Debiasing Masks: A New Framework for Shortcut Mitigation in NLU

Johannes Mario Meissner[†], **Saku Sugawara**[‡], **Akiko Aizawa**[†‡]
[†]The University of Tokyo, [‡]National Institute of Informatics
{meissner,saku,aizawa}@nii.ac.jp

Abstract

Debiasing language models from unwanted behaviors in Natural Language Understanding tasks is a topic with rapidly increasing interest in the NLP community. Spurious statistical correlations in the data allow models to perform shortcuts and avoid uncovering more advanced and desirable linguistic features. A multitude of effective debiasing approaches has been proposed, but flexibility remains a major issue. For the most part, models must be retrained to find a new set of weights with debiased behavior. We propose a new debiasing method in which we identify debiased pruning masks that can be applied to a finetuned model. This enables the selective and conditional application of debiasing behaviors. We assume that bias is caused by a certain subset of weights in the network; our method is, in essence, a mask search to identify and remove biased weights. Our masks show equivalent or superior performance to the standard counterparts, while offering important benefits. Pruning masks can be stored with high efficiency in memory, and it becomes possible to switch among several debiasing behaviors (or revert back to the original biased model) at inference time. Finally, it opens the doors to further research on how biases are acquired by studying the generated masks. For example, we observed that the early layers and attention heads were pruned more aggressively, possibly hinting towards the location in which biases may be encoded.

1 Introduction

The issue of spurious correlations in natural language understanding datasets has been extensively studied in recent years (Gururangan et al., 2018; McCoy et al., 2019; Gardner et al., 2021). In the MNLI dataset (Williams et al., 2018), negation words such as "not" are unintended hints for the *contradiction* label (Gururangan et al., 2018), while a high word overlap between the premise and the hypothesis often correlates with *entailment* (McCoy et al., 2019).

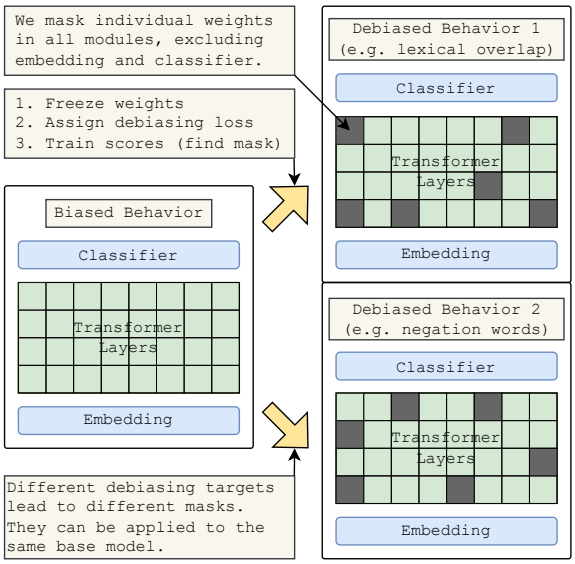

Figure 1: We find masks that remove bias from a finetuned model by using frozen-weight movement pruning.

A common way to prevent the acquisition of biases (in this paper defined as unintended shortcut behaviors) is to adopt a debiasing loss function, such as product-of-experts (Hinton, 2002), to discourage learning shortcuts through the use of the annotated level of bias in each sample. Because manual bias annotation is difficult and expensive, the predictions of a biased model are used as bias annotations (Clark et al., 2019). Commonly, models must be retrained to achieve debiased behavior; this limits our ability to target different biases separately, as well as choose varied levels of debiasing strength (which impacts the in- vs. out-of-distribution trade-off).

We propose a new debiasing framework that focuses on removing bias from an existing model, instead of the now-common approach of re-training from scratch. Our approach is depicted in Figure 1. We find a pruning mask that zeroes out the weights that cause biased outputs, producing the desired effect without altering the original model. This

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7607 - 7613
December 7-11, 2022

approach offers several clear advantages. First, pruning masks can be stored very efficiently, only occupying a fraction of the original model size. This enables creating multiple masks for varied debiasing behaviors. Secondly, masks can be effortlessly set or unset at inference time, as opposed to replacing the entire model. This allows for flexible application, as well as easy reversion to the original model when needed. Finally, re-framing the debiasing process as a mask search opens the doors towards future analysis directions, helping to more deeply understand how biases are learned, and how they can be eliminated.

2 Related Work

2.1 Shortcuts in NLU

The study of shortcut solutions in machine learning (Geirhos et al., 2020; D'Amour et al., 2020) has gained attention in recent years, including in the field of Natural Language Understanding. In SQuAD (Rajpurkar et al., 2016), Jia and Liang (2017) show that distractor sentences can be inserted in such a way that the frequently used spurious features mislead the model's answer. In MNLI, Gardner et al. (2021) discuss the idea that all simple feature-label correlations should be regarded as spurious. Their results go in line with Gururangan et al. (2018), who show that models are able to achieve strong performance just by training on the hypothesis, a scenario where the desirable advanced features are not available. Instead, simple word correlations are used to make the prediction. Finally, other kinds of features such as lexical overlap have been pointed out as important shortcuts (McCoy et al., 2019).

2.2 Debiasing

A wide range of debiasing approaches have been proposed to alleviate shortcut behavior. For example, perturbations in the model's embedding space can encourage robustness to shortcuts (Liu et al., 2020), and training on adversarial data (Wang and Bansal, 2018) has important benefits for generalization capabilities too. Removing a certain subset in the training data (filtering) can help to avoid learning spurious correlations. Bras et al. (2020) devise an algorithm that selects samples to filter out, reducing the training time and robustness of the target model.

We will focus our efforts on a family of approaches that rely on a debiasing loss function to discourage biased learning, instead of altering the original training data. A wide range of debiasing losses have been proposed to discourage shortcut learning. They all rely on having a measure of the level of bias for each training sample, which is commonly obtained via a biased model's predictions. Among the most common are product-of-experts and focal loss (Schuster et al., 2019). Other approaches introduce a higher level of complexity, such as confidence regularization (Utama et al., 2020a) requiring a teacher model, or learned-mixin (Clark et al., 2019) introducing an additional noise parameter.

2.3 Pruning

Pruning consists in masking out or completely removing weights in a neural network, such that they no longer contribute to the output. Common goals include reducing the model size or achieving an inference speed-up.

Magnitude pruning is a basic pruning method that removes weights based on their absolute value. It has proven to be an effective method to reduce model size without compromising on performance. Gordon et al. (2020) apply this technique on transformer language models. Movement pruning, on the other hand, was proposed by Sanh et al. (2020), and involves training a score value alongside each weight. Scores are updated as part of the optimization process, and weights with an associated score below the threshold are masked out.

Zhao et al. (2020) introduce an alternative to the usual finetuning process by finding a mask for the pretrained base model such that performance on the target task increases. The same base model can be used with several masks to perform multiple tasks. They show that this approach achieves similar performance to standard finetuning on the GLUE tasks (Wang et al., 2018).

Our work is inspired by Zhao et al. (2020); but we focus on removing biases from an already fine-tuned model. We will refer to this method as a mask search. It benefits from the same advantages: masks can be easily applied and removed from a shared base model, while additionally offering storage improvements.

3 Masked Debiasing

Our proposed approach to debiasing takes a unique point of view in the debiasing field by assuming that biased behavior is encoded in specific weights

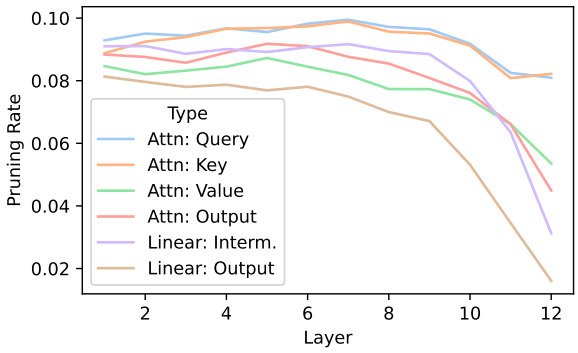

Figure 2: Pruning rate for each module in one of our debiasing masks.

of the network and can be removed without altering the remaining weights. Thus, we perform a mask search to identify and remove those weights, experiencing debiased behavior in the resulting model. Our approach comes together by combining a debiasing loss, a biased model, and a score-based pruning technique.

Debiasing Loss Among the debiasing losses mentioned in Section 2.2, none appear to be clearly superior, each offering certain strengths and weaknesses, and striking a balance between in- and out-of-distribution performance. We err on the side of simplicity and run our experiments with the focal loss. We follow Clark et al. (2019) for the implementation.

Bias Model To utilize a debiasing loss, we must obtain the predictions of a biased model on the training samples. We utilize weak learners, following the nomenclature introduced by Sanh et al. (2021); they do not require making any assumptions on the underlying biases. While our method works with any debiasing loss and bias level source, we choose this setting due to its flexibility and adaptability to new scenarios.

To date, two major learners have been proposed: undertrained (Utama et al., 2020b) and underparameterized (Sanh et al., 2021) learners. We train and extract the predictions of both of them for comparison. Undertrained learners are trained by selecting a very small subset of the training data, but keeping the target model's architecture and hyperparameters. This effectively translates into a model that overfits on the selected data subset. Features uncovered by the model in this manner are deemed spurious, as it is assumed that more advanced features require exploring larger quantities of data. We used 2000 samples for MNLI, and 500 samples for the other datasets. Underparameterized learners, on the other hand, restrict the model complexity by reducing layer size and count. The expectation is that this underparameterization is restrictive and limits the ability to find complex features. We use BERT-Tiny (Turc et al., 2019), a BERT model with 2 layers and an inner size of 128.

Frozen Movement Pruning Our approach is similar to Zhao et al. (2020); but we implement it by leveraging the framework provided by Sanh et al. (2020), with the addition of weight-freezing. We use unstructured pruning, which means that each individual weight is considered for exclusion. Further, we utilize a threshold approach: in each weight tensor we remove (mask out) those weights with an associated score that is lower than the threshold. A regularization term for the scores helps avoid all scores growing larger than the threshold. For further details, we refer the reader to Sanh et al. (2020). As the starting point, we load a model already finetuned on the target task.

4 Experimental Setup

We compare our approach against the standard approach of re-training the model with a debiasing loss. Our experiments are carried out with BERT-Base (Devlin et al., 2019).

4.1 Tasks and Evaluation Scenarios

We perform experiments on three tasks in NLU, and consider several biases and evaluation datasets.

Natural Language Inference (NLI) This task consists in classifying whether the relationship between a pair of sentences (premise and hypothesis) is entailed, neutral or contradicted. We evaluate on two well-known biases: lexical overlap bias (McCoy et al., 2019) and negation words bias (Gururangan et al., 2018). We train on MNLI, and evaluate on HANS (McCoy et al., 2019) and our own negation-words subset. HANS can be used to evaluate the lexical overlap bias. Additionally, we create our own negation-words anti-bias set by selecting *entailed* samples from the MNLI validation set that contain negation words in the hypothesis.[1]

[1]We select entailed samples with at least one of the following words in the hypothesis: no, not, don't, none, nothing, never, aren't, isn't, weren't, neither, don't, didn't, doesn't, cannot, hasn't won't.

	MNLI			QQP			FEVER	
	Val.	HANS	Neg.	Val.	PAWS	PAWS (¬)	Val.	Symm.
Baseline	**84.38**	61.97	78.47	**92.07**	44.28	23.37	85.67	64.08
Masking w/o Debiasing	83.15	64.16	77.87	90.91	37.61	16.38	84.71	63.26
Undertrained SD	81.02	67.06	74.82	–	–	–	**86.54**	65.06
Underparam. SD	83.03	67.75	72.36	88.79	43.50	27.16	85.70	**65.79**
Undertrained MD (Ours)	81.85	**68.69**	75.58	–	–	–	84.55	64.85
Underparam. MD (Ours)	82.24	67.86	73.95	89.61	**44.34**	**28.56**	84.96	63.37
HypOnly MD (Ours)	82.87	64.26	**79.03**	–	–	–	–	–

Table 1: Our experimental results comparing a BERT-Base baseline against several debiasing methods, along with our own mask debiasing approach. SD stands for standard debiasing, while MD stands for masked debiasing. MNLI Neg. is our split of the MNLI development set containing at least one negation word in the hypothesis. PAWS (¬) indicates the subset of PAWS with a negative label (the anti-bias set for lexical overlap in QQP).

Paraphrase Identification In this task the goal is to identify whether a pair of sentences are paraphrasing each other or not. We train on QQP (Quora Question Pairs)[2], and evaluate on PAWS (Zhang et al., 2019), which tests the model's reliability on the lexical overlap bias.

Fact Verification We train on FEVER (Thorne et al., 2018), and evaluate on FEVER Symmetric (Schuster et al., 2019). The task setup is similar to NLI; we classify a claim-evidence pair as either support, refutes or not-enough-information. FEVER Symmetric eliminates claim-only biases (clues in the claim that allow to guess the label), among which are negation words, in a similar fashion to MNLI.

4.2 Hyperparameters and Reproducibility

We aim for complete reproducibility by providing complete code and clear reproduction instructions in our repository.[3] In most cases, we follow the configurations indicated by the respective original papers. Appendix B is additionally provided for a report of important hyperparameters.

We run all of our experiments on five seeds. We not only seed the debiasing masks, but also the accompanying weak models and base models too.

5 Results

We compare our approach against our reproduction of the two weak model approaches, and compile our results in Table 1.

[2] https://www.kaggle.com/c/quora-question-pairs

[3] https://github.com/mariomeissner/shortcut-pruning

5.1 Masked Debiasing is Effective

First, we observe that our debiasing masks are very effective, surpassing the performance of their standard debiasing counterparts in two out of the three evaluated tasks, while keeping a competitive in-distribution performance. In the FEVER task, our masks provided a slight debiasing effect, but did not beat the standard debiasing baseline. We do not report undertrained debiasing results in the QQP setting due to their failure to converge to adequate results (in both standard and masked debiasing).

To confirm that the combination of masking with a debiasing loss is necessary to achieve our results, we provide an ablation experiment (Masking w/o Debiasing) where the debiasing loss is replaced with the standard cross-entropy loss, while using the same pruning parameters. Results suggest that pruning alone is not able to achieve the same debiasing effect.

5.2 Masks are Diverse

As a means to showcase the capacity of debiasing masks to offer different debiasing behaviors, we refer to each method's capacity to mitigate the negation bias, evaluated by our negation-words subset.

As the standard debiasing results reveal, both weak model approaches were unable to improve baseline performance on our subset. Therefore, we used the hypothesis-only bias model, as provided by Utama et al. (2020a). Negation words are strongly correlated with the hypothesis-only bias; accordingly, observe a performance improvement on our subset when leveraging this debiasing target.

6 Mask Analysis

Our pruning approach uses a score threshold, which allows for a dynamic pruning rate across modules. To gain a better understanding on our masks, we study the generated density distribution. Specifically, we study the rate at which weights were removed in different network regions. In each layer, we obtain the pruning rate of the attention layer's query, key, value, and output modules, as well as the two linear modules that follow. We plot the results in Figure 2.

We make two general observations. First, pruning rates are relatively higher in the early layers. It is known that the last few layers are very task specific (Tenney et al., 2019), which likely implies that their modification more directly impacts performance. Thus, our masks may be targeting early layers as a means to remove biases without causing a noticeable performance decrease. Secondly, the attention modules are more heavily pruned than the linear layers, which could suggest that attention heads play an important role in bias encoding.

7 Conclusion

We conclude that debiasing masks are an effective approach to mitigating shortcuts. Alongside providing surprisingly well-performing debiased behaviors, masks allow to shift the way we think about debiasing: no longer should biased and unbiased models be treated as two separate models; rather, it becomes possible to "remove" biases by simply eliminating certain weights from the network.

Limitations

An important limitation of this approach is that we found it necessary to follow Sanh et al. (2020) and run the movement pruning technique for 12 epochs, requiring longer training time. We hypothesize that in future work, it could become possible to drastically reduce training time for this approach.

Further, in this short paper we explored three tasks in NLU with a fixed pruning configuration, but it would be beneficial to explore its applicability to other domains, combine it with other debiasing approaches, or explore more varied configurations, such as structured pruning methods for improved inference efficiency.

Acknowledgments

The education period leading to this project has received funding from "la Caixa" Foundation (ID 100010434), under agreement LCF/BQ/AA19/11720042. This work was also supported by JST PRESTO Grant Number JPMJPR20C and JSPS KAKENHI Grant Number 21H03502.

References

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4069–4082, Hong Kong, China. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for*

NLP, pages 143–155, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv:2004.08994*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Prasetya Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9063–9074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. 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 575–581, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sen-

tence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2226–2241, Online. Association for Computational Linguistics.

A Negation Words

Within the MNLI Validation Matched set, we select those entailed samples where at least one of the following words is present in the hypothesis: no, not, don't, none, nothing, never, aren't, isn't, weren't, neither, don't, didn't, doesn't, cannot, hasn't won't.

B Hyperparameters

B.1 Baseline and Debiased Models

Our baseline models are trained with a batch size of 32, learning rate of 3e-5, weight decay of 0.1, and 20% warmup steps.

B.2 Weak Learners

Undertrained models are trained with the standard BERT architecture and 2000 samples (500 in the case of QQP and FEVER) of the dataset for 5 epochs. Utama et al. (2021) mention 3 in their paper, but train with 5 in their code repository. We confirm that 5 yields better results. For underparameterized models, we follow (Sanh et al., 2021) and use BERT-Tiny (Turc et al., 2019), a BERT model with 2 layers and an inner size of 128. This model is trained on the full training set for 3 epochs, and slightly adjusted hyperparameters.

We use focal loss (sample reweighting) in all mask debiasing experiments, with the exception of the HypOnly MD model, which uses product-of-experts (in an attempt to follow Utama et al. (2020a) as closely as possible).

B.3 Masked Debiasing Models

Our mask search is performed with a score learning rate of 0.1, batch size of 128, and 12 epochs of training.