

End-to-End Self-Debiasing Framework for Robust NLU Training

Abbas Ghaddar, Philippe Langlais[†], Mehdi Rezagholizadeh and Ahmad Rashid

Huawei Noah’s Ark Lab, Montreal Research Center, Canada

[†]RALI-DIRO, Université de Montréal, Montréal, Canada

abbas.ghaddar@huawei.com, felipe@iro.umontreal.ca

mehdi.rezagholizadeh@huawei.com, ahmad.rashid@huawei.com

Abstract

Existing Natural Language Understanding (NLU) models have been shown to incorporate dataset biases leading to strong performance on in-distribution (ID) test sets but poor performance on out-of-distribution (OOD) ones. We introduce a simple yet effective debiasing framework whereby the shallow representations of the main model are used to derive a bias model and both models are trained simultaneously. We demonstrate on three well studied NLU tasks that despite its simplicity, our method leads to competitive OOD results. It significantly outperforms other debiasing approaches on two tasks, while still delivering high in-distribution performance.

1 Introduction

Researchers have increasingly raised concerns about the tendency of recent NLU models (Devlin et al., 2019; Liu et al., 2019) to quickly leverage spurious surface lexical-syntactic features (Poliak et al., 2018; Gururangan et al., 2018; Dasgupta et al., 2018; Ghaddar and Langlais, 2017). These superficial properties, also referred as dataset biases (Shah et al., 2020; Utama et al., 2020b; Moosavi et al., 2020), result in significant performance drop on out-of-distribution (OOD) sets containing counterexamples to biases in the training data (McCoy et al., 2019; Schuster et al., 2019; Zhang et al., 2019; Ghaddar et al., 2021).

The most common approach to tackle the problem consists in training a *bias model* with hand-crafted features with the goal of identifying biased training examples. This information is used in a later stage to discourage the *main model* from adopting the naive strategy of the bias model. Several debiasing training paradigms have been proposed to adjust the importance of biased training samples, such as product of experts (Clark et al., 2019; He et al., 2019), learned-mixin (Clark et al.,

2019), example reweighting (Schuster et al., 2019), debiased focal loss (Mahabadi et al., 2020), and confidence regularization (Utama et al., 2020a).

Recently, there has been a number of endeavours to produce a bias model without prior knowledge on the targeted biases or without the need for manually designing features. Utama et al. (2020b) propose to use instead a model trained on a tiny fraction ($< 1\%$) of the training data for few epochs as a bias model; while Clark et al. (2020) and Sanh et al. (2020) trained a low capacity model on the full training set. These approaches target the training of the bias model alone, which is subsequently queried while training the main model of interest.

In this paper, we propose an end-to-end debiasing framework which does not require an extra training stage, or manual bias features engineering. The bias model is indeed a simple attention-based classification layer on top of the main model’s intermediate representations. Both models are trained simultaneously in an end-to-end manner as in (Mahabadi et al., 2020), where the importance of training samples for both models are adjusted using the example reweighting technique of (Schuster et al., 2019).¹ The idea of using intermediate classifiers has previously been explored to reduce the inference cost (Schwartz et al., 2020; Zhou et al., 2020; Xin et al., 2021) for Transformer-based (Vaswani et al., 2017) models.

In contrast to all previous works, our bias model helps locating lexico-syntactic bias features, inside the main model’s intermediate layers, whose importance is reduced by adding a noise vector; therefore preventing the main model to rely on them. During training, both the main and bias models interact by interchangeably re-weighting the importance of each others’ examples.

Our learning framework, when applied to a

¹The bias model for these two works is based on hand-crafted features.

vanilla BERT (Devlin et al., 2019) model leads to consistent and significant improvements on 3 NLU tasks, while maintaining a balanced performance between ID and OOD sets. It involves a single training stage, and only incurs a small number of extra parameters (0.5M) compared to other approaches. For instance, Utama et al. (2020a) used a copy of the main model (110M parameters) as the bias model, while (Sanh et al., 2020) used BERT-Tiny (Turc et al., 2019) that has 11M parameters and has been pre-trained from scratch using the masked LM objective (Devlin et al., 2019).

2 Method

We describe an end-to-end solution for debiasing a Transformer-based classification model. We note $x = \{x_1, x_2, \dots, x_n\}$ the input sequence of length n , and $y \in \{1, 2, \dots, T\}$ the gold label, where T is the number of classes. Figure 1 shows a diagram of our method.

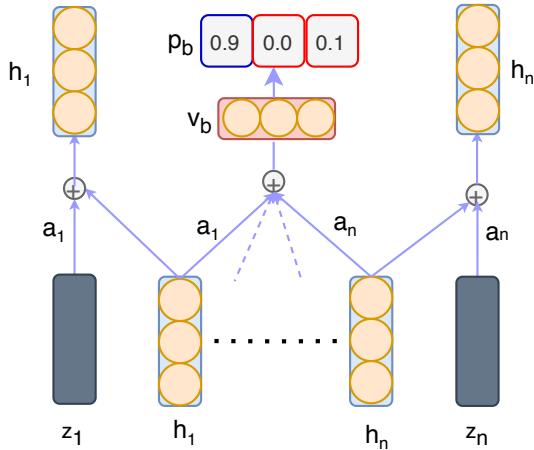


Figure 1: Illustration of our framework where the bias model is placed at the top of the k^{th} layer of the main model. Notation is introduced in the text, for simplicity we use h_i to refer to h_i^k . The blue rectangle indicates the index of of the ground truth class.

2.1 Main Model

The main model is a Transformer-based BERT encoder (Vaswani et al., 2017; Devlin et al., 2019) with a classifier on top of the classification [CLS] token of the last layer. This classifier generates the probability distribution $p_m \in \mathbb{R}^T$ over the output classes given an input x . At each layer k , BERT produces an internal hidden representation $h_i^k \in \mathbb{R}^d$ for each token x_i in the input.

2.2 Bias Model

We hypothesize that the bottom layers of the main model can serve as input for the bias model, thus avoiding the need for an external model or for manually designed features. Our intuition is built on the observation made by Jawahar et al. (2019) that bottom layers of BERT mainly encode lexical and syntactic information, and on the fact that such models tend to quickly overfit this type of information (Zellers et al., 2018; McCoy et al., 2019).

Our bias model is composed of an additive attention layer (Bahdanau et al., 2014), followed by a softmax one. This module uses the k^{th} layer of the main model as input in order to produce a probability distribution over classes. First, a scalar value $\tilde{a}_i \in \mathbb{R}$ is computed using a feed forward neural network with weight matrices $W_e \in \mathbb{R}^{d \times d}$ and $W_a \in \mathbb{R}^{1 \times d}$ such that:

$$\tilde{a}_i = W_a \tanh(W_e h_i^k) \quad (1)$$

Those scalar values are normalized and referred to as attentions a_i :

$$a_i = \frac{\exp(\tilde{a}_i)}{\sum_{i=1}^n \exp(\tilde{a}_i)} \quad (2)$$

The representation vector, v_b , for the bias model is computed as the weighted sum of the intermediate representations of the k^{th} layer:

$$v_b = \sum_{i=1}^n a_i h_i^k \quad (3)$$

We obtain the probability distribution p_b over the classes by applying a projection layer (parameterized by $W_y \in \mathbb{R}^{d \times T}$) followed by a softmax function: $p_b = \text{softmax}(W_y^\top v_b)$.

2.3 Biased Features Regularization

The attention weights of the bias model reveal which hidden representations are the most informative for the classification decision of that model. We propose to de-emphasize these ‘‘bias features’’ from the main model. We do so by adding a weighted noise vector to the hidden representations of the main model at layer k . Let $z \leftarrow \{z_1, \dots, z_n\}$ be a set of zero-mean Gaussian noise vectors corresponding to each token in the sequence, where $z_i \in \mathbb{R}^d$. The intermediate representations of the main model at layer k are updated as follows:

$$h_i^k \leftarrow a_i z_i + h_i^k \quad (4)$$

where a_i determines the relative amplitude of noise added to each token representation. A high value of a_i means that h_i^k most likely contains bias features. Consequently, a large amount of noise is added at this position, which will prevent the main model from overfitting to these features at subsequent layers.

2.4 Example Re-weighting

For both models, we adjust the importance of a training sample by directly assigning it a scalar weight that indicates whether the sample exhibits a bias or not. Let p_m^c and p_b^c be the predicted probabilities corresponding to the ground-truth class by the bias and main model respectively. The weight assigned to the main model training sample is calculated as follows:

$$w_m = \begin{cases} 1 - p_b^c & p_b^c > \gamma \\ 1 & \text{otherwise} \end{cases}$$

Differently from the reweighing method of (Clark et al., 2019; Schuster et al., 2019), we add a hard threshold $\gamma \in [0, 1]$ to control the number of samples to be re-weighted. The goal is to ensure that the importance of main model samples will be attenuated only if the bias model predictions fall into the high confidence bin ($\gamma > 0.8$) on biased samples. To further strengthen this constraint, we down-weight the importance of bias model training samples if the main model confidence is below a threshold β :

$$w_b = \begin{cases} p_m^c & p_m^c < \beta \\ 1 & \text{otherwise} \end{cases}$$

This ensures that the bias model is focusing on easy examples at early training stages, while challenging ones are gradually fed in the later training steps. Since the main and bias models are trained simultaneously, they interchangeably re-weight each others' examples. This is different from previous works where all samples have the same importance during the training phase of the bias model. For a single training instance, the individual loss term for the main and bias models are:

$$\mathcal{L}(\theta_B) = -w_b \log(p_b^c) \quad (5)$$

$$\mathcal{L}(\theta_M) = -w_m \log(p_m^c) \quad (6)$$

where $\theta_B = \{W_e, W_a, W_y\}$ are the trainable parameters for the bias model, while θ_M are the ones for the main model (BERT parameters). We train

both models by minimizing the aforementioned losses. At inference time, only the main model is used for prediction, and no noise is added to h^k . It is worth mentioning that our proposed extension has no impact on the overall training time.

3 Experiments

3.1 Datasets

We test our method on 3 sentence-pair classification tasks supported by ID training and validation sets as well as an OOD test set, which are specifically built to measure robustness to dataset bias.

For the **Natural Language Inference** task, we use the MNLI (Williams et al., 2017) benchmark as our ID data, and HANS (McCoy et al., 2019) as our OOD test set. For **Fact Verification**, we use the FEVER (Thorne et al., 2018) benchmark for ID evaluation, and FEVER Symmetric (Schuster et al., 2019) (version 1) as our OOD test set. For **Paraphrase identification**, we use QQP² as our ID data, and PAWS (Zhang et al., 2019) as our OOD test set.

3.2 Implementation

We use the 12-layer BERT-base model (Devlin et al., 2019) as our main model, thus our results can be compared with prior works. We adopt the standard setup of BERT and represent a pair of sentences as: [CLS] 1st sentence [SEP] 2nd sentence [SEP]. For BERT hyper-parameters, we use those of the baseline: a batch size of 64, learning rate of 1e-5 with the Adam (Kingma and Ba, 2014) optimizer.

Following (Clark et al., 2019; Grand and Belinkov, 2019; Clark et al., 2020; Sanh et al., 2020), we tune our method hyperparameters on the OOD sets. As pointed out by (Clark et al., 2019, 2020), this is not ideal since it assumes some prior knowledge of the OOD test sets. To best mitigate this impact, we follow the procedure of previous works and use the same hyper-parameters for all 3 tasks. We varied k ([2–5]), γ ([0.7–0.9]), β ([0.5–0.7]), and fix their values to 3, 0.8, 0.5 respectively.

However, different set of parameters performed roughly equally well, provided that k is 3 or 4, $\gamma \geq 0.8$ and β is set to 0.5. We use early stopping and report mean performance and standard deviation over 6 runs with different seeds.

²<https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question>

| Model | MNLI | | FEVER | | QQP | |
|------------------------------------|-------------|-----------------|------------------|-----------------|-------------|-------------|
| | Dev | HANS | Dev | Symm. | Dev | PAWS |
| BERT-base | 84.5 | 62.4 | 85.6 | 55.1 | 91.0 | 33.5 |
| (Clark et al., 2019) | 83.5 | 69.2 | - | - | - | - |
| (Schuster et al., 2019) | - | - | 85.4 | 61.7 | - | - |
| (Utama et al., 2020a) | 84.5 | 69.1 | 86.4 | 60.5 | 89.1 | 40.0 |
| (Utama et al., 2020b) [†] | 82.3 | 69.7 | - | - | 85.2 | 57.4 |
| (Utama et al., 2020b) [‡] | 84.3 | 67.1 | - | - | 89.0 | 43.0 |
| (Sanh et al., 2020) [♠] | 81.4 | 68.8 | 82.0 | 60.0 | - | - |
| (Sanh et al., 2020) [♣] | 83.3 | 67.9 | 85.3 | 57.9 | - | - |
| this work | 83.2±0.1 | 71.2±0.2 | 86.9 ±0.8 | 63.8±0.3 | 90.2±0.2 | 46.5±2.3 |
| w/o noise | 83.7±0.3 | 68.6±0.9 | 85.5±0.5 | 61.6±0.6 | 90.4±0.3 | 42.4±2.1 |
| w/o main reweighting | 84.0±0.4 | 62.8±1.1 | 85.1±0.4 | 57.4 ±0.8 | 91.0±0.2 | 36.7±1.6 |
| w/o bias reweighting | 81.5±0.9 | 64.6±0.6 | 83.9±0.4 | 60.3±0.5 | 89.3±0.7 | 39.7±2.4 |

Table 1: Model performance when evaluated on MNLI, Fever, QQP, and their corresponding challenge test sets. Ablation study of our method without adding noise, and without reweighting main or bias model’s training samples respectively. Symbols are used to distinguish variants from the same paper that use a different training technique: (†) example reweighting, (‡) confidence regularization, (♠) product of Experts (PoE), (♣) PoE+ cross-entropy.

3.3 Results

Table 1 reports accuracy scores on the development and OOD test sets of the 3 benchmarks we considered. The baseline is the vanilla BERT-base model which is used as a backbone for the main model in all the configurations reported. Also, the table shows the ablation results of our method without adding noise, and without reweighting the training instances of the bias and main models (setting β and γ to 0 or 1 respectively).

First, we notice high variance in performances between runs in the debiasing setting, which was also reported in (Clark et al., 2019; Mahabadi et al., 2020; Utama et al., 2020b; Sanh et al., 2020). Second, we observe that our method offers a good balance between gains on OOD test sets over the baseline, and losses on ID sets. More precisely, we report the best results on FEVER (both ID and OOD test sets), while we improve the HANS score on the MNLI task, but fail to maintain the baseline score on dev as Utama et al. (2020a) did.

On QQP, Utama et al. (2020b) reported a much higher score on PAWS (57.4% vs. 46.5%), but at the expense of an important drop on the ID dev set (85.2% vs. 90.2%). However, our method outperforms this particular model on both MNLI sets, and also shows better cross-task performances compared to prior works. The results are satisfactory, especially when considering the simplicity and ef-

ficiency of our approach. Moreover, the fact that a single configuration works well on 3 tasks is an indicator that our method has the potential to generalize on completely unknown OOD sets (Clark et al., 2020).

Expectedly, deactivating the main model reweighting mechanism results in near baseline performances. Solely adding noise signal leads to a modest gain of 2-3% on the OOD test sets and a slight drop ($< 1\%$) on the dev sets compared to the baseline. On one hand, without adding noise, our scores are comparable with previous works across the 3 tasks, that is, a significant drop on OOD test sets and minor gains on ID ones. These observations suggest that down-weighting biased examples is important, while de-emphasizing bias features further improves robustness.

| Task | Acc. | | $p_b^c > \gamma$ | |
|-------|------|------|------------------|-----|
| | w | w/o | w | w/o |
| MNLI | 67.9 | 69.7 | 18% | 23% |
| FEVER | 70.8 | 85.7 | 16% | 28% |
| QQP | 87.6 | 92.9 | 30% | 44% |

Table 2: Bias model training accuracy and percentage of training samples correctly classified with high probability, with and without reweighting bias model examples.

On the other hand, we observe that not reweight-

ing bias model examples results in the worst performance on the ID sets. We conduct an analysis on the bias model to better understand the impact of reweighting its examples. As shown in Table 2, its training accuracy as well as the percentage of high confidence predictions increases when examples are not re-weighted, further down-weighting the main model’s training examples, which leads to a significant drop of both the ID and OOD performances for all 3 tasks.

- a) **GT: C; Main: 0.81; Bias (w/o): 0.99 (0.94)**
H Yes, sir.
P No, not in particular Sir.
- b) **GT: E; Main: 0.74; Bias (w/o): 0.95 (0.92)**
H right after the war
P Just after the war ended.
- c) **GT: C; Main: 0.67 ; Bias (w/o): 0.0 (0.77)**
H well his knees were bothering him yeah
P He was in tip-top condition.
- d) **GT: C; Main: 0.39 ; Bias (w/o): 0.1 (0.89)**
H Even us if you needed,” said John.
P He told them not to ask him to lift a finger.
- e) **GT: N; Main: 0.72 ; Bias (w/o): 0.1 (0.92)**
H What changed?
P What was unique?

Figure 2: Ground Truth (GT) class probabilities of the main and bias models on examples from the MNLi dev set. Examples consist of an Hypothesis and a Premise, and valid labels are: Entailment, Contradiction, and Neutral. Probabilities of the bias model without reweighting its examples are placed within parenthesis.

We inspected the confident scores of the main and bias (with and without example reweighting) models on MNLi dev set, an excerpt of which is reported in Figure 2. On one hand, we observe that bias models successfully assign high probabilities to samples that can be easily classified via keywords (e.g. ”not” in example a) and those with high lexical overlap (example b). On the other hand, we noticed that the bias model is performs undesirably well on some challenging examples like (c) and (d) of Figure 2. However, the bias model probability always decreases when we re-weight its examples, which eventually leads to correct prediction of the main model as in example (c) and (e). Interestingly, this observation suggests that training a pure bias

model is as important (and challenging) as training the robust one.

4 Conclusion

Our key contribution is a framework that jointly identifies biased examples and features in an end-to-end manner. Our approach is geared towards addressing lexico-syntactic bias features for Transformer-based NLU models. Future work involves testing our approach on other tasks such as Question Answering (Rajpurkar et al., 2016; Agrawal et al., 2018), exploring methods to obtain proxy OOD data for hyper-parameters selection, and making our method hyper-parameter free.

Acknowledgments

We thank Mindspore³ for the partial support of this work, which is a new deep learning computing framework.

References

- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- 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 4060–4073.
- Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. *arXiv preprint arXiv:2011.03856*.
- Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.
- 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*

³<https://www.mindspore.cn/>

- of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.
- Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, and Mehdi Rezagholizadeh. 2021. Context-aware Adversarial Training for Name Regularity Bias in Named Entity Recognition. *Transactions of the Association for Computational Linguistics*, 9.
- Abbas Ghaddar and Phillippe Langlais. 2017. Winer: A wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422.
- Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 1–13.
- 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.
- He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. *EMNLP-IJCNLP 2019*, page 132.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716. Association for Computational Linguistics.
- 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.
- Nafise Sadat Moosavi, Marcel de Boer, Prasetya Ajie Utama, and Iryna Gurevych. 2020. Improving robustness by augmenting training sentences with predicate-argument structures. *arXiv preprint arXiv:2010.12510*.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.
- 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.
- Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2020. Learning from others’ mistakes: Avoiding dataset biases without modeling them. *arXiv preprint arXiv:2012.01300*.
- 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 3410–3416.
- Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651.
- Deven Santosh Shah, H Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 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.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing nlu

models without degrading the in-distribution performance. [arXiv preprint arXiv:2005.00315](#).

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In [Advances in Neural Information Processing Systems](#), pages 5998–6008.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. [arXiv preprint arXiv:1704.05426](#).

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. Berxit: Early exiting for bert with better fine-tuning and extension to regression. In [Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume](#), pages 91–104.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In [Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing](#), pages 93–104.

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.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. 2020. Bert loses patience: Fast and robust inference with early exit. [Advances in Neural Information Processing Systems](#), 33.