Robust Natural Language Understanding with Residual Attention Debiasing

Fei Wang,* **James Y. Huang,*** **Tianyi Yan, Wenxuan Zhou** and **Muhao Chen**
University of Southern California
{fwang598,huangjam,tianyiy,zhouwenx,muhaoche}@usc.edu

Abstract

Natural language understanding (NLU) models often suffer from unintended dataset biases. Among bias mitigation methods, ensemble-based debiasing methods, especially product-of-experts (PoE), have stood out for their impressive empirical success. However, previous ensemble-based debiasing methods typically apply debiasing on top-level logits without directly addressing biased attention patterns. Attention serves as the main media of feature interaction and aggregation in PLMs and plays a crucial role in providing robust prediction. In this paper, we propose REsidual Attention Debiasing (READ), an end-to-end debiasing method that mitigates unintended biases from attention. Experiments on three NLU tasks show that READ significantly improves the performance of BERT-based models on OOD data with shortcuts removed, including +12.9% accuracy on HANS, +11.0% accuracy on FEVER-Symmetric, and +2.7% F1 on PAWS. Detailed analyses demonstrate the crucial role of unbiased attention in robust NLU models and that READ effectively mitigates biases in attention.[1]

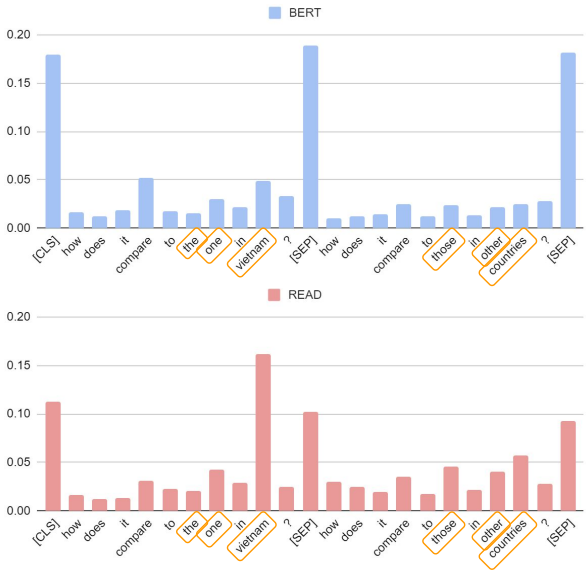

Figure 1: Attention distribution on a non-duplicated sentence pair. Red bars are debiased [CLS] attention from the last ensemble layer of READ and blue bars are corresponding attention from finetuned BERT. Distinct tokens in the two sentences are highlighted with orange borderlines. READ pays more attention to distinct tokens and is more robust to lexical overlap bias.

1 Introduction

Natural language understanding (NLU) models often suffer from unintended dataset biases (Jia and Liang, 2017; Gururangan et al., 2018; Poliak et al., 2018; Gardner et al., 2021; Rajaee et al., 2022), causing them to learn spurious shortcuts and make unfaithful or under-generalized prediction (McCoy et al., 2019; Schuster et al., 2019; Zhang et al., 2019b). While a number of methods have been proposed to tackle this problem based on prior knowledge of specific biasing features (Clark et al., 2019a; He et al., 2019; Mahabadi et al., 2020; Utama et al., 2020a; Liu et al., 2022), various unintended biases exist in NLU datasets, and not all of them are identifiable (Sanh et al., 2020; Utama et al., 2020b). More recent works start to focus on mitigating unknown biases (Sanh et al., 2020; Utama et al., 2020b; Xiong et al., 2021; Ghaddar et al., 2021; Meissner et al., 2022). Among them, ensemble-based debiasing methods, especially product-of-experts (PoE), have stood out for their impressive empirical success (Sanh et al., 2020; Utama et al., 2020b; Xiong et al., 2021; Ghaddar et al., 2021).

Although the attention mechanism (Vaswani et al., 2017) is essential to the success of Transformer-based pretrained language models (PLMs), attention can also capture potentially spurious shortcut features leading to prediction biases. For example, too much or too little attention across

*The first two authors contributed equally.
[1]Code is available at https://github.com/luka-group/READ.

Findings of the Association for Computational Linguistics: ACL 2023, pages 504–519
July 9-14, 2023

sentences in natural language inference may lead to the lexical overlap bias (McCoy et al., 2019; Rajaee et al., 2022) or the hypothesis-only bias (Poliak et al., 2018). Since attention serves as the main media for feature interactions in PLMs, many of the aforementioned biases can be associated with biased attention patterns. In fact, a number of recent studies have shown that appropriate attention plays a critical role in ensuring robust[2] prediction (Chen et al., 2020; Li et al., 2020; Stacey et al., 2022). However, existing ensemble-based debiasing methods typically apply debiasing on top-level logits (Clark et al., 2019a; He et al., 2019; Sanh et al., 2020; Utama et al., 2020b; Ghaddar et al., 2021). These methods do not proactively mitigate attention biases, but instead, rely on debiasing signals being propagated from final predictions to the attention modules in a top-down manner. Top-level logits are highly compressed and the propagation may suffer from information loss, thus providing limited debiasing signal to low-level attention. Instead, we seek for an effective attention debiasing method that prevents models from learning spurious shortcuts, especially those captured by the attention mechanism.

In this paper, we propose REsidual Attention Debiasing (READ), an end-to-end debiasing method that mitigates unintended biases from attention. Our method is inspired by the recent success of one-stage PoE (Ghaddar et al., 2021). As an ensemble-based debiasing method, it trains a biased model to capture spurious in-distribution shortcuts and trains the ensemble of the biased model and a main model to prevent the main model from relying on spurious shortcuts. To do this end-to-end, one-stage PoE trains the biased model and its ensemble with the main model simultaneously in a weight-sharing manner. In READ, we let the two models share all weights except attention modules and classification heads, allowing the main model to fit the unbiased attention residual with respect to the attention of the biased model. Intuitively, since they are trained on the same batch of data at each iteration, biased model attention and main model attention are likely to capture similar spurious features, making their residual free of such biases. Fig. 1 presents an example of the attention change. Given a non-duplicate sentence pair, BERT, which suffers from lexical overlap bias, does not aggregate much information from non-overlapping tokens. In contrast, READ learns to pay more attention to informative non-overlapping tokens.

Experiments on three NLU tasks show that READ significantly improves the performance of BERT-based models on OOD data where common types of shortcuts are removed, including +12.9% accuracy on HANS, +11.0% accuracy on FEVER-Symmetric, and +2.7% F1 on PAWS. We further examine the attention scores of the debiased main model and find that its distribution is more balanced (§4.1). These results indicate the crucial role of unbiased attention in robust NLU models. We also demonstrate that our method is still effective when using a biased model with the same parameter size as the main model (§4.2), which differs from the previous assumption that the biased model of unknown biases should be weaker [3] (Sanh et al., 2020; Utama et al., 2020b).

Our contributions are three-fold. First, we propose READ, an ensemble-based debiasing method for NLU models, mitigating attention biases through learning attention residual. Second, experiments on three NLU tasks consistently demonstrate that READ can significantly improve the OOD performance of different NLU tasks with various dataset biases. Third, detailed analyses provide useful insights for developing robust NLP models, including the importance and properties of unbiased attention, and the design of biased models in ensemble-based debiasing methods.

2 Method

Our method, READ, combines one-stage product-of-experts (PoE) with learning attention residual to mitigate unknown dataset biases for NLU tasks. Based on the problem definition, we introduce the two key components of our method, followed by the details of training and inference.

2.1 Problem Definition

For a discriminative task, given the dataset $D = \{x_i, y_i\}$, where x_i is the raw input and y_i is the gold label, our goal is to learn a robust function f with parameters θ, that can predict a probability distribution $\mathbf{p} = f(x_i; \theta)$ without relying on spurious features. in NLU tasks, x_i is typically a textual

[2] Robustness typically refers to the consistency of model behavior given original and (adversarially) perturbed inputs (Jia et al., 2019), or given in-distribution and out-of-distribution (OOD) data (Hendrycks et al., 2020). This paper focuses on OOD robustness, where OOD data do not share dataset biases with in-distribution data.

[3] Under-trained or under-parameterized.

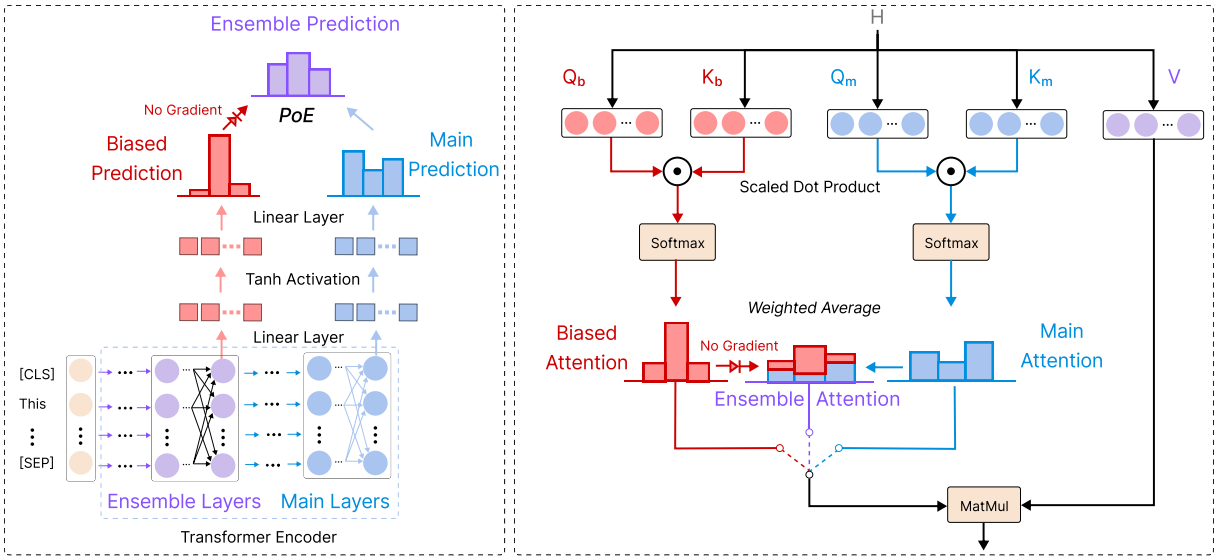

Figure 2: Illustration of one-stage PoE (left) and learning attention residual in ensemble layers (right). Dotted lines in the right figure are conditionally activated. During training, ensemble attention is activated to compute the main prediction and biased attention is activated to compute biased prediction. Through learning their residual, READ mitigates biases from the main attention. During inference, the debiased main attention is activated to compute robust main prediction.

sequence. As discussed by prior studies (Gardner et al., 2021; Eisenstein, 2022), spurious features captured by f, such as particular words (Gardner et al., 2021) and lexical overlap ratios (Rajaee et al., 2022), although may be statistically correlated with y_i due to dataset artifacts (Gururangan et al., 2018), should not be regarded as useful information for predicting y_i. In other words, the prediction of a robust and faithful model should be independent of these non-causal features. Since diverse spurious features may exist in NLU datasets, we focus on mitigating dataset biases without any assumption of the type or structure of bias, so that the proposed method is generalizable to most unknown biases.

2.2 One-Stage PoE

Due to the automated process of feature extraction in neural networks, it is impractical to train a robust model that directly identifies all robust features in the tremendous feature space. Considering spurious features are often simple features (also, of easy-to-learn data instances) that model tends to memorize in the first place (Shah et al., 2020), an ensemble-based debiasing method trains a biased model to collect biased prediction p_b and approximates an ensemble prediction p_e based on p_b and another prediction p_m from a main model towards the observations in training data. Considering both parts of the ensemble prediction p_e, since the biased model mainly captures spurious shortcuts, as its complement, the main model then focuses on capturing robust features. READ adopts PoE (Clark et al., 2019a) to obtain a multiplicative ensemble of the two models' predictions:

$$p_e \propto p_b p_m. \tag{1}$$

Specifically, READ follows the one-stage PoE framework (Ghaddar et al., 2021) that simultaneously optimizes the ensemble prediction and biased prediction, and shares weights between the main model and biased model, as shown in Fig. 2. When using a PLM as the main model, one-stage PoE typically uses one or a few bottom layers of PLMs stacked with an independent classification head as the biased model, because these low-level layers preserve rich surface features (Jawahar et al., 2019) which can easily cause unintended biases (McCoy et al., 2019; Gardner et al., 2021). The main model has shared encoder layers at the bottom followed by independent encoder layers and its classification head. This weight-sharing design makes it possible to debias the model end-to-end with a few additional parameters. However, shared layers result in shared biases in these layers. Although PoE mitigates biases from predictions, it preserves biases in shared layers.

2.3 Learning Attention Residual

Ensemble prediction with PoE cannot effectively mitigate unintended biases in attention, which is the major part of feature aggregation and interaction in PLMs. For example, the `[CLS]` representation aggregates information from all token representations according to the attention distribution, and all token representations interact with each other based on the attention values. Therefore, biased attention becomes the direct source of many spurious features, such as lexical overlap in natural language inference and semantic-neutral phrases in sentiment analysis (Friedman et al., 2022). To prevent the main model from learning biased attention, READ further conducts additive ensemble of the attention distributions of both the main and biased models. Similar to ensemble prediction, the attention ensemble here encourages the main model attention to learn from the residual of biased model attention, so as to mitigate the biases captured by the latter from the former.

Fig. 2 shows the workflow of learning attention residual. The self-attention mechanism (Vaswani et al., 2017) allows each vector in a matrix to interact with all vectors in the same matrix. Specifically, the input matrix H is first projected to a query matrix Q, a key matrix K, and a value matrix V. Attention scores of all vectors to each vector is a probability distribution computed based on the dot product between Q and K. With attention scores as weights, the self-attention module maps each vector in H to the weighted average of V. In READ, the main attention and biased attention use distinct projection weights for Q and K, but take the same H as inputs and share the same projection weights for V. Distinct Q and K allow the two models to have their own attention. Sharing H and V ensures the attention in the biased and main models are distributed in the same semantic space so that they are additive.[4]

The ensemble attention $\mathbf{a_e}$ combines main attention $\mathbf{a_m}$ and biased attention $\mathbf{a_b}$ with weighted average.[5] This additive ensemble is inspired by the success of using the probability difference for post-hoc debiasing (Niu et al., 2021; Qian et al., 2021; Wang et al., 2022c) and preventing overconfidence (Miao et al., 2021). In our case, the main attention is the difference between ensemble attention and biased attention. READ also adds a coefficient $\alpha \in (0, 1)$ to balance the ensemble ratio. An appropriate coefficient can prevent over- or under-debiasing. Finally, the ensemble attention becomes

$$\mathbf{a_e} = (1 - \alpha)\mathbf{a_m} + \alpha\mathbf{a_b}. \tag{2}$$

Now that we have three paths in the attention module, including ensemble attention, main attention, and biased attention. In each forward pass from the input to p_m or p_b, only one of them is activated as the final attention distribution. During training, READ adopts ensemble attention to compute p_m and biased attention to compute p_b, for mitigating biases from main attention by learning their residual. During inference, READ adopts main attention, which is free of bias, to compute robust prediction p_m.

2.4 Training and Inference

We train the ensemble model and the biased model on the same dataset batch B simultaneously with a cross-entropy loss

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_e + \mathcal{L}_b \\ &= -\frac{1}{|B|}\sum_{i=1}^{|B|} \log p_e(y_i|x_i) + \log p_b(y_i|x_i).\end{aligned} \tag{3}$$

When minimizing $\mathcal{L}_e$, gradients on p_b in Eq. 1 and gradients on $\mathbf{a_b}$ in Eq. 2 are disabled, because they serve as auxiliary values for computing p_e. Backward passes on p_b and $\mathbf{a_b}$ are only allowed when minimizing $\mathcal{L}_b$. During inference, only the main

[4] In contrast, if we use a completely independent attention module in the biased model, its attention will not be aligned with the main model attention.

[5] Multiplicative ensemble (e.g. PoE), although works well on ensemble prediction, is unstable during training for attention ensemble and causes models to fail according to our observation. This phenomenon is related to the plausibility problem in Li et al. (2022). The fluctuation of a tiny probability on an uninformative token (e.g. a stop word) may significantly influence the result of PoE. Assuming we have simple distributions over two candidates $p_e = [10^{-8}, 1 - 10^{-8}]$ and $p_b = [10^{-6}, 1 - 10^{-6}]$, then according to Eq. 1, the learned $p_m \approx [0.99, 0.01]$. Due to the probability change from 10^{-8} to 10^{-6} of the first candidate, the division between p_e and p_b maps the probability from extremely high (i.e. close to 1) to low (i.e. 0.01) and vice versa, i.e. over-debiasing. Such behavior is harmful to the learning process.

model is used to predict a label $\hat{y}_i$ from the label set $\mathcal{C}$:

$$\hat{y}_i = \underset{c=1}{\overset{|\mathcal{C}|}{\operatorname{argmax}}} p_m(c|x_i). \quad (4)$$

3 Experiment

In this section, we evaluate the debiasing performance of READ on three NLU tasks. We first provide an overview of the experimental settings (§3.1 and §3.2), followed by a brief description of baseline methods (§3.3). Finally, we present a detailed analysis of empirical results (§3.4).

3.1 Datasets

Following previous studies (Utama et al., 2020b; Ghaddar et al., 2021; Gao et al., 2022), we use three English NLU tasks for evaluation, namely natural language inference, fact verification, and paraphrase identification. Specifically, each of the tasks uses an out-of-distribution (OOD) test set where common types of shortcuts in the training data have been removed, in order to test the robustness of the debiased model. More details can be found in Appx. §A.

MNLI (Multi-Genre Natural Language Inference; Williams et al. (2018)) is a natural language inference dataset. The dataset contains 392k pairs of premises and hypotheses for training, which are annotated with textual entailment information (*entailment*, *neutral*, *contradiction*). For evaluation, we report accuracy on the MNLI dev set and the OOD challenge set HANS (McCoy et al., 2019). HANS contains premise-hypothesis pairs that have significant lexical overlap, and therefore models with lexical overlap bias would perform close to an entailment-only baseline.

FEVER (Thorne et al., 2018) is a fact verification dataset that contains 311k pairs of claims and evidence labeled with the validity of the claim with the given evidence as context. For OOD testing, we report accuracy on the FEVER-Symmetric[6] test set (Schuster et al., 2019) where each claim is paired with both positive and negative evidences to avoid claim-only bias[7] .

QQP is a paraphrase identification dataset consisting of pairs of questions that are labeled as either *duplicated* or *non-duplicate* depending on whether one sentence is a paraphrased version of the other. For testing, we report F1 score on PAWS (Zhang et al., 2019c), which represents a more challenging test set containing non-duplicate question pairs with high lexical overlap.

3.2 Implementation

Following previous works (Utama et al., 2020b; Ghaddar et al., 2021; Gao et al., 2022), we use BERT-base-uncased model (Devlin et al., 2019) as the backbone of the debiasing framework. All experiments are conducted on a single NVIDIA RTX A5000 GPU. We use the same set of hyperparameters across all three tasks, with the learning rate, batch size, and ensemble ratio (α) set to 2e-5, 32, and 0.1 respectively. We train all models for 5 epochs and pick the best checkpoint based on the main model performance on the in-distribution dev set. On each dataset, we report average results and standard deviations of five runs. More details can be found in Appx. §B.

3.3 Baseline

We include a vanilla BERT model and compare our method with a wide selection of previous debiasing methods for language models as follows:

- *Reweighting* (Clark et al., 2019a) first trains a biased model to identify biased instances. During main model training, the biased instances are down-weighted, which encourages the model to focus more on unbiased instances.
- *PoE* (Clark et al., 2019a) and *DRiFt* (He et al., 2019) both train an ensemble of the biased and main models to learn the unbiased residual logits. The biased model is trained on data observed with a specific type of bias. Unlike our proposed READ, these methods do not directly address biased attention patterns.
- *Conf-Reg* (Utama et al., 2020a) applies logit smoothing to a biased model to improve distillation. It prevents the model from making overly-confident predictions that are likely biased.
- *MoCaD* (Xiong et al., 2021) applies model calibration to improve the uncertainty estimations of a biased model. This method is generally complementary to a variety of ensemble-based methods.
- *PoE w/ Weak Learner* (Sanh et al., 2020) and *Self-Debias* (Utama et al., 2020b) propose to use under-parameterized and under-trained models as biased models for ensemble-based debiasing methods, such as PoE. Since these weak models tend to rely on spurious shortcuts in datasets, they are effective in mitigating unknown bias.

[6]Version 1.

[7]Models overly relying on misleading cues from the claims while ignoring evidence.

Model	MNLI (Acc.)		FEVER (Acc.)		QQP (F1)	
	Dev	HANS	Dev	Sym.	Dev	PAWS
BERT-base	84.8‡	60.2‡	87.0‡	57.7‡	88.4‡	44.0‡
Known Bias Mitigation						
Reweighting (Clark et al., 2019a)	83.5	69.2	-	-	-	-
PoE (Clark et al., 2019a)	83.0	67.9	-	-	-	-
DRiFt (He et al., 2019)	81.8†	66.5†	84.2†	62.3†	-	-
Conf-Reg (Utama et al., 2020a)	84.3	69.1	86.4	60.5	-	46.1*
MoCaD (Xiong et al., 2021)	84.1	70.7	87.1	65.9	-	-
Unknown Bias Mitigation						
PoE w/ Weak Learner (Sanh et al., 2020)	81.4	68.8*	82.0	60.0	-	-
Self-Debias (Utama et al., 2020b)	82.3	69.7	-	-	-	-
MoCaD (Xiong et al., 2021)	82.3	70.7	-	-	-	-
End2End (Ghaddar et al., 2021)	83.2	71.2	86.9	63.8	-	-
Masked Debiasing (Meissner et al., 2022)	82.2	67.9	-	-	89.6	44.3
DCT (Lyu et al., 2023)	84.2	68.3	87.1	63.3	-	-
Kernel-Whitening (Gao et al., 2022)	-	70.9	-	66.2	-	45.2*
READ	79.6 ± 0.7	**73.1** ± 0.7	79.2 ± 1.9	**68.7** ± 2.1	84.5 ± 0.3	**46.7** ± 1.7
READ (p_e)	83.6 ± 0.3	64.8 ± 1.2	84.3 ± 1.1	55.3 ± 1.8	87.7 ± 0.0	44.8 ± 0.7

Table 1: Model performance on MNLI, FEVER, and QQP. We report results on both the in-distribution dev set and the OOD challenge set (highlighted in blue). All baseline results are copied from the referenced paper unless marked otherwise. For methods that have multiple variants, we report the variant with the best average OOD performance. ‡ reproduced with our code base. * computed based on reported (subset) accuracy. † copied from Xiong et al. (2021).

- *End2End* (Ghaddar et al., 2021) is an ensemble-based debiasing method that shares the bottom layers of the main model as the whole encoder of the biased model. It reweights instances based on model predictions and regularizes intermediate representations by adding noise.
- *Masked Debiasing* (Meissner et al., 2022) searches and removes biased model parameters that contribute to biased model predictions, leading to a debiased subnetwork.
- *DCT* (Lyu et al., 2023) reduces biased latent features through contrastive learning with a specifically designed sampling strategy.
- *Kernel-Whitening* (Gao et al., 2022) transforms sentence representations into isotropic distribution with kernel approximation to eliminate nonlinear correlations between spurious features and model predictions.

In addition, previous methods can also be categorized based on whether prior knowledge of specific biased features, such as hypothesis-only and lexical overlap biases in NLI, is incorporated in the debiasing process. We accordingly group the compared methods when reporting the results (Tab. 1) in the following two categories:

- Methods for *known bias mitigation* have access to the biased features before debiasing and therefore can train a biased model that only takes known biased features as inputs. While each of the OOD test sets we use for evaluation is crafted to target one specific form of bias, biased features can be highly complex and implicit in real-world scenarios, which limits the applicability of these methods.
- Methods for *unknown bias mitigation* do not assume the form of bias in the dataset to be given. Our proposed method belongs to this category.

3.4 Results

As shown in Tab. 1, among all baselines, unknown bias mitigation methods can achieve comparable or better performance than those for mitigating known biases on OOD test sets of NLI and fact verification. Although all baseline methods improve OOD performance in comparison with vanilla BERT, there is not a single baseline method that outperforms others on all three tasks.

Overall, our proposed method, READ, significantly improves model robustness and outperforms baseline methods on all OOD test sets with different biases. On HANS, the challenging test set for MNLI, our method achieves an accuracy score of 73.1%, i.e. a 12.9% of absolute improvement from vanilla BERT and a 1.9% improvement from the best-performing baseline *End2End*. Compared to

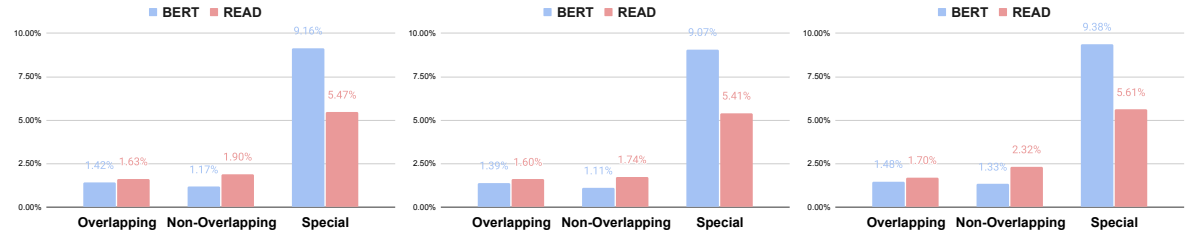

Figure 3: Average attention probability of each overlapping token, non-overlapping token, and special token per sentence pair on the PAWS test set for all instances (left), non-duplicated instances (middle), and duplicated instances (right). We present the [CLS] attention over all input tokens from the last ensemble layer of READ and the same attention layer of BERT. READ increases the attention over non-overlapping tokens to reduce lexical overlap bias.

End2End, residual debiasing on attention of READ directly debiases on the interactions of token-level features, leading to more effective mitigation of lexical overlap biases. On FEVER-Symmetric, READ outperforms vanilla BERT by 11.0% accuracy and outperforms the best-performing method *Kernel-Whitening* by 2.5%. On PAWS, the challenging test set for paraphrase identification, READ improves model performance by 2.7% F1, and outperforms the best-performing baseline method *Conf-Reg*, which relies on extra training data with lexical overlap bias. These results demonstrate the generalizability of READ for mitigating various biases in different NLU tasks.

We also observe that the in-distribution performance of READ is generally lower than baseline methods. In fact, almost all debiasing methods shown in Tab. 1 enhance OOD generalization performance at the cost of decreased in-distribution performance This aligns with the inherent trade-off between in-distribution performance and OOD robustness as shown by recent studies (Tsipras et al., 2018; Zhang et al., 2019a). The optimal in-distribution classifier and robust classifier rely on fundamentally different features, so not surprisingly, more robust classifiers with less distribution-dependent features perform worse on in-distribution dev sets. However, note that generalizability is even more critical to a learning-based system in real-world application scenarios where it often sees way more diverse OOD inputs than it uses in in-distribution training. Our method emphasizes the effectiveness and generalizability of debiasing on unknown OOD test sets and demonstrates the importance of learning unbiased attention patterns across different tasks. In the case where in-distribution performance is prioritized, the ensemble prediction p_e can always be used in place of the debiased main prediction p_m without requiring any additional training. Future work may also explore to further balance the trade-off between in-distribution and OOD performance (Raghunathan et al., 2020; Nam et al., 2020; Liu et al., 2021). It is also worth noting that our method only introduces a very small amount of additional parameters, thanks to the majority of shared parameters between biased and main models.

4 Analysis

To provide a comprehensive understanding of key techniques in READ, we further analyze the debiased attention distribution (§4.1) and the effect of number of ensemble layers (§4.2).

4.1 Debiased Attention Distribution

To understand the influence of READ on attention, we examine the attention distribution of BERT and READ on the PAWS test set. Specifically, we take the attention between [CLS], which serves as feature aggregation, and all other tokens as an example. We group tokens into three categories, including overlapping tokens (e.g. how and does in Fig. 1), non-overlapping tokens (e.g. one and those in Fig. 1), and special tokens (e.g. [CLS] and [SEP]). Since attention residual for attention debiasing exists in ensemble layers of READ, we compare the attention on the last ensemble layer of READ and the corresponding layer of BERT.

As discussed in §3.4, vanilla BERT finetuned on QQP suffers from the lexical overlap bias and does not generalize well on PAWS. This problem is reflected in the inner attention patterns. As shown in Fig. 3, BERT assigns less (-0.25%) attention to non-overlapping tokens than to overlapping tokens on average. In contrast, READ increases the attention on non-overlapping tokens to larger than (+0.27%) the attention on overlapping tokens. The same observation also appears in the subset of duplicate

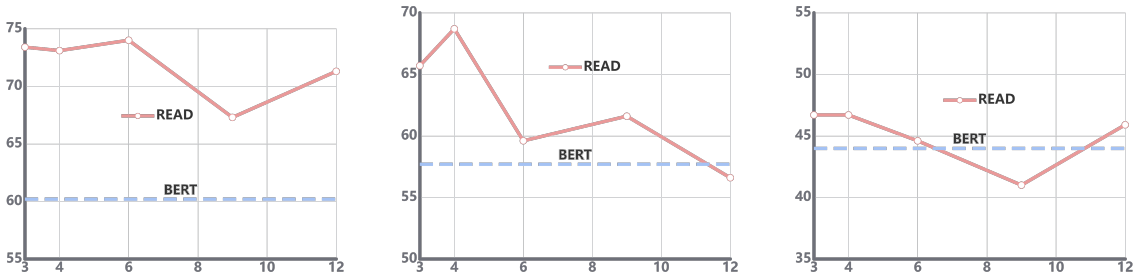

Figure 4: Performance of READ by the number of ensemble layers on HANS (left), FEVER-Symmetric (middle), and PAWS (right). READ is still effective when using twelve ensemble layers on HANS and PAWS.

sentence pairs and the subset of non-duplicate sentence pairs. This change in attention patterns reveals the inner behavior of READ for effectively preventing the model from overly relying on the lexical overlap feature.

4.2 Effect of Number of Ensemble Layers

Some previous studies assume that the biased model in PoE for unknown bias should be weaker (i.e. less trained or less parameterized) than the main model so as to focus on learning spurious features (Sanh et al., 2020; Utama et al., 2020b). One-stage PoE follows this assumption, using the bottom layers of the main model as the encoder of the biased model (Ghaddar et al., 2021). Since biased attention patterns may appear in any layer, including top layers, we examine whether this assumption holds for READ. Specifically, we evaluate READ with different numbers of ensemble layers on three OOD evaluation sets.

As shown in Fig. 4, although the best-performing READ variant has few ensemble layers, the configuration where the biased and main models share all encoder layers is still effective on HANS and PAWS. For example, on HANS, READ achieves comparable performance with the previous state-of-the-art method when the biased and main models share all encoder layers. This observation indicates that the shared encoder layer with distinct attention allows the biased model to focus on spurious attention patterns. Moreover, it is apart from the assumption that a biased model is necessarily a weak model, such as the bottom layers of the main model with a simple classification head. Future works on ensemble-based debiasing can explore a larger model space for the biased model.

5 Related Work

We present two lines of relevant research topics, each of which has a large body of work, so we can only provide a highly selected summary.

Debiasing NLU Models. Unintended dataset biases hinder the generalizability and reliability of NLU models (McCoy et al., 2019; Schuster et al., 2019; Zhang et al., 2019b). While a wide range of methods have been proposed to tackle this problem, such as knowledge distillation (Utama et al., 2020a; Du et al., 2021), neural network pruning (Meissner et al., 2022; Liu et al., 2022), and counterfactual inference (Udomcharoenchaikit et al., 2022), ensemble-based methods (Clark et al., 2019a; He et al., 2019; Lyu et al., 2023) stand out for their impressive empirical success. Recent works extend ensemble-based methods, such as PoE, to mitigate unknown biases by training a weak model to proactively capture the underlying data bias, then learn the residue between the captured biases and original task observations for debiasing (Sanh et al., 2020; Utama et al., 2020b; Ghaddar et al., 2021). Xiong et al. (2021) further improves the performance of these methods using a biased model with uncertainty calibration. Nevertheless, most prior works only mitigate unintended biases from top-level logits, ignoring biases in low-level attention.

Attention Intervention. In current language modeling technologies, the attention mechanism is widely used to characterize the focus, interactions and aggregations on features (Bahdanau et al., 2015; Vaswani et al., 2017). Although the interpretation of attention is under discussion (Li et al., 2016; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), it still provides useful clues about the internal behavior of deep, especially Transformer-based, language models (Clark et al., 2019b). Through attention intervention, which

seeks to re-parameterize the original attention to represent a conditioned or restricted structure, a number of works have successfully improved various model capabilities, such as long sequences understanding (Beltagy et al., 2020; Shi et al., 2021; Ma et al., 2022), contextualizing entity representation (Yamada et al., 2020), information retrieval (Jiang et al., 2022), and salient content selection (Hsu et al., 2018; Wang et al., 2022a). Some recent works also add attention constraints to improve model robustness towards specific distribution shifts, including identity biases (Pruthi et al., 2020; Attanasio et al., 2022; Gaci et al., 2022) and structural perturbations (Wang et al., 2022b).

6 Conclusion

In this paper, we propose READ, an end-to-end debiasing method that mitigates unintended feature biases through learning the attention residual of two models. Evaluation on OOD test sets of three NLU tasks demonstrates its effectiveness of unknown bias mitigation and reveals the crucial role of attention in robust NLU models. Future work can apply attention debiasing to mitigate dataset biases in generative tasks and multi-modality tasks, such as societal biases in language generation (Sheng et al., 2021) and language bias in visual question answering (Niu et al., 2021).

Acknowledgement

We appreciate the reviewers for their insightful comments and suggestions. This work was partially supported by the NSF Grant IIS 2105329, by the Air Force Research Laboratory under agreement number FA8750-20-2-10002, by an Amazon Research Award and a Cisco Research Award. Fei Wang was supported by the Annenberg Fellowship at USC. Tianyi Yan was supported by the Center for Undergraduate Research in Viterbi Engineering (CURVE) Fellowship. Muhao Chen was also supported by a subaward of the INFER Program through UMD ARLIS. Computing of this work was partly supported by a subaward of NSF Cloudbank 1925001 through UCSD.

Limitation

Although our experiments follow the setting of previous works, the experimented tasks, types of biases, languages, and backbone PLMs can be further increased. As we do not enforce additional constraints when learning attention residual, there is a potential risk of over-debiasing, which is currently controlled by ensemble ratio α. We implement the idea of residual attention debiasing based on the one-stage PoE framework because it is one of the most successful end-to-end debiasing methods for NLU models. However, the effectiveness of attention debiasing may not be limited to the specific debiasing framework. Since the proposed method focuses on mitigating attention biases, it cannot be directly applied to PLMs without attention modules, such as BiLSTM-based PLMs (Peters et al., 2018). Moreover, the proposed debiasing method may also be effective to generative PLMs, such as T5 (Raffel et al., 2020) and GPT-3 (Brown et al., 2020). We leave this for future work.

References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sizhe Chen, Zhengbao He, Chengjin Sun, Jie Yang, and Xiaolin Huang. 2020. Universal adversarial attack on attention and the resulting dataset damagenet. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019a. 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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019b. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

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.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of nlu models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929.

Jacob Eisenstein. 2022. Informativeness and invariance: Two perspectives on spurious correlations in natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4326–4331.

Dan Friedman, Alexander Wettig, and Danqi Chen. 2022. Finding dataset shortcuts with grammar induction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022. Debiasing pretrained text encoders by paying attention to paying attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Songyang Gao, Shihan Dou, Qi Zhang, and Xuanjing Huang. 2022. Kernel-whitening: Overcome dataset bias with isotropic sentence embedding.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E 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.

Abbas Ghaddar, Philippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021. End-to-end self-debiasing framework for robust nlu training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1923–1929.

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. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

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.

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.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. 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 4129–4142.

Zhengbao Jiang, Luyu Gao, Jun Araki, Haibo Ding, Zhiruo Wang, Jamie Callan, and Graham Neubig. 2022. Retrieval as attention: End-to-end learning of retrieval and reading within a single transformer. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. 2020. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

Yuanxin Liu, Fandong Meng, Zheng Lin, Jiangnan Li, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. A win-win deal: Towards sparse and robust pre-trained language models. In *Advances in Neural Information Processing Systems*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yougang Lyu, Piji Li, Yechang Yang, Maarten de Rijke, Pengjie Ren, Yukun Zhao, Dawei Yin, and Zhaochun Ren. 2023. Feature-level debiased natural language understanding. *Proceedings of AAAI*.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2022. Mega: moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*.

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.

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.

Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2022. Debiasing masks: A new framework for shortcut mitigation in nlu. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. 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 Papers)*, pages 3456–3468.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: Debiasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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.

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.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. 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 Papers)*, pages 5434–5445.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. 2020. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning*, pages 7909–7919. PMLR.

Sara Rajaee, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. Looking at the overlooked: An analysis on the word-overlap bias in natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2020. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

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.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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 Papers)*, pages 4275–4293.

Han Shi, Jiahui Gao, Xiaozhe Ren, Hang Xu, Xiaodan Liang, Zhenguo Li, and James T Kwok. 2021. Sparsebert: Rethinking the importance analysis in self-attention.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11349–11357.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Can Udomcharoenchaikit, Wuttikorn Ponwitayarat, Patomporn Payoungkhamdee, Kanruethai Masuk, Weerayut Buaphet, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Mitigating spurious correlation in natural language understanding with counterfactual inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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.

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. *Advances in neural information processing systems*, 30.

Fei Wang, Kaiqiang Song, Hongming Zhang, Lifeng Jin, Sangwoo Cho, Wenlin Yao, Xiaoyang Wang, Muhao Chen, and Dong Yu. 2022a. Salience allocation as guidance for abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Fei Wang, Zhewei Xu, Pedro Szekely, and Muhao Chen. 2022b. Robust (controlled) table-to-text generation with structure-aware equivariance learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022c. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *NAACL*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ruibin Xiong, Yimeng Chen, Liang Pang, Xueqi Cheng, Zhi-Ming Ma, and Yanyan Lan. 2021. Uncertainty calibration for ensemble-based debiasing methods. *Advances in Neural Information Processing Systems*, 34:13657–13669.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware

self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019a. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. 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.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019c. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

A Datasets

We use all the datasets in their intended ways.

MNLI dataset contains different subsets released under the OANC's license, Creative Commons Share-Alike 3.0 Unported License, and Creative Commons Attribution 3.0 Unported License, respectively. Among all the data entries, 392,702 samples are used for training. 9,815 and 9,832 samples from validation matched and validation mismatched subsets of MNLI respectively are used for evaluation.

HANS is released under MIT License. The validation subset of HANS contains 30,000 data entries, which are used for OOD evaluation of natural language inference.

FEVER follows the Wikipedia Copyright Policy, and Creative Commons Attribution-ShareAlike License 3.0 if the former is unavailable. 311,431 examples from the FEVER dataset are used to train the model.

FEVER-Symmetric test set with 717 samples is used as the OOD challenge set for fact verification.

QQP[8] consists of 363,846 samples for training, and 40430 samples for in-distribution evaluation.

PAWS dataset with 677 entries is used for OOD evaluation of paraphrase identification.

B Implementation

Our Implementation is based on HuggingFace's Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). Since the training sets of three tasks are of roughly the same size, it takes about 5 to 6 hours to finetune the BERT-base model, which has around 110 million parameters, on each task. Our ensemble model adds 5.3M parameters, a 4.8% increase from the BERT-base model. These additional parameters will be removed after the completion of training. During training, we use a linear learning rate scheduler and the AdamW optimizer (Loshchilov and Hutter, 2018). Models finetuned on the MNLI dataset will predict three labels, including *entailment*, *neutral*, and *contradiction*. During inference on the OOD test set, we map the latter two labels to the *non-entailment* label in HANS.

[8]The dataset is available at https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section at the end.

☑ A2. Did you discuss any potential risks of your work?
Limitation section at the end.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 at the beginning.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 and Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.