel loss that serves as a more intuitive interpretation of algorithmic bias and an application of Pfeiffer et al. (2020)'s Adapter.

3.1 Equivalence Loss

Motivation Since we would like to remove racial bias from contextual language models while preserving its language modeling ability from pretraining, we take the common approach of fine-tuning the base model and define the debiasing process as a downstream task. The method CDA as proposed by Zmigrod et al. (2019) leverages the fine-tuning approach and emulates the original masked language modeling objective used in training BERT to show the model data that equally represent all protected groups. However, we find that this approach remains very expensive computationally and does not perform well due to an unguided approach. More explicit approaches such as Kaneko and Bollegala (2021); Liang et al. (2020); Ravfogel et al. (2020) have put an emphasis on minimizing the bias space's overlap with the attribute space. Specifically, Kaneko and Bollegala (2021) defines a bias loss which drives the bias space to be completely orthogonal to the attribute space. However, we believe this penalization of bias is not accurate to true bias and potentially destructive to the model's understanding of language. For example, although the phrase "The black man is violent" is prone to bias, we should not completely annihilate the relationship between "black" and "violent," as the "black" man is the subject of the adjective. This effect is reflected in Meade et al. (2022) results.

In this study, we propose a new bias loss we call the *Equivalence Loss* to better interpret and remove bias and preserve model information. To achieve this, we return to the Implicit Association Test (IAT) defined by Greenwald et al. (1998) for motivation. His test, which is used to construct many word embedding bias measurements such as Word Embedding Association Test (WEAT) (Caliskan et al., 2017) and StereoSet (Nadeem et al., 2021), defines bias as the difference in the subject's association between an attribute (happiness) and two different protected groups (e.g. Black vs. White). We follow this definition of bias and propose the *Equivalence Loss* that penalizes these differences in reaction and rewards a word embedding's equal affinity to different races. We believe that this is a more intuitive objective for debiasing versus trying to completely divide racial and attribute information.

Loss Definition We compile three word sets A, B, T. A and B are each a set of words indicative of a certain protected group. Since we only investigate racial biases in this study, our word set A are words relating to people of African descent while B collectively represents people of European descent. T is the collection of attributes unrelated to race such as competency, occupation, and pleasantness. Let us set the training dataset as D. For each word set W we create sub-datasets S_w for word $w \in W$, which are all sentences in D that contain w. Therefore we create $S_a \forall a \in A$, $S_b \forall b \in B$, $S_t \forall t \in T$. We expand upon the specifics of these word and data sets in Section 4.2.

Overall, we define our *Equivalence Loss* to be Equation 1 and 2

$$\begin{aligned} L_{bias} =& \frac{1}{n}\sum_{n}^{i}\sum_{T}^{t}\sum_{S_t}^{x}(\vec{r_i}(A)^\top Emb_i(t,x) \\ &- \vec{r_i}(B)^\top Emb_i(t,x))^2 \end{aligned} \tag{1}$$

$$\vec{r_i}(W) = \frac{1}{|D \cap W|}\sum_{W}^{w}\sum_{S_w}^{x} Emb_i(w,x) \tag{2}$$

where $Emb_i(t,x)$ gets the embedding at layer i specific to the token t in a sentence x, n is the number of encoder/decoder modules, and $|D \cap W|$ is the number of examples in D that contain a word from word set W.

Seen from the multi-layer sum, we follow Lauscher et al. (2021)'s proposal to debias throughout all hidden states present in the model to fully excavate its biases. Since contextual word embeddings can procure multiple different word representations for the same word under different contexts, we implement a $\vec{r_i}$ function that estimates a general word representation at layer i of each protected group (racial bias - Black vs. White). We iterate through all sentences containing words relevant to the protected group and extract the token-level representation of the relevant word. These varied representations of the protected group are then coalesced as an average to form the general word representation.

In L_{bias}, we subtract the dot product between the representation of B and the embeddings of T from the dot product between the representation of A and the embeddings of T. This loss term is minimized when the embedding of T is equidistant to the representation of A and B. Visualized in Figure 2, the vector space evolves to satisfy the loss term, creating more balanced embeddings. In the context of racial bias, race A and race B have an equal affinity to attribute T in the model's understanding of language, meaning the model treats both races fairly in relation to attributes such as competency, occupation, happiness, etc.

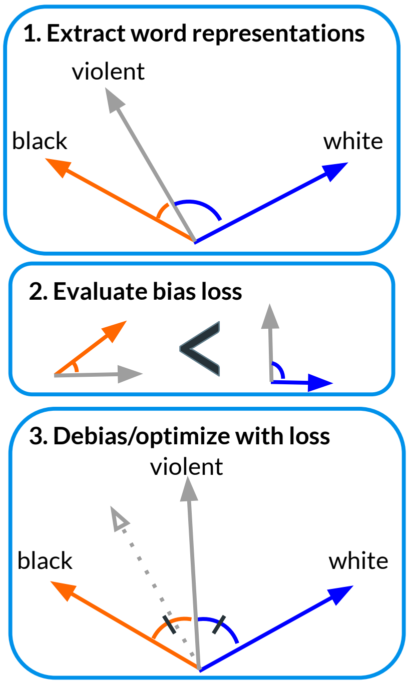

Figure 2: This is a conceptual visualization of our proposed *Equivalence Loss*. Since we define, bias to be disproportionate association to one protected-group, the loss compares the cosine distance between the vectors. The loss is then used to push the model towards the vector states in Step 3.

3.2 Model Preservation

To further hedge against catastrophic forgetting during our fine-tuning process, we adopt Pfeiffer et al. (2020)'s Adapter modules.

Originally proposed to preserve model information during downstream task fine-tuning, the Adapter functions by appending additional feed-forward networks to each encoder or decoder module. The injected layers are fine-tuned while the original weights are frozen. Due to the comparatively small amount of weights present within the Adapter layers, its addition significantly increases training efficiency. These properties make the Adapter very lucrative for the debiasing process. Lauscher et al. (2021) proposed to use the Adapter in combination with the debiasing technique CDA. However, we believe the Adapter's capabilities can be better utilized with a more aggressive debiasing process using our proposed loss function.

In our work, we employ the Bottleneck Adapter variation of Pfeiffer et al. (2020)'s Adapter. In each Bottleneck Adapter, there are two feed-forward layers: one a down-projection and another an up-projection. In our study, we configure the Bottleneck Adapters to have a parameter reduction factor of 12. This provides a further improvement in debiasing efficiency and has been shown to maintain comparable performance in fine-tuning. An additional benefit of the Adapter is its "plug-and-play" capability where once a set of debiasing Adapters have been trained, it can be directly injected into real-world contextual word embedding applications to remove pre-packaged bias, making debiasing accessible and widespread.

4 Results

In this section, we define our experimental setup and the metric we employ. We report our findings on not only debiasing efficacy but also other important criteria of model preservation and data and computation economy.

4.1 Metric

For our study, we employ the StereoSet Metric (Nadeem et al., 2021) to assess the performance of our debiasing technique against others. The StereoSet metric evaluates a debiased language in two ways: 1) contained stereotypical bias and 2) language modeling ability. As with many other metrics, its chief purpose is to evaluate the bias of the model but its ability to quantify the language modeling ability of an embedding model serves as a strong gauge.

At its core, StereoSet is a crowdsourced dataset. Each example contains a context sentence with a masked token, for which there exists three fillin-the-blank options: a stereotypically biased response, an anti-stereotypical response, and an unrelated one. As a real example, for the context sentence "People of African descent are [MASK]," a group of labelers collectively rated "savages" as the stereotype-conforming response, "cultured" as the anti-stereotype, and "red" as unrelated.

This dataset is then used to calculate Language Modeling Score (LMS) and Stereotype Score (SS), which measure language modeling ability and bias respectively. The Stereotype Score is obtained by comparing the predicted likelihood of the stereotypical response vs. the anti-stereotypical response; this measure signifies no bias at 50, growing anti-stereotypical bias $50 \rightarrow 0$ and growing stereotypical bias $50 \rightarrow 100$. These aforementioned probabilities are then compared to that of the unrelated response to measure the Language Modeling Score which lies between 0-100 with 100 as perfect and 0 as poor.

We note that we only use the intra-sentence case of StereoSet as past studies have. For our racial bias mitigation experiments, we use the Race category of the dataset, which contains 3,996 context sentences.

4.2 Training and Data

As we outlined in Section 3.1, our debiasing process follows a direct fine-tuning process. For our target model, we debias one of the most widely used contextual language models, the Bidirectional Encoder Representations from Transformers (BERT), introduced by Devlin et al. (2018).

Following the same naming convention, we source our wordsets A, B, and T from external datasets. For our experiments, A contains words relevant to or colloquially referencing people of African descent and those of European descent for B (e.g. *Black, Anglo-Saxon, African American*). As for T, we source our bias-vulnerable attributes from the Greenwald et al. (1998)'s seminal study on implicit biases, Implicit Association Test. The original test is composed of words from various attribute categories such as likeability, competency, and more. In addition to those words, we also develop our own occupation bias wordset, where the racial prejudices on low-income and high-income jobs are studied.

We note that the IAT is also used in WEAT to quantify bias in traditional word embeddings, making it a recognized pointer to bias. Moreover, since our benchmark is completely different from the datasets of WEAT and IAT, our method construction is independent of our method evaluation; this is so that we do not optimize for a specific metric, so we expect the results of our benchmark to be reflective of a general bias removal.

Although past techniques like Liang et al. (2020) have utilized sentence templates to find approximations of bias, we believe that natural context is critical to excavating true bias that would occur in real-world deployment. So we utilize *News-commentary-v15* (Tiedemann, 2012) corpus as our D dataset, which Kaneko and Bollegala (2021) also employed. It is a dataset of news commentary, which we believe to be prone to implicit racial bias, leading to a comprehensive debiasing. As previously described, we extract the S_a, S_b, and S_t sentences from sentences of the *News-commentary-v15* corpus that contain words from their respective word sets.

During our bias removal fine-tuning process, we freeze all weights outside of the Adapters, to prevent catastrophic forgetting and increase efficiency. We set our *Equivalence Loss* as the sole target function the model optimizes for using the Adam optimizer. The training process iterates over D for 3 epochs and the learning rate is set at 1e-5, which we obtain from experimentation detailed in Section 4.6.

Debias Method	SS	Δ SS	LMS
Vanilla BERT	57.45	-	84.65
Zmigrod et al. (2019)	56.26	-1.19	84.49*
Liang et al. (2020)	57.76	0.31	83.95
Ravfogel et al. (2020)	57.29	-0.16	80.63
Webster et al. (2020)	57.07	-0.22	83.04
Kaneko and Bollegala (2021)	52.85	-4.59	70.43
Our Method - DCED	52.30*	-5.14*	82.73

Table 1: This table contains the StereoSet metric results (SS-StereoSet Score, LMS - Language Modeling Score) of various language model bias removal techniques. We compare them to the results of our proposed framework in the last row.

4.3 Debiasing Effectiveness

As described in Section 4.1, we evaluate our proposed debiasing technique with the StereoSet metric, which gives a StereoSet Score (SS) indicating bias levels and Language Modeling Score (LMS) measuring the effectiveness of the output embeddings.

In Table 1 we benchmark our results against that of the other techniques. The Vanilla BERT entry is a baseline we set for the original levels of bias and

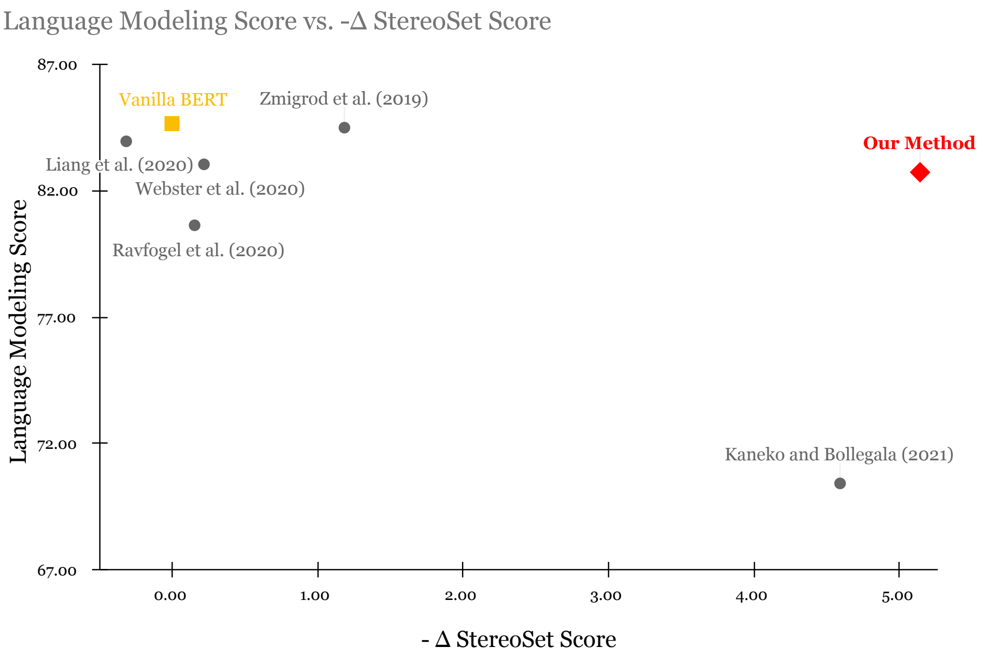

Figure 3: This figure shows a plot of LMS vs. - ΔSS (i.e. total reduction of bias) as a comparison of the trade-offs between bias removal effectiveness and language modeling preservation in various methods.

language modeling ability. From our experimentation, we find the SS to be 57.45 and the LMS 84.65, reflecting stereotypical bias and proficient language modeling ability. This baseline allows us to determine the severity of catastrophic forgetting and the effectiveness of bias removal.

Based on the ΔSS, we report that many of the methods proposed by past techniques show marginal improvement in reducing bias. After debiasing BERT with our proposed method, the metric yields 52.30 for SS, showing a significant improvement over past techniques. As a percent change in total bias, we remove 69.0% of the existing bias from the model. The only method that rivaled our debiasing effectiveness was another loss-based method by Kaneko and Bollegala (2021), which we will further analyze in the subsequent subsection.

4.4 Model Preservation

Though we observe comparable bias mitigation between Kaneko and Bollegala (2021)'s method and our proposed framework, our method demonstrates much more robust model information preservation. Referencing Table 1, we follow Kaneko's procedure and obtain an LMS of 70.43 for their method, which lies considerably lower than the unaltered BERT. Despite also being a loss-based method, our technique at 82.73 does not suffer the same drop-off in LMS, likely stemming from our more intuitive *Equivalence Loss* function and application of the Adapter.

Looking at Figure 3, we see this stark contrast reflected on the right side of the plot, where there is more bias removed. Directing our attention to the other methods, we see that they also lie clustered at the top left corner of the graph, demonstrating little change in ΔSS while mostly preserving the LMS. In other words, although these methods were able to maintain most of the language modeling ability they provided only minimal debiasing.

We see a game of trade-offs between debiasing effectiveness and model preservation. When the aforementioned Kaneko and Bollegala (2021) took aggressive measures of debiasing, they made large modifications to the language model, leading to poor model preservation. On the opposite side, some methods were too passive or ineffective, leaving the model mostly the same. Our technique breaks this constraint and can obtain the unique intersection of powerful debiasing and robust model information retention, placing it in its own corner in the top right.

4.5 Data and Computation Resources

In our study, we place a strong emphasis on efficiency in addition to debiasing efficacy and model preservation. Since language models like BERT are "foundation" models, they serve as the base of many downstream applications created by various independent developers. Therefore, for widespread debiasing to occur, there need to be low barriers in computational and data resources. For this study, we fine-tune BERT with a single GPU: Nvidia RTX 3090 Ti. In total, our method iterates over 150k sentences of text data and the total debiasing process takes 15 minutes. Our aggressive loss-based approach circumvents the heavy data and training requirements of methods like CDA that repeat the original training process, which potentially takes days with large GPU clusters.

4.6 Learning Rate Ablations

We experiment with the bias removal process' learning rate. As with traditional machine learning practice, a higher learning rate corresponds with a more aggressive pursuit of our loss function. We find that 1e-5 obtains the desired balance of debiasing effect and language modeling ability.

Learning Rate	SS	LMS
5e-6	55.10	80.22
1e-5	52.30	82.73
5e-5	51.61	81.56
1e-4	56.20	81.21

Table 2: This table contains the results of the different learning rates that were experimented with. SS - StereoSet Score, LMS - Language Modeling Score

5 Conclusions and Discussion

In this work, we present a novel contextual word embedding bias mitigation technique. Observing the shortcomings of past methods, our proposed methodology aims to achieve all three critical aspects of effective debiasing, language modeling preservation, and data-/computation- efficiency.

To achieve this, we proposed a novel *Equivalence Loss* that takes a more intuitive mathematical interpretation of bias and equality, as opposed to former methods attempting to completely divide the gender or race subspace from the attribute subspace. Moreover, we apply the Adapter, used for lightweight non-destructive, downstream fine-tuning. All in all, our method is designed to be a very natural integration into the original language model.

Tested on the StereoSet benchmark, our innovations achieve comprehensive debiasing without significant sacrifices in language modeling ability while using low levels of computing resources. The satisfaction of all three criteria makes our method an ideal candidate to remove bias from contextual word embeddings.

Though we only experiment with racial bias removal on the BERT model in this study, our method is versatile and easily adapted for other language models as well as different types of biases. As Zhang et al. (2022) reports, new language models are larger and larger—in turn more powerful—but are accumulating more bias. Our low-cost but highly effective solution makes debiasing not only widely accessible but also scalable to the most advanced models containing hundreds of billions of parameters.

6 Including acknowledgments

Acknowledgments appear immediately before the references. Do not number this section.[1] If you found the reviewers' or Action Editor's comments helpful, consider acknowledging them.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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. *Advances in neural information processing systems*, 29.

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

Robert Dale. 2019. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217.

[1] In LaTeX, one can use \section* instead of \section.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

C Hansen, M Tosik, G Goossen, C Li, L Bayeva, F Berbain, and M Rotaru. 2015. How to get the best word vectors for resume parsing. In *SNN Adaptive Intelligence/Symposium: Machine Learning*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *EACL*, pages 1256–1266.

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. Association for Computational Linguistics.

Anne Lauscher, Tobias Lüken, and Goran Glavas. 2021. Sustainable modular debiasing of language models. *ArXiv*, abs/2109.03646.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Long and Short Papers*, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, pages 622–628. Association for Computational Linguistics (ACL). Funding Information: CM is funded by IARPA MATERIAL; RR is funded by DARPA AIDA; AW is funded by an NSF fellowship. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of IARPA, DARPA, NSF, or the U.S. Government. Publisher Copyright: © 2019 Association for Computational Linguistics; 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019 ; Conference date: 02-06-2019 Through 07-06-2019.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Pa-*

pers), pages 5356–5371, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Matthew Stevenson, Christophe Mues, and Cristián Bravo. 2021. The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, 295(2):758–771.

Yi Chern Tan and L. Elisa Celis. 2019. *Assessing Social and Intersectional Biases in Contextualized Word Representations*. Curran Associates Inc., Red Hook, NY, USA.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pretrained models. Technical report.

Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. 2022. The ai index 2022 annual report.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.